ONLINE RADICALISATION AND ALGORITHMS

BY ISAAC KFIR

HJS
Henry
Jackson
Society

DEMOCRACY | FREEDOM | HUMAN RIGHTS

CENTRE ON RADICALISATION & TERRORISM

December 2021

Published in 2021 by The Henry Jackson Society

The Henry Jackson Society
Millbank Tower
21-24 Millbank
London SW1P 4QP

Registered charity no. 1140489
Tel: +44 (0)20 7340 4520

www.henryjacksonsociety.org

Title: "ONLINE RADICALISATION AND ALGORITHMS"
By Isaac Kfir

ISBN: 978-1-909035-70-6

£9.95 where sold

Front Cover image by Markus Spiske.

ONLINE RADICALISATION AND ALGORITHMS

BY ISAAC KFIR

HJS
Henry
Jackson
Society

DEMOCRACY | FREEDOM | HUMAN RIGHTS

CENTRE ON RADICALISATION & TERRORISM

December 2021

About the Author

Isaac Kfir is a Research Fellow at the Henry Jackson Society's Centre on Radicalisation and Terrorism. He received his PhD in International Relations at the London School of Economics in 1999 and has taught courses on 'Terrorism and Radicalisation' at the Reichman University (Israel), Syracuse University (New York) and Tokyo International University (Japan). Between 2017 and 2020 he was the Deputy Director of Defence, Strategy and National Security and the Head of the Counterterrorism Policy Centre at the Australian Strategic Policy Institute. He currently works as an independent consultant and an adjunct professor at the Charles Sturt University in Canberra, Australia.

Contents

About Us

HJS
Henry
Jackson
Society

DEMOCRACY | FREEDOM | HUMAN RIGHTS

About The Henry Jackson Society

The Henry Jackson Society is a think-tank and policy-shaping force that fights for the principles and alliances which keep societies free, working across borders and party lines to combat extremism, advance democracy and real human rights, and make a stand in an increasingly uncertain world.

CENTRE ON
RADICALISATION
& TERRORISM

About The Centre on Radicalisation and Terrorism

The Centre on Radicalisation and Terrorism (CRT) at the Henry Jackson Society is unique in addressing violent and non-violent extremism. By coupling high-quality, in-depth research with targeted and impactful policy recommendations, we aim to combat the threat of radicalisation and terrorism in our society.

Executive Summary

The role social media and technology companies play in people's lives and in society is enormous. Their services allow individuals to obtain news, connect with family and friends, shop online and interact with the world on unprecedented levels. This integrated ecosystem relies on continuing technological innovation, including the drastic change in mobile phone capabilities such as 5G and the growing ability of tech giants such as Facebook and Google to better mine, store and use personal information. The situation is affected by the fact that the technology is light years ahead of policy.

Governments and public bodies have demanded that social media and technology companies develop mechanisms to detect and remove extremist content. Some have tried to achieve this by reshaping their filters, revising their community and users' standards (Terms of Service), and adopting predictive algorithm modules. Others focus on content moderation, which means that, in their attempts to remove and limit extremist content, platforms operate as gatekeepers and organisers of content. In doing this, they assess content by using algorithms [1] to deprioritise what users see or to prevent them seeing content altogether (known as 'soft control'). If they determine the content does not meet specific standards, they may simply ban it – exercising 'hard control'.[2] The measures are generally temporary, as extremist actors adapt. For example, supporters of the Islamic State have recognised that Telegram's algorithms to detect extremist materials tend to analyse only one type of file for malicious content, and accordingly use the platform's flexibility to distribute files in different formats.[3] What the industry has sought to do has had limited efficacy and does not address the business model that continues to search for more content and user interaction.

