Wilfried
Martens Centre
for European Studies

Political Subversion

in the Age of Social Media

Edward Hunter Christie

Policy Brief

October 2018

The aim of this report is to identify some of the main vulnerabilities of Western information spaces with respect to current forms of political subversion, and to propose a set of policy principles to guide ongoing reflections on how best to respond to that challenge. Four areas of vulnerability are identified, namely individualised political messaging, group dynamics and political polarisation, platform algorithms and self-radicalisation, and falsehood dissemination dynamics. This leads to the formulation of four proposed policy principles, followed by a discussion of the extent to which recent measures, in selected Western nations and at EU level, are sufficient to address the challenge at hand.

Keywords Disinformation – Social media – Political polarisation – Fake news

[1] The views in this report are solely those of its author and do not necessarily represent those of NATO or of Allied governments. A previous version of this report was published as E. H. Christie, 'Countering Subversion Online: What Role for Public Policy?', in G. Bertolin (ed.), Hacking Humans: Responding to Cognitive Security Challenges. Riga: NATO StratCom COE (in print). The author wishes to thank Armand De Mets, Vineta Mekone, Janis Sarts, Giorgio Bertolin, Ulf Ehlert, Laura Brent, Christian Liflander, Chelsey Slack, Neil Robinson, Holly Vare and James Reynolds-Brown for comments on earlier versions of the text.

Introduction

The information space that is used by voters, politicians and interest groups in Western nations is being contested and challenged by new risks and threats, both from within and from without. Ubiquitous Internet platforms such as Facebook, Twitter and YouTube once held the promise of greater democratic participation and pluralism. But this has been tempered by concerns about the misuse of personal data and by new forms of political polarisation. This has occurred in tandem with an erosion of the balancing effect of trusted sources of information and a steep rise in the production and dissemination of false news ('fake news'). This report focuses on the threats posed to the normally intended functioning of democratic political systems by hostile actors who seek to subvert them for political and/or strategic purposes.

Systemic vulnerabilities in Western political information spaces have been avidly exploited by the Russian state. It has done this through the deployment of intentionally divisive and polarising false-flag content. This includes disinformation, which may be defined as 'deliberately distorted or manipulated information that is leaked into the communication system of the opponent, with the expectation that it will be accepted as genuine information, and influence either the decision-making process or public opinion.'[2]

Hostile non-state actors have also exploited existing vulnerabilities to spread extremist narratives and support their recruitment drives. However, a recent study by two French governmental research institutes reports an estimate according to which 80% of hostile foreign political influencing efforts in the EU can be attributed to the Russian Federation, and just 20% to other states and non-state actors combined.[3] For the specific case of the 2017 French presidential election, the study notes that all of the foreign influencing efforts that were detected were from Russia.[4]

This report identifies four areas of vulnerability, namely individualised political messaging, group dynamics and political polarisation, platform algorithms and self-

[2] This definition was reported in 1984 by Ladislav Bittman, a Cold War-era defector from the former Czechoslovak intelligence service. See *Soviet Active Measures* (film) (US Information Agency, 1984).

[3] J.-B. Jeangène Vilmer et al., *Les manipulations de l'information : un défi pour nos démocraties,* Centre d'analyse, de prévision et de stratégie and l'Institut de recherche stratégique de l'École militaire (Paris, August 2018), 50.

[4] Ibid.

radicalisation, and false news dissemination. The following four sections provide brief overviews of each of these areas. Policy implications and developments are addressed in a subsequent section, where four policy principles are proposed, one for each of the four areas of vulnerability. These principles are then used as the basis for a discussion of selected recent policy initiatives. Particular focus is given to the proposed Honest Ads Act in the US, the proposed law on the fight against false information in France and the European Commission's Communication on tackling disinformation online.

Individualised political messaging

The use of multiple computer-based and/or mobile applications by billions of users worldwide is leading to the generation and collection of very large volumes of very granular data ('Big Data'). This includes records of purchases of goods and services, search engine queries and emotional responses to a large array of online content—from news stories, through popular memes, entertainment and leisure activities, to commercial advertising and political campaigns. The average individual in a Western nation today has already voluntarily released thousands, if not millions, of data points into various software applications, almost always without any awareness of how such data might be used and what insights might be gained from cutting-edge analysis.

