



ISBN: 978-9934-564-31-4

Authors: Emily Taylor, Stacie Walsh and Samantha Bradshaw

Project manager: Sebastian Bay

Text editor: Anna Reynolds

Design: Kārlis Ulmanis

Riga, November 2018

NATO STRATCOM COE

11b Kalciema Iela

Riga LV1048, Latvia

[www.stratcomcoe.org](http://www.stratcomcoe.org)

Facebook/[stratcomcoe](https://www.facebook.com/stratcomcoe)

Twitter: [@stratcomcoe](https://twitter.com/stratcomcoe)



**Emily Taylor** is CEO of Oxford Information Labs, an Associate Fellow of Chatham House and is Editor of the Journal of Cyber Policy. Emily has been involved in the internet environment since 1999. Previous roles include Chair, ICANN WHOIS Review Team, Internet Governance Forum MAG, Director of Legal and Policy for Nominet. A regular commentator on technology for news outlets. she has written for the Guardian, Wired, New Statesman and the Slate.

**Stacie Walsh** is Internet Policy and Cybersecurity Consultant at Oxford Information Labs. Stacie is an experienced researcher, data analyst, writer, presenter and project manager, focusing primarily on the Internet addressing (DNS) ecosystem, Internet of Things (IoT), Artificial Intelligence (AI), Over-the-Top (OTT) services, and cybersecurity. Stacie is a CESSG certified Cyber Security/Information Assurance Auditor Practitioner and holds a certificate in ISO/IEC 27001 Information Security Management Principles. In 2015 Stacie was an ICANN NextGen participant.

**Samantha Bradshaw** is a D.Phil. Candidate at the Oxford Internet Institute and a researcher on the Computational Propaganda Project at Oxford University. Samantha has published several journal articles and public policy papers at the nexus of technology, governance and politics. Samantha holds an MA in Global Governance and a BA (joint-honours) in political science and legal studies from the University of Waterloo.

This publication does not represent the opinions or policies of NATO or NATO StratCom COE.

© All rights reserved by the NATO StratCom COE. Reports may not be copied, reproduced, distributed or publicly displayed without reference to the NATO StratCom COE. The views expressed here are solely those of the author in his private capacity and do not in any way represent the views of NATO StratCom COE. NATO StratCom COE does not take responsibility for the views of authors expressed in their articles.

” After the election, I made a comment that I thought the idea misinformation on Facebook changed the outcome of the election was a crazy idea. Calling that crazy was dismissive and I regret it.’

*Mark Zuckerberg, 27 September 2017*

## INTRODUCTION

In November 2016, the idea that one of the world’s most sophisticated democracies could have been gamed by hostile states using social media platforms seemed ‘crazy’ to many people, not just Mark Zuckerberg. But as evidence continues to emerge, whether from FBI investigations, disclosures by whistle-blowers, or digitally-distributed disinformation campaigns in other countries, an uncomfortable picture takes shape—Silicon Valley’s leading technology platforms have found themselves at the centre of a perfect storm.

What have social media companies done to combat the malicious use of their platforms? What do the leading players’ initiatives tell us about their coping strategies? How are their actions supported by their current terms and policies? And have there been any substantial policy changes as a result of the proliferation of malicious use of social media? We examined the company announcements and terms of service (ToS) agreements of Facebook, Google, and Twitter between November 2016 and September 2018 and found:

- In 2016–18 the platforms made 125 announcements about initiatives aimed at addressing disinformation to various degrees, including:
  - Changes to the algorithms underlying newsfeeds or ad targeting
  - New partnerships with third-party fact-checkers
  - Investment and support for professional journalism
  - ‘Ads centres’ and greater transparency about electoral advertising, including reporting, labelling, and enforcement
  - Greater transparency in internal content moderation practices, as well as additional investments in both automated and human moderation
  - Heightened security and education for ‘high risk’ targets
  - Changes to third-party access to user data



- The companies' official blogs indicate that 'enforcement of current terms' is the most prominent response currently being undertaken, often through a combination of automation/ AI, ads centres, and human content moderation.
- The platforms' responses seem heavily influenced by news events. Official announcements often reference current reporting and the companies' actions suggest that their coping strategies are emergent at best, reactive at worst.
- The initiatives taken show differences between the strategies of the three largest platforms as they search for effective self-regulatory responses amid a firestorm of public and political opprobrium.
- Overall, we observed no major changes to terms and policies directly related to disinformation, leading to the conclusion that existing terms and policies provide platforms with levers to address these issues.
- It is apparent that new and impending regulations are impacting company policies. Over the course of this study all three platforms have updated their terms and policies in May 2018, largely reflecting the General Data Protection Regulation (GDPR).

