The Public Accountability of Social Platforms:

Lessons from a Study on Bots and Trolls in the Brexit Campaign

Marco Bastos* Dan Mercea

(City, University of London)

Accepted for publication in Philosophical Transactions A

(pre-publication version: some changes still possible)

Abstract

In this article we review our study of 13,493 bot-like Twitter accounts that tweeted the U.K. European Union membership referendum and disappeared from the platform after the ballot. We discuss the methodological challenges and lessons learned from a study that emerged in a period of increasing weaponization of social media and mounting concerns about information warfare. We address the challenges and shortcomings involved in bot detection, the extent to which disinformation campaigns on social media are effective, valid metrics for user exposure, activation, and engagement in the context of disinformation campaigns, unsupervised and supervised posting protocols, along with infrastructure and ethical issues associated with social sciences research based on large-scale social media data. We argue for improving researchers' access to data associated with contentious issues and suggest that social media platforms should offer public Application Programming Interfaces to allow researchers access to content generated on their networks. We conclude with reflections on the relevance of this research agenda to public policy.

Main Text

The Brexit Botnet

In October 2017 we published an article detailing the activity patterns of a large cohort (13,493) of bot-like Twitter accounts that tweeted the U.K. European Union membership referendum and disappeared from the platform shortly after the ballot. The analysis of the Brexit Botnet [1] was part of a project to map the expression of ideological positions regarding Brexit on Twitter onto parliamentary constituencies. We had not envisioned inspecting bot activity in the Brexit debate; instead, we stumbled upon the problem when realizing that over 5% of the userbase that tweeted referendum-related hashtags had disappeared after the vote, a rate of account deletion that much exceeded patterns observed in previous studies implementing similar research designs.

The article on the Brexit bots emerged in a context of increasing weaponization of social media platforms and a growing scrutiny of algorithms in which bots feature as the most simple, cost-effective, and flexible approach [2] to gaming the social media attention economy [3]. Between the end of 2017 and early 2018, social media platforms and intelligence agencies worldwide issued several reports detailing the state of information warfare on social media [4-7].

In this reflexive paper, we contemplate the aftermath of the publication of our article as a vehicle for public discussion intended to make the invisible visible [8]. We chart the path from publication of the analysis to the public attention it received and the engagement it generated against the backdrop of congressional and parliamentary inquiries in the U.S. and the U.K. into foreign interference in national elections. The findings of the study were reported in over 250 news articles and were referenced in a House of Commons Briefing Paper.

The study was also cited as evidence in the House of Commons Debate on Russian Interference in U.K. Politics as well as in the House of Commons Fake News Inquiry led by the Digital, Culture, Media and Sport Committee. The communication between the Committee Chair Damian Collins MP and Twitter, Inc. likewise dwelled on the topic of the botnet and the extent to which Twitter could corroborate its activity and pinpoint its provenance [9].

In what follows, we discuss the most salient challenges to this type of research, what we learned from the study, the extant hindrances impinging on advancing this research agenda and its relevance to public policy. We revisit the strategic use of bots, trolls, and sockpuppets in the context of social media warfare and unpack metrics of exposure and activation in a media ecosystem largely dependent on networking technology. We thus undertake a broader discussion of the original study along with suggestions for future research on the weaponization of social networking sites. While we are cognizant that social scientists may often be shunned by policy makers [10], we believe there are important lessons to be learned from a positive attempt at public engagement.

The Art of Bot Detection

Bots are automatic posting protocols used to relay content in a programmatic fashion. As such, bots are simple algorithms programmed to scrape data from internet sources and post them via social media platforms. Twitter is a relatively bot-friendly platform [11] and a number of prominent Twitter accounts are openly bots, including those of established news outlets relaying breaking news, earthquake and tsunami warning systems, or Twitter accounts operated by the Vatican offering regular reflections on Catholic devotion. In the political sphere, bots can be leveraged to impersonate a third-party and are associated with sockpuppets, which are false online identities used to voice opinions and manipulate public opinion while pretending to be another person [12].

Although bots rely on trivial computing routines, bot detection is not an exact science and neither human annotators nor machine-learning algorithms perform flawlessly [13]. While human coders are better at generalizing and learning new features from observed data, machine learning algorithms are scalable and regularly outperform human annotators in searching for and detecting complex patterns. Training a machine learning algorithm is however a trade-off between recall, the number of correct results divided by the number of possible results, and precision, the ratio of positive and relevant matches [14].

