LSE Media and Communications

Media@LSE MSc Dissertation Series

Editors: Bart Cammaerts and Nick Anstead

POLARFLATION

The Inflationary Effect of Attention-Optimising Algorithms on Polarisation in the Public Sphere

Samuel Caveen

‘Polarflation’

The inflationary effect of attention-optimising algorithms on polarisation in the public sphere

SAMUEL M. CAVEEN[1]

[1] samuel@caveen.com

Published by Media@LSE, London School of Economics and Political Science ("LSE"), Houghton Street, London WC2A 2AE. The LSE is a School of the University of London. It is a Charity and is incorporated in England as a company limited by guarantee under the Companies Act (Reg number 70527).

In the interests of providing a free flow of debate, views expressed in this paper are not necessarily those of the compilers or the LSE.

Abstract

Digital platforms are hegemonic in the modern media ecology and represent the dominant access point to today's public sphere. Fundamentally advertising tools, platforms harvest and broker users' attention, and so must maximise it. They leverage the surfeit of data they collect on user preferences through advanced artificial intelligence algorithms to predict the content most likely to hold a user's attention. Research reveals the role affect, outrage and incivility play in fuelling engagement with online content; I posit "attention optimisation" promotes these. A theoretical framework for this perverse feature of platforms is proposed, conceived as ANTI-DELIBERATIVE OPTIMISATION (ADO)*. This process is demonstrated, including how it can lead to inflationary polarisation, or '*POLARFLATION*'. The hypothesis is tested on a large, longitudinal Twitter dataset covering every tweet (N = 22,593,965) from a panel of UK politically engaged users (n = 1,228) over the period of a decade. Twitter's introduction of algorithmic curation in 2016 is used as a discontinuity. Supervised machine learning classifies the tweets for incivility, while a retweet typology is used as a novel instrument for algorithmic influence. I find incivility on Twitter has increased in concurrence with the introduction of algorithmic curation, estimating a 42% increase. Additionally, retweets I believe are most indicative of algorithmic influence were between 5 and 11 p.p. more uncivil than other tweets. These findings represent a significant contribution to the understanding of platforms and their consequences for the digital public sphere.*

Glossary

ABBR./SYMBOL	DEFINITION
ADO	ANTI-DELIBERATIVE OPTIMISATION
API	APPLICATION PROGRAMMING INTERFACE
AV	ALTERNATIVE VOTE (ELECTORAL SYSETM)
CCA	COMPUTATIONAL CONTENT ANALYSIS
ENRT(s)	EXTRA-NETWORK RETWEET(S)
EU	EUROPEAN UNION
FE	FIXED EFFECTS
FRT(s)	FOLLOWING RETWEET(S)
ICTs	INFORMATION AND COMMUNICATION TECHNOLOGIES
ITS	INTERRUPTED TIME SERIES
INRT(s)	INTRA-NETWORK RETWEET(S)
ME	MIXED EFFECTS
NFRT(s)	NOT-FOLLOWING RETWEET(S)
OLS	ORDINARY LEAST SQUARES
OR	ODDS RATIO
p	PROBABILITY
p.p.	PERCENTAGE POINT(S)
$P(x)$	PROBABILITY OF x
QCA	QUANTITATIVE CONTENT ANALYSIS
RE	RANDOM EFFECTS
SD	STANDARD DEVIATION
SE	STANDARD ERROR
SEO	SEARCH ENGINE OPTIMISATION
SNSs	SOCIAL NETWORK SITES
β	BETA COEFFICIENT

INTRODUCTION

> The whole thing just felt to me like it was the strangest and darkest version of this thing that I'm becoming more and more consumed with worry about all the time, which is like… **the internet is doing something to us that is profoundly changing who we are**.
>
> **Andy Mills**, *Rabbit Hole*
> *The New York Times* podcast
>
> Reflecting on the 2019 Christchurch shooting,
> which the gunman livestreamed on Facebook.

A handful of online platforms — where, among other things, users view, share and comment on content — have become almost hegemonic in constituting the digital public sphere (Moore & Tambini, 2018). Their near ubiquity makes them the dominant access point to democratic deliberation, while obsessive use by activists, politicians and journalists gives them growing influence over political outcomes (Behr, 2018). Their affordances, practices and tendencies have become a matter of great debate and concern among scholars, particularly because several pathologies have been identified that plague platforms.

At the same time, political polarisation heightens apace in the developed world (Hobolt et al., 2020; Klein, 2020), while democratic gains once thought indelible now look fragile in the face of a rising tide of populism (Bermeo, 2016; Inglehart & Norris, 2016). As these tendencies have coincided with the dawn of platforms, many inevitably draw a connection, although the theoretical and empirical basis for this is contested (Flew, 2019; Schroeder, 2019).

The business models of these platforms require the maximation of their users' attention so it can be sold to advertisers. To achieve this goal, they employ increasingly sophisticated and effective algorithms to determine which content a user is most likely to engage with, so it can be displayed in a feed to hold their attention (Stöcker, 2020). This machine learning technique is based on a user's connections, preferences and behaviour on the platform, creating a positive feedback loop whereby further engagement with suggested content generates content suggestions better still at eliciting a user's engagement (Andersson Schwarz, 2017). This marketplace has been termed the "attention economy" (Wu, 2017a, 2017b), and the methods of competing within it, "surveillance capitalism" (Mosco, 2015; Zuboff, 2019). This study, in part, explores and integrates this process into theory.

In the political context, disquiet has been well debated over three proposed platform pathologies in particular: their role as a vector for the spread of misinformation (Dahlgren, 2018b; Del Vicario et al., 2016; Horowitz, 2019); their observation and manipulation of users' behaviour (Boucher, 2019; Cadwalladr & Graham-Harrison, 2018; J. Cohen, 2018); and their segmentation of users into "echo chambers" containing only attitudinally concordant content (J. Cohen, 2018; Pariser, 2011; Sunstein, 2008, 2017, 2018). Empirical evidence for this homophilic "filter bubble" effect is underwhelming (Bakshy et al., 2015; Cardenal et al., 2019; Dubois & Blank, 2018; Eady et al., 2019; Haim et al., 2018; Halberstam & Knight, 2014; Quattrociocchi et al., 2016), while the first two issues overstate the abuse of platforms at the expense of scrutinising the detrimental effects of their routine operation.

Comparatively less theoretical analysis and empirical study has focused on the effect of this attention optimisation on the deliberative character of these platforms. Research reveals the role affect, outrage and incivility play in fuelling engagement with online content and propelling its spread (Berger & Milkman, 2012; Brady et al., 2017; Crockett, 2017; Dahlgren, 2018a; Kormelink & Meijer, 2017; Kravetz, 2017; Rayson, 2017). It is reasonable to posit that content algorithms reproduce and propagate these qualities (Tufekci, 2015, 2018a, 2018b).

I contend that flaws in the three aforementioned diagnoses stem from an insufficient grounding in these nuances of the attention economy's commercial logic, and a meagre appreciation of the technological affordances platforms bring to bear in that logic's pursuit. This study attempts to codify that appreciation theoretically before testing it empirically.

Literature on the public sphere and the internet's impact upon it is first reviewed, while a typology distinguishes features of different web eras and how they relate to contemporary concerns. The attention economy and the technological supremacy which enables it are explicated, before content algorithms and their attention-optimising ends and means are examined in depth. The role incivility plays in this media ecology is then expounded.

The findings of this exploration and review are synthesised into a cogent theoretical framework wherein the concept of ANTI-DELIBERATIVE OPTIMISATION (ADO) is proposed: a perverse consequence of attention economy logic with dire ramifications for the public sphere. This process is demonstrated, including how it can lead to inflationary polarisation, or 'POLARFLATION': an exaggeration of the 'true' level of polarisation in the public sphere.

These concepts and proposed mechanisms are operationalised as a hypothesised increase in incivility on Twitter since the platform introduced algorithmic curation in 2016. While Twitter datasets have decreased over time due to prohibitive API costs (Zimmer & Proferes, 2014), I access a proprietary archive of the Twitter "Firehose API", allowing this hypothesis to be tested on a large, longitudinal dataset covering a decade of activity. I take a panel of more than a thousand UK politically engaged users (*n* = 1,228), recording every tweet over the 10 year period, totalling almost 23 million. I then train a supervised learning classifier to identify incivility in this corpus and, using a novel instrument for algorithmic influence, find strong evidence that ADO is underway on Twitter, suggesting its content algorithm is inflating polarisation in the public sphere. The implications of these findings for future research are discussed, as well as possible solutions to the issues evinced.

The study provides an original theoretical synthesis, substantial empirical findings and a novel methodological approach. Taken together, it represents a significant contribution to the understanding of platforms and their consequences for the digital public sphere.

Platforms, polarisation and the public sphere

Few concepts at the nexus of sociology, political theory and communication sciences are as resilient as Habermas's public sphere. Subject to ceaseless waves of scholarly critique since its first appearance in *The Structural Transformation of the Public Sphere* (1962/2015), the conceptual site of civic discussion and organisation — situated between the state and its citizens' private lives — is a notion that refuses to dissipate (Iosifidis, 2011). Whether due to Habermas's own receptivity to criticism (Calhoun, 1992) — much of which followed his seminal work's 1989 English translation — or the underlying flexibility of the idea itself, many theorists take for granted the normative premise that a space for the free and open exchange of ideas is a precondition of democracy (Dahlgren, 2005; Fraser, 1990; Pfister, 2018).

In fact, it is this normative account that allows the public sphere conception to endure. While much of the criticism directed at Habermas concerns his historical portrayal — too nostalgic for a bourgeois past (Hall et al., 2003); not critical enough of its exclusionary origins (Calhoun, 1992); too singular in its expression (Cunningham, 2016; Fraser, 1990) — his critics often integrate his theoretical analysis into their own, as Garnham (1986) did with his public service broadcasting argument, or Fraser (1990) with her "counterpublics". A constant across these reframings and adaptations is the foundational principle that the public sphere is constructed and facilitated by the communications practices and media technology of the day, and thus structural transformations therein have deep ramifications for democracy. Naturally then, a shift in the practices and technology of the day will raise vexed questions from a Habermassian perspective, and indeed the emergence of the internet was steroidal for public sphere debate.

The internet and the public sphere

Cyber-optimism vs. cyber-pessimism

Initially scholars asked whether the internet contributes to, or "impacts" the public sphere, before considering whether it constitutes a public sphere in its own right, all while the technology was still a niche curiosity (Dahlberg, 2001; Dahlgren, 2000; Galston, 2002; Masip et al., 2019; Papacharissi, 2002; Schäfer, 2016). The debate could broadly be divided between *cyber-optimist* and *cyber-pessimist* perspectives (Schäfer, 2016), but both positions drew from the critical discourse surrounding Habermas's original public sphere conception.

Optimists hoped the internet would counter the issues of access, diversity and capital interest in Habermas's account (Benkler, 2006; Murru, 2009; Shirky, 2011), while pessimists feared it would exacerbate the fragmentation of publics and commodification of civic discourse already underway in the offline sphere (Dahlgren, 2005; Iosifidis, 2011; Papacharissi, 2002; Pariser, 2011; Sunstein, 2001). As the technology's adoption reached saturation, the question was no longer *whether* it constituted a venue for democratic deliberation, but *what effect* it was having on democratic deliberation at large:

> It is there [the internet] that we find the real "vanguard" of the public sphere, the domain where the most intense developments are taking place—what we might call the cyber transformation of the public sphere. (Dahlgren, 2005)

The advent of social media brought the posited discursive properties of the internet to a mass audience (Stumpel, 2009), allowing for the first empirical tests of these theoretical exchanges. Early momentum supported the optimists, with the successful grassroots campaign of Barack Obama and uprisings in the Arab world demonstrating the democratising, decentralising and participatory capacity of platforms like Facebook and Twitter (Iosifidis, 2011; Shirky, 2011). But with the 2016 election of Donald Trump (Tufekci, 2018c), deepening polarisation (Narayanan et al., 2018) and rampant online misinformation (Marwick & Lewis, 2017), the mood has soured as many of the pessimists' warnings appear to have come to pass.

> *Academic literature has oscillated between an initial optimism about the potential for strengthening democracy of communication technologies to a critical scepticism. (Masip et al., 2019)*

Dahlgren, who had previously struck a cautiously optimistic note just 13 years previously about "new" politics imbued with a more grassroots character, now sounds the alarm about the "epistemic crisis of democracy" (Dahlgren, 2005, 2018b). Services whose bypassing of gatekeepers was once believed emancipatory are increasingly seen as megaphones for authoritarian populists (McNair, 2018). Deibert (2019) summarises the fall from grace well:

> Many thought that social media would empower transnational civil society, but now it seems that social media may be contributing to civil society's slow demise.

New information and communication technologies (ICTs) are now regularly linked to retreats around the world in the march of democracy once proclaimed inevitable (Bermeo, 2016; Fukuyama, 1993), but this concern was also raised in the past. Iosifidis (2011) argued that the "democratising and empowering functions of the Internet and the new social media are being exaggerated" by selective references to "Twitter revolutions" that discounted larger participatory issues. While Dahlgren (2005) acknowledged setbacks to democracy were visible as early as the 90s and the internet's amelioration of, or contribution to, the phenomenon were ambiguous.

Grim diagnoses of the contemporary public sphere now include concerns over privacy and surveillance (Blank & Duton, 2019; Zuboff, 2019), the power of microtargeted advertising (Nadler et al., 2018; Tufekci, 2016, 2017), diminished gatekeeping and responsibility (Bruns, 2017; Bruns & Highfield, 2015; Coddington & Holton, 2014; Singer, 2014), and harassment and hate speech (Barnes, 2018; Hsueh et al., 2015; Phillips, 2015). How then was there such disjunction in predictions of the internet's democratic impact?

Table 1: Typology of eras of online content.

WEB 1.0	WEB 2.0	WEB 3.0
	CONTENT FLOW	
One-way flow of content from publishers to users	Two-way flow of content between publishers and users	Constant, confluent flow of content between publishers and users, synergistically mediated by artificial intelligence
	CONTENT FORMAT	
Not dissimilar to print media, but operating with digital speed and flexibility	User-generated content and the dawn of social media, enabling online communities at scale	Algorithmically determined infinite feeds and recommendation engines
	EXAMPLES	
Yahoo! News, *NYTimes.com*, *BBC News website*	*MySpace*, *Tumblr*, message boards, comments, blogs	*Facebook News Feed*, *YouTube Up next*

A tale of two webs

I contend that the now apparently naïve optimism about the internet's democratising potential and the seemingly prescient pessimism share similar features in not fully presaging the actually deleterious force behind the internet's dark *volte-face*. To illustrate this, I first present a rudimentary historic framework for eras of online content in Table 1. In the *Web 1.0* era, the cost and expertise required for publishing content to the internet was prohibitive for most but established media institutions and early internet companies. As such, the content output was largely indistinguishable from print media, but for operating at the speed of 'bits not atoms' (Negroponte, 1995). *Web 2.0* saw the arrival of publishing tools such as blogging services, online comments systems and early social network sites (SNSs) that allowed a far larger audience to participate in the creation and sharing of content. This transition was not necessarily linear, as bulletin board systems (BBS) predated even the web itself (Driscoll, 2016; Lanier, 2018), but in the main there was a moment when, for the vast majority, the web went from a 'read-only' medium to a participatory one.

It is in this *Web 2.0* context that early digital public sphere debates were grounded, and many of the proposed promises and perils stem from the same *Web 2.0* features. Consider that, whereas *Web 1.0* was more an evolution of the 'one-to-many' communication enabled by legacy mass media, for Shirky (2008) *Web 2.0* represented the dawn of 'many-to-many' communication, which in his positive view would allow subaltern communities to bypass the establishment and access media power for themselves (Murru, 2009). To pessimists, this same circumvention of 'gatekeepers' would proliferate low quality information in the absence of journalistic vetting (Singer, 2010; Stöcker, 2020). For Lessig (2006), these frictionless connections would enable communities to transcend borders and enable the grassroots organising of the "new" politics Dahlgren (2005) hinted at. The more dour view saw dangerous actors taking advantage of the same transnational organisational capacity (Andersen & Sandberg, 2018; Guan & Liu, 2019; Karatas & Saka, 2017). For optimists, online anonymity was a way for marginalised groups to discover and connect with likeminded individuals without fear of reprisals (Christopherson, 2007; McKenna & Bargh, 1998), whereas for others, its disinhibitory effect would encourage toxicity (Haines et al., 2014; Rains, 2007; Suler, 2004).

There is evidence to support either side's predictions and balancing them against one another will be a matter of interpretation. In my view, the enablement of grassroots movements like #MeToo and Black Lives Matter — that can each claim to have quantifiably shifted attitudes (Cohn & Quealy, 2020; Keplinger et al., 2019) — are on balance more positive developments than, for instance, the distributed propaganda networks of the Islamic State are negative (Andersen & Sandberg, 2018). The calculations of others will differ.

Regardless, these amount to what Dahlgren (2005) termed the *representational* and *interactional* components of the public sphere, and I argue that it is Dahlgren's third, *structural*, dimension — the governing economic framework — that is responsible for the ills highlighted at the end of §1.1.1, a different phenomenon than either cyber-optimists or cyber-pessimists identified. Put differently, while the aforementioned promises and perils pertain to the "affordances" of social media for *users* (boyd & Ellison, 2007), they do not consider the affordances for the actual *customers* of these platforms: advertisers. Here the shift to *Web 3.0* is illuminating, and its confluence between users, publishers, advertisers and platforms — enabled by advances in artificial intelligence and big data collection — is the background against which the bleak lamentations of the contemporary public sphere take place.

For McEwan et al. (2018), scholarly theoretical perspectives lag behind the development of this modern media ecology. What follows in the next sections (§§1.2–1.4) is an explication of this *Web 3.0* shift, laying groundwork and context for an attempt in the subsequent chapter (§2) to provide one such theoretical framework.

The "attention economy": Business model of platforms

Deibert (2019) recalls with irony that it was once questioned whether turning a profit online was even possible. Today, of the ten most valuable companies in the world by market capitalisation, four — Amazon, Google/Alphabet, Facebook and Alibaba — are internet companies founded since 1995, two — Apple and Microsoft — are legacy technology companies that operate instrumental services in the internet economy, and a seventh — Tencent — is a conglomerate with significant holdings in internet services (Statista, 2020b).

All seven, particularly the first four, are considered 'platforms': digital infrastructures that act as intermediaries, facilitating interactions between consumers, advertisers, producers and suppliers (J. Cohen, 2018; Srnicek, 2017). Social media platforms like Facebook, Twitter and YouTube[2] are advertising tools at their core — they enact "multisided markets" (Andersson Schwarz, 2017) whereby user-to-user connections are facilitated as well as the exchange of content between publishers and users, while simultaneously permitting advertisers to access the same users. It is this latter transaction that is monetised, allowing — nay, *necessitating* — users free access to the service in order for their attention to be harvested and brokered by the platform, as Wu (2017b) illustrates by analogy:

> The spender of attention [has] a large supply of gold dust leaking from his pocket at a constant rate. This is the consumer, the spender of attention, whose very presence is valuable. As he walks down the street, merchants (the Attention Brokers) might offer him free … drinks … and then might charge the other patrons (the advertisers) extra for the opportunity to pick up some of the dust that falls as the man enjoys his drinks. That, in a nutshell, is the business model of the Attention Broker.

Scholars describe this as an "attention economy", in which platforms must maximise the attention of their users in order to increase revenue (Davenport, 2001; Goldhaber, 1997; Lanham, 2006; Wu, 2017a). As a user's interactions on the platform are digital, they are *inherently quantified*, and therefore enable a surfeit of information on their preferences and habits to be obtained by processing the data their usage generates. This data is useful both to the clients of platforms — who wish to use it to target potential consumer segments for marketing purposes — and the platforms themselves, who can use the data to provide increasingly relevant content, personalised to an individual user's tastes, increasing the amount of time they spend on the platform, and thus the amount of their attention available to sell. This process of monitoring users and capturing their data, in turn to make it more likely that they will purchase products and services, has appropriately earned the label "surveillance capitalism", coined by Mosco (2015) and popularised by Zuboff (2019).

[2] The Google subsidiary is not only the world's largest video platform (Stöcker, 2020), it shares several similarities with social networks (B. Lewis, 2018), enough to be considered a form of social media itself.

Algorithms, attention optimisation and “superlative efficacy”

Both the inherent quantification of platform usage and the ease of content creation in the digital age mean the sheer volume of information for platforms to process, organise and present to users is monumental (Deibert, 2019), making an automated sorting mechanism “inevitable” (Stöcker, 2020). However, beyond implementing an ‘internet Dewey Decimal System’, the demands of the attention economy compel platforms to utilise algorithms which create for users “their own unique immersive media environments.” (J. Cohen, 2018)

Content algorithms: “You thought you were alone in the universe”

Variously called “recommendation systems”, “automated decision making (ADM) systems” (Stöcker, 2020), “automated media” (Napoli, 2014), “news recommenders” (Helberger, 2019) and “algorithmic personalisation” (Perra & Rocha, 2019), I will use the terms “content algorithms” and “algorithmic media” to refer to the process by which platforms present content to a user, in the form of personalised timelines or recommendations, based on the previous usage of themselves and others. These systems rely on machine learning, a form of computing which makes predictions or decisions by optimising inputs to desired outputs through the identification of statistical patterns in a dataset, “without being explicitly programmed” (Koza et al., 1996).

For content algorithms, this means anticipating the content a user is likely to consume based on the behaviour of others who have consumed similar content in the past (J. Cohen, 2018; Stöcker, 2020). A particularly illustrative example is the music service Spotify, which uses a content algorithm to recommend new music a user may like in the form of its Discover Weekly playlist. It populates your playlist by finding new tracks other users who listen to similar music to you are currently playing often. Operating at a scale of almost 300 million users (Music Business Worldwide, 2020) this procedure can recommend songs with “uncanny” or “scary” accuracy, because someone somewhere has near identical taste in music to you, as a Spotify engineer told Pasick (2015):

> When I was young [music] helped identify who you were: I am this type of music. … You thought you were alone in the universe until you realise there’s a guy just like you, musically at least.

"Superlative efficacy": The virtuous cycle of Web 3.0

As with any statistical model, the more datapoints available to this process, the greater the confidence with which it can make predictions. Naturally then, the more one uses a service mediated by content algorithms — and thus provides more datapoints on which to base predictions — the better calibrated its content recommendations will be, simply by dint of the closed loop of positive feedback.

This is the defining feature of digital platforms in the *Web 3.0* era (Table 1). *Web 2.0* platforms benefitted from traditional network effects which were particularly pronounced given the typically zero financial cost of joining (Coyle, 2018; Mansell, 2015). *Web 3.0* platforms though, through the seamless interchange between a user's past behaviour and future experience, benefits from an "entirely novel form of synergy" (Andersson Schwarz, 2017). While a telephone — the classic network effect — becomes more valuable to its users with every new person who possesses one, it does not provide *better quality conversations* on phone calls simply as a product of each marginal call connected. Thompson (2018) calls the platform model "internalised network effects", and its continuous and self-perpetuating process forms a virtuous cycle, or what Andersson Schwarz (2017) calls "superlative efficacy", which has allowed platforms to establish hegemony in their respective markets (Iosifidis, 2011).

Attention optimisation: The social validation casino in your pocket

On platforms where users pay for access, like Spotify or Netflix, superlative efficacy simply provides an outstandingly useful service, notwithstanding severe disruption to their respective industries (Prey et al., 2020; Turner, 2019). Things go awry when this technological supremacy is brought to bear on the "free" business model of the attention economy.

Particularly following the election of Donald Trump (Madrigal, 2017) and the UK's surprise decision to leave the European Union (Cadwalladr, 2019), three aspects of attention economy platforms have received a lot of scholarly and popular scrutiny:

- the manipulation and hijacking of platform features — including content algorithms — by outside actors to promote deliberately false information or "fake news" (Bradshaw, 2019;

Dahlgren, 2018b; Del Vicario et al., 2016; Horowitz, 2019, 2019; Lewandowsky et al., 2017; Narayanan et al., 2018; Stöcker, 2020; Vosoughi et al., 2018), either for pecuniary gain (Hughes & Waismel-Manor, 2020), or as a propaganda tool of geopolitical strategy (Khaldarova & Pantti, 2016; Linvill & Warren, 2020)

- the potential for the unprecedented volume of behavioural data platforms collect to be used for "psychographic profiling" (J. Cohen, 2018), enabling users to be subconsciously and emotionally influenced (Kramer et al., 2014), in a manner encapsulated by the Cambridge Analytica scandal (Boucher, 2019; Cadwalladr & Graham-Harrison, 2018)
- the tendency of content algorithms to only present users with content concordant with their own views, erecting a "filter bubble" around them (Pariser, 2011), keeping them in an "echo chamber" (Sunstein, 2008, 2017, 2018), or sorting them into homophilous and separate strata (Farrell, 2012).

I explore and challenge each of these critiques in more detail than space allows here in Appendices 1 and 2. In summary:

- while weaponised fake news is certainly an issue for platforms, focusing on it elides the far more pernicious symbiosis platforms share with content which cannot be 'factchecked away'; so-called "borderline content", which goes right up to the edge of policy without crossing the line (Deibert, 2019; Kastrenakes, 2018), is fostered by attention economy logic
- the claims of Cambridge Analytica's shadowy operation were fraudulent (Halper, 2018), but the same principles of psychological manipulation are endogenous to the everyday operation of platform algorithms (J. Cohen, 2018; Deibert, 2019; Lanier, 2018)
- the "filter bubble" has been empirically rejected (Bakshy et al., 2015; Bechmann & Nielbo, 2018; Cardenal et al., 2019; Dubois & Blank, 2018; Eady et al., 2019; Haim et al., 2018; Scharkow et al., 2020; Shore et al., 2018) because, while some features of attention economy logic encourage and reinforce selective exposure, others require diverse, surprising and shocking content (Anzieu, 2019; J. Cohen, 2018); an exclusively pro-attitudinal environment would not be conducive to increasing attention.

According to McEwan et al. (2018), where these critiques fail — the "filter bubble" thesis in particular — is in a "binary perspective" which cannot "account for the complexity of online media ecologies". By beginning with an appreciation of the criteria content algorithms are optimising to, this complexity can be accommodated. Both "attention" and "engagement" are used in the literature, sometimes interchangeably (Crockett, 2017; Deibert, 2019; McEwan et al., 2018; Stöcker, 2020). Here I will delineate: "attention" is the resource platforms monetise and thus must maximise; "engagement" is a combination of factors describing a user's interaction with the platform from which attention is inferred, i.e. it is merely an instrument, a means to the platform's ends. In this regard, I refer to the process as 'attention optimisation', and 'engagement' as what is practically measured.

Several engagement metrics are available to platforms. In 2016, Facebook redesigned its 'Like' button with five additional emotional 'Reactions' a user can choose when engaging with content (Stinson, 2016). This drastically increased the diversity of data the platform collected, joining comments, shares, video views and dwell time (reading a post rather than scrolling) to compile an intricate, but quantifiable, picture of a user's behaviour on the platform (Stöcker, 2020). These inform the platform what content is likely to hold a user's attention, but they are also presented to the user as a form of feedback.

Controlling the administration of this feedback is an equally vital component of attention optimisation. Inspired by the pioneering methods of psychologist B. F. Skinner, such as operant conditioning, platforms create a "compulsion loop" through "variable-rate reinforcement", where the unpredictable nature of rewards and punishment can generate addiction (Crockett, 2017; Deibert, 2019). Feedback such as notifications of likes, comments or shares of a user's post are communicated as "triggering stimuli", through red dots, pop-ups and vibrations (Crockett, 2017; Deibert, 2019). The 'Like' button was expressly devised to give users "a little dopamine hit" ex-Facebook president Sean Parker admitted:

> It's a social-validation feedback loop … exactly the kind of thing that a hacker like myself would come up with, because you're exploiting a vulnerability in human psychology. (Solon, 2017)

Platform interface design reinforces this compulsion: just as casinos limit natural light and eschew clocks (Lane, 2006), video sites play the next recommended video automatically, while social media feeds scroll endlessly. Even the 'pull down to refresh' gesture has been likened to a slot machine's mechanism (Haubursin, 2018).

With the combination of a bottomless resource of content statistically determined to maximise their attention, and a system of social validatory feedback calibrated for maximum psychological impact, it is unsurprising that a study of social media users found tell-tale symptoms of addiction (Deibert, 2019).

Affect, outrage and incivility: "Easier than thinking"

The mechanisms of the attention optimisation system are intended to activate Daniel Kahneman's "System 1" cognitive reasoning, which Stöcker (2020) summarises as "quick, automatic, effortless, without voluntary control". Nir Eyal, a Silicon Valley product designer, writes in *Hooked: How to Build Habit-Forming Products* (2014), that to maximise a user's engagement with the platform, "doing must be easier than thinking". Put plainly, these principles are incompatible with a rationally deliberative public sphere.

Moral outrage: Engagement rocket fuel

Not coincidentally, deliberative reasoning is, instead, a feature of Kahneman's "System 2", whereas "System 1" is characterised by confirmation bias which favours "uncritical acceptance of suggestions and exaggeration of the likelihood of extreme and improbable events." (Kahneman, 2011) This helps explain the overwhelming evidence that affect, emotion and outrage fuel engagement with online content and propel its spread (Crockett, 2017; Dahlgren, 2018a; Kormelink & Meijer, 2017; Kravetz, 2017; Rayson, 2017). Berger & Milkman (2012) identified news content that evoked high emotional arousal was more viral, and Brady et al. (2017) found the use of moral-emotional words robustly increased the diffusion of political social media posts, while Hofmann et al. (2014) showed that the experience of moral outrage was far more common online than in other media or in person.

As Lanier (2018) points out, not only are outrage and emotion more engaging in their own right, but given the rapid method through which attention is judged, the most impulsive and reflexive users have the largest influence over what attention optimisation deems valuable (Klein, 2018a). As users respond to posts in the spur of the moment, content that provokes thought over a longer, more reflective period is less likely to be rewarded.

An abundance of incivility: Driving polarisation

Importantly for this paper, a public sphere perspective has prompted researchers to investigate what impact this is having on civic discourse and the prevalence of online incivility. Gervais & Chin (2018) found tweets that incorporated incivility received more attention, while Theocharis et al. (2020) discovered political incivility on Twitter is widespread among users, rather than only originating from an small section of trolls.

Crucially, exposure to incivility has been shown to have anti-deliberative effects (A. A. Anderson et al., 2014; Massaro & Stryker, 2012). Hwang et al. (2016) found uncivil discussion increased negative emotions and closed-mindedness, results that corroborate Borah's (2012) findings, while Theocharis et al. (2016) showed that incivility directed at politicians on Twitter curtailed civic engagement. In an experiment by Gervais (2013), exposure to "uncivil political talk" lead to "reduced satisfaction and willingness to compromise", while increasing uncivil behaviour in return.

This reciprocity indicates a relationship between incivility and polarisation, as shown by another Gervais (2019) experiment where exposure to counter-attitudinal incivility resulted in greater affective polarisation. Previous research supports this (A. A. Anderson et al., 2014; Lyons & Veenstra, 2016), including a Twitter study which found users became more polarised after exposure to counter-attitudinal messaging, contrary to a filter bubble expectation that polarisation would be ameliorated if differing views interacted more (Bail et al., 2018).

Anti-deliberative optimisation and 'polarflation'

This study proceeds with the concept of democratic deliberation to examine what impact attention optimisation, and the content algorithms which enact it, are having on the public sphere. It is concerned with the *deliberative character* of platforms, and the degree this is detrimentally shaped by the market logic directing information and interaction therein.

By deliberative character, I mean the necessary conditions for deliberation, invoking both Habermas's ideal speech situation (Habermas, 1981/2004) and Dahlgren's (2005) "civic culture" concept, which comprises five parameters: values, affinity, knowledge, identities and practices. Previous concerns regarding digital platforms as a public sphere can each be understood as a challenge to a different parameter: the proliferation of "fake news" brings the capacity of platforms as a venue for the exchange of *knowledge* into disrepute; the potential of reducing individuals to "psychographic profiles" violates their *identities* as citizens; and the division of groups into impermeable "filter bubbles" precludes finding *affinity* between different communities. This inquiry instead focuses on the *practices* parameter, specifically what Dahlgren calls "the most fundamental and ubiquitous practice": civic interaction and discussion.

Does attention optimisation — and its preoccupation with engagement, narrowly conceived — cultivate or impede the civic interaction that is foundational to a true public sphere? Computer programmer-cum-sociologist Zeynep Tufekci (2015, 2016, 2018a) thinks not, describing platforms as a "phantom public sphere":

> What does this algorithmic public sphere tend to feed us? … Glimmers of novelty, messages of affirmation and belonging, and messages of outrage toward perceived enemies. These kinds of messages are to human community what salt, sugar, and fat are to the human appetite. … Today's engagement algorithms … espouse no ideals about a healthy public sphere.