## METHODOLOGY

This paper provides an inventory of the self-regulatory initiatives taken by three Internet platforms between November 2016 and September 2018 in response to disinformation activities. Internet companies have a variety of terms and policies, ranging from high-level user-oriented community standards to detailed legal terms. We limited our analysis to three sources of primary and secondary documentation: (1) official announcements and company blogs; (2) ToS agreements, Community Guidelines, and Privacy Policies; and (3) selected news reports relating to company self-regulatory responses.

In total, 125 company announcements, policies and news articles were reviewed. Of those, we identified 10 categories of interventions described in detail below. We then analysed the Terms of Service, community guidelines, and privacy policies of the three companies to determine whether company announcements resulted in changes to the rules that govern the platforms



and their use. Table 1 summarises the terms and policies we analysed for this paper. Oral and written evidence of official inquiries, the FBI's indictment of the Internet Research Agency, and the majority of news reports published during the research period were not included in the scope of this study, but provide a rich source of contemporaneous information for future inquiry. As comparing terms and policies across jurisdictions was not a focal point of this study, only terms relevant to Europe and the UK are included here. Some terms are universal, but many companies have additional or different terms for those living within and outside of the USA.

**Table 1: Terms and Policies Analysed**

Platform	Terms and Policies	Versions since November 2016
	Statement of Rights and Responsibilities/ Terms of Service	31 January 2018 19 April 2018
	Facebook Principles	
	Community Standards	
	Data Policy	29 September 2016
	Google Terms of Service	25 October 2017
	Privacy Policy	18 December 2017 2 October 2017 17 April 2017 1 March 2017 25 May 2018
	Twitter Terms of Service (living outside the US)	2 October 2017 25 May 2018
	Twitter Rules	
	Twitter Privacy Policy	18 June 2017 25 May 2018



# AN OVERVIEW OF PLATFORM ANNOUNCEMENTS

Between November 2016 and September 2018, the platforms made several announcements about interventions they were undertaking to combat the malicious use of social media. After evaluating 125 announcements, we identified 10 broad categories of interventions, which are summarized in Table 2:

1. **Algorithmic changes / AI** – adjusting algorithms to demote the visibility of poor quality news stories, to display a range of viewpoints, and to identify abusive behaviour, spam, or other kinds of harmful, illegal, or unwanted content
2. **Fact checking and flagging** – establishing partnerships with NGOs or news organisations to fact-check information in real time
3. **Enhanced reporting** – implementing user-friendly reporting mechanisms to improve misinformation detection
4. **Human content moderation** – hiring or training staff for content moderation
5. **Enforcement** – enforcing existing terms and advertisement policies
6. **Partnerships and research** – establishing partnerships with other social media platforms, as well as news organisations, universities, and civil society organisations, to conduct research on disinformation
7. **Media literacy programmes** – funding education programs for primary school-age children and other groups to improve media literacy
8. **Supporting quality journalism** – providing financial support to train journalists, creating tools for online subscriptions, and designing interventions to encourage newsroom diversity
9. **Improved transparency** – creating searchable archives for political advertisements, labelling advertisers, publicising internal policies and content moderation guidelines, and notifying users if they have interacted with foreign operatives
10. **Data protection measures** – strengthening internal data and privacy protections by limiting access to data by third-party apps, APIs, and even some of their own services



The initiatives announced by the platforms continue to be reactive and heavily influenced by news events, especially the events of 2016 and other prominent concerns that have drawn public attention. For example, Google announced adjustments to its autocomplete function and ranking algorithms following revelations that Holocaust denial websites were among top search results;<sup>1</sup> Twitter deleted millions of followers from celebrity accounts days after revelations surrounding the proliferation of fake accounts on the platform;<sup>2</sup> and following the Cambridge Analytica scandal, Facebook adopted several interventions including banning individuals who break their terms of service and locking down third-party access to apps and its API.<sup>3</sup>