Despite these challenges, significant efforts have been made to detect patterns that pertain to bot activity and a growing catalogue of metrics exists for pinpointing political bots [15-17]. While bot detection was originally an enterprise devoted to the identification, demotion, and prevention of spam [18], it has since evolved to mitigate the detrimental impact of malicious activity on electoral politics, policy discussions, and the deliberation of contentious issues. There is compelling evidence that political bots produce systematically more positive content in support of a candidate [19] and only tweet hashtags used by opponents to disrupt their communication flow [20]. Bots identified in the 2016 U.S. Presidential Elections were effective information disseminators [19]. Similarly, during the E.U. referendum they focused on retweeting content from a selection of users [21], a marker of their potential to disseminate content.

In our study, we identified at least two false positives flagged as bots that turned out to be human users, including the very central accounts of @nero, which was operated by the alt-right controversialist and professional troll Milo Yiannopoulos, and @steveemmensUKIP, a Norwich UKIP and Brexit campaigner. Both accounts were highly connected to the remainder of the botnet and disappeared within the same timeframe, thereby triggering our classifier which relied on thresholding and filtering methods. The number of false negatives in our study most certainly extends beyond these accounts, with similar studies estimating false positives and false negatives in bot detection to hover at around 26% of the data, or 11% and 15%, respectively [13]. These figures are comparable to the results of our study, as Twitter acknowledge having removed 71% of the accounts identified in our study due to violations of its spam policies [22].

The opportunities to leverage bots in disinformation campaigns by strategically amplifying divisive content began to be publicly debated in the wake of the 2016 U.S. Presidential elections. The discussion rapidly escalated to a public outcry against fake accounts and bots on social media. The ensuing uproar regarding bot activity is rooted in the perception that social networking sites are opaque platforms unaccountable to regular users and governments alike, with sentiments towards Twitter bots quickly taking a negative turn [23]. This development prompted a question that remains largely uncharted in the literature discussing the weaponization of social media information, namely, how impactful are bots in disinformation campaigns?

Megaphone or Microphone

One important component of our study was that it sought to measure the impact of bot-like accounts in the broader Brexit debate. We sought to identify whether bots were used to increase the reach of a given user's message, much like a microphone, or deployed in a concerted effort to amplify a political narrative toward a targeted direction, much like a megaphone. After identifying accounts that both presented bot-like features and that disappeared shortly after the vote, we inspected the data to determine whether bots could generate greater or faster message cascades compared with active users. While it is not possible to rebuild every step of a retweet cascade, with independent entry points [24] not being reported by Twitter, we managed to identify seed messages and the temporal diffusion of the information. Using this method, we managed to rebuild retweet cascades from original to subsequent users that replicated the information. A variable time-to-retweet was calculated in a similar fashion, with the timestamp attached to each tweet offering the necessary information to estimate the time elapsed between the original tweet and the ith retweet for cascade of size S.

The conclusions presented in our study were far less alarming than one would think by following the public debate about bots. Our study indicated that while bot-like accounts exhibited clear patterns of specialization that allowed them to trigger small to medium-sized cascades in a fraction of the time required by active users to start cascades of comparable size, there was no evidence supporting the notion that bots had substantively altered the Brexit debate on Twitter. Indeed, by all measures employed in our study, the activity of bot-like accounts was relatively minor with respect to the larger conversation about the referendum. Our findings indicated that bots can potentially amplify a subset of accounts, but that their influence in the network is limited and falls short of a megaphone, a result consistent with literature on cognitive dissonance reporting that political persuasion seems to have little effect in attitude change at the individual level [25, 26].

The abovementioned results presented in the original study were emphasized by Twitter, Inc. [27] in its response to the Digital, Culture, Media and Sport Committee, but they contributed little to offsetting public concerns about Russian trolls operating on social platforms. Facebook, a platform averse to bots and requiring real identities [28], resisted publicly releasing any data, thereby increasing concerns that the opaqueness of social media platforms shields disinformation campaigns. In this climate, a rapidly evolving public debate seems to have eschewed fundamental differences between bots, trolls, and sockpuppets [12]. One important caveat about the above assessment is that in distributed systems the exposure and activation triggered by messages, whether relayed by bots or otherwise, likely exceeds the population of users directly exposed to them.

Activation and Network Effects

Concerns raised by political bots often stem from the belief that targeted advertisement affects users in a uniform, atomized fashion. This understanding of media effects dates from early communication research asserting that the media exerts a powerful and persuasive influence on audiences who were believed to be volatile, alienated, and inherently susceptible to manipulation. This framework is known as the hypodermic needle model [29] on account of portraying all-powerful messages being "injected" into easy-prey and suggestible individuals [30]. In the post-war period this framework was rebuked and eventually replaced by the two-step flow of communication, a model emphasizing the importance of opinion leaders and interpersonal communication in the flow of personal influence leading to the promotion of ideas and products [31].