Widespread social media use and the availability of the corresponding data have led to three interrelated developments. First, new ground has been broken in the academic field of psychometrics and in the corresponding applied field of psychographics. Recent analyses have revealed the close connection between individual preferences and behaviour, on the one hand, and private characteristics, on the other. As early as 2013, academics had demonstrated that Facebook 'likes' could be used to automatically and (largely) accurately predict an individual's sexual orientation, ethnicity, religious and political views, personality traits, intelligence level, state of happiness, use of addictive substances, age and gender.[5] It is important to note that these results are not dependent on uncovering overt self-reporting of any of these characteristics. Rather, thanks to Big Data, psychometric research has revealed hitherto poorly understood

[5] M. Kosinski, D. Stillwell and T. Graepel, 'Private Traits and Attributes Are Predictable from Digital Records of Human Behavior', *Proceedings of the National Academy of Sciences* 110/15 (April 2013), 5802–5.

correlations between overt online behaviour and intimate private information. The second development is that recent advances in psychometrics have revolutionised both marketing and political campaigning, based on improvements in predictive analytics, that is, the use of data, statistical algorithms and machine-learning techniques for purposes of prediction. The third development is that differentiated political messages can now be delivered much more easily and cheaply down to the level of the individual voter through social media. Based on access both to individual-level data and to the deployment of new insights from psychographics and from predictive analytics, political messaging can be differentiated according to individual characteristics and personality traits in order to have the greatest psychological impact. In addition, each campaign advertisement can use a presentation format (e.g. colours, text size, choice of words and visual illustrations) whose emotional appeal has been optimised for its target audience, thanks to machine-learning techniques.[6] Thus the cutting edge in political campaigning relies on a trinity of psychographics, predictive analytics and individualised political messaging.

This new structure can be weaponised by hostile actors—leading to more effective campaigns of political subversion. At an October 2017 US Senate sub-committee hearing,[7] it was revealed that Russian operatives had, for example, specifically targeted patriotically minded adult Texans with a political advertisement purporting to be from a Texas-based organisation, which contained false claims against then candidate Hillary Clinton. These advertisements used Facebook's own targeting technology. The company delivered sophisticated and targeted political messages to audiences within the US, in the context of a US election, in exchange for payment from an organisation under the ultimate control of the Kremlin. Compared to Cold War-era Soviet disinformation campaigns, it is striking how easily, quickly and cheaply Russia was able to reach audiences in a Western country.

[6] This is notably based on a method called 'A/B testing', whereby small variations in online advertising materials (version A versus version B) are selectively shown to large numbers of potential customers for the purpose of identifying the most effective version. The process is repeated multiple times with successive variations, leading to particularly effective materials.

[7] US Senate, 'Extremist Content and Russian Disinformation Online: Working with Tech to Find Solutions', Sub-committee hearing, 31 October 2017.

Psychometrics and mass persuasion: the stunning power of social media

In a study published in 2015, leading researchers in the field of psychometrics demonstrated that computer-based personality assessments based on an individual's Facebook likes are more accurate than those made by human acquaintances of that individual.[8] For the average individual the computer models required the input of just 10 likes to outperform an average work colleague, 70 likes to outperform a (real-life) friend or cohabitant, 150 likes to outperform a family member and 300 likes to outperform the individual's spouse. A follow-up study, based on large-scale online experiments, showed that matching the content of persuasive appeals to individuals' psychological characteristics, as estimated by their online actions, significantly altered these individuals' behaviour as measured by clicks and purchases.[9] These findings suggest that psychological targeting makes it possible to influence the behaviour of large groups of people by tailoring persuasive appeals to their psychological needs.

Group dynamics and political polarisation

A key question is whether social media increases political polarisation. Overall, societal or political polarisation is driven by many top-down and bottom-up factors, from the attitudes and statements of politicians and the editorial lines of influential media sources to real-life socio-economic and societal shifts. Group polarisation, the phenomenon whereby joining a group of like-minded people tends to entrench and sharpen pre-existing views, appears to be natural and

[8] W. Youyou, M. Kosinski and D. Stillwell, 'Computer-Based Personality Judgments are More Accurate than Those Made by Humans', *Proceedings of the National Academy of Sciences* 112/4 (January 2015), 1036–40.

[9] S. Matz et al., 'Psychological Targeting as an Effective Approach to Digital Mass Persuasion', *Proceedings of the National Academy of Sciences* 114/48 (November 2017), 12714–19.

was documented long before the advent of social media.[10] Likewise, seeking out information that conforms to pre-existing views (selective exposure) and finding such information more persuasive than contrary information (confirmation bias) are not new phenomena.[11] That group polarisation also occurs on social media, as shown in recent research,[12] is thus no surprise. But it is not immediately obvious that social media would necessarily lead to greater polarisation for societies as a whole: individuals face a wide range of groups to choose from, including many moderate ones within which individuals would become entrenched moderates. If the choice and visibility of social media groups reflected the pre-existing or underlying distribution of opinion in society, social media might merely mirror that distribution.