If attention-optimising algorithms operate as Tufekci suggests, I propose they are engaged in ANTI-DELIBERATIVE OPTIMISATION (ADO), in that they favour the promotion of content inconducive to deliberation and thus foster an environment antithetical to the public sphere. Furthermore, because of the cyclical nature of the attention optimisation process, ADO will form a feedback loop which, as McEwan et al. (2018) suggest, intensifies over time. Finally, given the tendency of anti-deliberative attitudes to increase and reinforce polarisation, ADO will have an inflationary effect on polarisation. Since this inflationary polarisation, or 'POLARFLATION', is, at least in part, the product of algorithmic curation, I contend it represents an exaggeration of the 'true' level of polarisation in the purported public sphere. Just as economic inflation is the increase in price over time relative to the same concrete basket of goods, POLARFLATION is the *appearance* of increasing polarisation between groups, relative to the actual change in polarisation of positions and feeling between them.

Conceptual framework

The premises of this study’s conceptual framework can be formalised like so:

(I) Content algorithms optimise for attention

(II) Attention is disproportionately garnered by affect, outrage and incivility

(III) Affect, outrage and incivility debilitate deliberation and increase polarisation

If (I) and (II) then:

(IV) Content algorithms disproportionately promote affect, outrage and incivility

And if (III) and (IV) then:

(V) Content algorithms debilitate deliberation and increase polarisation

To “debilitate deliberation” is to violate the practice of civil interaction Dahlgren outlined. If premises (I), (II) and (III) hold, then ANTI-DELIBERATIVE OPTIMISATION and POLARFLATION are in effect.

A simplified demonstration of how this process is proposed to occur is provided in the form of a spiral in Figure 1, which I will briefly expound. For illustrative purposes, all posts are considered to have either exclusively deliberative (D) or anti-deliberative (AD) properties, and the initial bundle of posts at the bottom of the spiral is imagined to be presented to the user before algorithmic intervention. In stage 1, the user engages with two anti-deliberative posts compared to one deliberative post, despite being presented with an even number of each, reflecting the psychological tendency detailed in §6.2.1. Based on this engagement, the algorithm provides more anti-deliberative posts in a 2:1 ratio. In the next stages, the user is presented as the same as during stage 1, but in reality, the *listening* stage of the process involves all other users on the platform as well, and it takes place simultaneously with the other stages, providing the algorithm with engagement data.

In stage 2, the user makes a considered, deliberative post, but few other users engage with it as they are preoccupied with more impulsive content. The algorithm recognises this and does not promote the user’s post into the feeds of others as a result. The user notices his post received few likes or comments. Lanier (2018) explains this lack of positive feedback approximates the 'punishment' of operant conditioning — the fear of being ignored is a powerful motivator in attention optimisation.

Frustrated by his last post's failure and provoked by further anti-deliberative posts in his feed, in stage 3 the user this time vents with an anti-deliberative post. Recognising the faster engagement, this time the algorithm promotes the user's post, causing it to garner even more engagement. The likes and shares send the user little dopamine hits. The algorithmic feed has been progressively optimised for anti-deliberative posts, while the user has begun to associate producing anti-deliberative content with a pleasure response.

In stage 4, this combination of exposure and validation have coalesced to simultaneously make the user more polaris*ed* and polaris*ing*.

Research questions

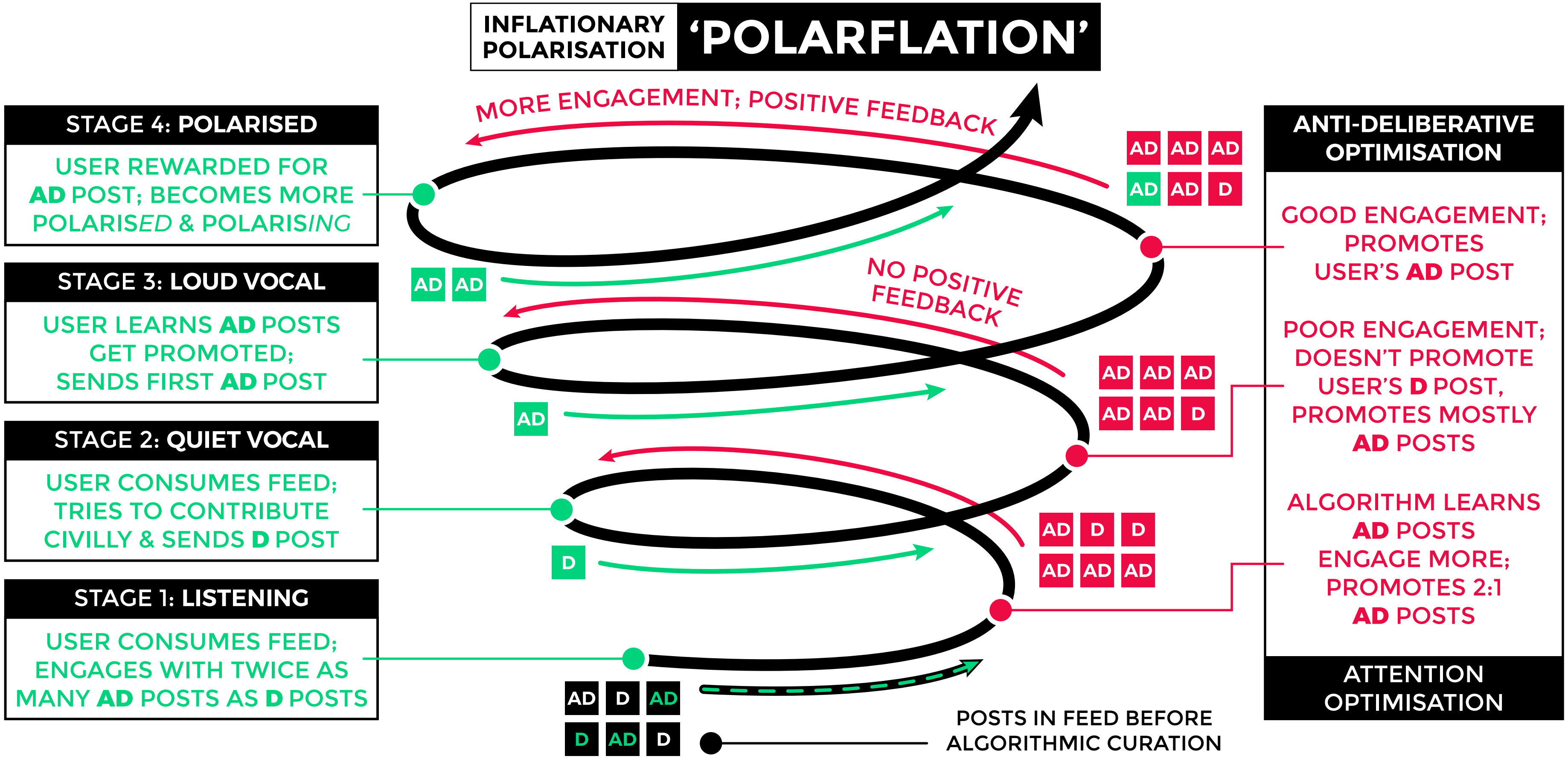
This research is guided by two questions. The first, informed by the ADO concept and Farrell's (2012) question over whether the internet is contributing to increasing political polarisation (Hobolt et al., 2020; Klein, 2020; The Policy Institute, 2019), asks:

> **RQ1**: Do attention-optimising algorithms promote anti-deliberative content, with the potential to further polarise the public sphere?

The second, with an understanding of the reinforcing feedback loop that powers attention optimisation (Andersson Schwarz, 2017; McEwan et al., 2018), asks:

> **RQ2**: Does this anti-deliberative promotion self-perpetuate, with the potential to inflate public sphere polarisation?

These research questions move empirical inquiry of digital platforms beyond content diversity and selective exposure to the deliberative character of the discourse they promote. They advance existing research on the abundance of anti-deliberative attitudes online and its connection to the affordances of digital platforms by tying this prevalence to the commercial logic of the attention economy. The research is grounded in the effects — insofar as these can be ascertained — of content algorithms, technological tools platforms utilise in pursuit of this logic. The following chapter (§3) details how this is to be achieved by offering a methodological innovation in the still-nascent research design of "algorithmic audits”. Finally, this study offers a theoretical contribution in the formalisation of the posited relationship between attention optimisation and polarisation, through the framework of ANTI-DELIBERATIVE OPTIMISATION and concept of POLARFLATION.


INFLATIONARY POLARISATION
'POLARFLATION'
STAGE 4: POLARISED
USER REWARDED FOR AD POST; BECOMES MORE POLARISED & POLARISING
STAGE 3: LOUD VOCAL
USER LEARNS AD POSTS GET PROMOTED; SENDS FIRST AD POST
STAGE 2: QUIET VOCAL
USER CONSUMES FEED; TRIES TO CONTRIBUTE CIVILLY & SENDS D POST
STAGE 1: LISTENING
USER CONSUMES FEED; ENGAGES WITH TWICE AS MANY AD POSTS AS D POSTS
MORE ENGAGEMENT; POSITIVE FEEDBACK
NO POSITIVE FEEDBACK
AD AD AD
AD AD D
AD AD
AD AD AD
AD AD D
AD
AD D D
AD AD AD
D
AD D AD
D AD D
POSTS IN FEED BEFORE ALGORITHMIC CURATION
ANTI-DELIBERATIVE OPTIMISATION
GOOD ENGAGEMENT; PROMOTES USER'S AD POST
POOR ENGAGEMENT; DOESN'T PROMOTE USER'S D POST, PROMOTES MOSTLY AD POSTS
ALGORITHM LEARNS AD POSTS ENGAGE MORE; PROMOTES 2:1 AD POSTS
ATTENTION OPTIMISATION

NOTE: D = deliberative; AD = anti-deliberative.

Figure 1: Illustration of how attention optimisation becomes ANTI-DELIBERATIVE OPTIMISATION (ADO) and leads to inflationary polarisation.

RESEARCH DESIGN

"Algorithmic audits": Observing the black box

The mechanisms of algorithms are opaque to the point they are described as a "black box" (Pasquale, 2015). This raises concerns that their governing principles do not comport with the public's interest (Mansell, 2015), but this opacity also poses an intractable challenge to platform research. As arcane constructions, algorithms require a great deal of technical knowledge to explain and interpret (J. Cohen, 2018; O'Neil, 2017; Pasquale, 2015). This might lend itself to elite interview research methods (Jupp, 2006), however as highly valued intellectual property, platforms are secretive about the workings of their systems. Even if they were more open, the nature of so-called "deep learning" neural network algorithms — which approximate the connections of neural synapses and operate through levels of mathematical abstraction — is such that even the people who design them do not fully understand their operation (Knight, 2017).

Researchers have utilised digital ethnography or surveys to understand how individuals perceive and use social media (Karatas & Saka, 2017; Neheli, 2018; Wojcik & Hughes, 2019). However, algorithms function in the background of the platform experience and have a manipulative quality (q.v. §1.3.3). In one study of Facebook, the majority of participants were not aware of algorithmic involvement at all (Kulshrestha et al., 2017).

These methodological challenges have given rise to a new class of research design: the "algorithmic audit". Inspired by studies the US government conducted to detect housing discrimination, the algorithmic audit approach understands an unobservable mechanism by its observable outputs (Mittelstadt, 2016; Sandvig et al., 2014). One such study was able to quantify ideological bias in search results (Kulshrestha et al., 2017), while experimental implementations have demonstrated the differences between users' algorithmically mediated content and their unfiltered feeds (Eslami et al., 2015). Experiments like this and others (Coletto et al., 2016; Gillani et al., 2018; Oz et al., 2018; Rudat et al., 2014; Steinfeld et al., 2016) would be the ideal approach for this study's research questions: an application would be installed on participants' devices to record the differences between content promoted by algorithmic feeds and the 'universe' of content available in the participants' networks. Unfortunately, the resources this would require are beyond this project's scope.

Instead an observational approach was chosen. As Lanier (2018) points out, attention optimisation operates on a probabilistic rather than deterministic basis. There will be users who produce a slew of anti-deliberative content regardless of the platform's subtle direction, whereas others will remain stoically deliberative despite cues to transgress. Both are consistent with the ADO thesis, which argues content algorithms will increase anti-deliberative content *on average*.

Quantitative content analysis (QCA) systematically identifies content features, and in combination with statistical methods, can infer the average tendency of a dataset (Hansen et al., 1998; Krippendorff, 2004). With computational content analysis (CCA) the coding of such features is automated, allowing an amount of content to be classified that would be manually infeasible (Grimmer & Stewart, 2013; Manning, 2008). As Theocharis et al. (2020) note in a study of similar design, this allows for a "bird's-eye view" that can uncover "temporal levels" which, as the next section (§3.2) describes, is central to this study's design.

Twitter's algorithmic turn

This research takes Twitter as its field of study. I will not argue that Twitter is the most consequential platform for the proposed phenomenon because it is not. Facebook, approaching three billion users, has effects an order of magnitude larger than Twitter

(Statista, 2020a, 2020c), while YouTube, with the potence of video content and its capacity as a platform of discovery (B. Lewis, 2018), is perhaps the most important venue of ADO, having been described as a "radicalisation engine" (Ingram, 2018). However, academic Facebook access is severely curtailed (Bastos & Walker, 2018), while video content is difficult to analyse at scale (M. Anderson & Jiang, 2018; Klein, 2018b).

Twitter, by contrast, provides relatively easy access to rich datasets, and the tweet format is well-adjusted to CCA, as several studies attest (Barberá et al., 2015; Demszky et al., 2019; Muddiman & Stroud, 2017; Parveen & Pandey, 2016; Theocharis et al., 2016, 2020), explaining the platform's overrepresentation in social research relative to its userbase (Bruns & Weller, 2014; Cihon & Yasseri, 2016).

Even more importantly for this study, unlike comparable platforms, Twitter did not previously employ an attention-optimising algorithm on its main timeline (Kantrowitz, 2016). Unlike the Facebook News Feed which has always been algorithmic, the Twitter timeline had formerly displayed tweets in simple reverse-chronological order. Facing user decline and struggling to turn a profit, Twitter introduced algorithmic curation to the timeline in 2016, immediately sparking a backlash with over a million users tweeting "#RIPTwitter" (Kantrowitz, 2018; Twitter, 2016). Two years later though, the algorithm was credited with increasing Twitter's usage and profits, a testament to the efficacy of attention optimisation (Hern, 2019; Kastrenakes, 2020).

Despite this, to my knowledge researchers have yet to study the timeline algorithm's introduction, which opens up the application of regression discontinuity designs like interrupted time series (ITS). These use average differences at the margins of a threshold to infer the effects of an intervention (Kontopantelis et al., 2015). To be clear, my implementation does not meet this quasi-experimental standard because it is not possible to know which users receive the 'treatment' of the algorithm and which do not. But by taking the introduction of the algorithm as an interruption in the regression analysis, and using proxy covariates to control for other plausible causes of increased anti-deliberative content over time, an average effect of the algorithm's introduction can be estimated.

Operationalisation: Incivility as anti-deliberative

Like other researchers in this area, I operationalise anti-deliberative content as tweets containing incivility (Gervais, 2013; Gervais & Chin, 2018; Oz et al., 2018; Theocharis et al., 2016, 2020). My research questions are first assessed purely on the temporal dimension, before and after February 2016, when the timeline algorithm was introduced (Twitter, 2016). **H1a** states that there will be more incivility after algorithmic introduction, answering **RQ1**; while **H1b** posits that incivility will increase at a higher rate following the algorithm's arrival, responding to **RQ2**.

> **H1**: The proportion of tweets containing incivility and the introduction of Twitter's timeline algorithm are associated.
>
> > **H1a**: The mean proportion of tweets containing incivility is larger than before Twitter introduced the timeline algorithm.
> >
> > **H1b**: The proportion of tweets containing incivility has increased at a higher rate since Twitter introduced the timeline algorithm.

I do not intend this temporal analysis alone as compelling evidence of an algorithmic effect — proxy variables can only take us so far in controlling for an unobserved variable as diffuse as heightened political tension, and a measurement as heterogenous as incivility (Theocharis et al., 2020). I intend these and the next hypotheses to be additive. Here I implement a new measure for algorithmic influence, one inherent to the Twitter platform and, I propose, less vulnerable to endogeneity: NOT-FOLLOWING RETWEETS (NFRTs).

One of the controversial features of the timeline algorithm is it promotes tweets into users' timelines from accounts they do not follow (Darcy, 2019; Kantrowitz, 2017). This was a feature prior to the algorithm (Dredge, 2014), but now with curated tweets appearing at the top of the timeline (Nemeth, 2020), I propose tweets from not-followed accounts will be more prevalent and prominent in user feeds, and therefore retweets of this type will be appreciably more common since algorithmic introduction (**H2a**) and, due to attention optimisation's virtuous cycle, increasingly common since (**H2b**).

> **H2**: The probability of tweets being NOT-FOLLOWING/EXTRA-NETWORK RETWEETS (NFRTs/ENRTs) and the introduction of Twitter's timeline algorithm are associated.
>
>> **H2a**: The mean probability of tweets being NFRTs is larger since Twitter introduced the timeline algorithm.
>>
>> **H2b**: The probability of tweets being NFRTs has increased at a higher rate since Twitter introduced the timeline algorithm.

Of course, the timeline algorithm is not the only way an NFRT could be produced. Indeed retweets themselves first emerged somewhat organically as a way to propagate messages beyond a single friendship network (Rocherolle, 2019). Many of the NFRTs in this sample may be the product of organic virality, i.e. retweeted by other accounts in a user's network. A more precise indicator of algorithmic influence, then, would be EXTRA-NETWORK RETWEETS (ENRTs), where the retweeter follows neither the original author nor anyone else who retweeted it first.

>> **H2c**: The mean probability of tweets being ENRTs is larger since Twitter introduced the timeline algorithm.
>>
>> **H2d**: The probability of tweets being ENRTs has increased at a higher rate since Twitter introduced the timeline algorithm.

H2c and **H2d** proposes a similar increase over time as NFRTs, while **H2e** posits ENRTs will be proportionally more common since the algorithm than those produced by organic virality.

>> **H2e**: The mean probability of tweets being ENRTs has increased by a larger percentage than the mean probability of tweets being INTRA-NETWORK RETWEETS (INRTs) since Twitter introduced the timeline algorithm.

Using ENRTs as a proxy for algorithmic influence allows **RQ1** to be answered from a different angle, as **H3** states incivility will be highest among this tweet type.

> **H3**: The probability of an EXTRA-NETWORK RETWEET (ENRT) containing incivility is higher than other tweet types.

The temporal and tweet type dimensions are then combined in **H4**, providing a layered answer to both research questions.

H4: The probability of an EXTRA-NETWORK RETWEET (ENRT) containing incivility and the introduction of Twitter's timeline algorithm are associated.

H4a: The mean probability of an ENRT containing incivility is higher since Twitter introduced the timeline algorithm.

H4b: The probability of an ENRT containing incivility has increased at a higher rate since Twitter introduced the timeline algorithm.

Finally, by proposing a lagged relationship between incivility across tweet type, **H5** provides further evidence to answer **RQ2**.

H5: The monthly change in proportion of a user's tweets containing incivility and the proportion of their NOT-FOLLOWING RETWEETS (NFRTs) containing incivility in previous months are associated.

DATA COLLECTION AND CASE SELECTION

Observational studies of social media are vulnerable to data collection criteria. This study followed a thorough collection and case selection process. The following is an expeditious synopsis, the process is detailed extensively in Appendix III, including references to precedent in the literature for decisions made.

Access to historical tweets

The field was accessed through the proprietary database of Brandwatch, a social listening tool which provides an addressable archive of every tweet since 2006, indexed directly from the Twitter Firehose API (Jaume, 2014; Morstatter et al., 2013). This allowed the circumvention of Twitter's prohibitive API fees and made this large scale historical study possible when these are usually precluded (Morstatter et al., 2013).

Sampling strategy

As Dahlgren (2005) noted, only a "small degree" of online interaction can be considered deliberative and so researchers observing political deliberation on Twitter need to identify it. I first defined a universe of UK *political talk* for the last 10 years by assembling a complex search string of political parties, individuals and topics. The former two comprised the 15 political parties receiving at least 0.1% of votes in the 2019 general election, and each of those parties' leaders over the last decade. While the latter topics were a list of 263 hashtags derived from contemporaneous sources reporting the most popular hashtags in each of the general elections and referenda held during the period of study.

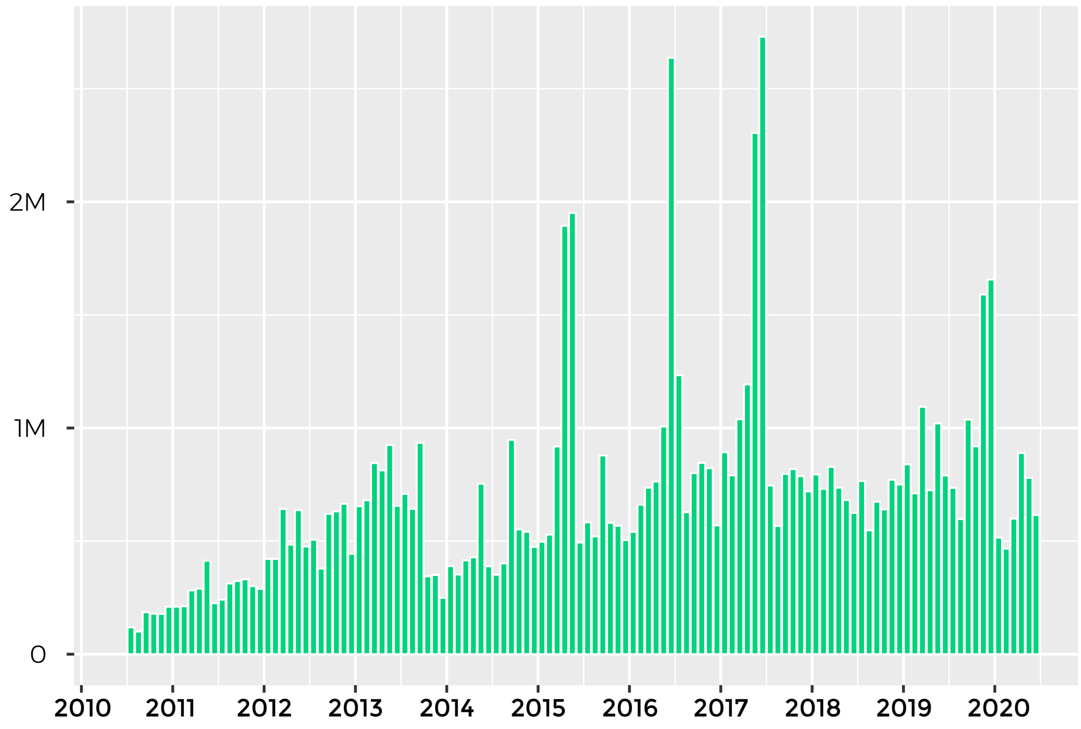


Figure 2: Histogram of tweets collected by political talk criteria, in millions.

This search string returned almost 83 million matches (Figure 2), from which a 5% random sample was downloaded, comprising tweets from 788,231 unique users.

Case selection

To assemble a panel of cases with a history of activity against which structural platform shifts can be examined over time, the sampling frame was therefore limited to users who both:

- had tweets in at least four different years in the political talk results
- registered accounts before July 2010, and have tweeted since June 2020.

Further steps were taken to identify genuine active accounts. The final 1,228 users were randomly sampled from this pool, and every tweet by these accounts from 1 July, 2010 to 30 June, 2020 was downloaded, comprising a final corpus of 22,593,965 tweets.

METHOD

I assess my five hypotheses by applying statistical methods to a matrix of covariates associated with the corpus documents. These variables were acquired from various existing sources and original research, combining computational text analysis with light network analysis. In the interests of time and space, what follows is an executive summary of the methods used. An exhaustive accounting of these methods is provided in Appendices IV–VI.

Computational classification of incivility

Hand-coded training set

To identify incivility at the scale of this study's corpus I use supervised machine learning. This method relies on a "training" set of 5,000 randomly sampled tweets, labelled by two human coders. This was stratified evenly either side of algorithmic introduction (Figure 3). 600 tweets were randomly sampled from equally from six crosscutting strata — before and after algorithmic introduction; across original tweets, retweets and replies — exceeding the recommended sample for intercoder reliability (ICR) for each substrata. A coding scheme for incivility heavily inspired by Theocharis et al. (2020) was implemented (Table 2) for a dichotomous variable: CIVIL or UNCIVIL. The confusion matrix from human coding is presented in Table 3, demonstrating the 88% agreement achieved. Cohen's kappa was calculated at .64, considered "very good" (Regier et al., 2013). 21% were labelled UNCIVIL.

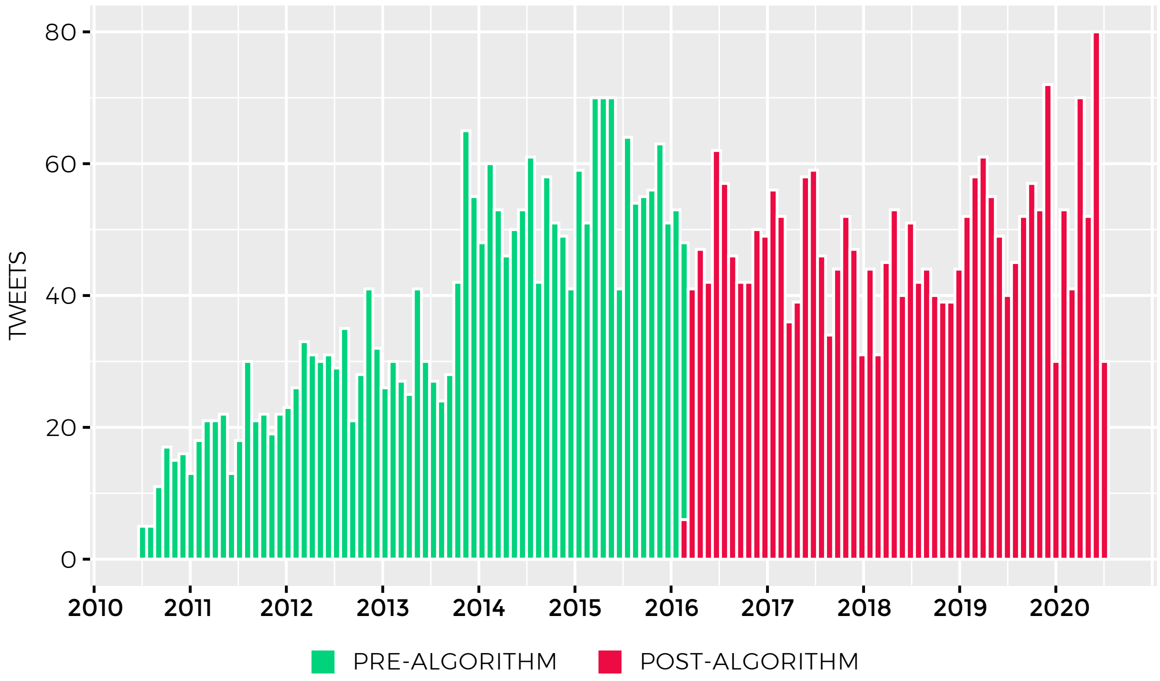


Figure 3: Histogram of tweets in validation sample, by date.

Table 2: Coding scheme used in human labelling, adapted from Theocharis et al. (2020).

EACH TWEET EXCLUSIVELY FALLS INTO ONE OF THE FOLLOWING CATEGORIES:

CIVIL	UNCIVIL
A tweet that adheres to politeness standards, that is, written in a well-mannered and non-offensive way. Even if it is critical of its object, it does so in a respectful way. Provides evidence or makes reference to supporting information when making assertions; otherwise is clear of its conjectural or subjective nature, without precluding the possibility of reasonable disagreement.	An ill-mannered, disrespectful tweet that may contain offensive language. This includes threatening one's rights (freedom to speak, life preferences); assigning stereotypes or hate speech; name-calling ("moron", "idiot"); casting aspersion ("liar", "traitor"); pejorative speak; wilful misrepresentation or distortion; vulgarity; sarcasm; ridicule; ALL CAPS; incendiary, obscene, and/or humiliating language.
EXAMPLES	
"I don't doubt your good intentions, I question your priorities. We'll have to agree to disagree." *"Thank you [@username] for your message."* *"Very apt for the times we are living in."*	*"Even you scum must see that this can only be solved on an international scale, you nasty group off selfish pricks."* *"You are a FUCKING disgrace."* *"Every one of them was a moron."*

Supervised learning

The training set was synthetically balanced using Google's Perspective API, increasing its size to 8,050 tweets.[3] Stemmed unigrams were then passed through a logistic "lasso" regression, using five-fold cross-validation, to build the classifier.

Table 3: Confusion matrices for intercoder reliability (ICR) and classifier validity.

HUMAN CODED

		CODER 2	
		CIVIL	**UNCIVIL**
CODER 1	**CIVIL**	435	59
	UNCIVIL	14	92

% AGREEMENT = 87.8

COHEN'S KAPPA = .64

SUPERVISED CLASSIFIER

		PREDICTIONS	
		CIVIL	**UNCIVIL**
HUMAN CODED	**CIVIL**	805	39
	UNCIVIL	257	509

% ACCURACY = 81.6

F1 = .84 (CIVIL) / .78 (UNCIVIL)

Validation

1,610 of the hand-coded tweets — 20% of the expanded training set — were withheld at random from the estimation to appraise the classifier's performance. Total accuracy was 82%, with precision and recall for the CIVIL class at 76% and 95%, while for UNCIVIL they were 93% and 67% respectively. This is a level of classification similar to previously published studies utilising the same process (Davidson et al., 2017; Gervais & Chin, 2018; Theocharis et al., 2016, 2020) and approaches human interrater agreement. If it can be criticised, it is for erring on the conservative side in *under*estimating the extent of incivility.

[3] See Appendix IV for details. This, along with the preceding steps for the supervised method, closely follows Theocharis et al. (2020) — I am indebted to them and to Blake Miller for suggesting their research.

To further ensure this process captured the desired concept, unigrams the model found most predictive of civility and incivility are presented in Table 4. The array of insults and profane language under incivility contrasted with “love”, “outstanding”, “appreciate” and the heart emoji confirms the classifier identified my dimension of interest.

Table 4: The most predictive unigrams for civil and uncivil tweets.

CLASSIFICATION	PREDICTIVE UNIGRAMS
UNCIVIL	fuck, shit, stupid, embarrass, idiot, ars, piss, racist, pathet, shite, cock, disgust, moron, bullshit, useless, liar, vile, arrog, betray, cunt, dick, familiar, twat, fuckin, privileg, fart, savil, dickhead, fascist, batshit, clown, hancock, tit, smug, scum, sicken, shitti, wank, #savil, gobshit, cock-up, snowflak, cockroach, scumbag, clusterfuck, awkward, shitpea, #fuckoffroy, anti-racist, aint, #pathet, explan, striker, missil, shambl, arsehol, 🤔, nadin, #libdem, fat, ethnic, dipshit, anti-fascist, wanker, ken, destruct, woke, elit, lie, cunti, hezbollah, fucker, bunch, bum, barnet, stupidest, #tori, irrelev, immigr, brexshitt
CIVIL	patriarchi, froth, hike, gregg, mansfield, no-on, pic, denier, saw, petit, reproduct, magnitud, paul, moment, b, solar, sunshin, appreci, potato, 💚, provid, diet, 😘, 7, love, outstand, first, close, mar, chariti, hous, research, fantast, array, #springwatch, scenario, via, includ, agent, novel, ht, garden, @, scene, announc, follow, latest, school, beauti, thank, extens, local, ohh, #london, wild, justic, birthday, pls, arm, chat, invest, excit, event, model, heart, comic, strong, #heartnew, wind, necessari, shall, toilet, chang, construct, common, vinc, congrat, recommend, comfort, awesom, soldier

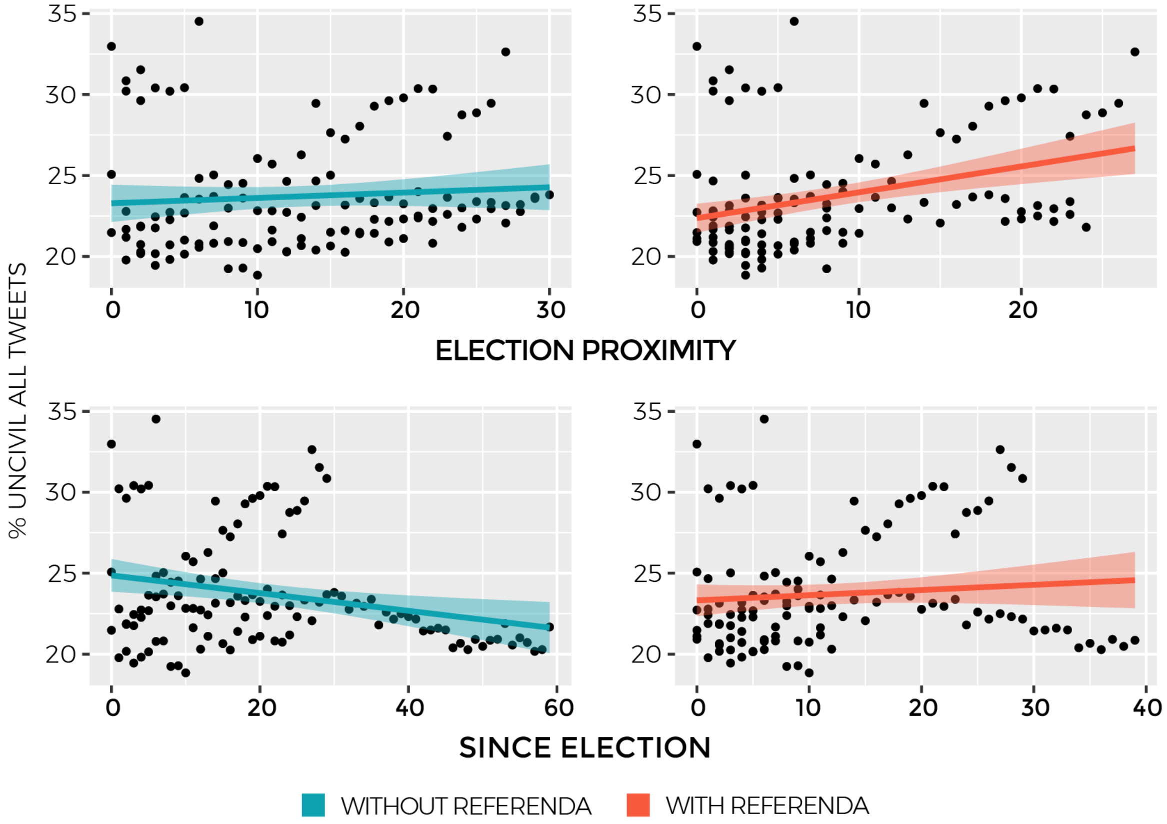
Construction of control variables

Theoretically, the explanatory factor I am interested in is the Twitter timeline algorithm, however this cannot be observed directly, and so time before and after its introduction operates as my independent variable. To make a compelling case that any temporal correlation with the dependent variable is genuine rather than spurious, I need a comprehensive accounting of other plausible covariates. Regression analysis can achieve this through control variables. If any relationship between time and incivility is robust to the inclusion of convincing measures of other potentially related factors, then I will have a persuasive case. Below I specify such measures.