The most popular response to the malicious use of social media by Facebook and Twitter was the enforcement of existing terms. Over the 22-month period of the study, more than half of Facebook's announcements referenced the enforcement of existing terms and policies, affecting both users and advertisers. For example, Facebook took action against over 30,000 fake accounts in the run up to the French Presidential Elections in 2017.<sup>4</sup> Facebook also increased its human content moderation team by approximately 5,000 in 2018.<sup>5</sup> In the same time period, more than 80 per cent of the announcements made by Twitter referenced enforcement of its terms and policies, primarily invoking anti-spam mechanisms to combat malicious automation on its

platform. In contrast, Google focused its announcements on partnership and research initiatives, highlighting cooperation with experts and third-party organizations to improve media literacy and strengthen quality journalism.

All three companies emphasized algorithmic changes to combat disinformation, as well as transparency initiatives around political advertisements. Google, Facebook, and Twitter are labelling political ads on their respective platforms, and have introduced location and user verification requirements upon purchasing. Google and Facebook are also taking a more global approach to combating the malicious use of social media. For example, both companies have announced several country-specific initiatives, such as partnering with local fact checkers in Brazil, or providing media literacy programs to help Canadian students spot 'fake news'.<sup>6</sup> To combat the spread of misinformation and limit the visibility of low-quality content, Facebook has announced several adjustments to its algorithms, such as downgrading content that has been fact-checked and flagged as false.<sup>7</sup> Twitter has focused on the use of machine learning and AI to identify bots and remove them from its platform. And Google has been developing new products to solve specific problems related to dis- and misinformation, such as tools to help legitimate newspapers gain subscribers.<sup>8</sup>

All three platforms have signed the EU's voluntary, non-binding Code of Conduct/



” The EU Code of Practice on Disinformation is a noteworthy step toward countering disinformation online. The Code reflects the four pillars of the ‘European approach’ outlined in the Commission’s report.

Best Practices document regarding hate speech and disinformation, and have pledged membership to organisations supporting transparency reporting such as the Global Networking Initiative (GNI). Facebook and Google—both members of GNI—have expanded their transparency reporting to include aspects seen as directly related to the malicious use of social media.<sup>9</sup>

The EU Code of Practice on Disinformation is a noteworthy step toward countering disinformation online. The Code reflects the four pillars of the ‘European approach’ outlined in the Commission’s report (i.e. transparency, diversity, credibility, and inclusivity).<sup>10</sup> However, the Code was created by the platform signatories themselves, reflecting their existing policies (i.e. the Annex on best practices) and current initiatives (i.e. ad transparency, support of high-quality journalism, and media literacy) without considering next steps, engaging in a critical analysis of lingering issues (i.e. the user notification and appeals processes), or guaranteeing human oversight.<sup>11</sup>

While there are some similarities across all three platforms, Google and Twitter’s responses to the malicious use of social media contrast markedly with that of Facebook. Facebook is active across all ten categories in this study, while Google and Twitter appear to position themselves as neutral conduits, focusing their efforts on creating technical tools or using AI to tweak algorithms or identify fake accounts. Unlike Facebook, Twitter and Google announcements rarely referred to user reporting, human content moderation, or data protection issues. However, proactive steps, such as modifying Terms of Service, changing business models, and addressing concerns of antitrust and market competition, have yet to be addressed by any of the three platforms. As a result, platform initiatives tend to focus on enhancing user experience and improving trust while working with third parties to improve education and journalism, rather than addressing the underlying business models built upon data collection and the attention economy, which can incentivise the spread of disinformation.