It is unsurprising that the framework used in propaganda studies continues to rely on the two-step flow of communication model [31], thereby foregrounding the role of opinion leaders whose influence in their community is a vector of social persuasion. This model continues to provide a dependable framework for studying Twitter [32], a social network largely populated by opinion leaders and in particular by the digiterati, and Instagram, a mobile photo-sharing app particularly suited to influencer marketing [33]. But the topology of social media platforms, which can accommodate various network formations, has modified the relatively simple equation in which persuasion is a function of activation, reinforcement, and conversion, as secondary network effects are not sufficiently developed in classic models of interpersonal communication [34] and information diffusion theory [35].

In other words, the assessment of campaign effects continues to rely on the assumption of independence to estimate the impact of exposure to partisan content on voting preferences. Political campaigns are expected to result in activation, when unmotivated actors confirm their support to the campaign; conversion, when motivated actors shift their vote to the opposing party; and reinforcement, when the initial vote preference is strengthened due to the campaign [34]. The assumption of independence underpinning campaign effects is correspondingly explored with fixed effects models [36], with exposure to campaign materials leading individuals to activate subsequent actors within their reach. Lazarsfeld resorted to the metaphors of photographic development and the rubbing of a coin to describe activation as the emergence of an ideological alignment that existed in latent form but only crystallised because of campaign propaganda, thereby tracing a linear path from voter's latent tendencies to activation or conversion.

The two step flow model is remarkably nuanced and downplays the power of propaganda epitomized by the hypodermic needle model [29]. It explores the subtle relationship between political communications broadcast by mass media and the direct personal influences exerted by activated individuals. It also describes how successful activation leads to increasing exposure in a continuous process, with propaganda leading to increased interest which in turn makes individuals more willing to expose themselves to further propaganda. The model nonetheless assumes exposure to be linear: either it flows from the media or is acquired through personal contacts. Relationships and interactions are thereby defined within the constraints of one's personal ego network, with no way of accounting for hundreds of millions of dynamic new ties forged and reinforced online through social platforms.

The problem is compounded by social networking sites whose internal topology is constantly shifting due to a growing userbase and successive modifications to the underlying technologies underpinning them. The small-world properties of physical social networks are one of many topologies found on social platforms, which allow for multiple secondary exposures and network effects drawing from single exposure points. As such, the potential impact of messages circulating in social networks cannot be benchmarked against the number of users exposed to the content, as activation might be achieved through subsequent steps through which information cascades extrapolate the assumptions of mass communication and propaganda models [31].

Cumulative Exposure and Disengagement

While Facebook is largely structured as a social network with reciprocal ties and overlapping clusters similar to physical social networks [37], Twitter is a mixed system that can rapidly shift from decentralized, horizontal networks to highly-centralized network formations, with few accounts sourcing information to communities of users [38]. The constantly changing topology of social platforms imposes considerable challenges to studying disinformation campaigns, but these constraints could be offset by incorporating network sciences methods to the task of identifying activation thresholds [39]. Network science can trace the processes through which disinformation navigates centralized and small-world networks to maximize the effects of information dissemination [40].

But even metrics of persuasion such as activation and conversion have limited heuristic power for understanding information warfare. Not only can activation occur due to secondary network effects, but propaganda campaigns can successfully employ psychological warfare techniques that bypass activation altogether. While researchers can track activation times of individuals recruited to political causes up to the

moment when critical mass is attained [41], psychological warfare techniques do not require a given threshold of actors to be activated, as the target is shaping perceptions and manipulating cognitions which can be achieved without change being registered in the public discourse [42, 43]. The objective of psychological warfare is not to move public opinion, but to create confusion, disorder, and distrust [42, 44].

The potential reach of propaganda resorting to broadcast channels such as radio and television is restricted to the population exposed to it and the interactions between activated individuals and their social networks. Social platforms however incorporate network externalities [45], so that users subjected to microtargeted propaganda are also likely to be embedded in cliques or communities equally exposed to the campaign, thereby snowballing the cumulative impression garnered by the piece. In addition to bandwagon effects, network externalities also impinge on an individual's ability to evaluate the extent to which an opinion is prevailing or dissenting relative to the broader population. These externalities can play a pivotal role in breaking the critical mass threshold after which social diffusion of new styles of behaviour grows rapidly [46].

These problems have long been studied in research of opinion evolution that foregrounds the non-linear patterns through which opinions and social change emerge from system interactions. While opinion dynamics are relatively simple, they often lead to nonlinearities and complex dynamic behaviour, of which clustering (i.e., "bubbles") and the polarization of opinions are common outcomes [47, 48]. Disinformation campaigns thrive on polarized discourses by mobilizing supporters in opposing clusters, but clusters do not have to convince each other of a prevailing or a minority opinion. In other words, disinformation campaigns are not intended to change the prevailing opinion, and therefore metrics of persuasion, including activation and conversion, provide limited heuristic guidance regarding such operations. For social issues requiring engagement such as voting and public deliberation, disengagement is as important an objective as engagement and can be achieved with limited activation.