However, if more extreme groups could benefit from undue advantages in the online world, then polarisation dynamics could be stronger than initial conditions in the offline world. And these dynamics could lead to greater polarisation both online and offline. One area of investigation is the rise—and partial mainstreaming—of anti-establishment and anti-liberal populism. It can involve either far-right or far-left views, and is often accompanied by sympathies or connections with the Kremlin. While it is not surprising that the Great Recession of 2009 and its aftermath have fuelled greater 'demand' for such views (without their Kremlin component), social media have not only facilitated the 'supply' of relevant content to receptive audiences but have also allowed such content to appear to be more popular than it is. Indeed, it has been shown that populist politicians and parties throughout the Western world enjoy considerably higher levels of support in the online world than they do at the ballot box. While some of that gap could be explained by the unwillingness of some voters to express discontent except online, it is likely that fictitious online support is mostly to blame.

Fictitious online support is based on a combination of home-grown and foreign astroturfing, the practice of masking the sponsors of a message or organisation to make it appear as though it originates from grass-roots participants. Online political astroturfing may combine both human agents (trolls) and automated agents (bots). It is not hard to see how online astroturfing can lead to distortions that could adversely affect the political process. Real-life swing voters could be

[10] See e.g. D. G. Meyers and H. Lamm, 'The Group Polarization Phenomenon', *Psychological Bulletin* 86/4 (1976), 602–27.

[11] See e.g. C. G. Lord, L. Ross and M. R. Lepper, 'Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence', *Journal of Personality and Social Psychology* 37/11 (1979), 2098–109.

[12] See e.g. M. Del Vicario et al., *Echo Chambers: Emotional Contagion and Group Polarization on Facebook, Nature,* Scientific Reports 6:37825 (2016).

swayed by the apparent prevalence, popularity and normalisation of extreme views and content. In addition, traditional media sources, and individual journalists, authors and commentators, will typically be influenced by the number of likes and shares their works receive. If extreme content is artificially rewarded, this will create an (apparent) incentive to produce more of it, leading to (further) polarisation in traditional media and to further polarisation of the electorate. This increase in polarisation takes place both directly, as real-life voters receive more polarised content, and indirectly, as extreme networks and groups on social media are able to further legitimise their views by pointing to supporting content from traditional media, rather than only from fringe sources.

Scale and attribution of political astroturfing: some preliminary evidence

According to an April 2017 report,[13] the Alternative for Germany party had a Facebook following twice the size of that of Germany's Christian Democratic Union party, although the latter obtained more than double the number of votes the former received in the 2017 election. Members of the European Parliament from the far right have their tweets shared on average almost five times more than those from mainstream parties. Far-left Members of the European Parliament account for 30% of the Twitter followers of all Members of the European Parliament, despite holding only 4% of the seats. Overall, the apparent online popularity of extremist politicians exceeds their electoral popularity by a factor of at least four. Relatedly, a 2018 study documents the case of an Austrian far-right activist instructing his followers to practice astroturfing: 'if you don't have any friends, set up multiple accounts and run them in parallel . . . People are herd animals, they are more inclined to follow a group than a single individual."[14]

[13] C. Hendrickson and W. A. Galston, 'Why Are Populists Winning Online? Social Media Reinforces Their Anti-Establishment Message', Brookings Institution (April 2017).

[14] Cited in P. Kreissel et al., *Hass auf Knopfdruck. Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz,* Institute for Strategic Dialogue (July 2018), 14. (Author's translation.)

Platform algorithms and self-radicalisation

In 2017, 45% of Americans reported obtaining news stories from Facebook, 18% from YouTube, 11% from Twitter and 7% from Instagram.[15] The news content on these platforms originates from third parties, notably from the websites of television, radio and print media outlets. In addition, there is a large volume of user-generated content which is political in nature. It deals with particular politicians and, perhaps more importantly, current political, social and cultural issues.

Social media platforms aggregate and filter content according to user preferences. Much of that filtering can be traced back to conscious choices by users—which they make by, for example, choosing to 'follow' or 'friend' specific opinion leaders and self-selecting into specific groups. But the algorithms used by the platforms generate individualised 'news feeds' as well as suggestions to follow or like or join additional opinion leaders or groups. These algorithms are partly based on machine-learning techniques and lead to filtering and selection criteria that are not transparently known. From the perspective of the platform operators, the goal is to retain the attention of users for as long as possible, given that longer attention time translates into greater exposure to advertising and thus into greater revenues. Insights into human psychology, including users' individual personality traits, can be harnessed to maximise revenue-generation. This may incentivise platform operators to seek to generate obsessive, compulsive or addictive emotional states—even beyond the platform operators' own explicit awareness.