The siloed, reactive approaches that have been taken to counter online radicalisation efforts underlie ongoing discussions about social media culpability in the process, though, as research suggests, what is missing is a more nuanced assessment, as technology itself plays a different role in the radicalisation of individuals. [4]

To address the phenomenon of online radicalisation, we need a better understanding of the systems, psychology and sociology that power the online world. This begins with algorithms – the computer programmes that identify and track users' online actions to help promote content. Principally, algorithms are meant to help users find what they want faster. However, over time, as more content has been uploaded, algorithms have become more sophisticated, serving to either recommend or hide content based on a user's personal browser history. [5] In other words, algorithms – which are key to search engines and social media – rank and determine who and what is visible and where. Through the collection of data, specifically a person's likes and dislikes, savvy individuals and companies can promote content that will appeal to the individual; this can provide enormous commercial value, but may also be used to foster forms of hatred and encourage violent acts. The challenge is that users are not necessarily appreciative that algorithms are the product of ontological decisions made by

1 It is worth noting that there is a typology of algorithms – some centre on search and trending while others focus on content filtering and ranking. There are also reputation, search and finance algorithms, as well as spam filtering algorithms and machine learning algorithms.

2 Soroush Vosoughi, Deb Roy and Sinan Aral, "The spread of true and false news online", *Science* 359, no. 6380 (March 2018): 1146-1151.

3 Bennett Clifford, "'Trucks, Knives, Bombs, Whatever:' Exploring Pro-Islamic State Instructional Material on Telegram", *CTC Sentinel* 11, no. 5 (2018): 27.

4 This study draws upon a body of scholarship, including information science, sociology, psychology and political science, that has looked to address how and why radicalisation occurs and what role social media and technology companies play in the process.

programmers and users, which means that the results they produce are not neutral as many believe but rather outcomes based on their programming.

This is an exploratory, qualitative, inductive piece of research aimed at encouraging further quantitative studies on whether there is a link between algorithms and radicalisation. It draws on a nascent body of quantitative studies looking at the link between internet use and extremism.[6] One of its goals is to underline the need for an effective regulatory regime that protects the public and supports business innovation. Such a result could happen if policymakers understood the platforms' business models and the social media and technology companies understood what good governance is.

An important caveat to this study is that information about social media and technology algorithms is hard to obtain because they are proprietary, and companies strive to keep them secret as these programmes drive their business model. Therefore, the assessment is largely deductive, aimed at generating an understanding of the social forces that drive and shape the dissemination of toxic information and why a reactive approach is counterproductive. Extremists have also built a parallel online world, one that operates as a self-contained, reinforcing echo-chamber. They have done this because of pressure from governments and platforms to remove toxic content. It is also vital to recognise the army of human-content moderators who spend countless hours reading, assessing and analysing content to help make the online space safer and more enjoyable.

Report Structure

This report comprises five chapters. The first chapter sets out the problem we seek to address and sets it in context. The second lays out the social media ecosystem, highlighting the power of algorithms and the role they play in collecting, assessing and disseminating information. The third chapter examines how users can 'game the system', highlighting how algorithms feed the development of toxic content and how nefarious entities, who understand the system, can use algorithms to radicalise. The fourth chapter reviews how the sector has responded to radicalisation, focusing on content moderation and de-platforming. Finally, the fifth chapter provides policy recommendations, ranging from quick wins to thinking outside the box.

[5] Joëlle Swart, "Experiencing Algorithms: How Young People Understand, Feel About, and Engage with Algorithmic News Selection on Social Media", *Social Media & Society* 7, no. 2 (2021): 2; A.J.A.M. van Deursen and J.A.G.M. van Dijk, "Measuring Internet Skills", *International Journal of Human–Computer Interaction* 26, no. 10 (2010): 891-916; Tarleton Gillespie, "The Relevance of Algorithms", in Tarleton Gillespie and Pablo Boczkowski (eds) *Media Technologies: Essays on Communication, Materiality, and Society* (Cambridge: MIT Press, 2014): 167-194.