In summary, while changing public opinion is a process governed by intrinsic dynamics, the transition to a new prevailing opinion is likely linked to changes in extrinsic control factors that affect intrinsic dynamics [49]. If an organization seeks to optimize the reach of their campaign, they might resort to a social platform as an information diffusion system and target "influentials"—i.e., users that are central to the network and perform the role of hubs relaying information to the periphery of the network. Trolls and botmasters do not necessarily have to convince "influentials" of their political agenda. For disinformation purposes, it might suffice that opinion leaders inhabiting the network perceive one side of the public debate as contentious and potentially damaging. The result is not a change in the prevailing opinion, but a change in public support for a cause that can well translate to disenchantment, apathy, and lower voter turnout by an ill-informed electorate. Originally derived from intelligence operations but employed in electoral campaigns worldwide [50], these adaptive strategies can skew public opinion without reaching the critical threshold for opinion formation.

Unsupervised and Supervised Automation

Equally important, the distinction between automated and supervised information warfare has remained peripheral to public deliberations. Surpassing bots in complexity and capillarity in the communities they operate, supervised accounts (e.g., trolls) were pivotal in the successful disinformation campaign led by the Kremlin-linked and St Petersburg based Internet Research Agency. This campaign relied primarily on supervised accounts operating on Facebook [51], a sharp contrast to the desolate life of Twitterbots communicating with each other and with modest impact outside their bubbles. The contrast between the two covert strategies raises topical questions about human-driven, curated, and supervised high-volume posting and conversely, automated, unsupervised, and scripted machine bots. Supervised high-volume posting encapsulates a new agent in the political arena to which little attention has been given beyond Reddit forums and the toxic corners of internet culture [52].

While bots and trolls continue to be described in the press as comparable forces undermining and reshaping political campaigning, there are fundamental differences we first identified in our studies of serial activists, who exhibited extraordinary levels of posting activity combined with a savvy strategy for activating opinion leaders such as journalists while at the same time assisting activists to coordinate across national boundaries

and protest sites [53, 54]. This pattern of activity foreshadowed a complex modality of engagement that bridged actions online and onsite at multiple protest locations, an astute and publicly visible modus operandi that may have been readily repurposed for sophisticated covert disinformation campaigns. This is in line with early reports of the effective disinformation campaign led by the Internet Research Agency, whose operatives galvanized partisan communication online and agitated for rallies across the U.S., often contacting campaign staff members in various U.S. states and appealing to individuals to take their grievances to the streets [55].

Conceivably, those operatives may well have relied on bots in their operations. We believe further research is necessary to determine the ramifications of the bifurcated communication modality fathered by serial activists, a strategy that sits alongside but often counters automated, unsupervised, and scripted posting protocols typified by bots by employing transparent operational tactics. One objective common to both automated and supervised activity is increasing the likelihood of activation. While bots might present potential for generating larger exposure, the benefits of network effects can only be achieved when multiple agents are coordinated through endogenous activation [56].

The Infrastructure of Social Platforms

Disinformation strategies centred around inflammatory social media messaging constitute a pressing research agenda for social scientists, notwithstanding the methodological challenges discussed in the previous sections. Social media platforms have however largely refused to share publicly data related to disinformation campaigns that could provide fundamental insights into the strategies of botmasters, trolls, and sockpuppets alike. While continuing to reject any role as a media company or content provider, social media platforms rarely offer access to data of public interest and disavow public and academic expectations about the release of data. In the meantime, social platforms and Facebook in particular are continuously building a complex apparatus of content moderation and user governance that enforce a set of opaque guidelines to public discourse with no public or external expert supervision, a development increasingly adopted by other social platforms seeking more control at the cost of incentivising innovation [57].

Our study of the Brexit botnet was only possible because Twitter operates three well documented, public Application Programming Interfaces in addition to their Premium and Enterprise APIs. But public and open APIs remain an exception in the social media ecosystem that largely operates in secrecy, with Facebook's Public Feed API being restricted to a limited set of media publishers that still require prior approval by Facebook [58]. Facebook's secrecy also extends to the algorithms used to feed information to users in their network. Despite scholarly efforts in algorithm auditing [59], social media algorithms remain largely opaque to public scrutiny; similarly, the criteria underpinning algorithmic decisions on what news stories are distributed to users are intellectual property and therefore unaccountable to the public [60].