An under-researched area of concern is the YouTube algorithm. As noted above, YouTube is a more important source of news for Americans than Twitter. Furthermore, surveys on sources of news likely underestimate the platform's true importance in indirectly shaping political perceptions through issues-based content that strongly correlates with political positioning—that is, issues such as gender, climate change and immigration. The YouTube algorithm tends to automatically suggest content that goes in the same general direction as a viewer's interests, as expressed by the viewer's choices of videos. The algorithm

[15] E. Shearer and J. Gottfried, 'News Use Across Social Media Platforms 2017', Pew Research Center (September 2017), 6.

is clearly oriented towards keeping users emotionally engaged, regardless of the factual accuracy of the content that is suggested to them. This is particularly problematic when it comes to politically charged issues, given that the platform contains a large volume of tendentious content, ranging quite seamlessly from subtly oriented and mostly true content to content that is emotionally dark, alarming or outright false. This notably applies to 'anti-establishment' views from the far right or the far left. As noted by some critics, this can result in some users being led down 'hateful rabbit holes'.[16] Recent academic research also notes the high visibility of controversial content, while researchers face difficulties elucidating how the algorithm works.[17]

Risks to political stability arise from such algorithms for two main reasons. The first is self-radicalisation: users with merely a slight predisposition towards radical views are likely to consume far larger quantities of tendentious or false content than in the offline world. The second is that hostile actors may study how the algorithms promote certain types of content and design their information operations accordingly.

Falsehood dissemination dynamics

The traditional toolbox of Soviet political subversion included disinformation operations. The latter involved the production of carefully fabricated falsehoods and their injection into the information space of the adversary. A variety of channels and relays were used, ideally including ones that appeared far removed from the originator. In some cases this included the production of forgeries. For example, official letters were forged to falsely attribute objectionable intentions to Western governments, or fake scientific research was produced on the origins of the AIDS virus.[18] Sometimes it was simply a matter of disseminating false rumours. Contemporary Russian disinformation is not essentially different,

[16] P. Lewis, '"Fiction is outperforming reality": How YouTube's Algorithm Distorts Truth', *The Guardian*, 2 February 2018.

[17] See e.g. B. Rieder, A. Matamoros-Fernandez and O. Coromina, 'From Ranking Algorithms to "Ranking Cultures": Investigating the Modulation of Visibility in YouTube Search Results', *Convergence: The International Journal of Research into New Media Technologies* 24/1 (2018), 50–68.

[18] See e.g. T. Boghardt, 'Operation INFEKTION: Soviet Bloc Intelligence and Its AIDS Disinformation Campaign', *Studies in Intelligence* 53/4 (December 2009), 1–24.

except that online technologies have greatly multiplied the speed and potential reach and depth of such operations.

Based on a large-scale analysis of 126,000 news stories that appeared on the Twitter platform in 2006–17, a recent study by MIT researchers found that falsehood diffused significantly farther, faster, deeper and more broadly than the truth.[19] The MIT researchers found that bots did not spread false news significantly more than true news over the particular sample they analysed, though one should bear in mind that the study did not distinguish between misinformation and disinformation. Other research and analyses make clear that bots are an important component of hostile disinformation campaigns.[20] Most importantly, the aforementioned study by MIT researchers found that human users are more likely to spread falsehood than to spread the truth.[21] As with the other issues highlighted in this report, core vulnerabilities in the political information space stem from natural human psychological traits, which can be exacerbated and exploited in online environments. People's natural tendency to spread false news, the group polarisation issues discussed previously, and the phenomena of confirmation bias and selective exposure are sobering reminders that any official efforts to compete with falsehood by publishing the truth are bound to have only limited success.

Proposed public policy principles

This report has highlighted four areas of vulnerability in contemporary liberal democratic systems. On the basis of these observations, four public policy principles are formulated below. A discussion then follows with examples of steps taken so far in selected jurisdictions. In many cases, the responses imply a need for a range of actions, including new legislation at national level and new (or strengthened) programmes of activities on the part of governmental and/or intergovernmental organisations. The active participation of the major platforms is a necessary condition for the success of certain measures. In some

[19] S. Vosoughi, D. Roy and S. Aral, 'The Spread of True and False News Online', *Science* 459 (2018), 1146–51.

[20] See e.g. R. Fredheim and J. Gallacher, *Robotrolling* 2018/3, NATO Strategic Communications Centre of Excellence (August 2018).

[21] S. Vosoughi, D. Roy and S. Aral, 'The Spread of True and False News Online', 1149–50.

cases voluntary actions by platforms may be sufficient. In others new legal obligations will prove necessary to ensure that the public interest is protected.

The four principles

Preventing false-flag individualised political messaging

Voters should know who is addressing political messages to them. False-flag messaging should be reduced as much as possible. Moreover, hostile foreign actors should not be permitted to promote any kind of political messaging in the context of domestic political campaigns.

Defeating political astroturfing operations

The fairness of the political process is endangered if malign actors are able to tip the scales in favour of any particular political actor or group of actors. Measures should be taken to prevent or block political astroturfing operations, or to render them ineffective.

Regulating major content-selection algorithms

Without prejudice to principles of free enterprise and to the promotion of technical innovation, governance mechanisms should be introduced to ensure that major content-selection algorithms do not generate individualised information spaces that entertain unduly extreme, delusional, obsessive or paranoid mental states.