Political atmosphere

To control for the heightened UK political atmosphere after Brexit (Evans & Schaffner, 2019; Hobolt et al., 2020; Meredith & Richardson, 2019; The UK in a Changing Europe, 2019), I provide three possible instruments.

The first two use the electoral schedule as a proxy. ELECTION PROXIMITY is the smallest of either the amount of time since the most recent election, or the amount of time until the next election; while the SINCE ELECTION variable increases linearly and resets to zero when an election takes place. Both were calculated with and without referenda.



NOTE: ELECTION PROXIMITY = in any given month, the smallest of either: months since last/until next election.

Figure 4: Proxy variables for political atmosphere, in months, against percentage incivility.

The expectation, consistent with Theocharis et al. (2020) is these variables will be negatively correlated with incivility, with moments of high political tension (and thus a low ELECTION value) related to more uncivil tweets, but as Figure 4 shows, this only holds in one of the four cases. This is discussed in more depth in §6; for now, another measure is needed.

The BBC PARLIAMENT variable tracks weekly viewing figures for the channel. They show a significant increase since algorithmic introduction and the Brexit crisis. Crucially, as Figure 5 shows, they exhibit the expected correlation between rising political tension and increased incivility. This makes BBC PARLIAMENT a suitable proxy to statistically control for the heightened political atmosphere since 2016.

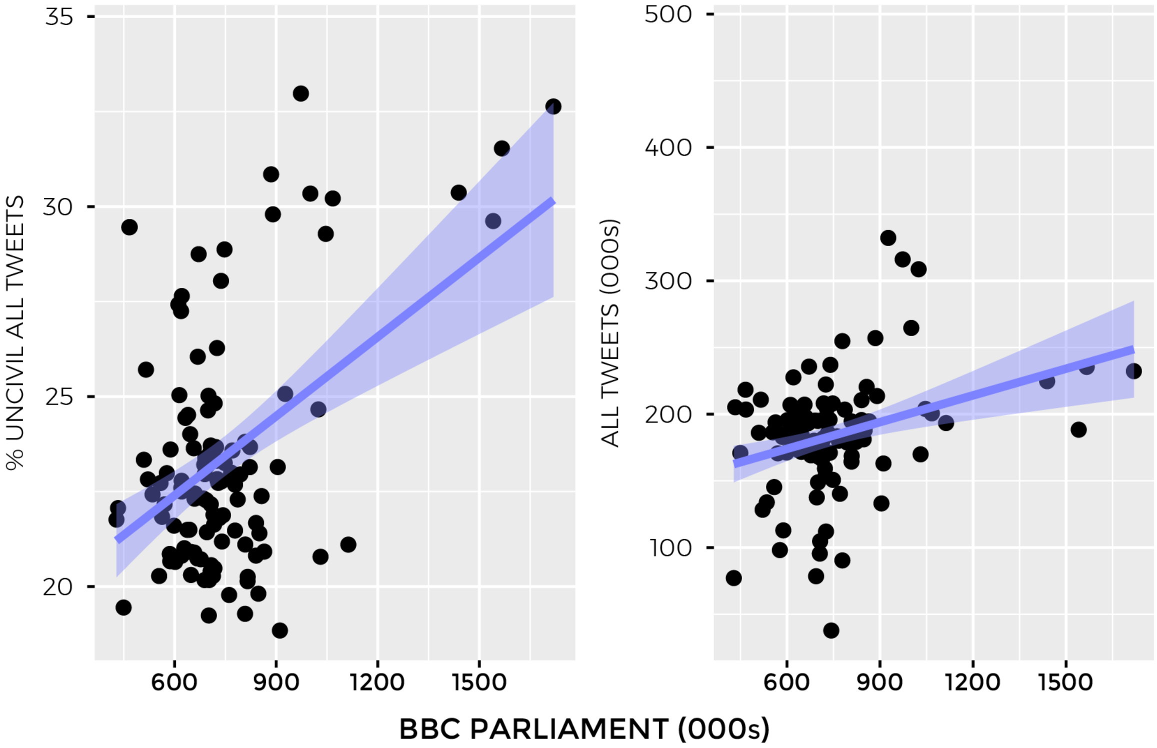


Figure 5: Monthly BBC Parliament viewing figures against % uncivil and total tweets.

Partisanship and media diet

I estimate the partisanship and ideological media diet of participants on the basis of the accounts they follow in a simplified method to Barberá et al. (2015). User follow lists were compared to a register of 972 former and current politicians, and an additional 128 offical party and affiliated accounts. Each were assigned a left–right score, derived from the Chapel Hill Expert Survey (Bakker et al., 2020). A composite score for each user based on which of these accounts they follow was then calculated as the PARTISAN SCALE variable. To check its validity, a stratified sample was drawn from users in the sampling frame with one of the political parties' names in their bio, taken as a declaration of support. Boxplots of these results by supported party are presented in Figure 6, where the positions of users adhere quite closely to those of their respective parties, particularly relative to others.

A similar process was followed for the MEDIA SCALE, based on a comprehensive list of 406 news media accounts from 182 organisations. Scores analogous to the PARTISAN SCALE were taken from several established media bias databases, and validated the same way, with similarly satisfactory results (Figure 6).

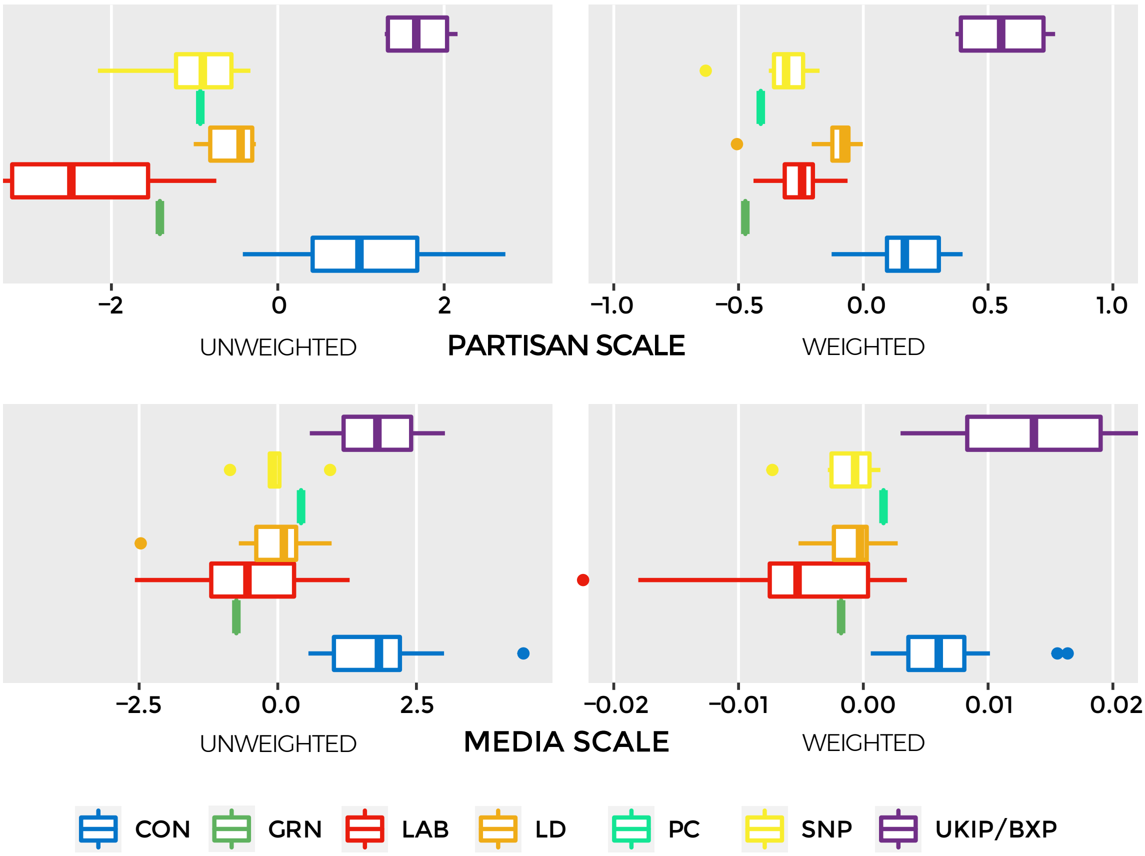


Figure 6: Boxplots of partisan and media scales, by inferred party support.

Scatterplots and generalised additive model (GAM) trendlines in Figure 7 confirm the non-linear, U-like relationship between these scales and incivility that theory dictates.

Political engagement

To account for levels of political engagement independent of ideological valence, the PARTISANS FOLLOWED and MEDIA FOLLOWED variables are simply the absolute number of accounts followed that were used to infer the previous scales. NON-PARTISANS FOLLOWED is the number of neutral political accounts followed (e.g. official governmental and parliamentary accounts, local councils, etc.). It should be noted that while users following fewer than two partisan or media accounts were rejected from the sample; no such minimum was placed on non-partisan accounts.

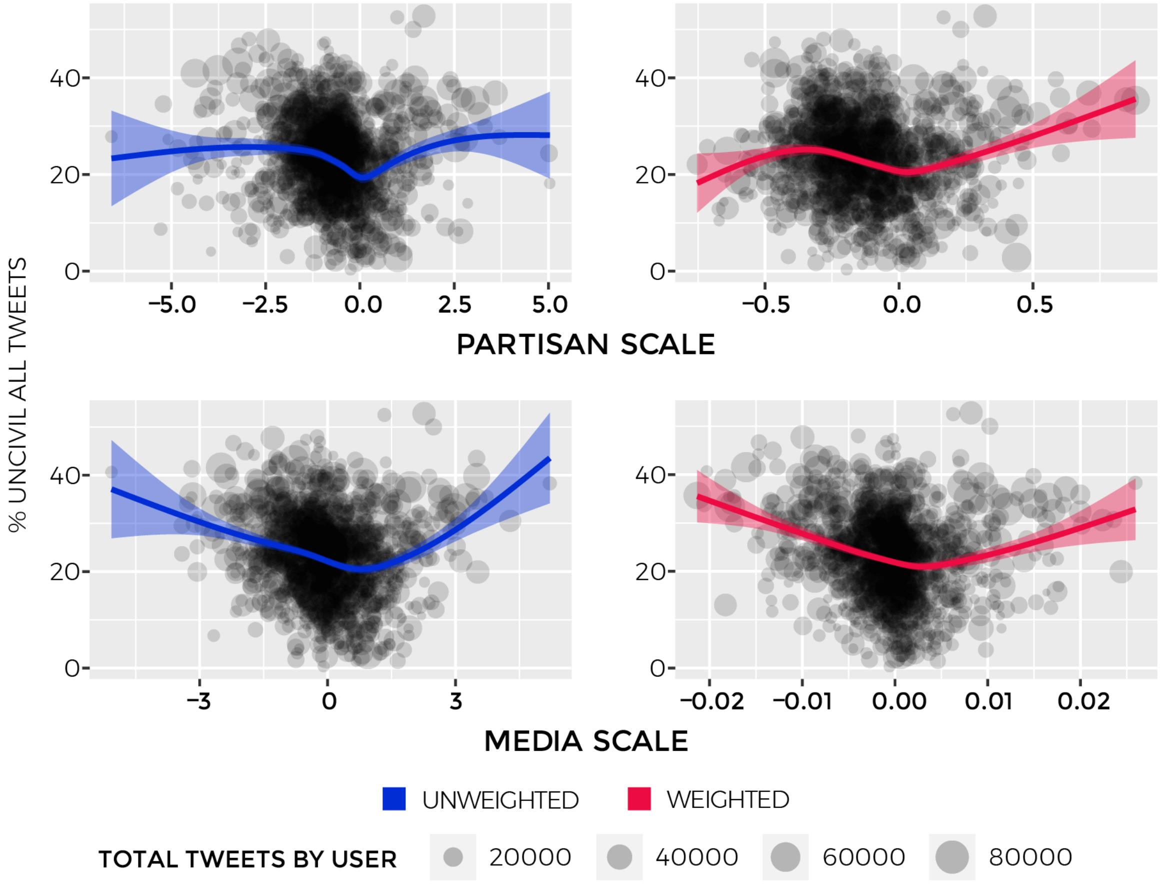


Figure 7: User partisan and media scores against their percentage of uncivil tweets.

Twitter usage

Finally, level of Twitter usage is controlled through a user's absolute number of FOLLOWERS (at time of sampling) and total number of in-sample tweets (ALL TWEETS) each month.

Calculation of tweet type

This analysis does not rely solely on time as a measure of the hypothesised algorithmic phenomenon (q.v. §3.3). I propose certain categories within a typology of tweet type can also be used as a proxy for algorithmic influence on Twitter. Below I detail their calculation.

Main sample

Retweets were divided into FOLLOWING RETWEETS (FRTS), when the retweeter follows the original tweet author; and NOT-FOLLOWING RETWEETS (NFRTS), when they do not. Their distributions are compared to other tweet types in the first column of Table 5.

Table 5: Tweet type distribution compared across three samples of analysis.

TWEET TYPE	MAIN SAMPLE*		SUBSAMPLE		REAL-TIME SAMPLE	
	n	%	*n*	%	*n*	%
TWEETS	5,162,891	31.2	43,826	31.4	852	10
REPLIES	4,909,679	29.6	41,638	29.8	3,958	46.7
RTS	6,499,899	39.2	54,271	38.8	3,670	43.3
FRTS	3,461,958	20.9	29,452	21.1	1,871	21.1
NFRTS	3,037,941	18.3	24,819	17.8	1799	21.2
INRTS			16,591	11.9	1,341	15.8
ENRTS			8,228	5.9	458	5.4
N	**16,572,469**		**139,735**		**8,480**	

NOTE: Percentages represent tweet type as a percentage of the whole sample; may not sum to 100 due to rounding. *Tweet type not reliably available before November 2013, tallies exclude tweets from this period. RTS wholly comprise FRTS and NFRTS; NFRTS wholly comprise INRTS and ENRTS. Subsample tweets drawn from main sample; real-time sample users drawn from same sampling frame as main and subsample.

This is an imperfect measure. Three control measures of individual follow behaviour: FOLLOWING GROWTH, FOLLOWING GROWTH (*SE*), and FOLLOWING VOLATILITY, account for known measurement error.

Subsample

A further retweet type distinction was made *within* NFRTs: between INTRA-NETWORK RETWEETS (INRTs), where the retweeting user does not follow the author of the original tweet, but does follow someone who has retweeted it; and EXTRA-NETWORK RETWEETS (ENRTs), where the retweeting user follows neither. The distinctions between these tweet types are illustrated in Figure 8. Calculating these for the main sample was unrealistic with the available resources, and so a representative subsample of tweets (N = 139,735) were randomly selected, including 24,819 NFRTs. 33.2% of the sampled retweets were labelled ENRTs, while the rest were designated INRTs. Their distributions are compared to other tweet types in the second column of Table 5.

1.1.1 *Real-time sample*

As a further robustness check of the tweet type classification, a secondary data collection was conducted on a new random sample of 200 users. These tweets were collected in real-time between 4–11 August, 2020, and checked against follow lists that were updated daily, providing a far more precise measure. The tweet type distribution is compared to the other samples in Table 5, to which they are markedly different, but consistent with the trend over time for each type, as shown later in Figure 11 (§6).

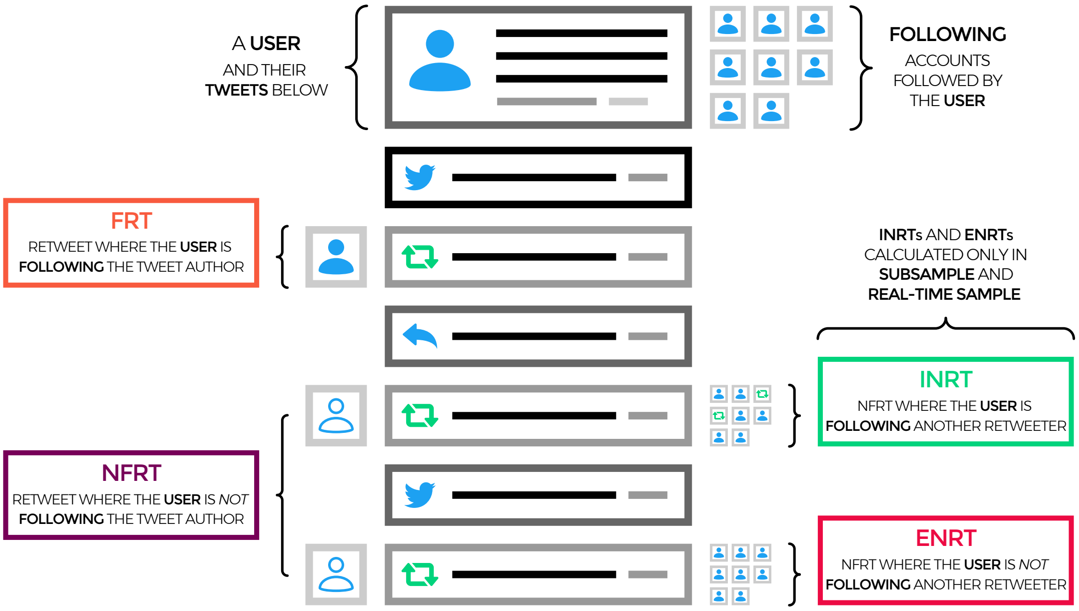

A USER
AND THEIR
TWEETS BELOW
FOLLOWING
ACCOUNTS
FOLLOWED BY
THE USER
FRT
RETWEET WHERE THE USER IS
FOLLOWING THE TWEET AUTHOR
INRTs AND ENRTs
CALCULATED ONLY IN
SUBSAMPLE AND
REAL-TIME SAMPLE
INRT
NFRT WHERE THE USER IS
FOLLOWING ANOTHER RETWEETER
NFRT
RETWEET WHERE THE USER IS NOT
FOLLOWING THE TWEET AUTHOR
ENRT
NFRT WHERE THE USER IS NOT
FOLLOWING ANOTHER RETWEETER

Figure 8: Diagram of the different retweet types identified in the analysis.

Ethics and reflexivity

All social research raises ethical issues and for ethical practice to be achieved, reflection needs to be ongoing throughout the research process (Guillemin & Gillam, 2004). Social media data pose particular questions over privacy (Zimmer & Proferes, 2014), as even though only a small portion of Twitter users protect their profiles (Liu et al., 2014), public users do not explicitly consent to being observed for research. To protect users, all retained data has been anonymised, consistent with other research (Bruns & Weller, 2014), only aggregate level data is published, and any illustrative examples of tweets have been altered to not be uniquely identifiable through Twitter search. All research was undertaken in concordance with LSE's Research Ethics Policy.

As a matter of reflexivity, I cannot claim to be a solely dispassionate observer in this study. Both my own personal experience with these platforms, and my professional knowledge as a digital strategist, provide strong anecdotal evidence for the phenomena I propose and seek to test. In particular, I have a background in the methods detailed by Bradshaw (2019) — such as search engine optimisation (SEO) — used to influence algorithms.

I have taken every effort, to the best of my ability, to ensure this study is falsifiable, robust and consistent with research best practices as exemplified in similar literature. I hope this transparency lends it further credibility.

RESULTS AND ANALYSIS

Hypothesis 1: Incivility, time and the algorithm

To test **H1**, I first aggregate the dataset into a monthly time series and estimate three basic ordinary least squares regressions of the total percentage of uncivil tweets[4] on time and algorithmic introduction (Table 6). In Model 1 the TIME (t) trend alone has a significant positive association ($\beta = .06$, $p < .01$) which is maintained with Model 2's inclusion of the ALGORITHM dummy, and this itself becomes significant when its interaction with t is

[4] "Uncivil tweets", "tweets containing incivility", and simply "incivility" are used interchangeably from here on.

introduced in Model 3. Here the *t* effect becomes negative ($\beta = -.05$, $p < .01$), as the interaction term ($\beta = .31$, $p < .01$) models the interruption in *t*'s association with incivility, illustrating that the percentage of incivility had been decreasing over time until early 2016 when the trend reverses, equivalent to a 3.12 p.p. increase in incivility each year.

This represents the hypothesised relationship between algorithmic introduction and increasing incivility, however, to be confident this relationship is not spurious, I need to show it is robust to individual variation and other plausible covariates. To do so, I first disaggregate the time series by individual, returning it to a panel structure (Table 7). With this dataset, I use the `plm` package in `R` (Croissant et al., 2020) to estimate several fixed effects regressions that control for variation across individuals (Table 9). Models 4 and 5 are the equivalent of the OLS Models 2 and 3, and they show a similarly significant relationship, albeit with the net effect of the *t* × ALGORITHM term slightly reduced ($\beta = .23$, $p < .01$), now equivalent to +2.04 p.p. each year.

Table 6: OLS regression models of incivility on time and algorithmic introduction.

	DEPENDENT VARIABLE		
	% UNCIVIL ALL TWEETS		
COVARIATE	1	2	3
TIME (*t*)	.06*** (0.01)	.05*** (.01)	-.05*** (.01)
ALGORITHM		.85 (1.02)	-22.04*** (1.01)
***t* × ALGORITHM**			.31*** (.01)
CONSTANT	20.17*** (.52)	20.43*** (.61)	23.90*** (.28)
OBSERVATIONS (*N*)	120	120	120
***R*^{2}**	.34	.35	.90
ADJUSTED *R*^{2}	.34	.34	.89
RESIDUAL *SE*	2.83 (df = 118)	2.83 (df = 117)	1.13 (df = 116)
***F*-STATISTIC**	62*** (DF = 1; 118)	31*** (DF = 2; 117)	336*** (DF = 3; 116)

NOTE: Standard errors in parentheses. Time in months. ALL TWEETS include retweets and replies. ALGORITHM dummy coded '1' after January 2016.

* $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$

To control for the intensifying political atmosphere over the period of analysis, I next include three variables in sequence. The ELECTION PROXIMITY term (Model 6) is significant (*p* < .01) but with an opposite to hypothesised effect; incivility increases by .03 p.p. per month *further* from an election. This remains unchanged whether or not referenda are included, and might be explained by an increase in positive, get-out-the-vote messaging by supporters during the short campaign, in comparison to more vitriolic posts in the middle of a parliament. The SINCE ELECTION term in Model 7 is insignificant both with and without referenda, indicating incivility is no more intense in the wake of an election than in the "quiet" periods predicted by Theocharis et al. (2020). Finally, BBC PARLIAMENT viewing figures, included in Model 8, are significant (*β* = .002, *p* < .01), with a five *SD*s increase (occurring nine times since algorithmic introduction; just once before) associated with +2 p.p. incivility. As established in §5.2.1, this is a suitable proxy for the heightened political environment since 2016, and while it is strongly and positively associated with incivility, the *t* × ALGORITHM term is robust to its inclusion, only suffering a slight reduction in effect to an annual 1.92 p.p. increase in incivility.

Next I switch to a random effects implementation which allows the inclusion of time-invariant individual characteristics research (q.v. Appendix V) suggests are associated with increased incivility.[5] The PARTISAN and MEDIA SCALE terms account for the left–right ideological orientation of individual users. They are included in Model 9 along with their squared terms, as their relationship with incivility is quadratic (Figure 7, §5.2.2). Both PARTISAN SCALE^2 (*β* = 7.64, *p* < .05) and MEDIA SCALE^2 (*β* = 19,613, *p* < .01) are significant.[6] A followed partisan political network 3 *SD*s left-of-centre is associated with +1.83 p.p. incivility, whereas a network the same distance to the right is associated with +4.39 p.p. incivility. This asymmetry is reversed for followed news media, where a network 3 *SD*s left-of-centre is associated with +9 p.p. incivility, compared to +2.79 p.p. for an equal rightward composition. This evidence that partisanship and media diet are associated with incivility is consistent with theory and previous findings.

[5] The ALL TWEETS measure varies over time as well as units.

[6] Both the weighted and non-weighted terms were significant with similar effects. The weighted version is used from here as its estimation is more conservative.

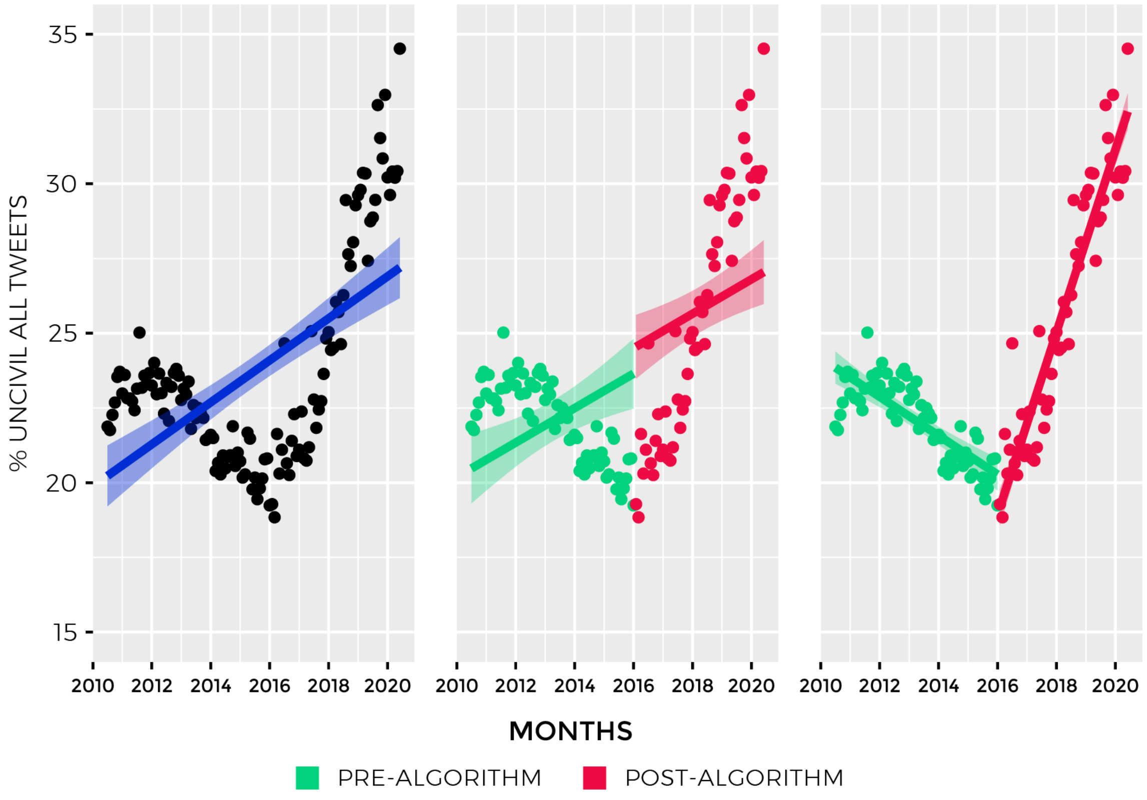


Figure 9: Aggregate incivility over time, with general trendline (left), ALGORITHM dummy trendlines (centre), and *t* × ALGORITHM trendlines (right).

Table 7: Illustration of panel data structure with correct totals.

i	*t*	**ALL TWEETS**	**UNCIVIL TWEETS**	**% UNCIVIL**	**…**
USER 1	MONTH 1	350	92	26.3	…
USER 1	MONTH 2	300	76	25.3	…
USER 1	MONTH 3	552	118	21.4	…
…	…	…	…	…	…
USER 2	MONTH 1	330	84	25.5	…
USER 2	MONTH 2	184	66	35.9	…
USER 2	MONTH 3	218	83	38.1	…
…	…	…	…	…	…
***n* = 1,228**	***T* = 120**	**22,593,965**	**5,470,714**	**24.2**	**…**

***N* = 147,360**

The next three variables (Model 10) estimate an individual's level of political engagement. PARTISANS FOLLOWED is significant with a positive effect (β = .02, $p < .01$), while MEDIA FOLLOWED is insignificant. NON-PARTISANS FOLLOWED has a significant negative effect (β = -.35, $p < .01$), with a *SD* increase associated with -2.79 p.p. incivility. This perhaps indicates an institutional, in contrast to partisan, interest in politics dampens incivility.

Finally, the last two control variables (Model 11) capture the level of Twitter usage. Both are significant, with FOLLOWERS negatively (β = -.001, $p < .01$) and ALL TWEETS positively (β = .004, $p < .01$) associated with the dependent variable. The latter finding, consistent with previous evidence, shows heavier social media usage is correlated with higher incivility.

Crucially, the significance and effect of the t × ALGORITHM term remains unchanged despite the inclusion and (in most cases) statistical significance of each control variable. Models 9–11 have progressively higher *F*-statistics, indicating the added variables each improve the model's fit without reducing the explanatory power of the proposed algorithm–incivility relationship. On this basis, controlling for changes in political atmosphere, along with partisanship, news media diet, political engagement, Twitter usage and unobserved individual variation, I find support for **H1**: in Model 12, where the ALGORITHM dummy term is included without interaction, its β = 1 ($p < .01$), thus the mean proportion of tweets containing incivility is 1 p.p. larger after the algorithm's introduction (**H1a**); in Model 11, the interaction term has a larger positive effect than t alone, which is actually negative, ($\beta_{t \times \text{ALGORITHM}}$ = .22, $p < .01$; β_t = -.06, $p < .01$), thus the 52 months since algorithmic introduction are associated with an 8.53 p.p. increase in incivility, or a 41.9% increase over the pre-algorithm average, the period of which is associated with a 4.08 p.p. decrease (**H1b**).

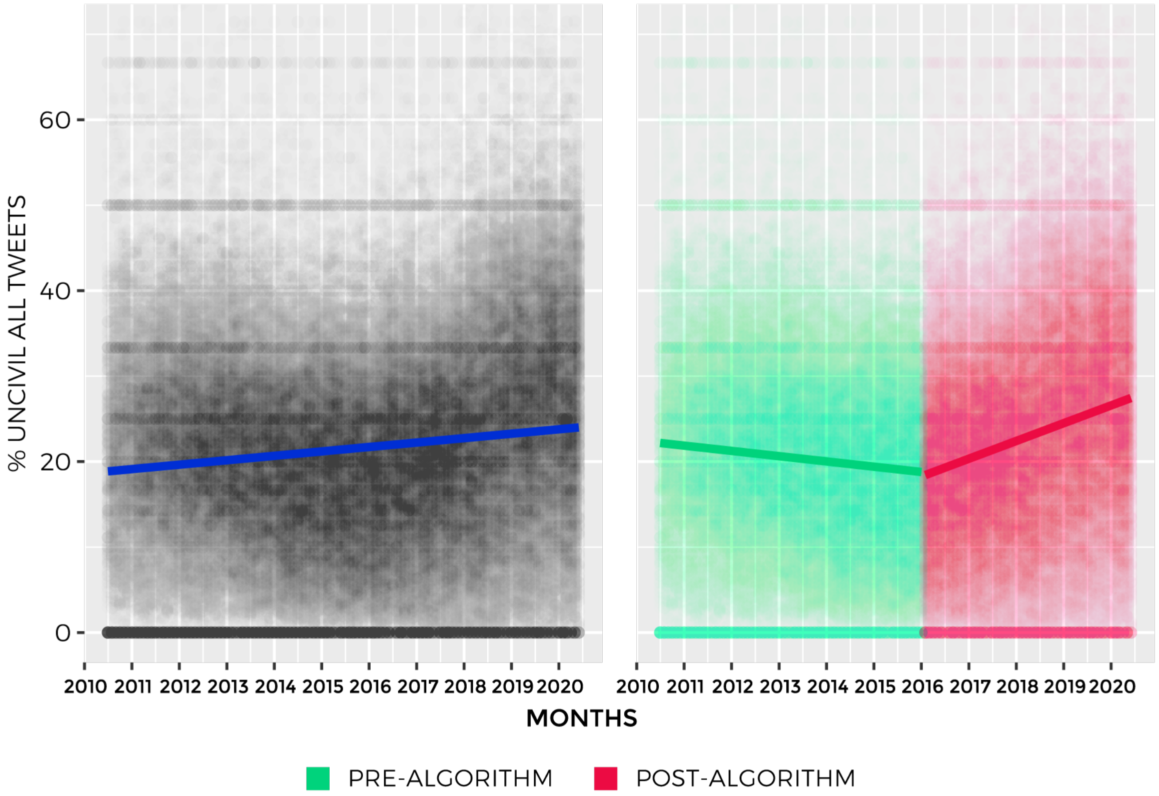


Figure 10: Individuals' incivility over time, with general trendline (left) and *t* × ALGORITHM trendlines (right).