[6] Tiana Gaudette, Ryan Scrivens and Vivek Venkatesh, "The Role of the Internet in Facilitating Violent Extremism: Insights from Former Right-Wing Extremists", *Terrorism and Political Violence* (July 2020), DOI: 10.1080/09546553.2020.1784147; Ryan Scrivens, "Exploring Radical Right-Wing Posting Behaviors Online", *Deviant Behavior* (April 2020): 1-15, DOI: 10.1080/01639625.2020.1756391; Maura Conway, Ryan Scrivens and Logan McNair, "Right-wing extremists' persistent online presence: History and contemporary trends", *ICCT Policy Brief* (October 2019): 1-24; The Editorial Board, "The New Radicalization of the Internet", *The New York Times*, 24 November 2018, https://www.nytimes.com/2018/11/24/opinion/sunday/facebook-twitter-terrorism-extremism.html.

Chapter 1: Introduction

Digitisation and social media have reshaped society, impacting every facet of human life: through technology, individuals can now market, communicate, collaborate, consume and create like never before. These forces have birthed a unique ecosystem committed to enhancing interaction and making life and the acquisition of information easier. [7]

Social media platforms have been identified as prime facilitators and locations for content and 'communities of practice' to emerge, [8] with some of the content aiding anti-social behaviour and extremist violence. [9] Extremists have adapted their communication model to be audience-based, as they construct toxic socio-political, economic and cultural narratives to project their ideological goals. One explanation for the spread of toxic content is simply that the content is untrue and evidence suggests that false information spreads faster than truthful information. That is, people respond more to stories that seem to inspire disgust, fear and surprise, as opposed to truthful stories that evoke sadness, joy, anticipation or trust. [10]

The ubiquitousness of social media and technology has metastasised the social world. Initially, the goal was to connect the world and create a global community that transcends geographical boundaries, but nefarious actors recognised the potential of the tools to aid them to achieve their goals. John Perry Barlow, the American poet, cyberlibertarian and essayist, idealistically portrayed cyberspace as a realm "without privilege or prejudice accorded by race, economic power, military force, or station of birth." [11] Barlow concluded that governance was unnecessary because in this space, "anyone, anywhere may express his or her beliefs, no matter how singular, without fear of being coerced into silence or conformity." [12] However, technological and social media innovation has enabled companies and individuals to acquire the ability to track what people are reading, watching and doing, including what they have 'liked' and 'disliked'. [13] By recognising that humans are predictable, and by tracking their actions, activities and behaviours, and what they like and dislike, it becomes possible to predict their actions and facilitate desired outcomes. This is even more so because of explicit actions (i.e., the manual personalisation tools a platform offers) and implicit actions (e.g., adjusting browsing behaviour). [14]

For a while, the platforms and public bodies did not know how to deal with these nefarious actors, possibly hoping that they would magically disappear. Yet the persistence of extremists

[7] It is a common misconception that one is acquiring knowledge when undertaking research. Search engines provide information, raw data, that over time could become knowledge if the user spends time critically evaluating the information.

[8] Jean Lave and Etienne Wenger, *Situated learning: Legitimate peripheral participation* (Cambridge: Cambridge University Press, 1991).

[9] See, for example, Charlie Edwards and Luke Gribbon, "Pathways to Violent Extremism in the Digital Era", *The RUSI Journal* 158, no. 5 (2013): 40-47; Loo Seng Neo, Leevia Dillon and Majeed Khader, "Identifying individuals at risk of being radicalised via the Internet", *Security Journal* 30, no. 4 (2017): 1112-1133. It is also worth noting that investigations by *The New York Times*, *The Washington Post* and *The Wall Street Journal* have highlighted how YouTube can facilitate radicalisation, polarisation and misinformation.

[10] Vosoughi, Roy and Aral, "The Spread of True and False News Online", 1146-1151.

[11] John Perry Barlow, "A Declaration of the Independence of Cyberspace", Electronic Frontier Foundation, 8 February 1996, https://www.eff.org/cyberspace-independence.

[12] Ibid.