Countering disinformation

Liberal democracies must actively defend the integrity of their domestic political discourse against disinformation. Without prejudice to the freedom of expression and conscience of ordinary members of the public, the ability and capacity of hostile foreign actors to successfully carry out such operations should be countered.

Implementing the four principles: recent developments and recommendations

Preventing false-flag individualised political messaging

Major platforms should systematically add clear labels on to paid political messaging. These labels should identify not only the formal name of the sponsoring person(s) or organisation(s) but also, given the widespread phenomenon of front organisations, the identity of the ultimate sponsor. If the ultimate sponsor is a foreign entity, paid political messaging should not be accepted at all. These considerations have been reflected in new legislative proposals in the US and in France.

In the US a bill for an Honest Ads Act was introduced in the Senate in October 2017. The desired legal principles on clearly labelling political campaign messages and banning campaigning by foreign entities already exist in US electoral law, namely in the Federal Election Campaign Act of 1971. However, these principles are not clearly applicable to online platforms. The Honest Ads Act is thus a long-overdue measure to ensure that political campaign rules that apply to radio, television and print media also apply to online platforms. The Act would also obligate platforms to maintain a publically accessible file of all electioneering communications purchased by a person or group that spends more than $500 in total on advertisements published on the platform in question. In April 2018 both Facebook and Twitter publicly stated their support for the bill, although some media reports have suggested that Facebook lobbyists have attempted to convince lawmakers to trust their platform with a purely voluntary approach. The bill has yet to go through the legislative process.

In France a proposal for a new law on the fight against false information was submitted by members of President Macron's party in the National Assembly in March 2018.[22] The law would notably amend France's electoral code, as well as its broadcasting law. At the time of writing, it remains to be seen what

[22] France, Assemblée Nationale, Proposition de Loi relative à la lutte contre les fausses informations, Proposition no. 799 (21 March 2018).

the final outcome will be.[23] With respect to false-flag political messaging, under the French proposal platforms would be obligated to provide users with accurate, clear and transparent information on the identity of the people or organisations (and on whose behalf the latter operate, if applicable) that have paid the platform to promote informational content. Furthermore, platforms would have to publish the amounts received and the identity of the sponsors of informational content. In terms of substance, the proposed amendments to France's electoral law are very similar to those proposed for the corresponding US legislation.

Defeating political astroturfing operations

Ideally, astroturfing should be illegal as a matter of principle and prevented from occurring. Activities such as open political debate (on social media platforms or on the websites of major media outlets), online petitions and online public policy consultations ought to be fair and transparent mechanisms that allow citizens to contribute genuinely held views to the political process. But while astroturfing could easily be made illegal in principle, in practice enforcement could prove challenging unless quite stringent identity checks were put in place. Even if this is not done, certain steps could be taken while retaining largely open avenues for voluntary expression.

First, public authorities should fully accept that whenever hostile foreign actors contest a policy issue, there is a risk that these actors will launch astroturfing operations. For policymaking processes that require public consultation, governments and institutions such as the European Commission should make greater use of more reliable mechanisms, for example focus groups and polls based on random sampling. Statutory requirements and guidelines should be amended accordingly.

Second, for political campaigns public authorities should work in partnership with industry and independent researchers to improve the detection of fake online identities and of astroturfing campaigns, and to develop and test a range of options to reduce the impact of such campaigns. On the legislative side, the proposed French law on the fight against false information includes

[23] The National Assembly (the lower chamber) adopted its first reading position on 3 July 2018, essentially supporting the original text with relatively minor amendments. A minor setback occurred when the Senate rejected the text outright on 26 July 2018, rather than produce a rival first reading position. However, France's legislative process gives the last word to the lower chamber when the two chambers cannot reach a consensus.

an emergency mechanism, for electoral campaign periods only, which would allow a designated judge to take any measures necessary, including shutting down websites, if state authorities detected a case where false information 'that would alter the fairness of the upcoming vote is disseminated artificially and massively to the public through an online platform'.[24] The intention is to provide a legal basis for stopping a pre-planned operation that would rely on networks of trolls and bots seeking to make a false story go viral. While this provision would certainly be useful, it would not address long-term astroturfing campaigns.

Third, to deter the adversary, nations may choose to retaliate or to use threats of future retaliation. In response to Russian meddling in its 2016 election, the US imposed sanctions against the FSB, the GRU and other Russian entities in December 2016 (Executive Order 13757).[25] In March 2018 the US Treasury imposed additional sanctions on named senior directors of the GRU for election-related cyberattacks, and on named employees and associates of Russia's so-called Internet Research Agency, because the latter had 'created and managed a vast number of fake online personas that posed as legitimate U.S. persons to include grassroots organizations, interest groups, and a state political party on social media . . . [and] posted thousands of ads that reached millions of people online'.[26] Sanctions were thus imposed in response to both false-flag political messaging and political astroturfing.