Social platforms consequently occupy the centre of a media ecosystem that allows hyper segmentation of social groups and highly targeted political communication with customizable messages that are invisible to the broader public [61]. This infrastructure emerged from a context in which social platforms remain private enclosures even in the aftermath of the weaponization of social networking sites [51]. There have so far been no efforts from their part to create publicly-accessible data repositories for researchers studying public communication associated with contentious issues or disinformation campaigns deployed by state actors and affiliated organizations. Conversely, the scholarly debate over access to social media data is largely focused on individuals' rights not to have their information harvested by corporations, with comparatively less thought given to the corporate ownership of information of public interest. Indeed, researchers continue to face mounting obstacles erected by social media companies that at times have actively blocked access even to publicly available data [62]. On the rare occasions when data were made available to the public, the dataset was anonymously released due to being in breach of Twitter's Developer Policy [63].

Notwithstanding data grants previously offered to research institutions [64], exiting public APIs are in fact not designed for the academic community. These endpoints for data collection are intended for programmers building application software that add to the growing ecosystem of services offered by social platforms, whose business model remains focused on selling users' data by making them available to advertisers and

campaigners targeting individuals with specifically tailored content, a vulnerability under close scrutiny in the wake of the Cambridge Analytica data scandal [50]. Social platforms thus balance the untenable task of convincing regulators that rampant propaganda on social platforms is ineffective while telling advertisers the very opposite. Concurrently, social platforms refuse being classified as media companies because they are in the business of distributing rather than producing content [6]. In other words, the business model of social platforms asserts that they simultaneously own users' information while not being responsible for it, thereby evading concerns regarding both the publicness and the private nature of social data.

Ethical Dilemmas in Social Media Research

The ethics debate over issues surrounding the use of research data collected from social platforms are largely conceived as a struggle to protect realms of private life from the burgeoning technologies of surveillance and control, a surprising development that united fundamentally different sensibilities about what privacy means in the United States and the countries of Western Europe [65]. While Twitter Privacy Policy states its services are public and that private accounts are removed from data streamed through Twitter's Streaming API [66], critical data studies have highlighted the risks that public trace data pose to the subjectivity of individual users. Although public, users might hold a reasonable expectation that the publicness of their activities will not infringe on their privacy or make them vulnerable to unintended scrutiny and even abuse [67], an expectation only likely to grow in the aftermath of the Cambridge Analytica data scandal [50].

The potential vulnerability of users engaging in politically-charged debates poses important ethical challenges. Many of the Twitter handles flagged as Brexit bots included important information to the story we sought to tell, including political slogans, party or campaign affiliation, and ideological leanings. The semantics of usernames was thus an important part of the story leading us to identify 11 handles of suspected bots (e.g., EuFear and @no_eusssr_thx). After putting in balance the risks of false positives and the fact that the accounts had disappeared from the platform shortly after the vote, we decided to disclose the Twitter handles in the interest of accountability. We did so whenever there was a reasonable level of confidence that we were dealing with Twitterbots, to which ethical considerations of privacy are immaterial. Yet, we are cognizant of at least two false positives in our study. These accounts were nonetheless important to the story due to their central position in the bot network and their role in sourcing content to bots. We also recognise that the claim of account automation was not intrinsically harmful as it is not in breach of Twitter Terms of Service.

These considerations can only be properly managed within the context of the research, namely the public debate unfolding in the period leading up to the U.K. E.U. membership referendum. In that context, we faced the challenge of analysing data that could potentially reveal activities detrimental to the functioning of democratic institutions. In our case, specifically, it was only in January 2018 that we learned from the Parliamentary investigators that Twitter had suspended over 70 percent of the accounts in the botnet we identified. While not constituting direct proof of the automation of those accounts, the statement by Twitter, Inc. provided a partial external validation of our research findings, testifying to the disruptive character of the communication they instigated. Indeed, Twitter's response to the Parliamentary enquiry states that a large section of those accounts was suspended because their conduct was in breach of Twitter's Spam Policy.

In the end, the most pressing ethical issue faced in our original study was the obligation not to display deleted Tweets. After long deliberation we decided that the content of the study was of public and scholarly interest. It shed light on a large botnet that participated in a politically contentious debate. Therefore, the social benefits of the research superseded user rights to not have their deleted tweets made public [68], which we deemed immaterial in the case of bots. In our case, we identified users positioned at the core of the botnet and only quoted retweets verbatim if the original tweet had amassed a minimum of 500 retweets, thus avoiding the risk of exposing potentially unnoticed content. We also removed the username that authored the content and identified retweets as exchanges between active users and/or bots. For this cohort of accounts, and regardless of the level of automation involved, we expected users to have a clear sense of the publicness of their quoted posts.

Conclusions

Research exploring the weaponization of social platforms is an embryonic field advancing multiple metrics of networked information warfare. This emerging field continues to struggle with scant and incomplete data due to corporate regulations governing access to public communication data. So long as data of public interest are held by private companies embodied by social platforms, researchers will continue to face considerable challenges to cultivate practices of open data and replicability [69], tenets of high-quality data-driven research without which the field is likely to be diminished in scope and depth while also failing to engage independent academic researchers devoted to an issue of growing public interest.