Hypothesis 2: Tweet type, time and the algorithm

For **H2** I turn to tweet type as different measures for the hypothesised algorithmic phenomenon, beginning with NOT-FOLLOWING RETWEETS (NFRTs). I regress each user's monthly percentage of NFRTs on time, algorithmic introduction, and measures related to variable Twitter usage across users (Table 10). Controlling for Twitter usage, follow behaviour and unobserved individual variation, I find initial, partial support for **H2**: in Model 13, where the ALGORITHM dummy term is included without interaction, its $\beta = 1.53$ ($p < .01$), thus the mean NFRTs is 1.53 p.p. larger after the algorithm's introduction (**H2a**); however in Model 14, the interaction term is negative ($\beta = -.062$, $p < .01$), and while the net effect with *t* ($\beta = .091$, $p < .01$) is still positive, this is evidence the rate of increase in NFRTs has *reduced* since algorithmic introduction, contrary to **H2b**.

Table 8: Crosstabs of main sample incivility, by tweet type and algorithmic introduction.

TWEET TYPE	**PRE-ALGORITHM**			**POST-ALGORITHM**			**COMBINED**		
	UNCIVIL	**%**	***n***	**UNCIVIL**	**%**	***n***	**UNCIVIL**	**%**	***n***
TWEETS*	1,386,169	23	6,021,496				1,386,169	23	6,021,496
TWEETS	406,473	21.3	1,906,980	763,132	23.4	3,255,911	1,169,605	22.7	5,162,891
REPLIES	370,567	23.6	1,572,905	968,861	29	3,336,774	1,339,428	27.3	4,909,679
RTs	247,330	16.9	1,466,785	132,8182	26.4	5,033,114	1,575,512	24.2	6,499,899
FRTs	118,787	15.3	777,727	662,183	24.7	2,684,231	780,970	22.6	3,461,958
NFRTs	128,543	18.7	689,058	665,999	28.4	2,348,883	794,542	26.2	3,037,941
n	**2,410,539**	**22**	**10,968,166**	**3,060,175**	**26.3**	**11,625,799**	**5,470,714**	**24.2**	**22,593,965**

NOTE: Percentages represent percentage of incivility within tweet type; may not sum to 100 due to rounding. *Tweet type not reliably available before November 2013, top row is an aggregation for that period. RTs wholly comprise FRTs and NFRTs.

Table 9: Panel regression models with fixed (4–8) and random effects (9–12) of incivility on time, algorithmic introduction, political atmosphere, and individual characteristics.

	DEPENDENT VARIABLE								
	% UNCIVIL ALL TWEETS								
COVARIATE	4	5	6	7	8	9	10	11	12
CONSTANT						20.42***	21.79***	22.62***	19.04***
						(.35)	(.40)	(.40)	(.40)
TIME (*t*)	.03***	-.06***	-.05***	-.06***	-.06***	-.06***	-.06***	-.06***	.01***
	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)
ALGORITHM	.51***	-15.67***	-15.43***	-15.65***	-15.51***	-15.52***	-15.52***	-15.40***	1.00***
	(.12)	(.29)	(.29)	(.29)	(.33)	(.33)	(.33)	(.33)	(.12)
***t* × ALGORITHM**		.23***	.22***	.23***	.22***	.22***	.22***	.22***	
		(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	
POLITICAL ATMOSPHERE (TIME-VARIANT)									
ELECTION PROXIMITY			.03***						
			(.004)						
SINCE ELECTION				.003					
				(.002)					
BBC PARLIAMENT					.002***	.002***	.002***	.002***	.003***
					(.000)	(.000)	(.000)	(.000)	(.000)
PARTISANSHIP, MEDIA DIET, POLITICAL ENGAGMENT AND TWITTER USAGE (UNIT-VARIANT)									
PARTISAN SCALE						1.97	.71	.57	.59
						(1.49)	(1.43)	(1.35)	(1.35)
PARTISAN SCALE^2						7.64**	5.95*	4.23	4.13
						(3.25)	(3.14)	(2.96)	(2.96)
MEDIA SCALE						-180.9***	-47.14	-46.19	-47.49
						(50.59)	(50.27)	(47.36)	(47.39)
MEDIA SCALE^2						19,613***	18,585 ***	18,247***	18,195***
						(3,632)	(3,509)	(3,306)	(3,308)
PARTISANS FOLLOWED							.02***	.02***	.02***
							(.01)	(.01)	(.01)
MEDIA FOLLOWED							-.01	-.01	-.01
							(.01)	(.01)	(.01)
NON-PARTISANS FOLLOWED							-.35***	-.28***	-.28***
							(.03)	(.03)	(.03)
FOLLOWERS								-.001***	-.001***
								(.000)	(.000)
ALL TWEETS								.004***	.004***
								(.000)	(.000)
INDIVIDUALS (*n*)	1,228	1,228	1,228	1,228	1,228	1,228	1,228	1,228	1,228
OBSERVATIONS (*N*)	129,452	129,452	129,452	129,452	123,418	123,418	123,418	123,418	123,418
***R*^{2}**	.01	.04	.04	.04	.04	.04	.04	.05	.02
ADJUSTED *R*^{2}	.004	.03	.03	.03	.03	.04	.04	.05	.02
***F*-STATISTIC**	898***	1,871***	1,422***	1,404***	1,154***	4,871***	5,016***	5,808***	2,808***

NOTE: Standard errors in parentheses. Time in months. ALL TWEETS include retweets and replies. ALGORITHM dummy coded '1' after January 2016.

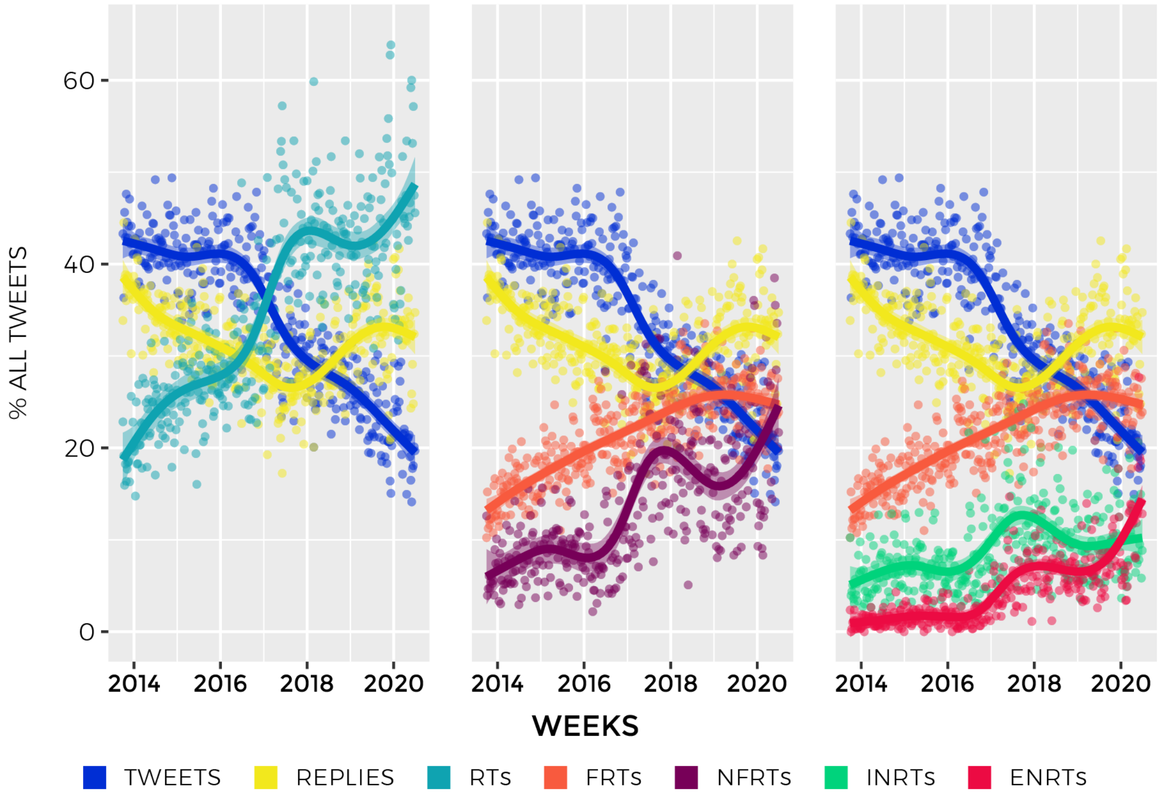
* **$p < 0.10$** ** **$p < 0.05$** *** **$p < 0.01$**

Table 10: Panel random effects (13–14) and binomial mixed effects (15–19) regression models of tweet type on time, algorithmic introduction and individual Twitter usage.

	DEPENDENT VARIABLE						
	% NFRTs		P(NFRT)		P(ENRT)		P(INRT)
COVARIATE	13	14	15	16	17	18	19
CONSTANT	13*** (.01)	10*** (.652)	-2.43*** (.06)	-2.51*** (.07)	-4.47*** (.07)	-4.48*** (.11)	-2.65*** (.07)
TIME (*t*)	.035*** (.003)	.091*** (.008)	.000*** (.000)	.001*** (.000)	.001*** (.000)	.001*** (.000)	.000*** (.000)
ALGORITHM	1.53*** (.128)	5.19*** (.504)	.28*** (.03)	.38*** (.05)	.21*** (.06)	.23** (.11)	.75*** (.06)
***t* × ALGORITHM**		-.062*** (.008)		-.000*** (.000)		-.000 (.000)	-.001*** (.000)
	TWITTER USAGE (UNIT-VARIANT)						
FOLLOWERS	-.001*** (.000)	-.001*** (.000)	-.000*** (.000)	-.000*** (.000)	-.000*** (.000)	-.000*** (.000)	-.000*** (.000)
ALL TWEETS	.003*** (.000)	.003*** (.000)					
FOLLOWING GROWTH	-.242 (1.38)	-.248 (1.38)	.14 (.16)	.15 (.16)	.19 (.16)	.18 (.16)	.15 (.15)
FOLLOWING GROWTH (*SE*)	59.7 (72.8)	59.4 (72.8)	.76 (5.39)	1.09 (1.86)	.70 (1.99)	2.86 (8.33)	4.46 (7.78)
FOLLOWING VOLATILITY	.005* (.003)	.005* (.003)	-.000 (.000)	-.000 (.000)	-.000 (.000)	-.000 (.000)	-.000 (.000)
FOLLOWING	-.002*** (.001)	-.002*** (.001)	-.000 (.000)	-.000 (.000)	-.000 (.000)	-.000 (.000)	-.000 (.000)
% FRTs	.057*** (.003)	.062*** (.003)					
INDIVIDUALS (*n*)	1,228	1,228	1,228	1,228	1,228	1,228	1,228
OBSERVATIONS (*N*)	93,350	93,350	135,184	135,184	135,184	135,184	135,184
***R*²**	.03	.03					
ADJUSTED *R*²	.03	.03					
PSEUDO *R*²			.28	.29	.32	.32	.22
***F*-STATISTIC**	3,089.18***	3,147.19***					

NOTE: Standard errors in parentheses. Time in months (13–14) and days (15–19). ALGORITHM dummy coded '1' after January 2016. Data points are monthly averages (13–14) and individual tweets (15–19). P(*x*) = probability of *x*.

* $p < 0.10$ ** $p < 0.05$ *** $p < 0.01$



NOTE: RTs wholly comprise FRTs and NFRTs; NFRTs wholly comprise INRTs and ENRTs. Trendlines from generalised additive model (GAM).

Figure 11: Relative volume by tweet type over time.

Table 11: Crosstabs of subsample incivility, by tweet type and algorithmic introduction.

TWEET TYPE	**PRE-ALGORITHM**			**POST-ALGORITHM**			**COMBINED**		
	UNCIVIL	**%**	***n***	**UNCIVIL**	**%**	***n***	**UNCIVIL**	**%**	***n***
TWEETS	3,667	21	17,440	6,225	23.6	26,386	9,892	22.6	43,826
REPLIES	3,316	23.3	14,247	7,877	28.8	27,391	11,193	26.9	41,638
FRTs	1,048	14.8	7,066	5,576	24.9	22,386	6,624	22.5	29,452
INRTs	1,107	17	6,529	2,191	21.8	10,062	3,298	19.9	16,591
ENRTs	386	25.4	1,517	2,393	35.7	6,711	2,779	33.8	8,228
n	**9,524**	**20.3**	**46,953**	**24,262**	**26**	**93,216**	**33,786**	**24.2**	**139,735**

NOTE: Percentages represent percentage of incivility within tweet type; may not sum to 100 due to rounding.

I now shift to the subsample (q.v. §5.3.2). Aggregating this smaller sample to a monthly panel structure would result in more erratic monthly proportions with questionable validity. Instead I move the unit of analysis to individual tweets, using the `lme4` package in `R` (Bates et al., 2020) to estimate a binomial mixed effects model, probabilistically regressing a dichotomous variable — whether or not a tweet is a certain type — on the same explanatory variables as Models 13–14, while controlling for individual variation.[7] I take the probability of a tweet being an NFRT as the first dependent variable in this approach (Models 15–16, Table 10) to check the results from this subsample are consistent with the main sample, which they are with a positive coefficient for the ALGORITHM dummy (β = .28, p < .01, *OR* = 1.32) and negative coefficient for the t × ALGORITHM term (β = -.000, p < .01, *OR* = .999).

Table 12: Mean probability (%) of incivility, by tweet type and algorithmic introduction.

TWEET TYPE	PRE-ALGO.	vs. ENRT	POST-ALGO.	vs. ENRT	Δ ALGO.
TWEETS	19.2	-4.6	22.8	-5.1	+3.7
REPLIES	20.2	-3.5	24.0	-3.9	+3.8
FRTs	16.0	-7.7	19.2	-8.8	+3.2
INRTs	14.2	-9.5	17.1	-10.8	+2.9
ENRTs	23.7		28.0		+4.3

NOTE: Probabilities calculated from fitted values with mean inputs for other covariates. Δ ALGO. = change between pre- and post-algorithm periods.

I next move the dependent variable to the probability of a tweet being an EXTRA-NETWORK RETWEET (ENRT) – notated P(ENRT) from here. Similarly to Models 11–14, I find partial support for **H2**: in Model 17, the ALGORITHM term without interaction is significant and positive (β = .21, p < .01, *OR* = 1.24), equivalent to a 3.54 p.p. increase in P(ENRT) since algorithmic introduction[8] (**H2c**); however in Model 18, the interaction term is insignificant

[7] With the exception of aggregated variables ALL TWEETS and % FRTs which are unavailable at the tweet level.

[8] Calculated as the change in odds of a tweet from an average account being an ENRT between the median dates of the pre- and post-algorithm periods.

(p > .10), providing no evidence that the rate of increase in P(ENRT) has accelerated with the algorithm's introduction, contrary to **H2d**.

Of the follow behaviour measures, FOLLOWING GROWTH, FOLLOWING GROWTH (*SE*), and FOLLOWING VOLATILITY, only the latter term is (marginally) significant (β = .005, p < .10), and only for the main sample, losing its significance for P(ENRT). Meanwhile, the rejections of **H2b** and **H2d** at least provide more confidence in the robustness of the retrospective tweet type measurement, since a systematic error here would manifest as a temporal correlation, i.e. either more or less NFRTs/ENRTs over time as follow lists become less accurate. Not noticing this sooner is an error on my part in the research design that fortunately has little impact for the interpretation of the results.

By running the same regression, this time with P(INRT) as the dependent variable (Model 19, Table 10), I find that while P(INRT) has risen by 45.6% since algorithmic introduction[9], P(ENRT) has seen a proportionally larger increase of 297%, consistent with **H2e**.

Hypothesis 3: Incivility and tweet type

To test **H3** and assess whether ENRTs — as a proxy for algorithmically favoured content — exhibit more incivility on average than other tweet types, I construct another binomial mixed effects model (Model 20, Table 14), regressing the probability of a tweet containing incivility – P(INCIVILITY) – on tweet type, time and algorithmic introduction, along with the control variables for political atmosphere and individual characteristics used in Models 12–19. The results are confirmatory, with all tweet types displaying statistically significant coefficients — negative for FRTs (β = -.22, p < .01, *OR* = -.8) and INRTs (β = -.359, p < .01, *OR* = -.7); positive for replies (β = .066, p < .01, *OR* = 1.07) and ENRTs (β = .272, p < .01, *OR* = 1.31) — relative to the base category of regular tweets. Table 12 displays the mean P(INCIVILITY)

[9] Calculated as the change in odds of a tweet from an average account being an INRT between the median dates of the pre- and post-algorithm periods (β_t = .000, p < .01, *OR* = 1; $\beta_{\text{ALGORITHM}}$ = .75, p < .01, *OR* = 2.11; $\beta_{t \times \text{ALGORITHM}}$ = -.001, p < .01, *OR* = .999).

for each tweet type before and after algorithmic introduction — ENRTS are highest in both cases (**H3**).

To ensure the ENRT measure is robust, I now turn to the real-time sample (q.v. §5.3.3). In Model 21 (Table 14), incivility is regressed on the same variables as Model 20, excluding the temporal dimension, to ensure the association with P(INCIVILITY) holds when ENRTS are more precisely measured, and it does (β = .584, p < .01, *OR* = 1.79) with ENRTS associated with between 4.1 and 11.1 p.p. higher P(INCIVILITY) than other tweet types. This provides confidence that support for **H3** is not due to measurement error.

Table 13: Crosstabs of incivility by tweet type compared across three samples of analysis.

TWEET TYPE	MAIN SAMPLE*			SUBSAMPLE			REAL-TIME SAMPLE		
	UNCIVIL	%	*n*	UNCIVIL	%	*n*	UNCIVIL	%	*n*
TWEETS	1,169,605	22.7	5,162,891	9,892	22.6	43,826	180	21.1	852
REPLIES	1,339,428	27.3	4,909,679	11,193	26.9	41,638	1,268	32	3,958
RTS	1,575,512	24.2	6,499,899	12,701	23.4	54,271	1,242	33.8	3,670
FRTS	780,970	22.6	3,461,958	6,624	22.5	29,452	624	33.4	1,871
NFRTS	794,542	26.2	3,037,941	6,077	24.5	24,819	618	34.4	1799
INRTS				3,298	19.9	16,591	439	32.7	1,341
ENRTS				2,779	33.8	8,228	179	39.1	458
N	**4,084,545**	**24.7**	**16,572,469**	**33,786**	**24.2**	**139,735**	**2,690**	**31.7**	**8,480**

NOTE: Percentages represent percentage of incivility within tweet type; may not sum to 100 due to rounding. *Tweet type not reliably available before November 2013, tallies exclude tweets from this period. RTS wholly comprise FRTS and NFRTS; NFRTS wholly comprise INRTS and ENRTS. Subsample tweets drawn from main sample; real-time sample users drawn from same sampling frame as main and subsample.

Hypothesis 4: Tweet type incivility, time and the algorithm

To test whether this association has changed over time (**H4**), in Models 22–23 I regress the probability in the subsample of an ENRT containing incivility — P(INCIVILITY | ENRT) — on time, algorithmic introduction and controls. The results are again confirmatory: in Model 22, where the ALGORITHM term is included alone, its β = .338 (p < .01, *OR* = 1.4), equivalent to a mean 7.9 p.p. increase in P(INCIVILITY | ENRT) after the algorithm's introduction (**H4a**); in Model 23, t is introduced, and while its coefficient is insignificant (p > .10), the interaction term is significantly positive ($\beta = .001$, $p < .05$, $OR = 1$), equivalent to an 18.8 p.p. increase in P(INCIVILITY | ENRT) over the interquartile range of the post-algorithm period, compared to a .01 p.p. *decrease* over the same time range in the pre-algorithm period (**H4b**). This is evidence that incivility in EXTRA-NETWORK RETWEETS is substantially increasing over time since the timeline algorithm was introduced, whereas it was effectively static before.

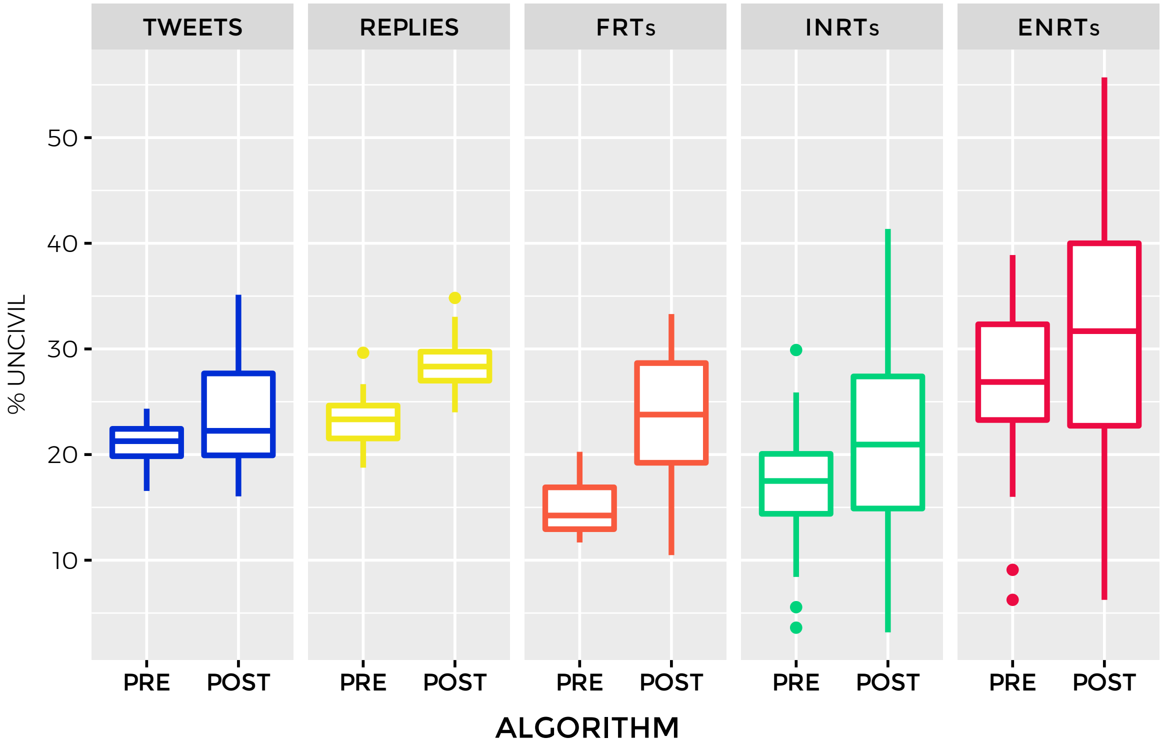


Figure 12: Boxplot of monthly incivility, by tweet type and algorithmic introduction.

Hypothesis 5: Incivility change & lagged tweet type incivility

Finally, in Models 24–29 (Table 15) I return to the panel data from the main sample to regress the change in percentage of incivility in a user's original tweets (i.e. not retweets) from the previous month on the lagged percentage of incivility in their retweets in each of the previous 12 months, as well as the lagged dependent variable as a control. For the change in regular tweet incivility (Model 24), the percentage of uncivil NFRTs from all 12 preceding months are significant predictors, with a net β of .227; for change in reply incivility (Model 26), nine months are significant, with a net β of .148. This compares to 10 and nine respectively for uncivil FRTs (Models 25 and 27), and net βs of .204 and .18. The difference in R^2 – .01 in both cases — is negligible. This presents mixed support for **H5**. NFRTs are an imperfect variable here. As I've shown, ENRTs are a more reliable measure, however I was not able to calculate a large enough sample to aggregate a panel structure that would allow this lagged design. This represents a potentially fruitful avenue for future research with greater resources, especially if a large and long enough sample can be collected in real-time, as I've shown on a small scale in Model 21.

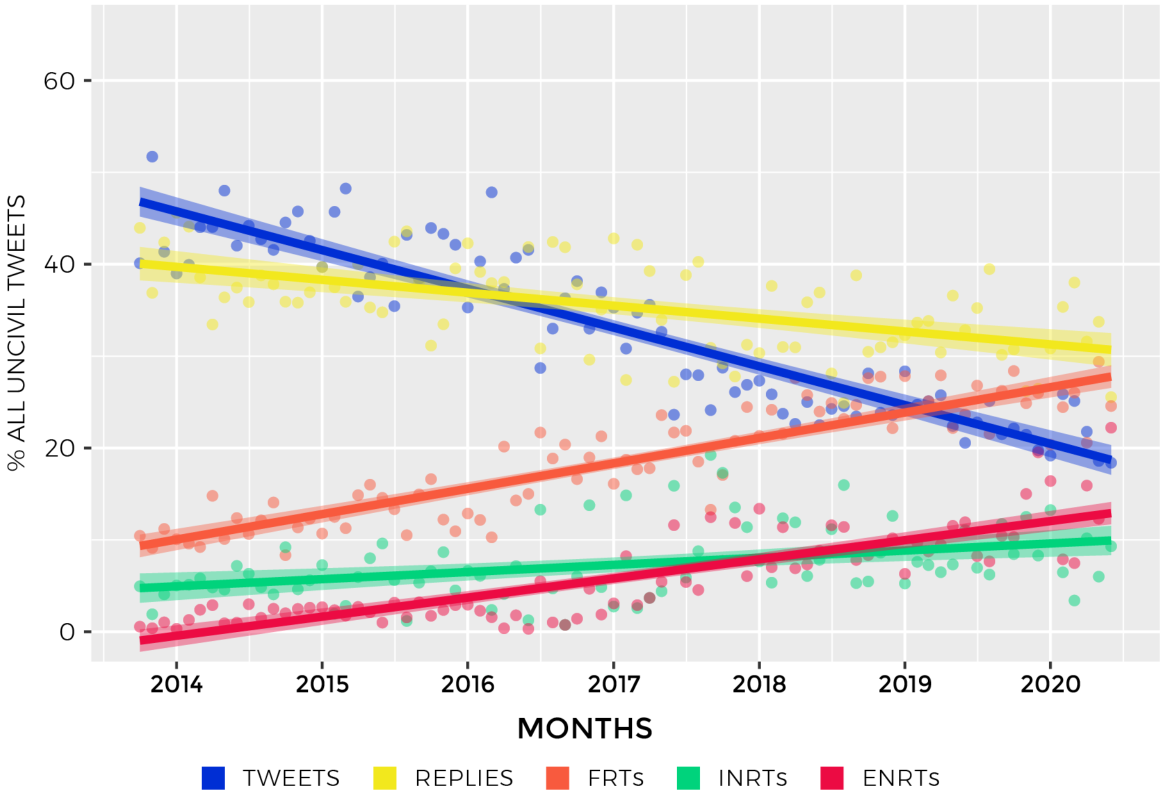


Figure 13: Share of all uncivil tweets over time, by tweet type.

Table 14: Binomial mixed effects regression models of incivility on time, algorithmic introduction, tweet type, political atmosphere and individual characteristics.

	DEPENDENT VARIABLE			
	P(INCIVILITY)			
COVARIATE	20	21	22	23
CONSTANT	-1.38*** (.05)	-1.30*** (.15)	-1.15*** (.14)	-.90*** (.25)
TIME (*t*)	-.000 (.000)			-0.000 (.000)
ALGORITHM	-.39*** (.05)		.34*** (.11)	-1.15*** (.29)
***t* × ALGORITHM**	.000*** (.000)			.001** (.000)
TWEET TYPE: REPLY	.07*** (.02)	.38*** (.10)		
TWEET TYPE: FRT	-.22*** (.02)	.36*** (.11)		
TWEET TYPE: INRT	-.36*** (.03)	.26** (.11)		
TWEET TYPE: ENRT	.27*** (.04)	.58*** (.14)		
POLITICAL ATMOSPHERE (TIME-VARIANT)				
BBC PARLIAMENT	.000*** (.000)		.000*** (.000)	0.000 (.000)
PARTISANSHIP, MEDIA DIET, POLITICAL ENGAGMENT AND TWITTER USAGE (UNIT-VARIANT)				
FOLLOWERS	-.000*** (.000)	-.000* (.000)	.000 (.000)	.000 (.000)
FOLLOWING GROWTH	.32*** (.09)		-.07 (.20)	-.15 (.20)
FOLLOWING GROWTH (*SE*)	7.63 (4.83)		4.38 (8.97)	2.00 (8.86)
FOLLOWING VOLATILITY	-.000 (.000)		.000 (.000)	.000 (.000)
FOLLOWING	-.000*** (.000)	-.000*** (.000)	-.000 (.000)	-.000 (.000)
ALL TWEETS	.000*** (.000)	.000** (.000)	-.000 (.000)	.000 (.000)
PARTISAN SCALE	.09 (.11)	.88** (.36)	.31 (.22)	.24 (.22)
PARTISAN SCALE^2	.34 (.23)	2.36*** (.83)	.30 (.41)	.16 (.40)
MEDIA SCALE	-8.55** (4.04)	-37.31*** (11.85)	-4.85 (8.77)	-1.24 (8.74)
MEDIA SCALE^2	858.59*** (272.39)	-424.63 (750.10)	-196.72 (583.89)	-110.17 (580.18)
PARTISANS FOLLOWED	.002*** (.000)	.002 (.002)	-0.000 (.001)	-0.000 (.001)
MEDIA FOLLOWED	.002 (.001)	.01*** (.004)	.002 (.002)	.002 (.002)
NON-PARTISANS FOLLOWED	-.01*** (.003)	-.02 (.01)	-.001 (.01)	-.001 (.01)
INDIVIDUALS (*n*)	1,228	184	957	957
OBSERVATIONS (*N*)	97,255	8,480	4,294	4,294
PSEUDO *R*^{2}	.09	.10	.01	.03

NOTE: Standard errors in parentheses. Time in days. ALGORITHM dummy coded '1' after January 2016. TWEET TYPE dummy base category = tweets. Data points are all tweets (20–21) and uncivil tweets only (22–23). P(*x*) = probability of *x*.

* **$p < 0.10$** ** **$p < 0.05$** *** **$p < 0.01$**

Table 15: FE regressions of monthly change in incivility on lagged incivility, by tweet type.

	DEPENDENT VARIABLE			
	Δ % UNCIVIL TWEETS		**Δ % UNCIVIL REPLIES**	
	INDEPENDENT VARIABLE			
	% UNCIVIL NFRTS	% UNCIVIL FRTS	% UNCIVIL NFRTS	% UNCIVIL FRTS
COVARIATE	24	25	26	27
	DEPENDENT VARIABLE LAGGED IN MONTHS			
LAG(DV) 1	.11*** (.005)	.11*** (.005)	.03*** (.01)	.03*** (.01)
LAG(DV) 2	.07*** (.005)	.07*** (.005)	.03*** (.01)	.02*** (.01)
LAG(DV) 3	.05*** (.005)	.05*** (.005)	.03*** (.01)	.04*** (.01)
LAG(DV) 4	.05*** (.005)	.05*** (.005)	.02*** (.01)	.02*** (.01)
LAG(DV) 5	.03*** (.005)	.04*** (.005)	.02*** (.01)	.02*** (.01)
LAG(DV) 6	.04*** (.01)	.04*** (.005)	.02*** (.01)	.02*** (.01)
LAG(DV) 7	.03*** (.01)	.03*** (.005)	.000 (.01)	.01 (.01)
LAG(DV) 8	.03*** (.01)	.03*** (.005)	.02*** (.01)	.02*** (.01)
LAG(DV) 9	.02*** (.005)	.01*** (.005)	.02*** (.01)	.01*** (.01)
LAG(DV) 10	.02*** (.005)	.03*** (.005)	.02*** (.01)	.01*** (.01)
LAG(DV) 11	.03*** (.005)	.02*** (.005)	.01** (.01)	.01* (.01)
LAG(DV) 12	.02*** (.005)	.03*** (.005)	.01** (.01)	.003 (.005)
	INDEPENDENT VARIABLE LAGGED IN MONTHS			
LAG(IV) 1	.02*** (.004)	.03*** (.004)	.02*** (.005)	.03*** (.005)
LAG(IV) 2	.02*** (.004)	.02*** (.004)	.02*** (.005)	.03*** (.005)
LAG(IV) 3	.02*** (.004)	.02*** (.004)	.01*** (.005)	.01** (.005)
LAG(IV) 4	.02*** (.004)	.02*** (.004)	.02*** (.005)	.02*** (.005)
LAG(IV) 5	.01*** (.004)	.003 (.004)	.01 (.005)	.004 (.01)
LAG(IV) 6	.03*** (.004)	.02*** (.004)	.01** (.005)	.02*** (.01)
LAG(IV) 7	.03*** (.004)	.02*** (.004)	.01* (.005)	.01* (.01)
LAG(IV) 8	.02*** (.004)	.02*** (.004)	.01* (.005)	.004 (.01)
LAG(IV) 9	.01* (.004)	.02*** (.004)	.01** (.005)	.01* (.01)
LAG(IV) 10	.02*** (.004)	.005 (.004)	.01 (.005)	.005 (.01)
LAG(IV) 11	.02*** (.004)	.01* (.004)	.004 (.005)	.02*** (.005)
LAG(IV) 12	.01** (.004)	.01*** (.004)	.01* (.005)	.01** (.005)
CASES (*n*)	1,103	1,139	1,021	1,037
OBSERVATIONS (*N*)	47,108	50,619	39,847	41,040
***R*^{2}**	.09	.08	.02	.01
ADJUSTED *R*^{2}	.07	.06	-.01	-.01
F STATISTIC	184.2*** (DF = 24; 45981)	190.7*** (DF = 24; 49456)	26.9*** (DF = 24; 38802)	25.4*** (DF = 24; 39979)

NOTE: IV = Independent variable; DV = Dependent variable. Δ = change. ***$^{*}p < 0.10$ $^{**}p < 0.05$ $^{***}p < 0.01$***

DISCUSSION

Taken together, I believe these results offer strong evidence of the association between incivility and content algorithms, and offer new insight into how the deliberative character of digital platforms changes as they adhere to attention economy logic. Of the almost 23 million tweets processed by my incivility classifier, 24% were labelled uncivil, not too far from the 18% Theocharis et al. (2020) found in a recent similar study. Where my findings differ is, instead of the relatively stable proportion of incivility they find over the period of around a year, my study of ten years of tweets finds a substantial increase in incivility over time.