Regulating major content-selection algorithms

While the algorithms of major platforms are proprietary, there is a public interest case for some form of independent scrutiny. The best approach would be to designate authorised independent bodies to audit systemically important algorithms. The objective would be to mitigate risks relating to political polarisation, extremism and self-radicalisation—and ultimately to individuals' mental health and to societal and political stability. This was first

[24] Art.1(I)(2). (Author's translation, emphasis added.)

[25] The FSB is the Federal Security Service of the Russian Federation. The abbreviation 'GRU' is commonly used to refer to the foreign military intelligence agency of the Russian armed forces, in line with the name it held from 1942 to 2010, i.e. Main Intelligence Directorate (GRU) of the General Staff of the Armed Forces (of the USSR up to 1992, of the Russian Federation from 1992). In 2010, the agency's name was shortened to Main Directorate (GU) of the General Staff of the Armed Forces.

[26] US Department of the Treasury, 'Treasury Sanctions Russian Cyber Actors for Interference with the 2016 U.S. Elections and Malicious Cyber-Attacks', Press release (15 March 2018).

proposed in an earlier version of this report.[27] The idea of auditing important algorithms without necessarily releasing them to the public was suggested more recently by author and mathematician Cathy O'Neil.[28] Building on these suggestions, states could draw inspiration from other cases of state regulatory bodies with governance and financing arrangements that ensure independence from both government and industry, and with a legal obligation to respect the confidentiality of proprietary information. Such a body should have the power to instruct a major platform both to modify algorithms it uses and to demonstrate that changes have been implemented in such a way as to achieve the designated outcomes. In the European context, the cross-border nature of the phenomenon means that it may be desirable to create a single EU-wide regulator under European law.

Countering disinformation

Of all the areas identified in this report, countering disinformation has generated the greatest response so far. This has included monitoring, flagging and debunking. However, this response has focused on reactive measures that implicitly accept the battlefield as it is, rather than trying to shape it (e.g. through new legislation). Many Western governments have created special inter-ministerial task forces.[29] The task forces typically involve the interior or justice ministries, alongside other ministries (most often defence and foreign affairs). In a smaller number of cases, dedicated centres are also involved. Both the EU's External Action Service and NATO's Public Diplomacy Division have dedicated programmes and budgets to monitor and respond to disinformation. A group of NATO Allies has also created the NATO Strategic Communications Centre of Excellence in Riga, which produces analyses and research, and provides expertise and training on countering hostile information activities by state and non-state actors.

Recent work at EU level calls for a series of measures to increase resilience in the face of disinformation. Particularly noteworthy are a major technical report

27 E. H. Christie, 'Artificial Intelligence in Tomorrow's Political Information Space', paper presented at the NATO STO Experts Meeting on Big Data and Artificial Intelligence for Military Decision Making, Bordeaux, 30 May – 1 June 2018.

28 C. O'Neil, 'Audit the Algorithms that Are Ruling Our Lives', Opinion, *The Financial Times,* 30 July 2018.

29 For a list of examples from specific countries, see J.-B. Jeangène Vilmer et al., *Les manipulations de l'information,* 119.

by the EU Joint Research Centre[30] and a Communication by the European Commission,[31] both published in April 2018. These two documents stress the need for greater transparency on the part of platforms in order to address challenges such as false-flag political messaging. To date the European Commission's chosen approach has been not to resort to new legislation, but to encourage platforms to develop their own solutions on the basis of a new EU-wide Code of Practice on Disinformation. In addition, the European Commission wishes to establish an independent network of European fact-checkers, to develop positive incentives to foster quality journalism and to encourage media literacy among the general public. A further area is the development of improved technologies, based on artificial intelligence, to identify, tag and verify disinformation.

The measures proposed in the European Commission's April 2018 Communication are well justified and should be supported with adequate resources. Two interrelated activities which the Commission has identified and assigned to voluntary actions by platforms are to 'facilitate users' assessment of content through indicators of the trustworthiness of content sources' and to 'dilute the visibility of disinformation by improving the findability of trustworthy content'.[32] This is a promising prospect which could be strengthened if new legislation were in place to ensure sufficiently strong incentives. The regulator for content-selection algorithms proposed earlier in this report could be the institutional actor who would ensure that the Commission's proposal is implemented appropriately. This regulator would be supported by a more robust independent fact-checking ecosystem (bearing in mind the permanent danger of political capture). In this context the Communication mentions the trustworthiness of both content and content sources, an important distinction which will require careful consideration.

[30] B. Martens et al., *The Digital Transformation of News Media and the Rise of Disinformation and Fake News: An Economic Perspective,* Digital Economy Working Paper 2018-02, JRC Technical Reports (April 2018).