The public reception of our analysis provides a counterfactual to the notion that scholarly insights into unfolding social transformations garner little traction in an attention economy marked by elevated elite competition between newsmakers, pundits, political actors, and academics [10]. The response to our paper was both meaningful and sustained, while also throwing into relief the importance of research that is both professionally and publicly accountable [8]. While anxieties regarding the relevance and impact of social science research to emerging societal issues are warranted [10], we have sought to address some of the challenges we encountered at the level of individual scholarship.

Academics are no exception in requiring a substantial amount of time to conduct research and summarize it in reports for public review. Both British and American legislators found themselves in the same position when investigating the charges of Russian interference in the democratic process in the two countries. We struggled with the lag between the real-time event, media coverage, public debate, and academic research, but the timeframes eventually overlapped and filled an expanding public conversation with journalists drawing the attention of policy makers to our study of the Brexit bots. The same conversations revealed a public appreciation for the peer-review process as an instrument for professional accountability while also prompting the additional clarifications and caveats presented in this commentary piece.

Ultimately, we would emphasize that public attention affords a new and broader cycle of scrutiny beyond one's community of peers. We have reviewed the benefits and challenges attendant to the increased visibility of a peer-reviewed paper examining a sensitive and politically contentious topic. We have put forward our experience with public engagement as a negotiated outcome where the academic paper is only one of multiple seeds in fast-evolving public deliberations. Lastly, we would highlight the extraordinary and, in our experience, little recognized outlay of time and labour necessary to advance the public role of social science in an already crowded field of myriad institutional commitments by academics [70], namely to teaching, pastoral care, administration, and finally to our research activities.

Additional Information

Acknowledgments
M.B. and D.M. acknowledge support from the Press Office of City, University of London, particularly from Ed Grover who authored the original press released and managed press inquiries related to the original study.

Funding Statement
M.B. acknowledges financial support from the Research Pump-Priming Fund of the School of Arts and Social Sciences of City, University of London.

Data Accessibility
The data collected and analysed in the original study are available from the corresponding author on reasonable request and within the constraints of Twitter's Terms and Conditions governing the sharing of Twitter data.

Competing Interests
The authors declare no competing financial interests.

Authors' Contributions

M.B. conceived and designed the original study. M.B. collected the data and carried out the statistical analyses of the original study. M.B. and D.M. wrote the original manuscript and the present paper. M.B. and D.M. discussed the results and implications and commented on multiple versions of this manuscript.

References

[1] Bastos, M. T. & Mercea, D. 2018 The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*. (DOI:10.1177/0894439317734157).

[2] Confessore, N., Dance, G. J. X., Harris, R. & Hansen, M. 2018 The Follower Factory. In *The New York Times*. New York.

[3] Tufekci, Z. 2013 "Not This One": Social Movements, the Attention Economy, and Microcelebrity Networked Activism. *American Behavioral Scientist*, 0002764213479369.

[4] Bertolin, G., Agarwal, N., Bandeli, K., Biteniece, N. & Sedova, K. 2017 Digital Hydra: Security Implications of False Information Online. ed. G. Bertolin), NATO.

[5] Marwick, A. & Lewis, R. 2017 Media Manipulation and Disinformation Online. Data & Society.

[6] Weedon, J., Nuland, W. & Stamos, A. 2017 Information Operations and Facebook. Facebook.

[7] CSIS. 2018 Who Said What? The Security Challenges of Modern Disinformation. In *World Watch: Expert Notes Series*. Ottawa, Canada, Canadian Security Intelligence Service.

[8] Burawoy, M. 2005 For public sociology. *Am Sociol Rev* **70**, 4-28. (DOI:10.1177/000312240507000102).

[9] Pickles, N. 2018 Letter to Damian Collins MP, Chair, Digital Culture, Media and Sport Select Committee. London, Twitter, Inc.

[10] Nielsen, R. K. 2017 No One Cares What We Know: Three Responsesto the Irrelevance of Political Communication Research. *Political Communication*, 1-5. (DOI:10.1080/10584609.2017.1406591).

[11] Twitter. 2017 Automation rules. Twitter, Inc.

[12] Gorwa, R. & Guilbeault, D. 2018 Understanding Bots for Policy and Research: Challenges, Methods, and Solutions. *arXiv preprint arXiv:1801.06863*.

[13] Varol, O., Ferrara, E., Davis, C. A., Menczer, F. & Flammini, A. 2017 Online Human-Bot Interactions: Detection, Estimation, and Characterization. In *11th International AAAI Conference on Weblogs and Social Media*. Motreal, Canada, AAAI.