This is particularly stark with regard to the period following Twitter's introduction of the timeline algorithm, as the support for **H1** attests. The right-hand scatterplot of Figure 9 makes this especially lucid, where separate trend lines for the pre- and post-algorithm periods point in opposite directions, a pattern repeated in Figure 10 where individual users are plotted. This indicates that Twitter incivility was actually in decline until the timeline algorithm's introduction. The results demonstrate the power of a large, longitudinal dataset to reveal long-term tendencies and shifts, with implications for other platform studies.

That this relationship remained robust to individual variation (fixed effects), time shocks (random effects), and the inclusion of an array of control variables, should increase confidence the association is genuine rather than spurious. In particular, the robustness to independently significant partisanship, media diet and political engagement measures, combined with the panel structure of the sample — a persistent group of users over the whole period of study — provides sufficient confidence to rule out an ideological compositional effect, whereby the most extreme, uncivil users tweet more over time while more moderate, civil users are cowed into silence or leave the platform altogether. This structure did result in a very unbalanced panel, as most users were not active in each and every of the study's 120 months. The regression models used can account for unbalanced panels, but to be certain this did not affect the results, additional regressions were run using only the 309 users with data for each month (Appendix VII), finding consistent results.

I appreciate readers will differ in how convincing they find the controls for changing political atmosphere — clearly a consequential omitted variable — which is why I do not solely rely on the temporal dimension for this analysis. While the use of retweets for modelling cascades or diffusion is commonplace (Cihon & Yasseri, 2016), delineating a typology of retweets is a novel method of inferring features of algorithmic influence. Admittedly, retrospective measurement requires a small leap of faith, but I believe my efforts to rule out systematic error combined with the consistent findings of the real-time measurement sample are strong reasons not to dismiss the support for **H2c**. This finds EXTRA-NETWORK RETWEETS are meaningfully more prevalent since algorithmic introduction, and since the rejection of **H2d** implies this was a categorical shift — rather than a continuous one more consistent with measurement bias — I contend ENRTS should be taken seriously as a product of algorithmic influence.

In this light, the support I find for **H3** and **H4** has serious implications. Even if there are unobserved political factors not accounted for in the analysis and correlated with increased incivility, it is difficult to find a reason incivility would increase within the tweet types I predict *relative* to other tweet types and *in conjunction* with the introduction of the algorithm. It stretches credulity to think these changes in relation to affordances exclusive to Twitter could result solely from an exogenous shock in the political culture. It is more reasonable, I think, to conclude these increases follow an endogenous shift in the platform's structure.

I also find lukewarm support for **H5**, but I believe the more robust ENRTS measure would need to be calculated at a scale beyond this study's capability to provide more confidence in a true influential effect of retweet incivility in previous months on present tweet incivility.

The final finding aside, I consider these results on the whole to represent compelling evidence that anti-deliberative optimisation has been taking place on Twitter since the timeline algorithm was introduced, answering **RQ1** in the affirmative, while the continuing rise in the amount of incivility since algorithmic introduction suggests this process may be self-perpetuating, as **RQ2** supposes.

Furthermore, these conclusions are corroborated by findings of the same phenomenon published in *The Economist* (2020) just a few weeks before this paper. They confirm Twitter's algorithmic timeline favours inflammatory language relative to a chronological feed.

Reconciling "filter bubbles" with the "hell site"

This research responds to calls from scholars like Crockett (2017) asking for studies which "investigate the extent to which digital media platforms intensify moral emotions [and] suppress productive social discourse", while it tentatively supports mechanisms from "Mediated Skewed Diffusion of Issues Information" theory proposed by McEwan et al. (2018).

In particular, the MSDII model helps resolve two apparently inconsistent observations about digital platforms:

(I) They have a tendency to reinforce the existing views of users

(II) They foster a disagreeable and hostile environment

(I) has prompted theorists to conceive of platforms as "echo chambers" which feed back to us what we want to hear (Sunstein, 2008, 2017, 2018), or "filter bubbles" which sift out counter-attitudinal messaging (Pariser, 2011); while (II) is commonly expressed as concern for the widespread toxicity on platforms (Behr, 2018; H. Lewis, 2018), with users describing Twitter as a "hell site" (Keck, 2019; Manavis, 2019). Dutton et al. (2017) explain the former view is "intuitively appealing" but unsupported empirically (q.v. §1.3.3 and Appendix I). Expounding this disjunction, McEwan et al. posit that platforms, through "weak ties", *do* expose users to oppositional messages, but they are actually *further* polarised as a result, due to the poor quality of platform discourse. Bail et al. (2018) find empirical support for this. Those interested in politics experience this process more intensely than most.

This paper theoretically and empirically contributes to this understanding by clarifying how the discussed phenomenon is 'a feature, not a bug'. When platforms promote oppositional messaging it is in the most incendiary light, as this maximises engagement, owing to the enticing appeal of moral outrage (Crockett, 2017). I call this a 'noisy neighbours' effect:

> You are not strictly in an echo chamber, sealed off from your next-door neighbours, when you can see them leaving the house for work or in the garden on a sunny weekend. But if the most notable thing you learn about them comes from hearing their shouting arguments through the walls, you do not build a positive, generous perception of them conducive to fruitful exchanges .

While the people most interested in politics may be more likely to come into contact with diverse and opposing views, in an environment mediated by attention optimisation, it accompanies an incentive structure to ridicule, mock and direct morally indignant invective.

"Twitter isn't real life"

Trust is a fickle thing

A critic may minimise these findings by pointing to the exceptionally dim view people take of social media as the "least trusted sources of news" (Newman et al., 2019), and thus the quality of discourse therein is of little consequence. However, the relationship with trust and news is complex — consider that low trust in *The Sun* and *Daily Mail* among *their own readers* accompanies the highest news reach in the UK (Bold, 2018). Regardless of trust, social media feeds act as neutral venues through which other news sources are discovered (Groshek & Koc-Michalska, 2017). Thanks to cognitive biases, these sources are still judged "on the merits", which in practice often means consistent with users' prior views (Stöcker, 2020).

'If it trends, it ascends'

Perhaps one is still unconvinced what happens on social media matters in its own right, owing to the unrepresentative demographic composition, evincing the oft-repeated declaration: "Twitter isn't real life" (Berkowitz, 2019). Even so, it is inarguable that platforms have influence beyond their putative bounds, and Twitter matters because it influences the influential (Behr, 2018). At this point, how Donald Trump tweets to manipulate established media coverage is a matter of record (Wells et al., 2020) — not just bypassing gatekeepers to speak directly to the people, but also to coerce coverage from those same gatekeepers. To the well-worn adage of media coverage '*if it bleeds, it leads*' (Stöcker, 2020), we might add '*if it trends, it ascends*'.[10] Here, the para-social connection social platforms create between influencers and their followers (Kim et al., 2015), and accompanying pseudo-voyeurism for bemused onlookers, is used to great effect by populists the world over (Giuffrida et al., 2018).

It's the Zuck's world, we just live in it

Journalism has been pulled into this new media ecology whether it likes it or not. Platforms not only wield near-monopoly power in their own domain, they are effectively *monopsonies* in related domains, most notably in the 'content creation arena', a.k.a. news media (Martínez, 2019). The logic and incentives of attention optimisation are unavoidably transferred to journalists, encouraging clickbait, sensationalism and emotive appeals to immutable identities (Klein, 2020; McEwan et al., 2018), to the detriment of the public sphere.

Gasoline on the polarisation fire

The symptom not the cause

From what we know about anti-deliberative messaging, its promotion — as identified in this study — by attention optimisation is likely increasing polarisation among politically engaged Twitter users (A. A. Anderson et al., 2014; Gervais, 2019; Lyons & Veenstra, 2016). Others, though, will point to longer term trends in the media landscape — such as fragmentation, corporate hegemony and the perverse incentives towards sensationalism which emerge as a result — as evidence this is nothing new (Flew, 2019; Schroeder, 2019).

Relatedly, sceptical scholars can add to these what Lewandowsky et al. (2017) call "societal mega-trends", such as declining social capital, growing economic inequality and disintegrating trust in institutions and expertise. For good measure they can include the increasing sociocultural heterogeneity and changing position of the nation state Blumler & Gurevitch (2000) recognised two decades ago. It is useful in this context also to recall that

[10] It is remarkable how few words rhyme with "trend".

recent fears over "democratic backsliding" (Bermeo, 2016) were also raised as far back as the early 1990s (Dahlgren, 2005), practically contiguously with Fukuyama's (1993) 'end of history' thesis. Against this background, it is understandable why some scholars see the internet's relationship with polarisation in reverse and conclude, as did Margolis & Resnick (2000), that while "there is extensive political life on the Net … it is mostly an extension of political life off the Net."

Table 16: A framework for comparing socio-technological developments by analogy to the hardware/software distinction.

TECHNOLOGICAL	**SOCIOLOGICAL**
HARDWARE CONCRETE AND TANGIBLE	
Devices, infrastructures, machinery	Bodies, settlements, resources
EXAMPLE DEVELOPMENTS	
Smartphones, microchips, TV, radio, cameras, aviation, internal combustion, printing press	Migration, urbanisation, globalisation, global trade, public health, climate change
SOFTWARE ABSTRACT AND INCORPOREAL	
Programs, services, content	Psychology, social relationships, culture
EXAMPLE DEVELOPMENTS	
Algorithms, social media, the web, the internet, operating systems, machine code	Sensationalism, addiction, identity politics, education, nationalism, hate, prejudice

Counterfactuals: A world without attention optimisation

I do not wish to challenge this sensible view, but instead complicate it. In Table 16 I present a basic framework which contrasts sociological developments with technological ones, by analogy to the hardware/software distinction in computing. Generally speaking, matters of 'hardware' can be considered concrete and tangible in contrast to 'software' which is abstract and incorporeal. Software developments tend to flow from hardware ones, e.g. the internet is dependent on telecommunications infrastructure. In the sociological realm, a similar example is the heightening of nationalist sentiment that accompanies national competition for resources. Importantly though, this causal relationship can occasionally operate in reverse, e.g. how the computational demands of artificial intelligence influence the development of processor architecture (R. Harris, 2019). A sociological, and grave, instance of this is when a reign of white supremacist terror in the American South drove the "Great Migration" of African Americans northward (Tolnay et al., 2018).[11] Crucially, these causal relationships can crossover from the sociological to the technological, and vice versa, like the association between telecommunications and globalisation, or how "selfie" culture led to the development of augmented reality "filters" on platforms like Snapchat.

These hardware and software components of technological and sociological conditions can be conceived as four threads or paths running in parallel, ostensibly independent of one another, but with causality and contingency weaving across and between. I emphasise this to aid the visualisation of counterfactuals. With all else equal, consider the absence of a single development along one of these threads at a pivotal point in the past. Imagine the conclusion of the Second World War without the Manhattan Project (Groves, 1975), or conversely, the development of nuclear energy without the United States entering the conflict, as transpires in Philip Roth's *The Plot Against America* (2004).

The three eras of web content I laid out in §1.2.2 each represent a development along the technological software path. My point is, to consider the impact of *Web 3.0* — the attention economy model — we must imagine its absence. Envisage today's moment, where all the forces that have led us here — industrial disruption, blurring of national borders, broadening cultural diversity, media fragmentation — still pertain; and the precursor technologies of the internet, smartphones and the social media of *Web 2.0* are still intact, with their accompanying network effects and selective exposure. The only difference: there is no attention economy. No attention-optimising content algorithms, no superlative efficacy, no infinitely scrolling feeds, no video autoplay. No Facebook. No YouTube. Only… MySpace.

[11] Clearly violence and murder are as concrete as the migration of people. However hate and prejudice are not, and while, to its victims, the consequences of hate are absolutely concrete, this serves to illustrate the point that abstract notions can have tangible, and sometimes horrifying, results.

Or Twitter before the timeline algorithm.

I argue, with the support of this study's empirical findings, that this slightly alternative reality would be less polarised than our own. Not absent polarisation, but meaningfully less polarised. Just as scholars have previously stressed the need to foreground the affordances of social media in understanding their role in the new media ecology (Halpern & Gibbs, 2013), this paper contributes to the discourse on platforms and polarisation by emphasising the need to interpret contemporary platform pathologies through the lens of their *technological* affordances in accruing greater amounts of attention for economic purposes. It presents the typology of web content eras (§1.1.2), the concept of ANTI-DELIBERATIVE OPTIMISATION (§2), and empirical evidence thereof (§6) to demonstrate that, while not the cause of political polarisation, digital platforms, and the ADO at their core, are a significant accelerant.

Polarflationary peril

I stress that if anti-deliberative content is increasing at a higher rate on Twitter than it otherwise would without an attention-optimising algorithm — which these results suggest but cannot provide confidence for as a matter of research validity — then this would be a paradigmatic example of POLARFLATION. This is to say, the resultant 'increase' in polarisation, rather than expressing a genuine widening in preferences or values, is instead a creation of presentation, i.e. the selective promotion of certain heightened views or positions from a possible pool otherwise equally polarised as before the nominal 'increase'.

This poses a problem above and beyond 'legitimate' polarisation. The economic *portmanteau* cousin of POLARFLATION, "stagflation", demonstrates how monetary inflation — on its own, a problem when out of control but useful when controlled — becomes an aggravating factor when paired with high unemployment, as it makes the latter problem harder to solve. POLARFLATION has similar features.

While polarisation on its own is typically framed as concerning for democracies, a certain amount of polarisation is deemed beneficial by political scientists. A famous issue of the *American Political Science Review* (1950) once called for a more ideologically sorted — i.e. polarised — party system in the US. While panned at the time, the suggestion was vindicated over the next two decades when the ideologically mixed parties contrived to block civil rights reform despite majoritarian popular support (Klein, 2020). In this scenario, polarisation creates tension that invites resolution. The peril of POLARFLATION is, when polarisation increases unmoored from actual differences in specific interests or demands — i.e. it inflates — it is unclear how to puncture the impasse. As polarisation militates against compromise, there is a risk that conflict continues unabated, to the detriment of consensual governance.

If, as these results suggest, this phenomenon is occurring as the result of platform attention optimisation, because of the inscrutable nature of content algorithms, we might not have certainty until the damage is irreversible. Which is why further studies like this are so vital.

CONCLUSIONS

> And behold the Shepard tone, a sound illusion, always sliding upward … yet never quite disappearing, like the red and blue stripes on a barber's poll. As this sonic mass moves up and up and up, the tension rises also until — nothing, no resolution, no catharsis. Here comes another rising line and another after that.
>
> **Brooke Gladstone**, *On the Media*
> *WNYC Studios* podcast

This study has provided a theoretical framework for the process by which the attention-optimising content algorithms of digital platforms debilitate democratic deliberation. ANTI-DELIBERATIVE OPTIMISATION (ADO) was expounded and grounded in the market logic of the attention economy, and technological supremacy through which this logic is pursued.

Empirical support for this phenomenon was found on Twitter in an original research design that takes inspiration from a new class of "algorithmic audit" studies. Accessing a proprietary archive of the Twitter "Firehose API", a large, longitudinal dataset covering a decade of tweets was obtained. A panel of more than a thousand UK politically engaged users was retrospectively constructed (n = 1,228), with every tweet over the 10 year period recorded, comprising a total corpus of almost 23 million. By taking the introduction of Twitter's timeline algorithm in 2016 as a discontinuity, the study assessed whether the degree of incivility on Twitter has changed following this intervention.

Computational methods were undertaken to classify tweets for incivility, while a retweet typology was defined and calculated, providing a novel measure of algorithmic influence. Statistical analyses revealed support for the majority of hypotheses, finding strong evidence that incivility on Twitter has increased in concurrence with the introduction of algorithmic curation, an effect equal to a 42% increase in incivility. Additionally, retweets I believe are most indicative of algorithmic influence were found to be between 5 and 11 p.p. more uncivil than other tweets. Results were robust to control estimates of political atmosphere, partisanship, media diet, political engagement and Twitter usage. Furthermore, this increase in incivility has continued over time, indicating the process is self-perpetuating. I claim this is reason to believe ADO is underway on Twitter. Consistent with theory and the findings of other studies (Gervais, 2019; McEwan et al., 2018), I propose this is contributing to POLARFLATION: polarisation inflated to appear greater than it is in any concrete terms. This supports the "painful truths" Deibert (2019) says platforms must face up to in regard to their detrimental impact on democratic ideals, and corroborates much of Stöcker's (2020) work.

Limitations and implications for future research

I do not pretend this study is without limitations, and would caution against generalising these results, either to platforms other than Twitter — which have different mechanisms, affordances, and have introduced attention optimisation in different ways — or groups other than politically engaged users who are likely to have both different levels of incivility and levels of discussion that would constitute public sphere contributions (Barberá et al., 2015). Beyond this, I have discussed other limitations regarding measurement throughout the paper, but I believe EXTRA-NETWORK RETWEETS are a fundamentally sound instrument for measuring algorithmic influence on Twitter. Future studies with more substantial financial and computational resources could calculate these at a greater scale and precision, bolstering statistical confidence. International comparisons would increase generalisability to other political contexts and go a long way toward settling omitted variable concerns.

Moreover, the study proves the formidable power large scale observation can provide for auditing algorithms. It should be more widely adopted and applied to other platforms. I have used incivility as an instrument for ANTI-DELIBERATIVE OPTIMISATION, consistent with other studies (Gervais, 2019; Gervais & Chin, 2018; Theocharis et al., 2016, 2020), but would urge the scholarly community to arrive at an agreed-upon standard for measuring this crucial phenomenon, so it can be identified and tracked at scale and over time. This could ultimately form a 'public sphere index'. Taking inspiration from the stellar work at Indiana University's Observatory of Social Media (2020), it could be maintained by the media and communications field and track the deliberative quality of platforms, similarly to the Doomsday Clock (Bulletin of the Atomic Scientists, 2020) or Freedom House (2019) rankings.

What can be done?

These findings are bleak for the digital public sphere, but there are some potential solutions. First, and simplest, media literacy curricula must immediately integrate and *foreground* an understanding of the attention economy and its aims manifested through algorithmic curation (J. Cohen, 2018). Second, governments need to assert their oversight authority and compel platforms to increase transparency about their algorithms' processes and effects. Until, and in lieu of this, researchers will need to continue and increase their research in the area so the scale of the problem can be comprehended. I hope this study can join others like it as a guide forward in this regard. Third, platforms need to be pressured to integrate optimisation metrics more conducive to deliberation into their algorithms. This is unlikely to occur with the current economic incentives, to wit… Fourth, platforms must have their hegemonic grip on their respective markets shaken to allow competing business models to flourish. Given the properties of superlative efficacy, this is likely to require intervention. Antitrust has been mooted and should be explored, although it remains unclear how this can be achieved without compromising service quality, which is substantial. The alternative — to regulate platforms as the public utilities they have effectively become —should be pursued if a suitable settlement cannot be reached. These measures are fraught, but so too is the process evidenced in these pages. If allowed to continue unabated, it threatens to hold the coarsened tenor of political debate at fever pitch, with little prospect of amelioration, and grave consequences for public sphere deliberation and the democracy it supports.

SUPPLEMENTARY MATERIALS

Replication data from this study, including the constitutive datasets for the calculation of the PARTISAN and MEDIA SCALE variables, are published online at caveen.com/polarflation.

ACKNOWLEDGEMENTS

It would not have been possible to complete this Master's degree without the support and understanding of family, friends, colleagues and the staff of LSE. I would like to express my gratitude to all these people. Special thanks go to my parents, for enabling me to undertake this challenge; to my friends, for still being my friends after I've ignored them for weeks; to my supervisor Robin Mansell, for her generous guidance and encouragement; and to Sarah, for her everything — I'm back in the real world now, I promise.

References

Aghababaei, S., & Makrehchi, M. (2017). Activity-based Twitter sampling for content-based and user-centric prediction models. *Human-Centric Computing and Information Sciences*, *7*(1), 3. https://doi.org/10.1186/s13673-016-0084-z

AllSides. (2020). *AllSides Media Bias Ratings*. AllSides. https://www.allsides.com/media-bias/media-bias-ratings

American Political Science Association. (1950). Toward a More Responsible Two-Party System: A Report of the Committee on Political Parties. *The American Political Science Review*, *44*(3), 1–14. JSTOR. https://doi.org/10.2307/1950998

Andersen, J. C., & Sandberg, S. (2018). Islamic State Propaganda: Between Social Movement Framing and Subcultural Provocation. *Terrorism and Political Violence*, 1–21. https://doi.org/10.1080/09546553.2018.1484356

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies*. *Journal of Computer-Mediated Communication*, *19*(3), 373–387. https://doi.org/10.1111/jcc4.12009

Anderson, M., & Jiang, J. (2018, May 31). Teens, Social Media & Technology 2018. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/

Andersson Schwarz, J. (2017). Platform Logic: An Interdisciplinary Approach to the Platform-Based Economy. *Policy & Internet*, *9*(4), 374–394. https://doi.org/10.1002/poi3.159

Anzieu, A. (2019, August 1). Introducing Serendipity into Recommendation Algorithms. *SSENSE TECH*. https://medium.com/ssense-tech/introducing-serendipity-into-recommendation-algorithms-fb92af88ee0b

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216. https://doi.org/10.1073/pnas.1804840115

Bakker, R., Hooghe, L., Jolly, S., Marks, G., Polk, J., Rovny, J., Steenbergen, M., & Vachudova, M. A. (2020). *Chapel Hill Expert Surveys 1999–2019*. University of North Carolina. https://www.chesdata.eu/2019-chapel-hill-expert-survey

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

BARB. (n.d.). *Weekly TV set viewing summary*. Broadcasters' Audience Research Board. Retrieved 24 August 2020, from https://www.barb.co.uk/viewing-data/weekly-viewing-summary/

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2020). Automated Text Classification of News Articles: A Practical Guide. *Political Analysis*, 1–24. https://doi.org/10.1017/pan.2020.8

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Barnes, R. (2018). Uncovering online commenting culture: Trolls, fanboys and lurkers. Cham, Switzerland : Palgrave Macmillan.

Bastos, M., & Walker, S. T. (2018, April 11). *Facebook's data lockdown is a disaster for academic researchers*. The Conversation. http://theconversation.com/facebooks-data-lockdown-is-a-disaster-for-academic-researchers-94533

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., & Fox, J. (2020). *lme4: Linear Mixed-Effects Models using 'Eigen' and S4* (1.1-23) [Computer software]. https://CRAN.R-project.org/package=lme4

Bechmann, A., & Nielbo, K. L. (2018). Are We Exposed to the Same "News" in the News Feed? *Digital Journalism*, *6*(8), 990–1002. https://doi.org/10.1080/21670811.2018.1510741

Behr, R. (2018). How Twitter poisoned politics. *Prospect Magazine*.

Benkler, Y. (2006). The wealth of networks: How social production transforms markets and freedom. Yale University Press.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Berger, J., & Milkman, K. L. (2012). What Makes Online Content Viral? *Journal of Marketing Research*, *49*(2), 192–205. https://doi.org/10.1509/jmr.10.0353

Berkowitz, J. (2019, April 24). Study confirms: Twitter is not real life. *Fast Company*. https://www.fastcompany.com/90339526/study-confirms-twitter-is-not-real-life

Bermeo, N. (2016). On Democratic Backsliding. *Journal of Democracy*, *27*(1), 5–19. https://doi.org/10.1353/jod.2016.0012

Blank, G., & Duton, W. H. (2019). *Perceived Threats to Privacy Online: The Internet in Britain*. Oxford Internet Institute.

Blumler, J. G., & Gurevitch, M. (2000). Rethinking the study of political communication. In J. Curran & M. Gurevitch (Eds.), *Mass media and society* (3rd ed., pp. 155–172). Arnold.

Bold, B. (2018, September 17). *The Guardian most trusted and The Sun least trusted online news brand, Pamco reveals*. PR Week. http://www.prweek.com/article/1492977?utm_source=website&utm_medium=social

Borah, P. (2012). Does It Matter Where You Read the News Story? Interaction of Incivility and News Frames in the Political Blogosphere: *Communication Research*. https://doi.org/10.1177/0093650212449353

Boucher, P. (2019). *Technology and social polarisation* (European Parliament Research Service PE 634.412). European Parliament Research Service.

boyd, danah, & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Bradshaw, S. (2019). Disinformation optimised: Gaming search engine algorithms to amplify junk news. *Internet Policy Review*, *8*(4). https://policyreview.info/articles/analysis/disinformation-optimised-gaming-search-engine-algorithms-amplify-junk-news

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Bavel, J. J. V. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Bruns, A. (2017). Gatewatching and news curation: Journalism, social media, and the public sphere. New York : Peter Lang.

Bruns, A., & Highfield, T. (2015). From news blogs to news on Twitter: Gatewatching and collaborative news curation. In S. Coleman & D. Freelon (Eds.), *Handbook of digital politics* (pp. 325–339). Cheltenham, UK : Edward Elgar Publishing.

Bruns, A., & Highfield, T. (2016). Is Habermas on Twitter? Social media and the public sphere. In A. Bruns, G. Enli, E. Skogerbø, A. O. Larsson, & C. Christensen (Eds.), *The Routledge companion to social media and politics* (pp. 56–73). Routledge.

Bruns, A., & Weller, K. (2014). Twitter data analytics – or: The pleasures and perils of studying Twitter. *Aslib Journal of Information Management*, *66*(3). https://doi.org/10.1108/AJIM-02-2014-0027

Bulletin of the Atomic Scientists. (2020, January 23). Doomsday Clock. *Bulletin of the Atomic Scientists*. https://thebulletin.org/doomsday-clock/

Bush, S. (2018, May 30). What does your political hashtag say about you? *New Statesman*. https://www.newstatesman.com/politics/elections/2018/05/what-does-your-political-hashtag-say-about-you

Cadwalladr, C. (2019). *Facebook's role in Brexit—And the threat to democracy*. TED. https://www.ted.com/talks/carole_cadwalladr_facebook_s_role_in_brexit_and_the_threat_to_democracy

Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Calhoun, C. J. (1992). *Habermas and the public sphere*. MIT Press. http://web.a.ebscohost.com.gate3.library.lse.ac.uk/ehost/detail/detail?vid=0&sid=87e16b8c-3f3f-4e29-8e71-5993fc9593c0%40sessionmgr4008&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=nlebk&AN=48445

Cardenal, A. S., Aguilar-Paredes, C., Cristancho, C., & Majó-Vázquez, S. (2019). Echo-chambers in online news consumption: Evidence from survey and navigation data in Spain. *European Journal of Communication*, *34*(4), 360–376. https://doi.org/10.1177/0267323119844409

Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions: 'On the Internet, Nobody Knows You're a Dog.' *Computers in Human Behavior*, *23*(6), 3038–3056. https://doi.org/10.1016/j.chb.2006.09.001

Cihon, P., & Yasseri, T. (2016). A biased review of biases in Twitter studies on political collective action. *Frontiers in Physics*, *4*, 34. https://doi.org/10.3389/fphy.2016.00034

Coddington, M., & Holton, A. E. (2014). When the Gates Swing Open: Examining Network Gatekeeping in a Social Media Setting. *Mass Communication and Society*, *17*(2), 236–257. https://doi.org/10.1080/15205436.2013.779717

Cohen, D. (2019, August 13). 'Loud, obsessive, tribal': The radicalisation of remain. *The Guardian*. https://www.theguardian.com/politics/2019/aug/13/brexit-remain-radicalisation-fbpe-peoples-vote

Cohen, J. (2018). Exploring Echo-Systems: How Algorithms Shape Immersive Media Environments. *Journal of Media Literacy Education*, *10*(2), 139–151. https://doi.org/10.23860/JMLE-2018-10-2-8

Cohn, N., & Quealy, K. (2020, June 10). How Public Opinion Has Moved on Black Lives Matter. *The New York Times*. https://www.nytimes.com/interactive/2020/06/10/upshot/black-lives-matter-attitudes.html

Coletto, M., Lucchese, C., Orlando, S., & Perego, R. (2016). Polarized User and Topic Tracking in Twitter. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 945–948. https://doi.org/10.1145/2911451.2914716

Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the Political Alignment of Twitter Users. *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 192–199. https://doi.org/10.1109/PASSAT/SocialCom.2011.34

Conover, M. D., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., & Flammini, A. (2011). *Political Polarization on Twitter*. Fifth International AAAI Conference on Web and Social Media.

Coyle, D. (2018). Platform Dominance: The Shortcomings of Antitrust Policy. In M. Moore & D. Tambini, *Digital dominance: The power of Google, Amazon, Facebook, and Apple* (pp. 50–70). Oxford University Press. http://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=5359105

Cram, L. (2015, November 17). How Remain and Leave camps use #hashtags. *UK in a Changing Europe*. https://ukandeu.ac.uk/how-remain-and-leave-camps-use-hashtags/

Cram, L., Llewellyn, C., Hill, R., & Magdy, W. (2017). *General Election 2017—A Twitter analysis*. The UK in a Changing Europe. https://ukandeu.ac.uk/research-papers/general-elections-2017-a-twitter-analysis/

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3

Croissant, Y., Millo, G., Tappe, K., Toomet, O., Kleiber, C., Zeileis, A., Henningsen, A., Andronic, L., & Schoenfelder, N. (2020). *plm: Linear Models for Panel Data* (2.2-3) [Computer software]. https://CRAN.R-project.org/package=plm

Cunningham, S. (2016). Popular media as public 'sphericules' for diasporic communities. *International Journal of Cultural Studies*, *4*(2), 131–147. https://doi.org/10.1177/136787790100400201

Dahlberg, L. (2001). The Internet and Democratic Discourse: Exploring The Prospects of Online Deliberative Forums Extending the Public Sphere. *Information, Communication & Society*, *4*(4), 615–633. https://doi.org/10.1080/13691180110097030

Dahlgren, P. (2000). The Internet and the Democratization of Civic Culture. *Political Communication*, *17*(4), 335–340. https://doi.org/10.1080/10584600050178933

Dahlgren, P. (2005). The Internet, Public Spheres, and Political Communication: Dispersion and Deliberation. *Political Communication*, *22*(2), 147–162. https://doi.org/10.1080/10584600590933160

Dahlgren, P. (2018a). Public Sphere Participation Online: The Ambiguities of Affect. *Les Enjeux de l'information et de La Communication*, *19/1*(1), 5–20. https://doi.org/10.3917/enic.024.0005

Dahlgren, P. (2018b). Media, Knowledge and Trust: The Deepening Epistemic Crisis of Democracy. *Javnost - The Public*, *25*(1–2), 20–27. https://doi.org/10.1080/13183222.2018.1418819

Darcy, O. (2019). How Twitter's algorithm is amplifying extreme political rhetoric. *CNN Business*.