**Table 2: An Overview of Platform Announcements to Combat Disinformation (11.2016 - 09.2018)**

	<b>facebook</b> n= 80	<b>Google</b> n=26	<b>twitter</b>  n=19	<b>TOTAL</b> n= 125
Algorithm changes /AI	28	17	8	<b>53</b>
Fact checking and flagging	19	9	1	<b>29</b>
Enhance reporting	16	1	1	<b>18</b>
Human content moderation	23	3	4	<b>30</b>
Enforcement	42	6	16	<b>64</b>
Partnerships and research	28	19	7	<b>54</b>
Media literacy programs	9	9	3	<b>21</b>
Supporting quality journalism	8	15	1	<b>24</b>
Improved transparency	31	11	9	<b>51</b>
Data protection measures	16	4	1	<b>21</b>

**Source:** Authors, 2018. Based on official company announcements posted on relevant company websites and blogs. Table 2 shows the types of initiatives announced by each of the platforms between November 2016 and September 2018. The numbers represent the number of times each type of initiative was mentioned (1 per announcement per category). Note that several initiatives are typically mentioned in each announcement, hence the numbers in the columns do not add up to the ‘n’ for each company.



# AN OVERVIEW OF CHANGES TO TERMS OF SERVICE

Content moderation, terms of service, and privacy policies govern user activities on the platforms.<sup>12</sup> Thus, when evaluating industry announcements, we also looked for changes to the platforms' core policies. Although several interventions were identified through company announcements, no major changes to user terms and policies were observed (Table 3 summarises these findings). For instance, one prominent announcement was made in regard to the algorithmic demotion of disinformation. However, to date this strategy has not been matched with a relevant policy change in the user terms; such a policy amendment might describe the kinds of content that will be demoted, how users will be notified, and stipulate opportunities for appeal.

Facebook was the only platform to add new language to its terms and policies. The Community Standards policy published in May 2018 introduced new vocabulary such as 'integrity and authenticity', 'false news', and 'financial gain'. False news is to be downgraded—not removed—from the News Feed, and the policy now prohibits 'fraudulent criminal behaviour' and the use of 'misleading or inaccurate information to collect likes, followers or shares'. However, these modifications have not inherently changed Facebook's terms. This leads us to conclude that the current rules provide the platform with adequate opportunity to address these issues, indicating that the previous lack of response has not been due to a lack of regulations, but rather to a lack of enforcement.

Of all the companies included in this study, Google's policies employ the broadest language relevant to user content. Google retains more rights than the others to remove user content, particularly content that is 'illegal' or 'violates [Google's] policies'.<sup>13</sup> Yet, the terms lack clear definitions as to what constitutes unacceptable content beyond what is clearly illegal. The documents make no reference to fake news, propaganda, or dis- or misinformation. However, unlike the other platforms, which usually name specific jurisdictions such as the EU, Google's policy is to abide by national laws regarding disputes, consumer rights, and applicability of national law wherever California law is not applicable. In addition, the company distances itself from content liability stating in the Terms of Service, 'we don't make any comments about the content within the services'.<sup>14</sup>



Twitter's policies directly reference 'freedom of expression' in support of open dialogue,<sup>15</sup> and its terms state that the company does not 'guarantee the completeness, truthfulness, accuracy, or reliability of any Content'.<sup>16</sup> Users are required to enter into an agreement whereby they 'understand' that content 'might be offensive, harmful, inaccurate or otherwise inappropriate, ... or are otherwise deceptive'.<sup>17</sup> Although exclusions of liability for content are a regular feature of Internet company terms and policies, few companies in this study have adopted language referencing 'truthful' content in any form.

In summary, language used in the existing policies is broad enough to enable companies to apply their policies to a range of issues related to disinformation. Commonly used terms include 'illegal', 'unlawful', 'deceitful', and 'misleading'. The platforms have been able to rely on terms which prohibit spam to combat bots, as both spam and bots exhibit similar characteristics: repeated posting or sending of unwanted messages. The vague language of the terms of service across all three platforms, and the lack of significant changes to their user policies, highlight the tension that Internet companies contend with—reluctant to become 'arbiters of truth', yet must reserve the right to remove content that may cause harm to its brand as a safe, trusted online space.<sup>18</sup>

Fears of future regulation are shaping company policies and practices. When the EU's General Data Protection Regulation (GDPR) came into force in May 2018, all platforms updated their terms of service agreements to reflect changes introduced by the GDPR. The platforms also announced changes to their advertising policies and added new ad transparency initiatives in response to government concerns of misuse and potential regulatory action. We expect that as current proposals for national legislation (discussed by Bradshaw, Neudert, and Howard in Chapter 6) pass into law, there will be continued updates to terms and policies, as well as to platform activities concerned with enforcement, content moderation, combatting spam, third-party data access and use, and combatting spam.