[14] Bastos, M. & Mercea, D. 2018 Parametrizing Brexit: Mapping Twitter Political Space to Parliamentary Constituencies. *Information, Communication & Society*. (DOI:10.1080/1369118X.2018.1433224).

[15] Ratkiewicz, J., Conover, M. D., Meiss, M., Gonçalves, B., Flammini, A. & Menczer, F. 2011 Detecting and Tracking Political Abuse in Social Media. In *5th International AAAI Conference on Weblogs and Social Media (ICWSM11)*. Barcelona.

[16] Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. 2016 The Rise of Social Bots. *Communications of the ACM* **59**, 96-104. (DOI:10.1145/2818717).

[17] Abokhodair, N., Yoo, D. & McDonald, D. W. 2015 Dissecting a Social Botnet: Growth, Content and Influence in Twitter. In *18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 839-851. Vancouver, BC, Canada, ACM.

[18] Heymann, P., Koutrika, G. & Garcia-Molina, H. 2007 Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing* **11**.

[19] Bessi, A. & Ferrara, E. 2016 Social bots distort the 2016 US Presidential election online discussion. *First Monday* **21**.

[20] Woolley, S. C. 2016 Automating power: Social bot interference in global politics. *First Monday* **21**.

[21] Howard, P. N. & Kollanyi, B. 2016 Bots, #StrongerIn, and #Brexit: computational propaganda during the UK-EU Referendum. In *SSRN*.

[22] Twitter. 2018 Letter from Nick Pickles, Twitter, to the Chair of Fake News Inquiry. ed. Chair of Fake News Inquiry). London, Parliament of the United Kingdom.

[23] Boyd, D. 2018 The Reality of Twitter Puffery. Or Why Does Everyone Now Hate Bots? , February 13, 2018 ed, NewCo Shift.

[24] Cheng, J., Adamic, L. A., Dow, P. A., Kleinberg, J. & Leskovec, J. 2014 Can cascades be predicted? In *23rd International Conference on World Wide Web (WWW'14)*. Seoul, Korea, ACM.

[25] Wood, W. 2000 Attitude change: Persuasion and social influence. *Annual review of psychology* **51**, 539-570.
[26] Mutz, D. C., Sniderman, P. M. & Brody, R. A. 1996 *Political persuasion and attitude change*, University of Michigan Press.
[27] Twitter. 2017 Letter from Nick Pickles, Twitter, to the Digital, Culture, Media and Sport Committee. ed. C. Digital, Media and Sport Committee,). London, Parliament of the United Kingdom.
[28] Facebook. 2018 Community Standards. Facebook, Inc.
[29] Lasswell, H. D. 1948 The structure and function of communication in society. *The communication of ideas* **37**, 215-228.
[30] Bineham, J. L. 1988 A historical account of the hypodermic model in mass communication. *Communication Monographs* **55**, 230-246. (DOI:10.1080/03637758809376169).
[31] Katz, E. 1957 The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis. *Public Opin Quart* **21**, 61-78. (DOI:10.1086/266687).
[32] Wu, S., Hofman, J. M., Mason, W. A. & Watts, D. J. 2011 Who Says What to Whom on Twitter. In *20th international conference on World Wide Web*, pp. 705-714. New York, ACM.
[33] De Veirman, M., Cauberghe, V. & Hudders, L. 2017 Marketing through Instagram influencers: the impact of number of followers and product divergence on brand attitude. *International Journal of Advertising* **36**, 798-828. (DOI:10.1080/02650487.2017.1348035).
[34] Lazarsfeld, P. F., Berelson, B. & Gaudet, H. 1948 *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*, 2nd ed. New York, Columbia University Press.
[35] Rogers, E. M. 1983 *Diffusion of innovations*. New York, The Free Press.
[36] Dilliplane, S. 2014 Activation, Conversion, or Reinforcement? The Impact of Partisan News Exposure on Vote Choice. *American Journal of Political Science* **58**, 79-94. (DOI:doi:10.1111/ajps.12046).
[37] Backstrom, L., Boldi, P., Rosa, M., Ugander, J. & Vigna, S. 2012 Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 33-42, ACM.
[38] Bastos, M. T., Piccardi, C., Levy, M., McRoberts, N. & Lubell, M. 2018 Core-periphery or decentralized? Topological shifts of specialized information on Twitter. *Soc Networks* **52**, 282-293. (DOI:10.1016/j.socnet.2017.09.006).
[39] Hilbert, M., Vásquez, J., Halpern, D., Valenzuela, S. & Arriagada, E. 2016 One Step, Two Step, Network Step? Complementary Perspectives on Communication Flows in Twittered Citizen Protests. *Social Science Computer Review*. (DOI:10.1177/0894439316639561).
[40] Myers, S. A., Sharma, A., Gupta, P. & Lin, J. 2014 Information network or social network? The structure of the twitter follow graph. In *23rd International Conference on World Wide Web*, pp. 493-498. Seoul, Korea, ACM.
[41] González-Bailón, S., Borge-Holthoefer, J., Rivero, A. & Moreno, Y. 2011 The Dynamics of Protest Recruitment through an Online Network. *Scientific Reports* **1**. (DOI:10.1038/srep00197).
[42] Jowett, G. S. & O'Donnell, V. 2014 *Propaganda & persuasion*, Sage.
[43] Linebarger, P. 1948 Psychological warfare. *Infantry Journal Press*.
[44] Taylor, P. M. 2003 *Munitions of the Mind: A history of propaganda from the ancient world to the present era*. Manchester, Manchester University Press.
[45] Katz, M. L. & Shapiro, C. 1985 Network externalities, competition, and compatibility. *The American economic review* **75**, 424-440.
[46] Bandura, A. 2001 Social Cognitive Theory of Mass Communication. *Media Psychology* **3**, 265-299. (DOI:10.1207/S1532785XMEP0303_03).
[47] Latané, B. 1996 Dynamic social impact: The creation of culture by communication. *J Commun* **46**, 13-25.
[48] Nowak, A., Szamrej, J. & Latané, B. 1990 From Private Attitude to Public Opinion: A Dynamic Theory of Social Impact. *Psychological Review* **97**, 362-376.
[49] Nowak, A., Lewenstein, M. & Frejlak, P. 1996 Dynamics of Public Opinion and Social Change. In *Modelle sozialer Dynamiken: Ordnung, Chaos und Komplexität* (eds. R. Hegselmann & H.-O. Peitgen). Wien, Hölder-Pichler-Tempsky.
[50] The Guardian. 2018 The Cambridge Analytica Files.
[51] US District Court. 2018 United States of America versus Internet Research Agency LLC. p. 37. Washington, DC, United States District Court for the District of Columbia.
[52] Massanari, A. 2015 #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*. (DOI:10.1177/1461444815608807).
[53] Bastos, M. T. & Mercea, D. 2016 Serial Activists: Political Twitter beyond Influentials and the Twittertariat. *New Media & Society* **18**. (DOI:10.1177/1461444815584764).
[54] Mercea, D. & Bastos, M. T. 2016 Being a Serial Transnational Activist. *J Comput-Mediat Comm* **21**, 140–155. (DOI:10.1111/jcc4.12150).