Davenport, T. H. (2001). The attention economy: Understanding the new economy of business. Harvard Business School Press.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *International AAAI Conference on Web and Social Media; Eleventh International AAAI Conference on Web and Social Media*. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665/14843

de Lima, A. P., & Peres, S. M. (2018). Limits to Surprise in Recommender Systems. *ArXiv:1807.03905 [Cs]*. http://arxiv.org/abs/1807.03905

Deibert, R. J. (2019). The Road to Digital Unfreedom: Three Painful Truths About Social Media. *Journal of Democracy*, *30*(1), 25–39. https://doi.org/10.1353/jod.2019.0002

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554. https://doi.org/10.1073/pnas.1517441113

Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., & Jurafsky, D. (2019). Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. *ArXiv:1904.01596 [Cs]*. http://arxiv.org/abs/1904.01596

Döring, M. (2018, December 7). Inference vs Prediction. *Data Science Blog*. https://www.datascienceblog.net/post/commentary/inference-vs-prediction/

Dredge, S. (2014, October 17). Yes, Twitter is putting tweets in your timeline from people you don't follow. *The Guardian*. https://www.theguardian.com/technology/2014/oct/17/twitter-tweets-timeline-dont-follow

Dreyfuss, E. (2019, February 5). Teens Don't Use Facebook, but They Can't Escape It, Either. *Wired*. https://www.wired.com/story/teens-cant-escape-facebook/

Driscoll, K. (2016). *Social Media's Dial-Up Ancestor: The Bulletin Board System*. IEEE Spectrum. https://spectrum.ieee.org/tech-history/cyberspace/social-medias-dialup-ancestor-the-bulletin-board-system

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, *21*(5), 729–745. https://doi.org/10.1080/1369118X.2018.1428656

Dutton, W. H., Reisdorf, B. C., Dubois, E., & Blank, G. (2017). Search and politics: The uses and impacts of search in Britain, France, Germany, Italy, Poland, Spain, and the United States. *Quello Center Working Paper*, *2944191*. https://ora.ox.ac.uk/objects/uuid:2cec8e9b-cce1-4339-9916-84715a62066c

Eady, G., Nagler, J., Guess, A., Zilinsky, J., & 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). https://doi.org/10.1177/2158244019832705

Erlanger, S. (2017, June 11). For Britain, Political Stability Is a Quaint Relic. *The New York Times*. https://www.nytimes.com/2017/06/11/world/europe/britain-politics-theresa-may.html

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). 'I always assumed that I wasn't really that close to [her]': Reasoning about Invisible Algorithms in News Feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 153–162. https://doi.org/10.1145/2702123.2702556

Evans, G., & Schaffner, F. (2019, January 23). Brexit identities: How Leave versus remain replaced Conservative versus Labour affiliations of British voters. *The UK in a Changing Europe*. https://ukandeu.ac.uk/brexit-identities-how-leave-versus-remain-replaced-conservative-versus-labour-affiliations-of-british-voters/

Eyal, N. (2014). *Hooked: How to Build Habit-Forming Products* (R. Hoover, Ed.). Penguin Books.

Facebook. (2020). *Facebook's Third-Party Fact-Checking Program*. Facebook's Third-Party Fact-Checking Program. https://www.facebook.com/journalismproject/programs/third-party-fact-checking

Farrell, H. (2012). The Consequences of the Internet for Politics. *Annual Review of Political Science*, *15*(1), 35–52. https://doi.org/10.1146/annurev-polisci-030810-110815

Filloux, F. (2017). The Facebook journalism project is nothing but a much-needed PR stunt. *Quartz*.

Flew, T. (2019). Digital communication, the crisis of trust, and the post-global. *Communication Research and Practice*, *5*(1), 4–22. https://doi.org/10.1080/22041451.2019.1561394

Flood, M. (2019). Fighting Fake: Promoting critical thinking and constructive public dialogue. https://www.fightingfake.org.uk/media-bias

Fraser, N. (1990). Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Social Text*, *25/26*(25/26), 56–80. JSTOR. https://doi.org/10.2307/466240

Freedom House. (2019). *Democracy in Retreat*. Freedom House. https://freedomhouse.org/report/freedom-world/2019/democracy-retreat

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., & Qian, J. (2020). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models* (4.0-2) [Computer software]. https://CRAN.R-project.org/package=glmnet

Fukuyama, F. (1993). *The End of History and the Last Man*. Penguin Books.

Galston, W. (2002). The impact of the Internet on civic life: An early assessment. In E. C. Kamarck & J. S. Nye (Eds.), *Governance.com: Democracy in the information age*. Brookings Institution Press.

Galston, W. (2003). If political fragmentation is the problem, is the Intemet the solution? In D. M. Anderson & M. Cornfield (Eds.), *The civic web: Online politics and democratic values*. Rowman & Littlefield.

Garnham, N. (1986). The media and the public sphere. *Intermedia, 14*(1), 28–33.

Gervais, B. T. (2013). Incivility in Online Political Discourse and Anti-Deliberative Attitudes: An Experimental Analysis. *Paper Prepared for Delivery at the Annual Meeting of the American Political Science Association*. 109th Annual Meeting of the American Political Science Association, Chicago, IL. https://papers.ssrn.com/abstract=2301194

Gervais, B. T. (2014). Following the News? Reception of Uncivil Partisan Media and the Use of Incivility in Political Expression. *Political Communication*, *31*(4), 564–583. https://doi.org/10.1080/10584609.2013.852640

Gervais, B. T. (2019). Rousing the Partisan Combatant: Elite Incivility, Anger, and Antideliberative Attitudes. *Political Psychology*, *40*(3), 637–655. https://doi.org/10.1111/pops.12532

Gervais, B. T., & Chin, A. (2018, August 31). Leveraging Metadata to Measure the Attention Grabbing Power of Incivility. *Paper Prepared for the APSA Annual Meeting & Exhibition*. 114th APSA Annual Meeting & Exhibition, Boston, MA. http://tinyurl.com/y9tpsltm

Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., & Roy, D. (2018). Me, My Echo Chamber, and I: Introspection on Social Media Polarization. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 823–831. https://doi.org/10.1145/3178876.3186130

Giuffrida, A., Safi, M., & Kalia, A. (2018, December 17). The populist social media playbook: The battle for Facebook, Twitter and Instagram. *The Guardian*. https://www.theguardian.com/world/2018/dec/17/populist-social-media-playbook-who-is-best-facebook-twitter-instagram-matteo-salvini-narendra-modi

Goldhaber, M. H. (1997). The attention economy and the Net. *First Monday*, *2*(4). https://doi.org/10.5210/fm.v2i4.519

González-Bailón, S., & Paltoglou, G. (2015). Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 95–107. https://doi.org/10.1177/0002716215569192

Google Jigsaw. (n.d.). *Perspective*. Perspective API. Retrieved 23 August 2020, from https://www.perspectiveapi.com/#/home

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Groshek, J., & Koc-Michalska, K. (2017). Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign. *Information, Communication & Society*, *20*(9), 1389–1407. https://doi.org/10.1080/1369118X.2017.1329334

Groves, L. R. (1975). Now it can be told: The story of the Manhattan Project. Da Capo Press.

Guan, T., & Liu, T. (2019). Globalized fears, localized securities: 'Terrorism' in political polarization in a one-party state. *Communist and Post-Communist Studies, 52*(4), 343–353. https://doi.org/10.1016/j.postcomstud.2019.10.008

Guillemin, M., & Gillam, L. (2004). Ethics, Reflexivity, and "Ethically Important Moments" in Research. *Qualitative Inquiry, 10*(2), 261–280. https://doi.org/10.1177/1077800403262360

Guzman, A. (2020). Interactive Media Bias Chart. *Ad Fontes Media*. https://www.adfontesmedia.com/interactive-media-bias-chart/

Habel, P., Ounis, I., Macdonald, C., Fang, A., McCreadie, R., & Birch, S. (2015, May 7). Tweeting Britain's #hashtag election. *Washington Post*. https://www.washingtonpost.com/news/monkey-cage/wp/2015/05/07/tweeting-britains-hashtag-election/

Habermas, J. (2004). *The Theory of Communicative Action: Reason and the Rationalization of Society, Volume 1*. Polity Press. https://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=4029981 (Original work published 1981)

Habermas, J. (2015). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Polity Press. https://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=2075499 (Original work published 1962)

Haim, M., Graefe, A., & Brosius, H.-B. (2018). Burst of the Filter Bubble?: Effects of personalization on the diversity of Google News. *Digital Journalism, 6*(3), 330–343. https://doi.org/10.1080/21670811.2017.1338145

Haines, R., Hough, J., Cao, L., & Haines, D. (2014). Anonymity in Computer-Mediated Communication: More Contrarian Ideas with Less Influence. *Group Decision and Negotiation, 23*(4), 765–786. https://doi.org/10.1007/s10726-012-9318-2

Halberstam, Y., & Knight, B. (2014). *Homophily, Group Size, and the Diffusion of Political Information in Social Networks: Evidence from Twitter* (Working Paper No. 20681; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w20681

Hall, S., Hobson, D., Lowe, A., & Willis, P. (2003). Culture, Media, Language: Working Papers in Cultural Studies, 1972-79. Routledge.

Halper, E. (2018, March 21). Was Cambridge Analytica a digital Svengali or snake-oil salesman? *Los Angeles Times*. https://www.latimes.com/politics/la-na-pol-cambridge-analytica-20180321-story.html

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior, 29*(3), 1159–1168. https://doi.org/10.1016/j.chb.2012.10.008

Hansen, A., Cottle, S., Negrine, R., & Newbold, C. (1998). Content analysis. In *Mass communication research methods* (pp. 91–129). New York University Press. https://contentstore.cla.co.uk/secure/link?id=f25ce840-2b53-e511-80bd-002590aca7cd

Harris, M., & Levene, M. (2019, December 18). *Twitter General Election 2019*. Department of Computer Science and Information Systems, Birkbeck, University of London. https://www.dcs.bbk.ac.uk/news/twitter-general-election-2019-2/

Harris, R. (2019, February 26). *Why AI will end Intel's processor dominance*. ZDNet. https://www.zdnet.com/article/why-ai-will-end-intels-processor-dominance/

Haubursin, C. (2018, February 23). *It's not you. Phones are designed to be addicting.* Vox. https://www.youtube.com/watch?v=NUMa0QkPzns

Helberger, N. (2019). On the Democratic Role of News Recommenders. *Digital Journalism*, *7*(8), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

Hern, A. (2019, July 26). Growth in number of users boosts Twitter revenue by a fifth. *The Guardian*. https://www.theguardian.com/technology/2019/jul/26/growth-in-daily-users-boosts-twitter-revenue-by-a-fifth

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter. *Journal of Computer-Mediated Communication*, *18*(2), 154–174. https://doi.org/10.1111/jcc4.12001

Hobolt, S. B., Leeper, T. J., & Tilley, J. (2020). Divided by the Vote: Affective Polarization in the Wake of the Brexit Referendum. *British Journal of Political Science*, 1–18. https://doi.org/10.1017/S0007123420000125

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560

Horowitz, M. A. (2019). Disinformation as warfare in the digital age: Dimensions, dilemmas, and solutions. *Journal of Vincentian Social Action*, *4*(2), 6–21.

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave Your Comment below": Can Biased Online Comments Influence Our Own Prejudicial Attitudes and Behaviors? *Human Communication Research*, *41*(4), 557–576. https://doi.org/10.1111/hcre.12059

Hughes, H. C., & Waismel-Manor, I. (2020). The Macedonian Fake News Industry and the 2016 US Election. *PS: Political Science & Politics*, 1–5. https://doi.org/10.1017/S1049096520000992

Hwang, H., Kim, Y., & Kim, Y. (2016). Influence of Discussion Incivility on Deliberation: An Examination of the Mediating Role of Moral Indignation. *Communication Research*, *45*(2), 213–240. https://doi.org/10.1177/0093650215616861

Inglehart, R. F., & Norris, P. (2016). *Trump, Brexit, and the Rise of Populism: Economic Have-Nots and Cultural Backlash* (HKS Working Paper No. RWP16-026; HKS Faculty Research Working Paper Series). Harvard Kennedy School. https://doi.org/10.2139/ssrn.2818659

Ingram, M. (2018, September 19). *YouTube's secret life as an engine for right-wing radicalization*. Columbia Journalism Review. https://www.cjr.org/the_media_today/youtube-conspiracy-radicalization.php

Iosifidis, P. (2011). The public sphere, social networks and public service media. *Information, Communication & Society*, *14*(5), 619–637. https://doi.org/10.1080/1369118X.2010.514356

Jaume, J. (2014, July 31). *Now Offering Historical Twitter Data Back to 2006*. Brandwatch. https://www.brandwatch.com/blog/now-offering-historical-twitter-data-back-2006/

Jupp, V. (2006). Elite interviewing. In *The SAGE Dictionary of Social Research Methods* (Vol. 1–0). https://methods.sagepub.com/reference/the-sage-dictionary-of-social-research-methods

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin Books.

Kantrowitz, A. (2016, February 10). *Twitter Confirms: Algorithm Coming To Your Timeline*. BuzzFeed News. https://www.buzzfeednews.com/article/alexkantrowitz/time-is-a-construct-anyway

Kantrowitz, A. (2017, January 11). *Why You're Seeing Tweets From People You Don't Follow In Your Timeline*. BuzzFeed News. https://www.buzzfeednews.com/article/alexkantrowitz/twitter-bug-is-inserting-tweets-into-peoples-timelines-from

Kantrowitz, A. (2018, June 21). *How Twitter Made The Tech World's Most Unlikely Comeback*. BuzzFeed News. https://www.buzzfeednews.com/article/alexkantrowitz/how-twitter-made-the-tech-worlds-most-unlikely-comeback

Karatas, D., & Saka, E. (2017). Online political trolling in the context of post-Gezi social media in Turkey. *International Journal of Digital Television, 8*(3), 383–401. https://doi.org/10.1386/jdtv.8.3.383_1

Kastrenakes, J. (2018). Facebook will reduce reach of 'sensationalist and provocative' content. *The Verge*.

Kastrenakes, J. (2020, February 6). *Twitter says AI tweet recommendations helped it add millions of users*. The Verge. https://www.theverge.com/2020/2/6/21125431/twitter-q4-2019-earnings-daily-user-growth-machine-learning

Keck, C. (2019, April 2). *Here's One Small Way We Can Try to Make Twitter Less of a Hell Site*. Gizmodo. https://gizmodo.com/heres-one-small-way-we-can-try-to-make-twitter-less-of-1833756806

Keplinger, K., Johnson, S. K., Kirk, J. F., & Barnes, L. Y. (2019). Women at work: Changes in sexual harassment between September 2016 and September 2018. *PLOS ONE, 14*(7), e0218313. https://doi.org/10.1371/journal.pone.0218313

Khaldarova, I., & Pantti, M. (2016). Fake News. *Journalism Practice, 10*(7), 891–901. https://doi.org/10.1080/17512786.2016.1163237

Kim, H., Ko, E., & Kim, J. (2015). SNS users' para-social relationships with celebrities: Social media effects on purchase intentions. *Journal of Global Scholars of Marketing Science, 25*(3), 279–294. https://doi.org/10.1080/21639159.2015.1043690

Klein, E. (2018a, July 9). *Jaron Lanier's case for deleting social media right now*. https://www.youtube.com/watch?v=o7XigVwjD1Y

Klein, E. (2018b, September 24). *The rise of YouTube's reactionary right*. Vox. https://www.vox.com/policy-and-politics/2018/9/24/17883330/dave-rubin-ben-shapiro-youtube-reactionary-right-peterson

Klein, E. (2020). *Why We're Polarized*. Simon & Schuster.

Knight, W. (2017, April 11). The Dark Secret at the Heart of AI. *MIT Technology Review*. https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/

Kontopantelis, E., Doran, T., Springate, D. A., Buchan, I., & Reeves, D. (2015). Regression based quasi-experimental approach when randomisation is not an option: Interrupted time series analysis. *BMJ, 350*. https://doi.org/10.1136/bmj.h2750

Kormelink, T. G., & Meijer, I. C. (2017). What clicks actually mean: Exploring digital news user practices. *Journalism, 19*(5), 668–683. https://doi.org/10.1177/1464884916688290

Koza, J. R., Bennett, F. H., Andre, D., & Keane, M. A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In J. S. Gero & F. Sudweeks (Eds.), *Artificial Intelligence in Design '96* (pp. 151–170). Springer Netherlands. https://doi.org/10.1007/978-94-009-0279-4_9

Kramer, A. D. I., Guillory, J. E., & 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. https://doi.org/10.1073/pnas.1320040111

Kravetz, L. D. (2017). Strange Contagion: Inside the Surprising Science of Infectious Behaviors and Viral Emotions and What They Tell Us About Ourselves. HarperCollins.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage. https://contentstore.cla.co.uk/secure/link?id=93deea53-2b53-e511-80bd-002590aca7cd

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2017). Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 417–432. https://doi.org/10.1145/2998181.2998321

Lacy, S., Fico, F., Watson, B., & Riffe, D. (2019). *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. Taylor & Francis Group. http://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=5732528

Lane, M. (2006, May 25). The psychology of super-casinos. *BBC News*. http://news.bbc.co.uk/1/hi/magazine/5013038.stm

Lanham, R. A. (2006). The economics of attention: Style and substance in the age of information. University of Chicago Press.

Lanier, J. (2018). Ten arguments for deleting your social media accounts right now. The Bodley Head.

Lantz, B. (2015). Machine learning with R: Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R (Second edition.). Packt Publishing. http://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=2122139

Lavery, D. M. (2020, May 2). Help! My Dad's Facebook Conspiracy Theory Posts Are Driving Me Insane. *Slate*. https://slate.com/human-interest/2020/05/dear-prudence-dad-facebook-conspiracy-theories.html

Lessig, L. (2006). Code: And Other Laws of Cyberspace, Version 2.0. Basic Books.

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lewis, B. (2018). *Alternative influence: Broadcasting the reactionary right on YouTube*. Data & Society. https://datasociety.net/library/alternative-influence/

Lewis, H. (2018, August 29). How Britain's political conversation turned toxic. *New Statesman*. https://www.newstatesman.com/politics/uk/2018/08/how-britain-political-conversation-turned-toxic

Lewis, P., Clarke, S., Barr, C., Kommenda, N., & Holder, J. (2018, November 20). Revealed: One in four Europeans vote populist. *The Guardian*. http://www.theguardian.com/world/ng-interactive/2018/nov/20/revealed-one-in-four-europeans-vote-populist

Ling, C. X., & Sheng, V. S. (2010). Class Imbalance Problem. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 171–171). Springer US. https://doi.org/10.1007/978-0-387-30164-8_110

Linvill, D. L., & Warren, P. L. (2020). Troll Factories: Manufacturing Specialized Disinformation on Twitter. *Political Communication*, *0*(0), 1–21. https://doi.org/10.1080/10584609.2020.1718257

Liu, Y., Kliman-Silver, C., & Mislove, A. (2014). The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior. *International AAAI Conference on Web and Social Media; Eighth International AAAI Conference on Weblogs and Social Media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8043

Lu, Y. Y. (2016, July 28). What Hashtags Reveal. *University of Oxford*. https://www.ox.ac.uk/news-and-events/oxford-and-brexit/brexit-analysis/what-hashtags-reveal

Lyons, B. A., & Veenstra, A. S. (2016). How (Not) to Talk on Twitter: Effects of Politicians' Tweets on Perceptions of the Twitter Environment. *Cyberpsychology, Behavior and Social Networking*, *19*(1), 8–15. PubMed. https://doi.org/10.1089/cyber.2015.0319

Macdowall, C. (2014, January 9). *How twitter is being used in the Scottish independence referendum debate*. Phys.Org. https://phys.org/news/2014-01-twitter-scottish-independence-referendum-debate.html

Madrigal, A. C. (2017, October 12). What Facebook Did to American Democracy. *The Atlantic*. https://www.theatlantic.com/technology/archive/2017/10/what-facebook-did/542502/

Makazhanov, A., & Rafiei, D. (2013). Predicting political preference of Twitter users. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 298–305. https://doi.org/10.1145/2492517.2492527

Manavis, S. (2019, August 13). Why can't we just quit Twitter? *New Statesman*. https://www.newstatesman.com/science-tech/social-media/2019/08/why-can-t-we-just-quit-twitter

Manning, C. D. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mansell, R. (2015). The public's interest in intermediaries. *Info - The Journal of Policy, Regulation and Strategy for Telecommunications*, *17*(6), 8–18. https://doi.org/10.1108/info-05-2015-0035

Margolis, M., & Resnick, D. (2000). *Politics as usual: The cyberspace 'revolution'*. Sage Publications.

Marshall, J., & Lilly, A. (2019). *Parliamentary Monitor 2019: Snapshot*. Institute for Government. https://www.instituteforgovernment.org.uk/publications/parliamentary-monitor-2019-snapshot

Martínez, A. G. (2017, November 18). Facebook Isn't Listening Through Your Phone's Microphone. It Doesn't Have To. *Wired*. https://www.wired.com/story/facebooks-listening-smartphone-microphone/

Martínez, A. G. (2018, February 23). How Trump Conquered Facebook Without Russian Ads. *Wired*. https://www.wired.com/story/how-trump-conquered-facebookwithout-russian-ads/

Martínez, A. G. (2019, March 14). Facebook Is Not a Monopoly, but It Should Be Broken Up. *Wired*. https://www.wired.com/story/facebook-not-monopoly-but-should-broken-up/

Marwick, A., & Lewis, R. (2017). *Media Manipulation and Disinformation Online*. Data & Society. https://datasociety.net/library/media-manipulation-and-disinfo-online/

Masip, P., Ruiz-Caballero, C., & Suau, J. (2019). Active audiences and social discussion on the digital public sphere. Review article. *El Profesional de La Información*, *28*(2). https://doi.org/10.3145/epi.2019.mar.04

Massaro, T. M., & Stryker, R. (2012). Freedom of speech, liberal democracy, and emerging evidence on civility and effective democratic engagement. *Arizona Law Review*, *54*(2), 375–411.

McEwan, B., Carpenter, C. J., & Hopke, J. E. (2018). Mediated Skewed Diffusion of Issues Information: A Theory. *Social Media + Society*, *4*(3). https://doi.org/10.1177/2056305118800319

McKenna, K. Y. A., & Bargh, J. A. (1998). Coming out in the age of the Internet: Identity 'demarginalization' through virtual group participation. *Journal of Personality and Social Psychology*, *75*(3), 681–694. https://doi.org/10.1037/0022-3514.75.3.681

McNair, B. (2018). From Control to Chaos, and Back Again: Journalism and the politics of populist authoritarianism. *Journalism Studies*, *19*(4), 499–511. https://doi.org/10.1080/1461670X.2017.1389297

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Media Bias/Fact Check. (2020). *Media Bias/Fact Check*. Media Bias/Fact Check. https://mediabiasfactcheck.com/

Meredith, J., & Richardson, E. (2019). The use of the political categories of Brexiter and Remainer in online comments about the EU referendum. *Journal of Community & Applied Social Psychology*, *29*(1), 43–55. https://doi.org/10.1002/casp.2384

Mittelstadt, B. (2016). Auditing for Transparency in Content Personalization Systems. *International Journal of Communication*, *10*.

Moore, M., & Tambini, D. (2018). *Digital dominance: The power of Google, Amazon, Facebook, and Apple*. Oxford University Press. http://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=5359105

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *ArXiv:1306.5204 [Physics]*. http://arxiv.org/abs/1306.5204

Mosco, V. (2015). To the Cloud: Big Data in a Turbulent World. Routledge.

MPs on Twitter. (n.d.). Retrieved 24 August 2020, from https://www.mpsontwitter.co.uk/

Muddiman, A., & Stroud, N. J. (2017). News Values, Cognitive Biases, and Partisan Incivility in Comment Sections. *Journal of Communication*, *67*(4), 586–609. https://doi.org/10.1111/jcom.12312

Murru, M. F. (2009). New media – new public spheres? An analysis of online shared spaces becoming public agoras. In N. Carpentier, P. Pruulmann-Vengerfeldt, R. Kilborn, T. Olsson, H. Nieminen, E. Sundin, & K. Nordenstreng (Eds.), *Communicative Approaches to Politics and Ethics in Europe* (pp. 141–153). Tartu University Press.

Music Business Worldwide. (2020, February 5). *Spotify now has 124m Premium subscribers, growing by over 3m per month*. Music Business Worldwide. https://www.musicbusinessworldwide.com/spotify-now-has-124m-subscribers-growing-by-over-3m-per-month/

Nadler, A., Crain, M., & Donovan, J. (2018). *Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech*. Data & Society. https://datasociety.net/library/weaponizing-the-digital-influence-machine/

Napoli, P. M. (2014). Automated Media: An Institutional Theory Perspective on Algorithmic Media Production and Consumption. *Communication Theory*, *24*(3), 340–360. https://doi.org/10.1111/comt.12039

Narayanan, V., Barash, V., Kelly, J., Kollanyi, B., Neudert, L.-M., & Howard, P. N. (2018). *Polarization, Partisanship and Junk News Consumption over Social Media in the US* (Data Memo 2018.1; Computational Propaganda Research Project). Oxford University Project on Computational Propaganda.

Nardelli, A. (2011, May 4). AV on Twitter: How Tweetminster sees it. *The Guardian*. https://www.theguardian.com/news/datablog/2011/may/04/av-twitter-tweetminster

Negroponte, N. (1995). *Being digital*. Vintage Books. https://books.google.co.uk/books?id=LcvR9WHvXmAC

Neheli, N. B. (2018). News By Numbers: The evolution of analytics in journalism. *Digital Journalism*, *6*(8), 1041–1051. https://doi.org/10.1080/21670811.2018.1504626

Nemeth, C. (2020, February 11). How the Twitter algorithm works in 2020. *Sprout Social*. https://sproutsocial.com/insights/twitter-algorithm/

Neuropolitics Research. (n.d.). *#ImagineEurope Twitter Demo*. Neuropolitics Research, University of Edinburgh. Retrieved 23 August 2020, from http://www.pol.ed.ac.uk/neuropoliticsresearch/sections/remote_content

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). *Reuters Institute Digital News Report 2019*. Reuters Institute for the Study of Journalism, Oxford University.

O'Neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Penguin Books.

OSoMe. (2020). *Observatory of Social Media*. Indiana University. https://osome.iuni.iu.edu/

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media &amp; Society*, *20*(9), 3400–3419. https://doi.org/10.1177/1461444817749516

Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. *New Media & Society*, *4*(1), 9. https://doi.org/10.1177/14614440222226244

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin Press.

Parveen, H., & Pandey, S. (2016). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (ICATccT)*, 416–419. https://doi.org/10.1109/ICATCCT.2016.7912034

Pasick, A. (2015, December 21). *The magic that makes Spotify's Discover Weekly playlists so damn good*. Quartz. https://qz.com/571007/the-magic-that-makes-spotifys-discover-weekly-playlists-so-damn-good/

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. https://www-jstor-org.gate3.library.lse.ac.uk/stable/j.ctt13x0hch

Perra, N., & Rocha, L. E. C. (2019). Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific Reports*, *9*(1), 1–11. https://doi.org/10.1038/s41598-019-43830-2

Persily, N. (2017). The 2016 U.S. Election: Can Democracy Survive the Internet? *Journal of Democracy*, *28*(2), 63–76. https://doi.org/10.1353/jod.2017.0025

Pew Research Center. (2017). *Partisan Conflict and Congressional Outreach*. Pew Research Center.

Pfister, D. S. (2018). Public Sphere(s), Publics, and Counterpublics. In *Oxford Research Encyclopedia of Communication*. Oxford University Press. https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-562

Phillips, W. (2015). This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture. Cambridge, Massachusetts : The MIT Press.

Piper, E., MacLellan, K., & James, W. (2019, March 13). Britain in Brexit chaos—Parliament crushes May's EU deal again. *Reuters*. https://www.reuters.com/article/uk-britain-eu-idUSKBN1QT0ZQ

Porter, M. (2006). *Porter Stemming Algorithm*. Tartarus.Org. https://tartarus.org/martin/PorterStemmer/

Posegga, O., & Jungherr, A. (2019, January 8). Characterizing Political Talk on Twitter: A Comparison Between Public Agenda, Media Agendas, and the Twitter Agenda with Regard to Topics and Dynamics. *Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019*. 52nd Hawaii International Conference on System Sciences | 2019. https://doi.org/10.24251/HICSS.2019.312

Prey, R., Valle, M. E. D., & Zwerwer, L. (2020). Platform pop: Disentangling Spotify's intermediary role in the music industry. *Information, Communication & Society*, *0*(0), 1–19. https://doi.org/10.1080/1369118X.2020.1761859

Quattrociocchi, W., Scala, A., & Sunstein, C. R. (2016). *Echo Chambers on Facebook* [SSRN Scholarly Paper]. https://doi.org/10.2139/ssrn.2795110

Rains, S. A. (2007). The Impact of Anonymity on Perceptions of Source Credibility and Influence in Computer-Mediated Group Communication: A Test of Two Competing Hypotheses. *Communication Research*, *34*(1), 100–125. https://doi.org/10.1177/0093650206296084

Rayson, S. (2017, June 26). We Analyzed 100 Million Headlines. Here's What We Learned (New Research). *BuzzSumo*. https://buzzsumo.com/blog/most-shared-headlines-study/

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 field trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *The American Journal of Psychiatry*, *170*(1), 59–70. https://doi.org/10.1176/appi.ajp.2012.12070999

Rocherolle, N. (2019, March 10). The Origin of the Retweet and other Twitter Arcana. *Medium*. https://medium.com/@narendra/the-origin-of-the-retweet-and-other-twitter-arcana-5c53289d9a47

Roth, P. (2004). *The Plot Against America*. Random House.

Rudat, A., Buder, J., & Hesse, F. W. (2014). Audience design in Twitter: Retweeting behavior between informational value and followers' interests. *Computers in Human Behavior*, *35*, 132–139. https://doi.org/10.1016/j.chb.2014.03.006

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *Paper Presented to "Data and Discrimination: Converting Critical Concerns into Productive Inquiry"*, 23.

Schäfer, M. S. (2016). Digital Public Sphere. In *The International Encyclopedia of Political Communication* (pp. 1–7). John Wiley & Sons. https://doi.org/10.1002/9781118541555.wbiepc087

Scharkow, M., Mangold, F., Stier, S., & Breuer, J. (2020). How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences*, *117*(6), 2761. https://doi.org/10.1073/pnas.1918279117

Schroeder, R. (2019). Digital Media and the Entrenchment of Right-Wing Populist Agendas. *Social Media + Society*, *5*(4). https://doi.org/10.1177/2056305119885328

Shirky, C. (2008). Here Comes Everybody: The Power of Organizing Without Organizations. Penguin Publishing Group.

Shirky, C. (2011). The Political Power of Social Media: Technology, the Public Sphere, and Political Change. *Foreign Affairs*, *90*(1), 28–41. JSTOR.

Shore, J., Baek, J., & Dellarocas, C. (2018). Network structure and patterns of information diversity on Twitter. *MIS Quarterly*, *42*(3), 849–872. https://doi.org/10.25300/MISQ/2018/14558

Silveira, T., Zhang, M., Lin, X., Liu, Y., & Ma, S. (2019). How good your recommender system is? A survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, *10*(5), 813–831. https://doi.org/10.1007/s13042-017-0762-9

Silverman, C. (2016, November 16). *This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook*. BuzzFeed News. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

Singer, J. B. (2010). Quality Control. *Journalism Practice*, *4*(2), 127–142. https://doi.org/10.1080/17512780903391979

Singer, J. B. (2014). User-generated visibility: Secondary gatekeeping in a shared media space. *New Media & Society*, *16*(1), 55–73. https://doi.org/10.1177/1461444813477833

Sîrbu, A., Pedreschi, D., Giannotti, F., & Kertész, J. (2019). Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. *PLoS ONE*, *14*(3), e0213246. https://doi.org/10.1371/journal.pone.0213246

Solon, O. (2017, November 9). Ex-Facebook president Sean Parker: Site made to exploit human 'vulnerability'. *The Guardian*. https://www.theguardian.com/technology/2017/nov/09/facebook-sean-parker-vulnerability-brain-psychology

Srnicek, N. (2017). *Platform capitalism*. Polity Press.

Statista. (2020a). *Facebook: Active users worldwide*. Statista. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Statista. (2020b). *The 100 largest companies in the world by market capitalization in 2020*. Statista. https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-capitalization/

Statista. (2020c). *Twitter: Monthly active users worldwide*. Statista. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Steinfeld, N., Samuel-Azran, T., & Lev-On, A. (2016). User comments and public opinion: Findings from an eye-tracking experiment. *Computers in Human Behavior*, *61*, 63–72. https://doi.org/10.1016/j.chb.2016.03.004

Stinson, L. (2016, February 24). Facebook Reactions, the Totally Redesigned Like Button, Is Here. *Wired*. https://www.wired.com/2016/02/facebook-reactions-totally-redesigned-like-button/

Stöcker, C. (2020). How Facebook and Google Accidentally Created a Perfect Ecosystem for Targeted Disinformation. In C. Grimme, M. Preuss, F. W. Takes, & A. Waldherr (Eds.), *Proceedings of MISDOOM 2019: Disinformation in Open Online Media* (pp. 129–149). Springer International Publishing. https://link-springer-com.gate3.library.lse.ac.uk/chapter/10.1007/978-3-030-39627-5_11

Stumpel, M. (2009, October 4). The Habermasian implications of the Twittersphere. *Marc Stumpel*. https://marcstumpel.wordpress.com/2009/10/04/the-habermasian-implications-of-the-twittersphere/

Suler, J. (2004). The Online Disinhibition Effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Sunstein, C. R. (2001). *Republic.com*. Princeton University Press.

Sunstein, C. R. (2008). Neither Hayek nor Habermas. *Public Choice; Dordrecht*, *134*(1–2), 87–95. http://dx.doi.org.gate3.library.lse.ac.uk/10.1007/s11127-007-9202-9

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton : Princeton University Press.

Sunstein, C. R. (2018). Is social media good or bad for democracy? *Sur International Journal on Human Rights*, *15*(27), 83.