[31] European Commission, *Tackling Online Disinformation: A European Approach,* COM (2018) 236 final (26 April 2018).

[32] Ibid., 8

Conclusions

The architecture that underpins political discourse in Western democracies has been utterly transformed in recent years. Social media platforms have become enormously powerful aggregators of news, opinion and debate. They also offer new forms of vulnerability that have been exploited by hostile actors in a fast-changing world. The central question for both national governments and the EU institutions is whether to accept the layout of this new battlefield of ideas, as it rapidly unfolds and changes with new technological advances, or, on the contrary, to lean forward and design new rules to protect the integrity of the democratic discourse. In both cases, policies and programmes to support activities such as fact-checking and media literacy will prove useful. In both cases, it may be possible to convince major platforms to take voluntary actions to mitigate some of the excesses that we have recently witnessed. It is, however, this author's view that efforts of a stronger nature are needed. The US and French proposals described in this report suggest that the legislative approach may be the best way forward for some of the policy principles that have been identified. In the bigger picture, there is no reason why legal restrictions on social media platforms should be off the table. Nor is it rational to expect platforms to go as far as may be needed to ensure an undistorted environment for our democratic discourse, unless clear incentives are put in place for them to do so.

Summary of recommendations

1. With regard to sponsored political messaging, drawing on the current US and French legislative proposals, national or EU legislation should ensure that major online platforms are obligated to clearly label and document the relevant sponsorship. Political messaging should not be hosted if its ultimate sponsors are not nationals of the state in which the targeted electoral process is scheduled to occur.

2. The phenomenon of political astroturfing should be studied and documented in greater detail. Its practice should be outlawed as well as actively discouraged, and approaches should be designed to mitigate its effects. Public authorities should work in partnership with industry and independent researchers to improve the detection of fake online identities and of astroturfing campaigns, and

to develop and test a range of options to reduce the impact of such campaigns. For policymaking processes that require public consultation, governments and institutions such as the European Commission should make greater use of mechanisms that cannot be distorted by astroturfing. These include focus groups and polls based on random sampling.

3. Systemically important content-selection algorithms should be subject to independent scrutiny by designated official bodies, with the aim of mitigating political and societal polarisation, extremism, self-radicalisation and other risks to individuals' mental health. Relevant new legislation could draw inspiration from other areas of industrial regulation and include governance mechanisms that would ensure confidentiality of information and independence from both government and industry. In the European context, the cross-border nature of the phenomenon entails that it may be desirable to create a single EU-wide regulator under European law.

4. Building on the European Commission's April 2018 Communication on tackling disinformation online, non-legislative activities to increase resilience should receive greater support and financing. Of special importance here are media literacy education, specialised training for journalists, fact-checking and debunking, and the development and piloting of artificial-intelligence solutions for detecting disinformation. The Commission's proposal to encourage platforms to develop indicators of content-source trustworthiness, and to make untrustworthy content or content sources less visible,[33] could be developed more decisively on the basis of incentives to be defined under new legislation. The regulator of content-selection algorithms proposed under Recommendation 3 could be tasked with ensuring that this proposal is implemented.

[33] The recently agreed EU Code of Practice on Disinformation, announced by the European Commission on 26 September 2018, describes the industry's proposed voluntary measures.

Bibliography

Boghardt, T., 'Operation INFEKTION: Soviet Bloc Intelligence and Its AIDS Disinformation Campaign', *Studies in Intelligence* 53/4 (December 2009), 1-24.

Christie, E. H., 'Artificial Intelligence in Tomorrow's Political Information Space', paper presented at the NATO STO Experts Meeting on Big Data and Artificial Intelligence for Military Decision Making, Bordeaux, 30 May – 1 June 2018, doi:10.14339/STO-MP-IST-160-PT-3-pdf.

Christie, E. H., 'Countering subversion online: what role for public policy?', in G. Bertolin (ed.), Hacking Humans: Responding to Cognitive Security Challenges. Riga: NATO StratCom COE (in print)

Del Vicario, M. et al., *Echo Chambers: Emotional Contagion and Group Polarization on Facebook*, *Nature*, Scientific Reports 6:37825 (2016).

European Commission, Tackling Online Disinformation: A European Approach, Communication, COM (2018) 236 final (26 April 2018).

France, Assemblée Nationale, République française, Proposition de Loi relative à la lutte contre les fausses informations, Proposition no. 799 (21 March 2018), accessed at http://www.assemblee-nationale.fr/15/propositions/pion0799.asp on 17 October 2018.

Fredheim, R. and Gallacher, J., *Robotrolling* 2018/3, NATO Strategic Communications Centre of Excellence (August 2018), accessed at https://www.stratcomcoe.org/robotrolling-20183 on 16 October 2018.