[55] Shane, S. & Mazzetti, M. 2018 Inside a 3-Year Russian Campaign to Influence U.S. Voters. In *The New York Times*. New York.
[56] Kempe, D., Kleinberg, J. & Tardos, E. 2005 Influential nodes in a diffusion model for social networks. *Lect Notes Comput Sc* **3580**, 1127-1138.
[57] Hogan, B. 2018 Social Media Giveth, Social Media Taketh Away: Facebook, Friendships, and APIs. *International Journal of Communication* **12**.
[58] Facebook. 2018 Public Feed API. Facebook, Inc.
[59] Sandvig, C., Hamilton, K., Karahalios, K. & Langbort, C. 2014 Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 1-23.
[60] DeVito, M. A. 2017 From Editors to Algorithms. *Digital Journalism* **5**, 753-773. (DOI:10.1080/21670811.2016.1178592).
[61] Ghosh, D. & Scott, B. 2018 Digital Deceit: The Technologies Behind Precision Propaganda on the Internet. In *Public Interest Technology*. Washington, DC, New America.
[62] Timberg, C. & Dwoskin, E. 2017 Facebook takes down data and thousands of posts, obscuring reach of Russian disinformation. In *The Washington Post*, October 12, 2017 ed. Washington, DC.
[63] Popken, B. 2018 Twitter deleted 200,000 Russian troll tweets. Read them here. In *NBC*.
[64] Twitter. 2014 Introducing Twitter Data Grants. Twitter, Inc.
[65] Whitman, J. Q. 2004 The two western cultures of privacy: dignity versus liberty. *Yale Law Journal*, 1151-1221.
[66] Twitter. 2018 Twitter Privacy Policy. Twitter, Inc.
[67] Metcalf, J. & Crawford, K. 2016 Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society* **3**, 2053951716650211.
[68] Markham, A. & Buchanan, E. 2012 Ethical Decision-Making and Internet Research: Recommendations from the AOIR Ethics Committee. Association of Internet Researchers.
[69] Hutson, M. 2018 Missing data hinder replication of artificial intelligence studies. *Science*. (DOI:10.1126/science.aat3298).
[70] Greer, C. M. 2018 Scholarly Engagement With the Public: The Risks and Benefitsof Engaging Outside of the Classroom. *Political Communication* **35**, 150-153. (DOI:10.1080/10584609.2017.1406589).