Taylor, T. (2018, September 13). *Help, my nan won't stop sharing dodgy posts on Facebook*. ABC News. https://www.abc.net.au/news/health/2018-09-14/health-claims-cancer-cures-talking-family-who-share-on-facebook/10237880

The Economist. (2020, August 1). Twitter's algorithm does not seem to silence conservatives. *The Economist*. https://www.economist.com/graphic-detail/2020/08/01/twitters-algorithm-does-not-seem-to-silence-conservatives

The Policy Institute. (2019). *Divided Britain?* King's College London. https://www.kcl.ac.uk/policy-institute/research-analysis/divided-britain

The UK in a Changing Europe. (2019). *Brexit and Public Opinion*. The UK in a Changing Europe. https://ukandeu.ac.uk/new-report-reveals-brexit-identities-stronger-than-party-identities/

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The Dynamics of Political Incivility on Twitter. *SAGE Open*, *10*(2), 2158244020919447. https://doi.org/10.1177/2158244020919447

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates. *Journal of Communication*, *66*(6), 1007–1031. https://doi.org/10.1111/jcom.12259

Thompson, B. (2018, May 15). The Moat Map. *Stratechery*. https://stratechery.com/2018/the-moat-map/

Toff, B., & Nielsen, R. K. (2018). "I Just Google It": Folk Theories of Distributed Discovery. *Journal of Communication*, *68*(3), 636–657. https://doi.org/10.1093/joc/jqy009

Tolnay, S. E., Beck, E. M., & Sass, V. (2018). Migration and protest in the Jim Crow South. *Social Science Research*, *73*, 13–30. https://doi.org/10.1016/j.ssresearch.2018.03.011

Tufekci, Z. (2015). Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency Symposium Essays. *Colorado Technology Law Journal*, *13*(2), 203–218.

Tufekci, Z. (2016). As the Pirates Become CEOs: The Closing of the Open Internet. *Daedalus*, *145*(1), 65–78. https://doi.org/10.1162/DAED_a_00366

Tufekci, Z. (2017). *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press. http://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=4849027

Tufekci, Z. (2018a, January 16). It's the (Democracy-Poisoning) Golden Age of Free Speech. *Wired*. https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship/

Tufekci, Z. (2018b, March 10). YouTube, the Great Radicalizer. *The New York Times*. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Tufekci, Z. (2018c, August 14). How social media took us from Tahrir Square to Donald Trump. *MIT Technology Review*. https://www.technologyreview.com/2018/08/14/240325/how-social-media-took-us-from-tahrir-square-to-donald-trump/

Turner, G. (2019). Approaching the cultures of use: Netflix, disruption and the audience: *Critical Studies in Television*. https://doi.org/10.1177/1749602019834554

Twitter. (2016, February 10). Never miss important Tweets from people you follow. *Twitter Blog*. https://blog.twitter.com/en_us/a/2016/never-miss-important-tweets-from-people-you-follow.html

van Dijck, J. (2011). Tracing Twitter: The rise of a microblogging platform. *International Journal of Media & Cultural Politics*, *7*(3), 333–348. https://doi.org/10.1386/macp.7.3.333_1

Vargo, C. J., & Hopp, T. (2017). Socioeconomic Status, Social Capital, and Partisan Polarity as Predictors of Political Incivility on Twitter: A Congressional District-Level Analysis. *Social Science Computer Review*, *35*(1), 10–32. https://doi.org/10.1177/0894439315602858

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Votta, F. (2020). *Favstats/peRspective* [R]. https://github.com/favstats/peRspective (Original work published 2019)

Wainwright, D. (2019, April 9). Why don't people vote in local elections? *BBC News*. https://www.bbc.com/news/uk-england-47666080

Waterson, J. (2019, January 23). Brexit boost for BBC Parliament as channel briefly outrates MTV. *The Guardian*. https://www.theguardian.com/media/2019/jan/23/brexit-boost-for-bbc-parliament-as-channel-briefly-outrates-mtv

Wells, C., Shah, D., Lukito, J., Pelled, A., Pevehouse, J. C., & Yang, J. (2020). Trump, Twitter, and news media responsiveness: A media systems approach. *New Media & Society*, *22*(4), 659–682. https://doi.org/10.1177/1461444819893987

Wojcik, S., & Hughes, A. (2019, April 24). Sizing Up Twitter Users. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Wong, F. M. F., Chee Wei, T., Sen, S., & Mung, C. (2016). Quantifying Political Leaning from Tweets, Retweets, and Retweeters. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 2158–2172. https://doi.org/10.1109/TKDE.2016.2553667

World Economic Forum. (2013). *Global Risks 2013 Eighth Edition*. World Economic Forum. http://wef.ch/GJKqei

Wu, T. (2017a). The attention merchants: From the daily newspaper to social media, how our time and attention is harvested and sold. London : Atlantic Books.

Wu, T. (2017b). Blind Spot: The Attention Economy and the Law. *Antitrust Law Journal*, *82*(3), 771–807. https://doi.org/10.2139/ssrn.2941094

Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250–261. https://doi.org/10.1108/AJIM-09-2013-0083

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power (First edition.). PublicAffairs.

Appendices

Appendix I: Fake news, shadowy analytics and filter bubbles

Three aspects of the surveillance capitalism model have received a lot of scholarly and popular scrutiny, particularly following the election of Donald Trump (Madrigal, 2017) and the UK's surprise decision to leave the European Union (Cadwalladr, 2019).

Fake news

Oxford Dictionaries's 2016 word of the year was 'post-truth': "a process whereby objective facts are less influential in shaping public opinion than emotional appeals." (Dahlgren, 2018b) The World Economic Forum (2013) described digital misinformation as one of the main threats facing society. In the closing days of the 2016 US presidential election, *BuzzFeed News* found that fake news stories were outperforming real stories on Facebook (Silverman, 2016).

This has prompted researchers to explore how digital platforms are vulnerable to the spread of falsities, with Del Vicario et al. (2016) showing how cascades of conspiracy rumours are allowed to grow over time on Facebook, and Vosoughi et al. (2018) finding that false news "diffused significantly farther, faster, deeper, and more broadly than the truth" on Twitter, with false political news receiving a more pronounced effect than any other news category.

This is not the same as the rumours, half-truths and outright lies spread by individuals on the internet since its inception — as noted in §1.1.2, due to the lack of journalistic factchecking in online fora — which we might call 'organic fake news'. By contrast, Narayanan et al. (2018) classify 'junk news' when "sources deliberately publish misleading, deceptive or incorrect information purporting to be real news about politics, economics or culture".

This could simply be motivated by pecuniary gain, as in the case of the now-famous teenagers of Veles, Macedonia (Hughes & Waismel-Manor, 2020), or as a matter of geopolitical strategy, as with the efforts of Russia's Internet Research Agency — which Linvill & Warren (2020) describe as a "troll factory" — to interfere with the 2016 US presidential election, or sow confusion during the Ukrainian conflict (Khaldarova & Pantti, 2016). In this context we might say 'junk news' becomes 'disinformation' or 'weaponised fake news'. Horowitz (2019) describes how platforms have become a tool of digital warfare and Stöcker (2020) says outside actors have co-opted them as staging grounds for targeted propaganda campaigns. This seemingly novel and potent tool for outside electoral interference motivates Persily (2017) to ask "can democracy survive the internet?"

Shadowy analytics

Fears of platform-enabled manipulation extend beyond spreading outright propaganda and falsehoods to more subtle and insidious efforts encapsulated by the Cambridge Analytica scandal (Boucher, 2019; Cadwalladr & Graham-Harrison, 2018). Here the unprecedented volume of data Facebook collects on its users' behaviour is capable, so is posited, of creating an accurate 'psychographic' profile of a user — based on the "big five" OCEAN personality traits — so they can be subconsciously manipulated through the platform.

As J. Cohen (2018) explains, this theory is given some credence by Edward Snowden's revelations and Facebook's own infamous study on "emotional contagion", where scientists, in collaboration with the platform, manipulated content in users' feeds to successfully influence their emotional state (Kramer et al., 2014).

Filter bubbles

Because surveillance capitalist platforms are said to be showing us what we want to see, there is much concern that they form digital "echo chambers", supercharging the "selective exposure" phenomenon whereby individuals choose to only consume media that comport with their previously held views. Sunstein (2008, 2017, 2018) is a proponent of the theory, which he ties to group polarisation — where like-minded people move to a more extreme position on an issue after discussing it with each other. A study he co-authored (Quattrociocchi et al., 2016) found evidence on Facebook of echo chambers, selective exposure and homophilous sorting — the propensity of similar individuals to cluster together (Farrell, 2012). Pariser (2011) coined a related term, the "filter bubble", which describes the effect of platform recommendation systems promoting only pro-attitudinal content while filtering out the rest, leaving users blissfully unaware of dissenting opinions.

As well as the aforementioned study, Halberstam & Knight (2014) also find some supporting evidence on Twitter, with users disproportionately exposed to like-minded information. However, these empirical results are the minority. A large Facebook study found individuals' choices of connections "played a stronger role in limiting exposure to cross-cutting content" than algorithmic filtration (Bakshy et al., 2015). A Danish study of a two-week snapshot of Facebook activity reported fewer than a third of users were in topical "filter bubbles" (Bechmann & Nielbo, 2018). Similar studies on Twitter recorded substantial overlap in ideological distributions of accounts followed (Eady et al., 2019) and news links tweeted (Shore et al., 2018), while research on Google News (Haim et al., 2018) and across all three services (Scharkow et al., 2020) found platforms actually increase exposure to diverse news. Finally, survey data from Spain (Cardenal et al., 2019) and the UK (Dubois & Blank, 2018) corroborate these discoveries.

Appendix II: Critiquing the critiques

While empirical evidence varies in support of the three platform dysfunctions in Appendix I, they all miss the mark in one way or another — either through underspecified theory or overemphasis of particular aspects — all of which can be augmented through a more thorough understanding of the attention optimisation logic, elaborated in §1.3. Here I briefly summarise these shortcomings.

Fake news, or: Saving democracy, one factcheck at a time

The hijacking of platforms by external forces to deliberately spread fake news is evidently a problem, but a focus on this issue — as platforms themselves have done with publicised factchecking units (Facebook, 2020; Filloux, 2017) — unwittingly paints a rosier picture of the situation than it at first seems, or, unfortunately, actually pertains. It suggests the issue can be controlled simply by removing demonstrable falsehoods or known bad actors from the platform, whereas in reality, there is a whole swathe of content that cannot earnestly be dismissed as untrue — either because its facts are open to interpretation, or it simply expresses opinion — that is nonetheless damaging to public discourse. The same goes for content that violates hate speech or harassment policies. Borderline content, which goes right up to the edge of policy without crossing the line, can be "most insidious" (Deibert, 2019; Kastrenakes, 2018).

As detailed in §1.3, attention optimisation actually encourages this content, while established media, far from immune, are pulled into the same ecosystem through "distributed discovery" (Toff & Nielsen, 2018) and thus subjected to the same logic. Meanwhile, it is always possible to find facts that support a worldview without lapsing into actual fake news, and the platform ecology supports this human trait of motivated reasoning through selective search (McEwan et al., 2018). Prioritising the abuse of these platforms by bad actors elides the far more pernicious symbiosis platforms share with sources of information that can never be 'factchecked away'.

Shadowy analytics, or: The call is coming from inside the house

Concern over the Cambridge Analytica scandal is misplaced, not because it was an exceptional case, but rather a mundane one. The firm's claims of psychographic profiling were snake oil (Halper, 2018); the services actually rendered to the Trump campaign used "the very tools that were originally built to help … Bed Bath & Beyond sell you towels." (Martínez, 2018) Like the abuse of platforms by Russian information warfare, the focus on Cambridge Analytica centres an anomalous breach in surveillance capitalism at the expense of its operation *by design* — though it's possible for an exogenous agent to use these tools for manipulation, the far more pressing issue is that same manipulation happens automatically, reflexively, without executive instruction, as an endogenous result of a system programmed, somewhat innocently, to maximise our intention. Psychographics and OCEAN personality groupings are only important if humans need to interface with and direct these tools as part of a strategy. The algorithm itself does not need to have any concept of these, it is indifferent to humans (J. Cohen, 2018). In data mining and prediction techniques, the underlying concepts do not need to be reproducible or valid in a diagnostic sense (Döring, 2018); in some cases they are not even visible or known to their designers (Knight, 2017). While Cambridge Analytica was ultimately fraudulent, the same principles can very well operate through the mindless dot connecting of deep learning AI, without any oversight, malevolent or otherwise.

Conspiratorial instincts are unwarranted. However unsympathetic Mark Zuckerberg may be, he is not pulling the levers of his platform with the deliberate aim of keeping a population divided, ineffectual and torpid; these are a logical result of the business model and the automated systems he's put in place to optimise it. Nor is Facebook listening to you through your smartphone — it's both technologically infeasible and entirely unnecessary; the far more unnerving reality is our pedestrian use of its platform is more than enough — in aggregate and at scale (Lanier, 2018) — to make meaningfully efficacious predictions about us to be an indispensable tool for advertisers (J. Cohen, 2018; Martínez, 2017). I emphasise this not to exonerate Silicon Valley leadership, but rather to underscore the responsibility they owe for the immensely powerful systems they have assembled in apparent good faith (J. Cohen, 2018; O'Neil, 2017).

Filter bubbles, or: With bubbles like these, who needs hell sites?

The filter bubble thesis errs in two ways. The first is an error of overdetermination. Fragmentation as a result of abundant media choice has been a fixture of the public sphere since at least the emergence of the internet, if not earlier (Galston, 2003; Papacharissi, 2002). Similarly, the ability to self-select into "ego-centric networks" based upon interest, theme or personal connection, creating "micro-publics" or "public sphericules" (Bruns & Highfield, 2016; Cunningham, 2016), was a feature of *Web 2.0*, acknowledged before the arrival of algorithmic media (Dahlgren, 2005). Finally, homophily is a well-known attribute of offline social networks (McPherson et al., 2001); the human capacity to cluster with those similar to themselves has deep sociological roots for which algorithms can hardly be blamed.

The second is an error of underdetermination. This is to say, while some of the mechanisms of surveillance capitalism operate as envisaged in the filter bubble, it is not the only — or even chief — mechanism. Both Perra & Rocha (2019) and Sîrbu et al. (2019) run simulations to test theories of opinion dynamics under algorithmic conditions, and they find qualified support for a filter bubble effect, but crucially it is dampened by "noise", indicating that other processes running parallel to a selective exposure tendency could lead to different results. This might explain the difficulty researchers have had in empirically verifying the effect. The requirement of holding someone's attention extends beyond simply giving them the expected; algorithm designers acknowledge the importance of serendipity and surprise in recommendations systems (Anzieu, 2019; de Lima & Peres, 2018; Silveira et al., 2019) as value is added by sometimes pushing users out of their "comfort zone" (J. Cohen, 2018). What is curious about the prevalence of the filter bubble theory is how it conflicts with other very common anecdotes of counter-attitudinal experience with social media, such as family members who share conspiracy theories on Facebook (Dreyfuss, 2019; Lavery, 2020; Taylor, 2018), or the description of Twitter as a "hell site" (Keck, 2019; Manavis, 2019). If there was a filter bubble, and only content tailored perfectly to you was allowed through it, social media would be a serene place, or at least an extremely predictable and boring one.

Appendix III: Data collection and case selection

Access to historical tweets

Researchers studying Twitter usually collect data for analysis through the platform's REST and streaming APIs, which provide fast, reliable and efficient access to content and metadata. Unfortunately, full access is charged at a high fee, while the free service is severely throttled in number of requests permitted and the historical period from which tweets can be downloaded (Morstatter et al., 2013). Such restrictions mostly forestall studies of a long time period as researchers typically need to collect data in real-time.

However, several commercial services broker access to Twitter data through economies of scale, and in this study I use the social listening tool Brandwatch which provides an addressable archive of every tweet since 2006, indexed directly from the Twitter Firehose API (Jaume, 2014; Morstatter et al., 2013). This allows me to analyse activity on the platform over a decade to identify any changes manifesting as shifting content patterns over time.

Sampling strategy

As Dahlgren (2005) noted, only a "small degree" of online interaction can be considered deliberative, and a large proportion of tweets in particular have been characterised as "pointless babble" (van Dijck, 2011). This means targeted sampling is unavoidable. Researchers observing political deliberation on Twitter have identified it in several ways: on a content level, by defining *political talk* — the medium of the public sphere (Fraser, 1990) — as discussion referencing certain political topics (Barberá et al., 2015), keywords (Himelboim et al., 2013; Posegga & Jungherr, 2019), hashtags (Conover, Ratkiewicz, et al., 2011), or addressing political figures/organisations (Oz et al., 2018; Theocharis et al., 2016, 2020); and on a user level, by defining *politically engaged individuals* as those who follow or retweet certain political accounts (Kulshrestha et al., 2017), or have explicit reference to their political affiliation in their bio/list membership (Kulshrestha et al., 2017).

Fewer studies have applied these methods to the context of British politics, nor implemented a definition of political talk or politically engaged users that can bear comparisons spanning several years. To do so for this study, I combine some of the methods above in stages. I first created a universe of UK political talk for the last 10 years by assembling a complex search string of political parties, individuals and topics. The former two comprised the 15 political parties receiving at least 0.1% of votes in the 2019 general election, and each of those parties' leaders over the last decade.[12] While the latter topics

[12] All permutations, synonyms and abbreviations for party names were included, e.g. "Conservatives" and "Tories"; "Liberal Democrats" and "LibDem"; "DUP" and "Democratic Unionist Party". To avoid excessive false positives, logical conjunctions with "Party" were

were a list of 263 hashtags derived from contemporaneous sources reporting the most popular hashtags in each of the general elections and referenda held during the period of study.[13]

This search string was then applied to all original tweets (i.e. excluding retweets and replies) in the Brandwatch database from 1 July, 2010 to 30 June, 2020, and returned almost 83 million matches (Figure 2), from which a 5% random sample was downloaded, comprising tweets from 788,231 unique users engaged in political deliberation.

Case selection

This study's longitudinal panel design necessitates cases with a history of activity against which structural changes on the platform can be examined over time, ensuring detected differences result from changes in user behaviour not composition. The sampling frame was therefore limited to users who both:

- had tweets in at least four different years in the political talk results, to ensure semi-regular use over the period study
- registered accounts before July 2010, and have tweeted since June 2020, to ensure activity that spans the period of study.

This narrowed the pool who could possibly be sampled to 25,316 users, which was reduced further by setting a floor of 100 tweets and followed accounts, and 25 followers — to identify genuine active accounts, as do Barberá et al. (2015) — and setting a ceiling of the 90th percentiles for the same metrics — to avoid sampling news organisations or celebrities, following Aghababaei & Makrehchi (2017). Following these measures, the finalised sampling frame stood at 18,968 users. Users needed to be checked upon sampling to verify their tweets were still public, and so for a target sample of 1,200 users, 1,300 were sampled, with 1,228 ultimately retained. Every tweet by these accounts from 1 July, 2010 to 30 June, 2020 was then downloaded from Brandwatch, comprising a final corpus of 22,593,965 tweets.

I have taken care to address the complaint of Cihon and Yasseri (2016) regarding "insufficiently defended" filtering decisions by justifying mine with reference to the research design and past literature.

exclusively used when necessary — e.g. "Green Party" and "Brexit Party" — while full names were used for party leaders.

[13] See Appendix VIII for the full list of hashtags and their sources.

Appendix IV: Computational classification of incivility

To identify incivility at the scale of this study's corpus required the use of computational methods. A prior pilot study implemented 'dictionary method' text analysis, whereby each document is assessed against a preselected lexicon — grouped by a variable of interest — producing a value based on the relative occurrence of instructive words (Grimmer & Stewart, 2013). This method benefits from being relatively simple to apply, providing a suitable, validated dictionary is available. However, there is an acute limit to how far it can be generalised beyond the precise domain of the dictionary's construction, and so it is perhaps unsurprising that for a concept as diffuse as incivility, and domain as dynamic as social media, the dictionary method had disappointing reliability measured against human coders.

For this study then, I instead opt for a supervised machine learning approach, where an algorithm is "trained" on a dataset labelled with the variable of interest by human coders. In this process, the algorithm "learns" patterns of cooccurrence in the text associated with different values of the given variable, then applies these to a much larger unseen dataset. As Grimmer & Stewart (2013) explain, supervised learning is "necessarily domain specific" since classification criteria are derived directly from the corpus. The methodological literature shows this approach outperforms dictionary methods on Twitter data (González-Bailón & Paltoglou, 2015), and has demonstrated identifying incivility at close to human coder accuracy (Gervais & Chin, 2018; Pew Research Center, 2017; Theocharis et al., 2016, 2020).

Hand-coded training set

To prepare a training dataset, 5,000 tweets were selected from the corpus, stratified evenly either side of algorithmic introduction (Figure 3), but otherwise randomly sampled. Barberá et al. (2020) identify this as the optimal sample size, beyond which gains in accuracy sharply diminish. Labelling of incivility in the training sample was divided between two human coders, who each coded 2,200 tweets independently, along with a subsample of 600 tweets which were double coded for validation. These 600 tweets were randomly sampled evenly from crosscutting strata — before and after algorithmic introduction; across original tweets, retweets and replies — exceeding the recommended sample for intercoder reliability (ICR) for each substrata (Krippendorff, 2004; Lacy et al., 2019).[14]

[14] For $N = 5{,}000$ documents, minimum reliability agreement of 85% and a 95% confidence level:

$$n = \frac{(N\text{-}1)(SE)^2 + PQN}{(N\text{-}1)(SE)^2 + PQ} = \frac{(5000\text{-}1)\left(\frac{.05}{1.64}\right)^2 + .09(5000)}{(5000\text{-}1)\left(\frac{.05}{1.64}\right)^2 + .09} = 95.99$$

A coding scheme for incivility heavily inspired by Theocharis et al. (2020) was implemented (Table 2), operationalising the concept as a dichotomous variable: a tweet is either CIVIL or UNCIVIL. The confusion matrix from human coding is presented in Table 3, demonstrating the 88% agreement achieved. Cohen's kappa — a suitable measure of intercoder reliability for two coders — was calculated at .64, considered within the "very good" range (Regier et al., 2013). During coding, 70 tweets were removed as they either included no text (only an image or URL), were not in English, or were otherwise unintelligible. Of the remaining 4,930 tweets, 21% were labelled UNCIVIL.

Supervised learning

Using `quanteda` — a quantitative text analysis package in the `R` statistical programming language (Benoit et al., 2018; Lantz, 2015; Porter, 2006) — the tweet corpus was tokenised, converted to lowercase and stemmed into unigrams, before extremely common "stop words" were removed. This allowed the creation of a document-feature matrix (dfm) — a quantitative representation of the corpus where each tweet (document) represents a row, and each feature — a variable tallying tokens of the same type — a column.

To build the machine learning classifier, this matrix was passed through the `glmnet` package (Friedman et al., 2020) to train a logistic "lasso" regression — or L1 regularisation — using five-fold cross-validation to establish the best lambda (λ) value. Regularised regression uses a penalty parameter, λ, to prevent a model from overfitting to training data and making poor predictions when applied to new inputs. By "folding" the training data into five groups, the algorithm can test different values for λ and select whichever produces the most accurate predictions against the human-coded variable.

Results from the classifier were initially disappointing, which is quite common when one value of a dichotomous training sets is significantly overrepresented — known as the "class imbalance problem" (Ling & Sheng, 2010). To synthetically balance the training set, I ran the existing 4,930 tweets, along with 15,000 new tweets identically sampled from the corpus, through Google's Perspective API via the `peRspective` wrapper for `R` (Votta, 2019/2020).[15] This provides scores for negative language features like "toxicity", "profanity" and "insult" (Google Jigsaw, n.d.). Using these Perspective scores as features, I fitted a new classifier to the 4,930 hand-coded labels, achieving 85% accuracy, before predicting labels for the other 15,000, returning 3,120 new uncivil tweets, which were added to the training set. Finally, using this newly balanced training set of 8,050 tweets and incivility labels, I retrained the initial classifier.

[15] This, along with the preceding steps for the supervised method, closely follows Theocharis et al. (2020) — I am indebted to them and to Blake Miller for suggesting their research.

Validation

1,610 of the hand-coded tweets — 20% of the expanded training set — were withheld at random from the estimation, providing a test set to appraise the classifier's performance. Total accuracy was 82%, with precision and recall for the CIVIL class at 76% and 95%, while for UNCIVIL they were 93% and 67% respectively. This is a level of classification similar to previously published studies utilising the same process (Davidson et al., 2017; Gervais & Chin, 2018; Theocharis et al., 2016, 2020) and approaches human interrater agreement. If it can be criticised, it is for erring on the conservative side in *under*estimating the extent of incivility. To further ensure this process captured the desired concept, unigrams the model found most predictive of civility and incivility are presented in Table 4. The array of insults and profane language under incivility contrasted with "love", "outstanding", "appreciate" and the heart emoji confirms the classifier identified my dimension of interest.

Appendix V: Construction of control variables

Theoretically, the explanatory factor I am interested in is the Twitter timeline algorithm, however this cannot be observed directly, and so time before and after its introduction operates as my independent variable. To make a compelling case that any temporal correlation with the dependent variable is genuine rather than spurious, I need a comprehensive accounting of other plausible covariates. Multiple linear regression — and its generalised cousins — can achieve this through control variables. If any relationship between time and incivility is robust to the inclusion of convincing measures of other potentially related factors, then I will have a persuasive case. Below I specify such measures.

Political atmosphere

British politics, like much of the democratic world, has been characterised by an increasingly febrile atmosphere over the last five years, among a rising tide of populism and deepening polarisation (Inglehart & Norris, 2016; Klein, 2020; P. Lewis et al., 2018). In the UK this is inextricably linked to Brexit, with sides in the EU referendum surpassing party affiliation in salience as political identities (Evans & Schaffner, 2019; Hobolt et al., 2020; Meredith & Richardson, 2019; The UK in a Changing Europe, 2019). Unhelpfully, the Brexit vote coincides with the intervention under analysis: introduction of Twitter's timeline algorithm. To control for the heightened political atmosphere after Brexit, I provide three possible instruments.

The first two use the electoral schedule as a proxy. The most concrete consequence so far of Brexit has been political instability, with a nation once known for stable government (Erlanger, 2017) thrown into two early elections in under three years. Furthermore, the cracks of political precariousness arguably predate the decision to leave the European Union, with a hung parliament, two UK-wide referenda in five years — in a country that had previously held just one, 36 years ago — and a third in Scotland of existential consequence to the polity as currently constituted. It stands to reason then that the unusual volume of electoral events will have a relationship — whether *ex ante* or *ex post* — with the nation's political tenor.

The ELECTION PROXIMITY variable attempts to capture this numerically: at any given point in time, it will be the smallest of either the amount of time since the most recent election, or the amount of time until the next election, meaning the variable is at its largest at the middle point between two elections — what Theocharis et al. (2020) call the "quieter" periods — and is equal to zero when an election takes place during that unit of time. A version of the variable was calculated with only the three general elections that took place during the period of study, and another version added the 2011 AV, 2014 Scottish independence and

2016 EU membership referenda.[16] The variable is calculated on the basis of the next *planned* election/referenda, and so changes when a referendum or snap election is called — I take this point to be when parliament passed legislation setting the date of the vote. This leads to some sharp changes in value which ought to estimate changes in political tension.

Importantly, ELECTION PROXIMITY treats the days immediately before and after an election homogenously, when it may be the case that the final days of a campaign and the first days in its aftermath are qualitatively different as regards levels of incivility, *ceteris paribus*. The SINCE ELECTION variable responds to this intuition by increasing linearly and only resetting to zero when an election takes place during that unit of time, thus the days before and after an election are of opposite magnitude. The expectation, consistent with Theocharis et al. (2020) is that each of these variables will be negatively correlated with incivility, with moments of high political tension (and thus a low ELECTION value) related to more uncivil tweets, but as Figure 4 shows, this only holds in one of the four cases. This is discussed in more depth in §6; for now, another measure is needed.

It was noted in early 2019, as parliamentary Brexit debate reached fever pitch, that the number of people watching the BBC Parliament channel briefly surpassed MTV (Waterson, 2019). Indeed, the Institute for Government (Marshall & Lilly, 2019) reported that record numbers watched parliamentary events on TV and online during the Brexit crisis. I was able to acquire weekly viewing figures for the channel — included in this analysis as the BBC PARLIAMENT variable — from 2010 until January, 2020 (BARB, n.d.). They show a 0.63 *SD*s increase since algorithmic introduction, averaging 1.84 *SD*s throughout 2019 and peaking at 13.3 *SD*s in the week of Boris Johnson's first prime minister's questions, having reached 10.9 *SD*s at the crescendo of the Brexit crisis (Piper et al., 2019). Crucially, as Figure 5 shows, the viewing figures exhibit the same correlation as expected between rising political tension and increased incivility, as well as a similar relationship to the absolute volume of tweets from politically engaged users. This makes BBC PARLIAMENT a suitable proxy to statistically control for the heightened political atmosphere since 2016.

Partisanship and media diet

The link between partisanship and political incivility has been well investigated by scholars, particularly the influence political elites have over supporter incivility (Gervais, 2019), the cooccurrence of partisan and uncivil language in online news comments (Muddiman & Stroud, 2017), and even the relationship between heighted partisan conflict and incivility on Twitter (Vargo & Hopp, 2017). Relatedly, research shows an association between an individual's partisan media diet and their "propensity to use incivility in textual political expression" (Gervais, 2014).

[16] Local elections were excluded because of low turnout (Wainwright, 2019).

This relationship could cause problems for my analysis if, for argument's sake, as UK political polarisation has coarsened over time, more partisan voices — with a greater uncivil inclination — have intensified their output, while more moderate, civil voices have been cowed into relative silence. This would present the appearance of a temporal correlation that would in fact be spurious, or at least underdetermined. Controls for partisanship are therefore necessary in building my case.

Thankfully there is a rich literature on inferring the partisan valence of Twitter users, from utilising semantic analyses of the content of users' tweets (Conover, Goncalves, et al., 2011; Makazhanov & Rafiei, 2013), their membership of partisan lists (Kulshrestha et al., 2017), the accounts they retweet (Wong et al., 2016), and the connections within their network (Barberá et al., 2015). These each represent fully-fledged research projects in their own right, and so under resource and time constraints, I implement a related but simplified method to Barberá et al. (2015), estimating the degree of partisanship and bias in media diet of my sample users on the basis of the accounts they follow.

First, a comprehensive register of the Twitter accounts of elected UK politicians, past and present, was assembled from sources such as *MPs on Twitter* (n.d.) and various Twitter lists. This accounting totalled 972 politicians, spanning several parliaments, and the years an MP first and last stood for election were recorded alongside.[17] All past and present party leaders, along with their years of leadership, were also included, whether they have been elected to parliament(s) or not.[18] Official party accounts were all included too, as well as those of affiliated campaign groups — like Momentum or Conservative Way Forward — with at least 5,000 followers, found through manual Twitter searches and researcher knowledge.

This total of 1,100 partisan accounts were each assigned a left–right score, derived from the Chapel Hill Expert Survey (Bakker et al., 2020) which has polled hundreds of political scientists since 1999 on the political positioning of parties across Europe. Survey scores were assigned to the nearest general election year with a different score for each election since 1997. The change in these scores over time is presented in Figure 14. These scores were assigned to the political accounts according to the following principles:

[17] UK MEPs from the 2009, 2014 and 2019 European Parliaments were also included, to ensure parties underrepresented in Westminster could still form part of the calculation.

[18] The Twitter accounts of the Tony Blair Institute for Global Change (@InstituteGC) and the Office of Gordon & Sarah Brown (@OfficeGSBrown), along with a well-followed Margaret Thatcher tribute account (@MrsMThatcher), were included as proxies for their respective politicians despite not being personal accounts, on the basis that, as former prime ministers, a user following their accounts will be particularly instructive.

- Politicians inherited the score of their party in the last year they stood for election as a member of that party, whether they were elected or not, with the following exceptions:
    - If the politician was previously leader of the party, they inherited the nearest score of the party during their leadership (taking an average score over the range of election years they were leader if more than one), e.g. Ed Miliband and Iain Duncan Smith inherited scores from 2015 and 2001 respectively despite standing for election in 2019
    - If the politician resigned from a party and/or defected to another party and then stood for election, they inherited the score of their new party in the year they last stood for election, e.g. Luciana Berger inherited the Liberal Democrat score from 2019 rather than the Labour score of 2017
    - If the politician resigned, or had the whip removed, from a party but either did not defect to another party, defected to a party without available scores, did not stand for election, or stood as an independent, they inherited the score of their previous party when they first stood for election, e.g. Chris Leslie and Dominic Grieve inherited scores from 1997 despite standing for election in 2017 under their original party
    - If the year a politician should have inherited a score from based on the above rules fell outside of the available range, an average score across all years for the party was used
- Party and affiliated group accounts all inherited the most recent score for their party, with the following exceptions:
    - If the group is a known supporter of a particular era of leadership from the party's past, it inherited the (average) score for the year(s) of that leadership, e.g. the Blairite Progress group and the Cameroon Tory Reform Group inherited scores between 1997–2005 and 2010–2015 respectively
    - If the group is a leadership campaign for a candidate associated with a particular era of leadership from the party, it inherited the (average) score for the year(s) of that leadership, even if it causes a conflict with the individual politician's latest score, e.g. As an MP who stood for Labour in 2019, Liz Kendall MP's account inherited a score from Jeremy Corbyn's leadership, whereas the Liz for Leader account, from her 2015 leadership bid as a recognised Blairite candidate, inherited a score between 1997–2005.