**Table 3: Summary of Platform’s User Content Terms and Policies (11.2016 - 09.2018)**

Term	facebook	Google	twitter 
User has sole responsibility for content	Yes	Yes	Yes
User has right to opt out of curated content	-	-	-
User has right to opt out of targeted advertising	Yes	No	Yes (via device & browser controls)
Policy banning user content that is illegal, misleading, shocking, spam etc.	illegal	illegal, misleading, inaccurate, impersonation, spam (11/16-4/18) unlawful, misleading, fraudulent criminal behaviour, objectionable content including hate speech, integrity and authenticity including spam, misrepresentation (5/18-9/18)	truth, illegal, unlawful, fake, mislead, confuse, deceive, spam, malware, phishing
Direct reference to false information, misinformation, or fake news	-	false news	-
Spam policy can be applied to combat bots	-	-	Yes
Policies include right of provider to review content	Yes	Yes	Yes
Policies include specific right of provider to do automated review of content	Yes	-	-
Policies include specific right of provider to do manual review of content	-	-	-
Policies include right to take down material that is not illegal in provider’s home territory	Yes	Yes	Yes
Right to take down content deemed threat to public safety (harm, violence, etc.)	-	Yes	Yes
Right to take down posts without notification	Yes	Yes	Yes
Right to suspend user accounts or services without notice	Yes	Yes	Yes
Right to limit visibility of content	-	Yes	Yes



# CONCLUSION

2016 was a defining moment for social media platforms. The ongoing shock relating to election interference, computational propaganda, and the Cambridge Analytica scandal, combined with deeper concerns about the viability of the business model for established news media, all conspired to undermine the confidence of citizens and of public authorities in social media platforms. Initially, major social media companies fell back on traditional postures—minimising the impact by quoting statistics about the number of accounts involved—but our inventory of industry responses identifies and tracks changing attitudes.

Since November 2016, there has been a raft of self-regulatory responses by all three of the platforms examined in this paper. A key area for intervention is enforcement of existing terms and policies, as well as taking steps towards increased collaboration with other actors, including news media, election committees and campaigns, fact-checkers, and civil society organisations. However, we found little evidence of major changes to the underlying user policy documents. This may change as pressure to regulate platforms continues to mount following formal government inquiries into Cambridge Analytica, the spread of ‘fake news’, and evidence of foreign interference.

There may be trouble ahead as Google, Twitter, and Facebook appear to be taking conflicting stances on their responsibility for content. As government regulation appears inevitable, the platforms have formulated numerous solutions to combat the malicious use of social media. Yet, despite more than 20 months of inquiries and bad press, there is little evidence of significant changes to the companies’ terms and policies, which grant extensive powers over users’ content, data, and behaviour. Thus far, most of the self-regulatory responses have been reactive, responding to media cycle concerns around Cambridge Analytica or foreign interference in elections. The platforms themselves have not taken any meaningful steps to get ahead of the problem and address the underlying structures that incentivise the malicious use of social media—whether for economic gain or political influence. For meaningful progress to be made, and trust to be restored, the relationships between platforms and people needs to be rebalanced and platforms need to proactively work alongside government and citizenry as responsible actors.