Hendrickson, C. and Galston, W. A., 'Why Are Populists Winning Online? Social Media Reinforces Their Anti-Establishment Message', Brookings Institution (April 2017), accessed at https://www.brookings.edu/blog/techtank/2017/04/28/why-are-populists-winning-online-social-media-reinforces-their-anti-establishment-message/ on 17 October 2018.

Jeangène Vilmer, J. B. et al., *Les manipulations de l'information : un défi pour nos démocraties*, Centre d'analyse, de prévision et de stratégie and l'Institut de recherche stratégique de l'École militaire (Paris, August 2018).

Kosinski, M., Stillwell, D. and Graepel, T., 'Private Traits and Attributes Are Predictable from Digital Records of Human Behavior', *Proceedings of the National Academy of Sciences* 110/15 (April 2013), 5802–5.

Kreissel, P. et al., '*Hass auf Knopfdruck. Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz*', Institute for Strategic Dialogue (July 2018), accessed at https://www.isdglobal.org/wp-content/uploads/2018/07/ISD_Ich_Bin_Hier_2.pdf on 17 October 2018.

'Lewis, P., '"Fiction is outperforming reality": How YouTube's Algorithm Distorts Truth', *The Guardian*, 2 February 2018, accessed at https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth on 17 October 2018.

Lord, C. G., Ross, L. and Lepper, M. R., 'Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence', *Journal of Personality and Social Psychology* 37/11 (1979), 2098–109.

Martens, B. et al., '*The Digital Transformation of News Media and the Rise of Disinformation and Fake News: An Economic Perspective*', Digital Economy Working Paper 2018-02, JRC Technical Reports (April 2018), accessed at https://ec.europa.eu/jrc/sites/jrcsh/files/jrc111529.pdf on 17 October 2018.

Matz, S. et al., 'Psychological Targeting as an Effective Approach to Digital Mass Persuasion', *Proceedings of the National Academy of Sciences* 114/48 (November 2017), 12714–19.

Meyers, D. G. and Lamm, H., 'The Group Polarization Phenomenon', *Psychological Bulletin* 86/4 (1976), 602–27.

O'Neil, C., 'Audit the Algorithms that Are Ruling Our Lives', Opinion, *The Financial Times*, 30 July 2018, accessed at https://www.ft.com/content/879d96d6-93db-11e8-95f8-8640db9060a7 on 17 October 2018.

Rieder, B., Matamoros-Fernandez, A. and Coromina, O., 'From Ranking Algorithms to "Ranking Cultures": Investigating the Modulation of Visibility in YouTube Search Results', *Convergence: The International Journal of Research into New Media Technologies* 24/1 (2018), 50–68.

Shearer, E. and Gottfried, J., 'News Use Across Social Media Platforms 2017', Pew Research Center (September 2017), accessed at http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/ on 16 October 2018.

US Department of the Treasury, 'Treasury Sanctions Russian Cyber Actors for Interference with the 2016 U.S. Elections and Malicious Cyber-Attacks', Press Release, 15 March 2018, accessed at https://home.treasury.gov/news/ press-releases/sm0312 on 17 October 2018.

US Senate, 'Extremist Content and Russian Disinformation Online: Working with Tech to Find Solutions', Sub-committee hearing, 31 October 2017, accessed at https://www.judiciary.senate.gov/meetings/extremist-content-and-russian-disinformation-online-working-with-tech-to-find-solutions on 16 October 2018.

Vosoughi, S., Roy, D. and Aral, S., 'The Spread of True and False News Online', *Science* 459 (2018), 1146–51.

Youyou, W., Kosinski, M. and Stillwell, D., 'Computer-Based Personality Judgments are More Accurate than Those Made by Humans', *Proceedings of the National Academy of Sciences* 112/4 (January 2015), 1036–40.

About the author

Edward Hunter Christie is an analyst at the Strategic Analysis Capability section, Emerging Security Challenges Division, NATO Headquarters. An economist by training, he worked as a research economist at the Vienna Institute for International Economic Studies from 2002 to 2010. After a period in EU affairs as Senior Policy Advisor and Chief Economist for a trade association, he joined NATO's International Staff in December 2014. He has published on a range of security- and defence-related topics, including energy security, economic sanctions, European defence spending, Russian defence spending and disinformation.

Credits

Wilfried Martens Centre for European Studies
Rue du Commerce 20
Brussels, BE 1000

The Wilfried Martens Centre for European Studies is the political foundation and think tank of the European People's Party (EPP), dedicated to the promotion of Christian Democrat, conservative and like-minded political values.

For more information please visit: www.martenscentre.eu

Editor: Dimitar Lilkov, Research Officer, Martens Centre
External Editor: Communicative English
Typesetting: Victoria Agency
Printed in Belgium by Drukkerij Jo Vandenbulcke

This publication receives funding from the European Parliament.