The Chapel Hill scores were expressed between 0 ("extreme left") and 10 ("extreme right"). By averaging the scores of all parties, weighted by their share of the 2019 general election vote, a UK "political centre" score was calculated of 4.78 with a quasi-*SD* of 2.41. From these, each political account's score was standardised as a *z*-score, where 0 = the "political centre".

The follow lists for the sampled users, as of 1 July, 2020, were checked against these political accounts, and for each user, the average score of their followed accounts was taken. In the manner of a slightly modified *t*-test[19], this average score was normalised by both the *SD* of the user's followed accounts, the UK quasi-*SD*, and the square root of the total of followed accounts, in order for single accounts to not skew a user's score too far, and for more followed accounts to increase confidence in a left or right position, while accounting for diminishing returns. A weighted version was also calculated with political account scores divided by the square root of the number of other accounts with the identical score, accounting for the fact there are simply far fewer Liberal Democrat or Brexit Party politicians to follow than Labour or Conservative, therefore a follow of a politician from a smaller cohort should carry more informative weight.

This forms the PARTISAN SCALE variable, and to check its validity, a stratified sample was drawn from users in the sampling frame with one of the political parties' names in their bio, taken as a declaration of support, and PARTISAN SCALE calculated for the sample. Boxplots of the results by supported party are presented in Figure 6, where the positions of users adhere quite closely to those of their respective parties, particularly relative to others.

A similar process was followed for the MEDIA SCALE. A comprehensive list of news media accounts was assembled in several stages. Firstly, every official account from all the UK's major broadcasters and national newspapers were included. This was achieved through manual search, and with the aid of Twitter lists that some news organisations provide of all their accounts. The BBC had far and away the most accounts with 71, followed by *The Guardian* with 17, ITV News with 16, through to — by way of example — *The Times* with 7, *Daily Mail* with 5 and *Daily Express* with 2. This process was followed for news magazines like *The Economist* and radio stations like LBC. Many foreign news organisations have international offerings, not least in the UK, and so this process was repeated for major global news brands such as *The New York Times*, CNN and Agence France-Presse. Then to ensure not only established media were included, new media organisations were collected using a "snowball sampling" method (Cihon & Yasseri, 2016), beginning with the accounts of *The Canary* on the left, and *Guido Fawkes* on the right, I was guided by Twitter's "You might like" feature which suggests other accounts based on similar network profiles. Finally, to avoid missing any significant accounts, the follow lists used to validate the PARTISAN SCALE were cross-referenced, with accounts followed by at least 5 users but not

[19] The quasi-*SD* is added to the *t*-test for a single mean to prevent the denominator from equalling zero when the *SD* of the scores of followed accounts equals zero:

$$\text{PARTISAN SCALE}_u = \frac{\overline{lr_u} - \overline{lr_{ge}}}{(s_{lr_u} + (s_{lr_{ge}})/\sqrt{n_{lr_u}}} \text{ where } \overline{lr_u} = \frac{\sum_{i=1}^{n} lr_{p,u}}{n}$$

and where *u* is a user; *p,u* is a political account *u* follows; *lr* is a left–right score, and *ge* is the general election.

yet included in the news media list manually validated and added if meeting the criteria of being a media organisation providing news and/or analysis or opinion on current affairs. In total, 406 news media accounts from 182 organisations were collected.

To establish scores analogous to the PARTISAN SCALE, several established media bias databases and sources were consulted (AllSides, 2020; Flood, 2019; Guzman, 2020; Media Bias/Fact Check, 2020). Scores were standardised to a five-point scale, -2 for "left-wing", +2 for "right-wing", and 0 for "centre" or "least biased". Where more than one source had scored a news organisation a rounded average was used, and where scores were available, researcher judgement was used. As with the PARTISAN SCALE, the scores were relativised, in this case by averaging the scores of all media accounts, weighted by their number of Twitter followers, calculating a Twitter "media centre" of -.47 with a quasi-*SD* of .93, demonstrating the large presence of left-of-centre news organisations on Twitter. From these, each media account's score was standardised as a *z*-score, where 0 = the "media centre". This meant organisations with the most balanced editorial stances, like the BBC or Reuters, had a score slightly right-of-centre in Twitter terms.

The MEDIA SCALE was then calculated by the same process as the PARTISAN SCALE[20], again including a weighted version, and the results were validated through the same method, with users showing similarly satisfactory positions relative to their parties (Figure 6).

Theory and existing evidence dictate these variables should have a positive relationship with incivility, but since partisanship of either polarity is associated with increased incivility, it ought to be a non-linear, U-like relationship. The scatterplots and generalised additive model (GAM) trendlines in Figure 7 confirm this is indeed the case, at least within the bulk of the data. The main exception is the weighted PARTISAN SCALE where a more right-wing follow network is unequivocally associated with higher incivility, whereas more left-wing followed accounts are at first associated with more uncivil tweets, but the relationship reverses after a point. This is at the outer range of the data, so firm conclusions should not be drawn, but a possible explanation is that following a lot of Labour MPs is simply not indicative of particularly left-wing partisans, who instead follow more figures who would not be captured by this model. Both versions of the MEDIA SCALE lend credence to this view, as they exhibit a more straightforward U-shape relationship with incivility. It may be that media diet is a more reliable measure of partisanship on Twitter.

[20] MEDIA SCALE$_u$ = $\frac{\overline{lr_u} - \overline{lr_{tw}}}{(s_{lr_u} + (s_{lr_{tw}})/\sqrt{n_{lr_u}}}$ where $\overline{lr_u} = \frac{\sum_{i=1}^{n} lr_{m,u}}{n}$

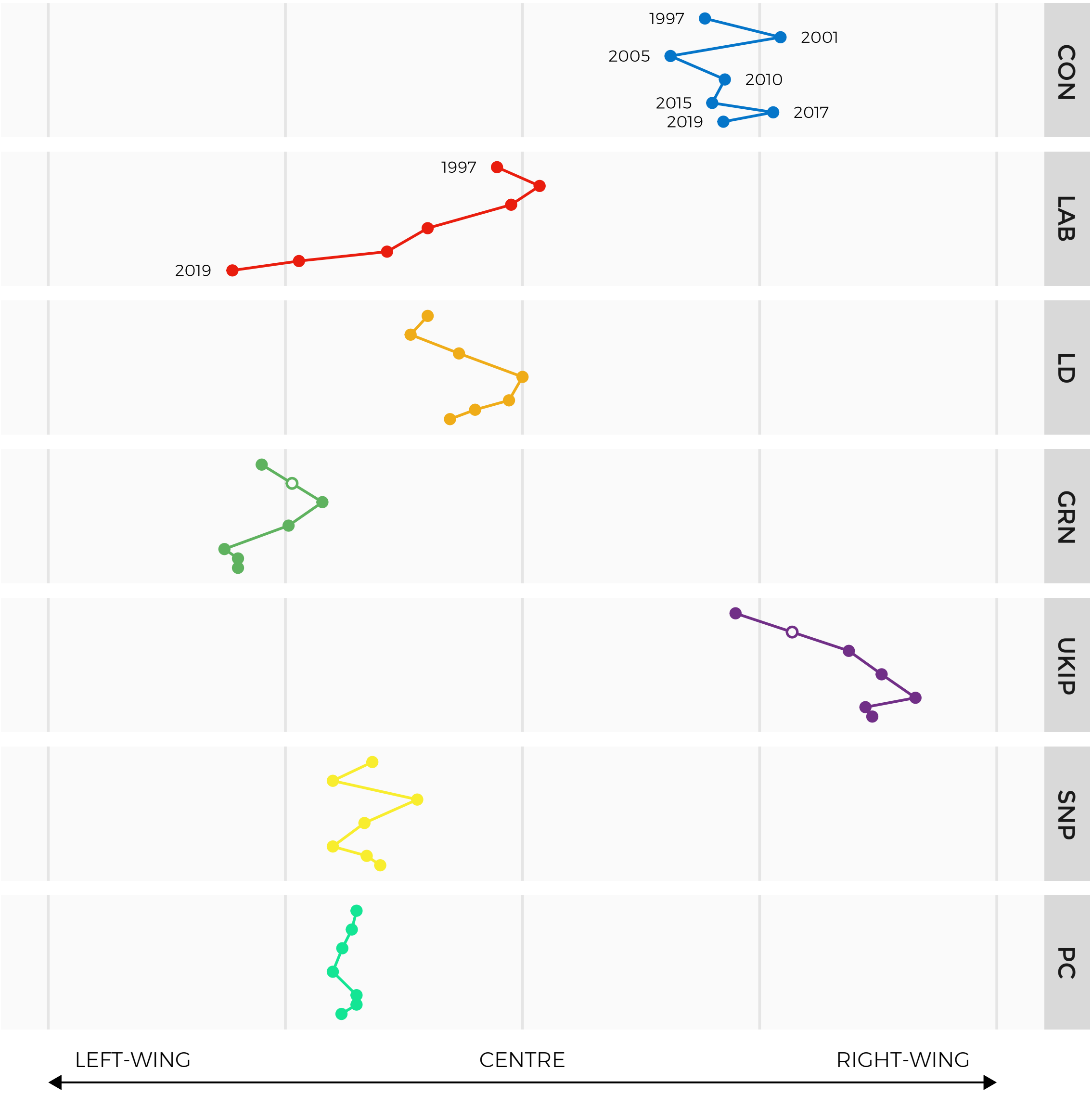
and where *u* is a user; *m,u* is a media account *u* follows; *lr* is a left–right score, and *tw* is Twitter.

Political engagement

Ideological valence is just one political characteristic that may influence an individual's level of incivility. It is quite plausible for someone moderately positioned on the political spectrum to be nonetheless aggressive in their civic comportment. Indeed, this has been observed about sections of online pro-EU support in resistance to Brexit (D. Cohen, 2019; H. Lewis, 2018; Meredith & Richardson, 2019), while Cardenal et al. (2019) find users with the highest news consumption are most polarised. The next three variables attempt to estimate an individual's level of political engagement aside from ideological orientation. The first two — PARTISANS FOLLOWED and MEDIA FOLLOWED — are simply the absolute number of accounts followed that were used to infer the previous scales, ignoring left–right alignment. The third — NON-PARTISANS FOLLOWED — is the number of neutral political accounts followed (e.g. official governmental and parliamentary accounts, local councils, etc.). It should be noted that while users following fewer than two partisan or media accounts were rejected from the sample; no such minimum was placed on non-partisan accounts.

Twitter usage

Finally, it may be that heavier usage of Twitter, or longer exposure to the platform, are associated with higher incivility. In such a scenario, it might be the case that rising incivility over time was simply the result of more users having longer to be influenced enough by the service in general to become more uncivil in their use of it. A user's absolute number of FOLLOWERS (at time of sampling) and total number of in-sample tweets (ALL TWEETS) each month are included to account for individual level of use.



NOTE: ○ = No score available, score imputed from surrounding values. Election year score taken from nearest survey year. Sinn Féin, DUP and the Brexit Party excluded from figure as scores only available for single year; available score used across all years to calculate PARTISAN SCALE, except for Sinn Féin where ROI scores were used. Labour scores used for SDLP, Conservatives scores for UUP, Liberal Democrats scores for the Alliance Party, and Green Party of England and Wales scores for Scottish Greens and Green Party in Northern Ireland.

Figure 14: UK political parties' left–right scores over time from Chapel Hill Expert Survey data (Bakker et al., 2020) used to calculate PARTISAN SCALE.

Appendix VI: Calculation of tweet type

This analysis will not rely solely on time — controlling for omitted variables — as a measure of the hypothesised algorithmic phenomenon (q.v. §3.3). I propose certain categories within a typology of tweet type can also be used as a proxy for algorithmic influence on Twitter. Below I detail their calculation.

Main sample

Retweets — where a user shares a tweet from another account with their own followers — are labelled as such in tweet metadata, allowing them to be easily filtered within a dataset. A retweet object contains both the user IDs of the retweeter and the author of the original tweet. Every retweet in the sample was processed by its retweeter — one of the 1,228 sampled users — checking the original tweet author ID against the follow list of the retweeter. This divided retweets into FOLLOWING RETWEETS (FRTs) when the original tweet author did appear on the retweeter's follow list, and NOT-FOLLOWING RETWEETS (NFRTs) when they did not, the distribution of which, along with other tweet types, is presented in the first column of Table 5.[21]

This is an imperfect measure. Since user follow lists could only be assessed retrospectively, after the period of analysis, it cannot be known whether an individual was following a user when retweeting them, only that they were not following them at the point of sampling. To account for this known measurement error, I include three control measures of individual follow behaviour: FOLLOWING GROWTH, the β-coefficient from a simple linear regression of a user's daily following count on time in days; FOLLOWING GROWTH (*SE*), that β's standard error; and FOLLOWING VOLATILITY, the standard deviation of a user's following count over time. The rate at which an individual follows other users might bias the NFRT measure since, following someone *after* retweeting causes an undercount. The statistical significance of any of these terms should indicate if this error is systematic, and if so, control for it.

Subsample

Acknowledging that the timeline algorithm is not the only way a user could come to retweet content from an account they do not follow, a further retweet type distinction was made *within* NFRTs: between INTRA-NETWORK RETWEETS (INRTs), where the retweeting user does not follow the author of the original tweet, but does follow someone who has retweeted it; and EXTRA-NETWORK RETWEETS (ENRTs), where the retweeting user follows neither the author of the original tweet, nor anyone else who has retweeted it. The distinctions between these tweet types are illustrated in Figure 8. To calculate these retweet types, every account

[21] Unfortunately, reliable retweet metadata is not available in my dataset before November 2013, due to changes in the way Brandwatch recorded tweets from the Twitter API. For the portion of the analysis involving tweet type, the dataset is limited to tweets from beyond this point, still comprising exactly two thirds of the whole data, including over two years prior to algorithmic introduction.

that has also retweeted an original tweet needs to be checked to see whether any are followed by the retweeting user. With the resources available this was unrealistic for the main sample. A representative subsample of tweets (*N* = 139,735) were randomly selected, including 24,819 NFRTs, and every other retweet of the same original tweets were retrieved from the Brandwatch database. These were cross-referenced with the follow lists of the sampled users, tallying the number of retweets of the same tweet by accounts the sampled user follows. On this metric, 33.2% of the sampled retweets totalled zero and were labelled ENRTs, while the rest were designated INRTs. The distribution of these tweet types, along with the other tweet types in the subsample, is presented in the second column of Table 5.

Real-time sample

While a more rigorous measure than NFRTs, ENRTs still suffer the same retrospective measurement error. As a further robustness check of the tweet type classification, a secondary data collection was conducted on a new sample of 200 users, randomly drawn from the same sampling frame as the main sample, however these tweets were collected in real-time between 4–11 August, 2020, and checked against follow lists that were updated daily between 12–4am BST, providing a far more precise measure of retweets from beyond a user's network. The tweet type distribution is compared to the other samples in Table 5, to which they are markedly different, but consistent with the trend over time for each type, as shown later in Figure 11 (§6).

Appendix VII: Comparison of balanced and unbalanced panels

Table 17: Comparison of unbalanced regression models (11–12) from Table 9 with balanced panel regression models (28–29) where individuals with missing data have been discarded.

	DEPENDENT VARIABLE			
	% UNCIVIL ALL TWEETS			
COVARIATE	12	11	28	29
CONSTANT	19.04*** (.40)	22.62*** (.40)	18.21*** (.73)	2.89*** (.73)
TIME (*t*)	.01*** (.002)	-.06*** (.002)	-.02*** (.002)	-.07*** (.003)
ALGORITHM	1.00*** (.12)	-15.40*** (.33)	1.16*** (.15)	-14.25*** (.42)
***t* × ALGORITHM**		.22*** (.004)		.20*** (.01)
	POLITICAL ATMOSPHERE (TIME-VARIANT)			
BBC PARLIAMENT	.003*** (.000)	.002*** (.000)	.004*** (.000)	.003*** (.000)
	PARTISANSHIP, MEDIA DIET, POLITICAL ENGAGMENT AND TWITTER USAGE (UNIT-VARIANT)			
PARTISAN SCALE	.59 (1.35)	.57 (1.35)	.84 (2.66)	.84 (2.66)
PARTISAN SCALE^2	4.13 (2.96)	4.23 (2.96)	4.64 (6.30)	4.61 (6.30)
MEDIA SCALE	-47.49 (47.39)	-46.19 (47.36)	-10.54 (86.92)	-10.13 (86.92)
MEDIA SCALE^2	18,195*** (3,308)	18,247*** (3,306)	13,474** (6,195)	13,463** (6,195)
POLITICIANS FOLLOWED	.02*** (.01)	.02*** (.01)	.02* (.01)	.02* (.01)
MEDIA FOLLOWED	-.01 (.01)	-.01 (.01)	-.02 (.02)	-.02 (.02)
NON-PARTISANS FOLLOWED	-.28*** (.03)	-.28*** (.03)	-.23*** (.06)	-.23*** (.06)
FOLLOWERS	-.001*** (.000)	-.001*** (.000)	-.001*** (.000)	-.001*** (.000)
ALL TWEETS	.004*** (.000)	.004*** (.000)	.01*** (.000)	.01*** (.000)
INDIVIDUALS (*n*)	1,228	1,228	309	309
OBSERVATIONS (*N*)	123,418	123,418	35,535	35,535
***R*^{2}**	.02	.04	.03	.07
ADJUSTED *R*^{2}	.02	.04	.03	.07
***F*-STATISTIC**	2,808***	5,808***	1,036***	2,623***

NOTE: Standard errors in parentheses. Time in months. ALL TWEETS includes retweets and replies. ALGORITHM dummy coded '1' after January 2016.

* $p < .10$ ** $p < .05$ *** $p < .01$

Appendix VIII: Hashtags used to define UK *political talk*

The most popular political hashtags used in the UK over the last decade was researched and queried, covering the 2015 (Habel et al., 2015), 2017 (Cram et al., 2017) and 2019 (M. Harris & Levene, 2019) general elections, as well as the 2011 AV (Nardelli, 2011), 2014 Scottish independence (Macdowall, 2014), and the 2016 EU membership referenda (Bush, 2018; D. Cohen, 2019; Cram, 2015; Lu, 2016; Neuropolitics Research, n.d.). All 263 are below:

#2ndRef
#2ndReferendum
#abtv
#andrewneil
#article50
#BackBoris
#BattleForNumber10
#bbcdebate
#BBCDP
#BBCElection
#bbcqt
#BBCSP
#beleave
#betteroffin
#betteroffout
#bettertogether
#BishopAucklandFarmer
#BollocksToBrexit
#BolloxToBrexit
#bregret
#Bremain
#brexit
#BrexitCountdown
#brexitfears
#brexitin5words
#BrexitIsBrexit
#BrexitJustice
#BrexitMeansBrexit
#BrexitParty
#BrexitReady
#brexitshambles
#BrexitThe
#brexitvote
#BrighterFuture
#britainout
#britin
#CameronGo
#CameronMustGo
#CameronResign
#CantTrustTheTories
#CatsAgainstBrexit
#ChangePoliticsForGood
#CleanBreakBrexit
#ClimateDebate
#conservative
#ConservativeManifesto
#conservatives
#ConsManifesto
#Corbyn
#CorbynOut
#CostOfCorbyn
#DebateHer
#DementiaTax
#Eleanor4Speaker
#election2015
#Election2017
#EqualityTownHall
#eu
#euinorout
#eunegotiation
#EUpol
#EUpoll
#euref
#eureferendum
#eureform
#eurefresults
#eurenegotiation
#europe
#europeanunion
#expelmetoo
#fbpe
#FinalSay
#FinalSayForAll
#ForTheMany
#ForTheManyNotTheFew
#ge15
#GE17
#ge2015
#ge2017
#GE2019
#GeneralElection
#GeneralElection2017
#GetBrexitDone
#grassrootsout
#Green
#greenerin
#greens
#Grenfell
#humanrights
#immigration
#ImVotingLabour
#IndependentScotland
#indyref
#indyref2
#indywales
#inorout
#InvestInTheRest
#IrishUnity
#ITVDebate
#IVotedBrexit
#iVotedLeave
#JC4PM
#JeremyCorbyn
#Johnson
#KeepCorbyn
#labdoorstep
#LabManifesto
#labour
#Labour2015
#Labour2017
#Labour2019
#labourcoup
#labourin
#LabourManifesto
#leadersdebate
#leadnotleave
#Leadsom4Leader
#leave
#leavechaos
#leaveeu
#LeaveMeansLeave
#LibDem
#libdems
#Londependence
#LondonIndependence
#loveeuropeleaveeu
#maga
#MakeUKGreatAgain
#MarchForEurope
#marr
#MarrShow
#May
#MayforPM
#MayvCorbyn
#merkel
#migrant
#migrants
#MoreInCommon
#Move4Europe
#MyImageOfTheEU
#nato
#Newsnight
#nhs
#NHScrisis
#niineurope
#no2AV
#no2eu
#NotForSale
#NotMyVote
#notoAV
#notoeu
#OFOC
#OnYourSide
#OurNHS
#OutIsOut
#OutMeansOut
#pcpeu
#peoplesvote
#peoplesvotemarch
#Peston
#plaid15
#plaid17
#plaid19
#PlaidCymru
#politics
#PostBrexitRacism
#PostRefRacism
#projectfact
#projectfear
#publicduty
#PutItToThePeople
#r4today
#RealChange
#ref
#referendum
#refugee
#refugeecrisis
#refugees
#refugeeswelcome
#Register2Vote
#RegisterToVote
#regrexit
#remain
#RemainerNow
#remaineu
#Revoke
#RevokeArticle50
#Ridge
#rjcob
#SaveCameron
#SaveDave
#SaveOurNHS
#scexit
#scotlandineu
#scotlandineurope
#scotref
#ScottishIndependence
#Scottxit
#Scotxit
#SecondReferendum
#snp
#snp15
#snp17
#snp19
#socialcare
#stopbrexit
#StopBrexitSaveBritain
#strongerin
#takebackcontrol
#takecontrol
#theresamay
#tories
#ToriesOut
#Tory
#ToryManifesto
#trump
#ttip
#UKandEU
#ukineu
#ukip
#UKRef
#UKreferendum
#un
#UnitedIreland
#UniteToRemain
#UnityRef
#VictoriaLIVE
#Vote
#VoteCons
#voteconservative
#VoteGreen
#VoteGreens
#votein
#VoteLab
#votelabour
#VoteLabour2015
#VoteLabour2017
#VoteLabour2019
#voteleave
#VoteLeaveLoseControl
#VoteLibDem
#VoteLibDems
#VoteNHS
#voteout
#votePlaid
#voteremain
#votesnp
#voteukip
#walesitsus
#WASPI
#waton
#WeAreReady
#wearethe48
#WeAreThe52
#wearewales
#WhyVote
#WomansHour
#WriteAPoemAboutBrexit
#yes
#yes2AV
#yes2eu
#yestoAV

Appendix IX: Summary statistics

Table 18: Summary statistics (monthly) for the continuous variables from the main sample.

VARIABLE	*N*	MEAN	*SD*	MIN	PCTL(25)	PCTL(75)	MAX
TIME (*t*)	147,360	60.5	34.64	1	30.8	90.2	120
POLITICAL ATMOSPHERE (TIME-VARIANT)							
ELECTION PROXIMITY	147,360	12.97	8.48	0	5	20	30
ELECTION PROXIMITY (REF)	147,360	8.44	7.51	0	2	13.2	27
SINCE ELECTION	147,360	21.04	15.98	0	8	29.2	59
SINCE ELECTION (REF)	147,360	12.28	10.37	0	4	19.2	39
BBC PARLIAMENT	141,220	743	204.67	428.4	632	808	1,718.6
PARTISANSHIP (UNIT-VARIANT)							
PARTISAN SCALE	147,360	-.72	1.21	-6.59	-1.32	-.13	5.03
PARTISAN SCALE^2	147,360	1.98	3.60	.000	.21	2.11	43.41
PARTISAN SCALE (WTD)	147,360	-.12	.21	-.75	-.26	-.01	.88
PARTISAN SCALE^2 (WTD)	147,360	.06	.08	.000	.01	.08	.78
MEDIA DIET (UNIT-VARIANT)							
MEDIA SCALE	147,360	.01	1.10	-5.07	-.62	.63	5.21
MEDIA SCALE^2	147,360	1.21	2.25	0.000	.07	1.34	27.15
MEDIA SCALE (WTD)	147,360	-.000	.01	-.02	-.003	.002	.03
MEDIA_SCALE^2 (WTD)	147,360	.000	.000	.00	.000	.000	.001
POLITICAL ENGAGMENT (UNIT-VARIANT)							
PARTISANS FOLLOWED	147,360	35.21	47.73	2	8	44	560
MEDIA FOLLOWED	147,360	19.86	19.68	2	6	26.2	138
NON-PARTISANS FOLLOWED	147,360	5.69	7.98	0	1	7	76
TWITTER USAGE (UNIT-VARIANT)							
FOLLOWING GROWTH	147,360	.36	.36	-.89	.13	.51	2.81
FOLLOWING GROWTH (*SE*)	147,360	.004	.01	.000	.001	.005	.06
FOLLOWING VOLATILITY	147,360	303.18	247.53	8.81	121.93	404.95	1,837.78
FOLLOWERS	147,360	1,963.62	2,509.69	40	487	2,205.5	13,807
FOLLOWING	147,360	1,493.50	956.02	104	750.8	2,104	4,311
TOTAL TWEETS	147,360	20,904.89	17,958.46	183	7,371.5	28,852	81,084
FAVOURITED TWEETS	147,360	12,383.76	18,377.01	0	1,752	14,568.5	177,029
NUMBER OF MONTHLY TWEETS, BY TWEET TYPE (TIME- & UNIT-VARIANT)							
ALL TWEETS	147,360	153.32	249.48	0	16	191	7,228
TWEETS	147,360	75.53	152.4	0	6	82	6,995
REPLIES	147,360	33.55	85.18	0	0	27	2,679

VARIABLE	*N*	MEAN	*SD*	MIN	PCTL(25)	PCTL(75)	MAX
RTs	147,360	44.24	135.64	0	0	37	6,958
FRTs	147,360	23.57	75.69	0	0	19	5,108
NFRTs	147,360	20.67	66.83	0	0	15	3,453
NUMBER OF MONTHLY UNCIVIL TWEETS, BY TWEET TYPE (TIME- & UNIT-VARIANT)							
UNCIVIL ALL TWEETS	147,360	37.12	75.71	0	2	41	3,014
UNCIVIL TWEETS	147,360	17.27	38.4	0	1	17	1,847
UNCIVIL REPLIES	147,360	9.14	26.26	0	0	6	781
UNCIVIL RTs	147,360	10.72	46.06	0	0	6	2,864
UNCIVIL FRTs	147,360	5.31	25.18	0	0	3	2,097
UNCIVIL NFRTs	147,360	5.41	23.03	0	0	3	1,726
PERCENTAGE OF MONTHLY TWEETS, BY TWEET TYPE (TIME- & UNIT-VARIANT)							
% TWEETS	129,452	55.88	34.48	.00	25.61	100	100
% REPLIES	129,452	18.73	22.19	.00	.00	32.31	100
% RTs	129,452	25.4	26.9	.00	.00	43.27	100
% FRTs	129,452	14.06	17.37	.00	.00	22.22	100
% NFRTs	129,452	11.34	14.52	.00	.00	17.85	100
PERCENTAGE OF MONTHLY TWEETS, BY TWEET TYPE, THAT ARE UNCIVIL (TIME- & UNIT-VARIANT)							
% UNCIVIL ALL TWEETS	129,452	21.61	13.88	.00	12.41	29.38	100
% UNCIVIL TWEETS	127,258	22.06	16.82	.00	10	31.21	100
% UNCIVIL REPLIES	82,938	23.87	19.16	.00	10.71	33.33	100
% UNCIVIL RTs	88,241	18.54	16.46	.00	5.88	27.27	100
% UNCIVIL FRTs	84,268	17.71	18.47	.00	.00	27.03	100
% UNCIVIL NFRTs	82,692	19.52	19.34	.00	.00	29.89	100
PERCENTAGE OF MONTHLY UNCIVIL TWEETS, BY TWEET TYPE (TIME- & UNIT-VARIANT)							
% TWEETS ALL UNCIVIL	119,882	55.25	36.78	.00	22.22	100	100
% REPLIES ALL UNCIVIL	119,882	21.13	26.01	.00	.00	37.5	100
% RTs ALL UNCIVIL	119,882	23.62	29.29	.00	.00	40	100
% FRTs ALL UNCIVIL	119,882	12.47	19.73	.00	.00	18.18	100
% NFRTs ALL UNCIVIL	119,882	11.14	17.37	.00	.00	17.24	100

NOTE: REF = including 2011 AV, 2014 Scottish independence and 2016 EU membership referendum. WTD = weighted by relative rarity of a partisan or media left–right score. RTs wholly comprise FRTs and NFRTs.

Table 19: Summary statistics (monthly) for the continuous variables from the subsample.

VARIABLE	*N*	MEAN	*SD*	MIN	PCTL(25)	PCTL(75)	MAX
TIME (*t*)	80	80.50	23.24	41	60.8	100.2	120
POLITICAL ATMOSPHERE (TIME-VARIANT)							
ELECTION PROXIMITY	80	10.36	7.42	0	4	16	27
ELECTION PROXIMITY (REF)	80	7.09	7.08	0	2	9	27
SINCE ELECTION	80	20.81	17.89	0	6	27.2	59
SINCE ELECTION (REF)	80	12.16	11.30	0	3.8	19.2	39
BBC PARLIAMENT	75	775.95	237.03	449.40	638.68	843.88	1,718.60
NUMBER OF MONTHLY TWEETS, BY TWEET TYPE (TIME- & UNIT-VARIANT)							
ALL TWEETS	80	1,680.33	458.14	1,291	1,440.2	1,754.2	4,428
TWEETS	80	543.88	96.98	358	463.2	603.5	856
REPLIES	80	516.72	125.16	361	441	564.5	1,243
FRTs	80	366.92	146.01	170	254	430.5	1,063
INRTs	80	160.93	99.00	55	96.5	204.5	549
ENRTs	80	91.88	114.21	7	23	121.2	815
NUMBER OF MONTHLY UNCIVIL TWEETS, BY TWEET TYPE (TIME- & UNIT-VARIANT)							
UNCIVIL ALL TWEETS	80	409.51	181.88	253	292.8	444.2	1,441
UNCIVIL TWEETS	80	122.74	27.31	86	104.8	133	265
UNCIVIL REPLIES	80	138.91	43.74	85	112.8	155.2	368
UNCIVIL FRTs	80	82.56	56.93	24	38	110	354
UNCIVIL INRTs	80	33.25	27.29	2	15	47.8	134
UNCIVIL ENRTs	80	32.05	45.46	1	6	46.2	320
PERCENTAGE OF MONTHLY TWEETS, BY TWEET TYPE (TIME- & UNIT-VARIANT)							
% TWEETS	80	33.70	7.95	17.62	26.71	41.09	44.40
% REPLIES	80	31.10	3.82	21.63	28.09	34.27	40.29
% FRTs	80	21.43	4.26	12.23	18.37	24.47	28.88
% INRTs	80	9.06	3.38	4.26	6.50	11.01	18.93
% ENRTs	80	4.71	3.76	.52	1.61	7.02	18.41
PERCENTAGE OF MONTHLY TWEETS, BY TWEET TYPE, THAT ARE UNCIVIL (TIME- & UNIT-VARIANT)							
% UNCIVIL ALL TWEETS	80	23.73	4.29	18.54	20.47	26.69	34.28
% UNCIVIL TWEETS	80	22.85	4.43	16.04	19.85	24.36	35.13
% UNCIVIL REPLIES	80	26.67	3.36	18.76	24.05	28.87	34.83
% UNCIVIL FRTs	80	20.63	6.54	10.48	14.58	26.47	33.30
% UNCIVIL INRTs	80	19.40	8.32	3.17	14.84	24.56	41.36
% UNCIVIL ENRTs	80	30.10	10.57	6.25	22.70	36.89	55.70
PERCENTAGE OF MONTHLY UNCIVIL TWEETS, BY TWEET TYPE (TIME- AND UNIT-VARIANT)							

VARIABLE	*N*	MEAN	*SD*	MIN	PCTL(25)	PCTL(75)	MAX
% TWEETS ALL UNCIVIL	80	32.67	9.06	18.39	24.48	40.87	51.71
% REPLIES ALL UNCIVIL	80	35.27	4.89	24.91	30.98	38.79	45.60
% FRTs ALL UNCIVIL	80	18.65	6.06	8.33	12.78	24.20	29.39
% INRTs ALL UNCIVIL	80	7.37	3.94	.74	4.84	9.22	19.24
% ENRTs ALL UNCIVIL	80	6.04	4.91	.31	2.09	9.00	22.21

NOTE: Figures are averages of aggregate monthly statistics rather than individuals as in the main sample. Analysis for this subsample is undertaken at the tweet level in the study. REF = including 2011 AV, 2014 Scottish independence and 2016 EU membership referendum. RTs wholly comprise FRTs and NFRTs.

ISSN: 1474-1938/1946