# Endnotes

- 1 Frank Pasquale, "From Holocaust Denial To Hitler Admiration, Google's Algorithm Is Dangerous," *Huffington Post* (blog), February 6, 2017, [https://www.huffingtonpost.com/entry/holocaust-google-algorithm\\_us\\_587e8628e4b0c147f0bb9893](https://www.huffingtonpost.com/entry/holocaust-google-algorithm_us_587e8628e4b0c147f0bb9893).
- 2 Nicholas Confessore, Gabriel J.X. Dance, and Rich Harris, "Twitter Followers Vanish Amid Inquiries Into Fake Accounts," *The New York Times*, January 31, 2018, <https://www.nytimes.com/interactive/2018/01/31/technology/social-media-bots-investigations.html>.
- 3 Mike Schroepfer, "An Update on Our Plans to Restrict Data Access on Facebook | Facebook Newsroom," April 4, 2018, <https://newsroom.fb.com/news/2018/04/restricting-data-access>; Paul Grewal, "Suspending Cambridge Analytica and SCL Group from Facebook," *Facebook Newsroom* (blog), March 16, 2018, <https://newsroom.fb.com/news/2018/03/suspending-cambridge-analytica>.
- 4 Facebook, "Improvements in Protecting the Integrity of Activity on Facebook," April 13, 2017, <https://www.facebook.com/notes/facebook-security/improvements-in-protecting-the-integrity-of-activity-on-facebook/10154323366590766>.
- 5 Dominic Rushe, "Zuckerberg's Testimony: CEO Will Defend Facebook as 'Positive Force,'" *The Guardian*, April 9, 2018, <https://www.theguardian.com/us-news/2018/apr/09/mark-zuckerberg-facebook-testimony-congress>.
- 6 Nathaniel Gleicher, "Protecting Our Community in Brazil | Facebook Newsroom," July 25, 2018, <https://newsroom.fb.com/news/2018/07/protecting-our-community-in-brazil>; Facebook Canada, "Facebook Canadian Election Integrity Initiative," Facebook, 2018, <http://facebookcanadianelectionintegrityinitiative.com>.
- 7 Mark Zuckerberg, "Mark Zuckerberg - One of Our Big Focus Areas for 2018 Is Making...", January 2018, <https://www.facebook.com/zuck/posts/10104413015393571>.
- 8 Jim Albrecht, "Introducing Subscribe with Google," Google News Initiative, March 20, 2018, <https://www.blog.google/topics/google-news-initiative/introducing-subscribe-google>.
- 9 Facebook Newsroom, "Reinforcing Our Commitment to Transparency | Facebook Newsroom," May 15, 2018, <https://newsroom.fb.com/news/2018/05/transparency-report-h2-2017>; Philippe de Lurand Pierre-Paul, "Greater Transparency and Control over Your Google Ad Experience," Google, June 14, 2018, <https://www.blog.google/technology/ads/greater-transparency-and-control-over-your-google-ad-experience>.
- 10 European Commission, "Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Tackling Online Disinformation: A European Approach.," 2018, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236>.
- 11 Global Partners Digital, "The EU's Code of Practice on Disinformation: First Thoughts - News," 2018, <https://www.gp-digital.org/news/the-eus-code-of-practice-on-disinformation-first-thoughts>.
- 12 Adrian Chen, "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed | WIRED," WIRED, October 23, 2014, <https://www.wired.com/2014/10/content-moderation>; Sarah T. Roberts, "Behind the Screen: The People and Politics of Commercial Content Moderation," Open Transcripts, May 2, 2016, <http://opentranscripts.org/transcript/politics-of-commercial-content-moderation>; Emily Taylor, "The Privatization of Human Rights," Global Commission on Internet Governance, January 2016, [https://ourinternet-files.s3.amazonaws.com/publications/no24\\_web\\_2.pdf](https://ourinternet-files.s3.amazonaws.com/publications/no24_web_2.pdf).
- 13 Google, "Google Terms of Service."
- 14 Google.
- 15 Twitter has two full sets of Terms of Service – one for those living in the USA and another for those living outside the USA Twitter, "The Twitter Rules.". The terms for those living *outside* the USA were analysed for the purpose of this study.
- 16 Twitter, "Twitter Terms of Service."
- 17 Twitter.
- 18 Samidh Chakrabarti, "Hard Questions: What Effect Does Social Media Have on Democracy?," Facebook Newsroom, January 22, 2018, <https://newsroom.fb.com/news/2018/01/effect-social-media-democracy>.





Prepared and published by the  
**NATO STRATEGIC COMMUNICATIONS  
CENTRE OF EXCELLENCE**

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

Operating since 2014, we have carried out significant research enhancing NATO nations' situational awareness of the information environment and have contributed to exercises and trainings with subject matter expertise.

[www.stratcomcoe.org](http://www.stratcomcoe.org) | [@stratcomcoe](https://twitter.com/stratcomcoe) | [info@stratcomcoe.org](mailto:info@stratcomcoe.org)