[13] For example, users upload pictures via such platforms as Instagram (owned by Facebook) or ask for advice regarding how to get from place to place quickly through such applications as Waze (owned by Google), while shopping loyalty cards provide insights into what one consumes. This data is mined and traded. For example, Qantas' most profitable division is its frequent flyer unit, as the airline sells the air miles to third parties who use them to reward their own customers. Angus Whitley, "Qantas frequent flyer program turning into airline's biggest money spinner", *The Sydney Morning Herald*, 12 May 2017, https://www.smh.com.au/business/companies/qantas-frequent-flyer-program-turning-into-airlines-biggest-money-spinner-20170512-gw34wq.html.

[14] Swart, "Experiencing Algorithms", 2; Mario Haim, Andreas Graefe and Hans-Bernd Brosius, "Burst of the filter bubble? Effects of personalization on the diversity of Google News", *Digital Journalism* 6, No. 3 (2018): 330-343; Seong Jae Min, "From Algorithmic Disengagement to Algorithmic Activism: Charting Social Media Users' Responses to News Filtering Algorithms", *Telematics and Informatics* 43 (October 2019): 1-9.

who saw the potential of social media and technology in actualising their agendas demanded that measures to limit and ideally remove their ability to use these tools be adopted. The platforms responded with a plethora of policies and measures, ranging from content moderation to de-platforming, delisting, re-directing and the like. The measures were tailored to the needs of each platform and varied in their application, particularly because there is limited cooperation between the platforms as they look to appeal to different users and have different functions, capabilities and ideologies.

Research into radicalisation is expansive. One area needing greater clarity is the role of algorithms, which effectively operate as digital intermediaries,[15] directing and pushing content to vulnerable people. The internet, when understood as a depository of information and as a mechanism to connect people, plays an important role in radicalisation. Extremist actors constantly look to innovate, and there is therefore a need to consider how they can circumvent some of the measures adopted to resist the dissemination of radical content.[16]

Detective Superintendent Jim Hall, the head of the Welsh Extremism and Counter-Terrorism Unit, has identified the challenge by pointing out that during the COVID-19 pandemic, recruiters have adapted "their narratives and methods" to create "discord and distrust within communities".[17] Because there is a greater understanding of how platforms work, even though owners provide little detail about their algorithms, those who understand how information is searched for, disseminated and collected can obtain desired outcomes. Nefarious actors look to exploit algorithms to promote content that gatekeepers cannot identify or remove, initiating the process of getting users into a frame of mind that makes them receptive to toxic and divisive content.

When thinking about social media and technology companies and their approach to countering online radicalisation, it is useful to be conscious of the language that these entities use in describing themselves and their functions as they look to protect their business model and grow. They recognise that their services are increasingly indispensable, inviting ever-greater scrutiny. It is unsurprising that many describe themselves as 'platforms', emphasising that their role is to provide access to information or services aimed at making users' lives easier. Their role is not to police what is shared.[18] Such descriptions create the illusion that they are neutral and ultimately uninvolved in determining who can use them and how.[19]

The mainstream platforms rely on algorithms and humans to identify toxic content and remove it. This may occur prior to the uploading of content, known as *ex-ante* content moderation: an example from a different field is the screening of content before it is uploaded in order to determine whether it contains copyright-infringing material. Similar outcomes are achieved by anti-child sexual abuse material tools such as PhotoDNA and ContentID, which help to assess whether the content involves sexual exploitation. There are also *ex-post proactive* content-moderation mechanisms. These involve the use of automated content-moderator programming that actively seeks out content deemed to breach community standards. *Ex-*

15 Rory Van Loo, "Rise of the Digital Regulator", *Duke Law Journal* 66, no. 6 (March 2017): 1269.

16 Gaudette, Scrivens and Venkatesh, "The Role of the Internet in Facilitating Violent Extremism"; Stijn Sieckelinck, Elga Sikkens, Marion van San, Sita Kotnis and Micha de Winter, "Transitional Journeys into and Out of Extremism. A Biographical Approach", *Studies in Conflict & Terrorism* 42, no. 7 (2019): 662–82; Isaac Kfir, "Innovating to Survive, a Look at How Extremists Adapt to Counterterrorism", *Studies in Conflict & Terrorism* (May 2021): 1-19, DOI: 10.1080/1057610X.2021.1926069.

17 Caleb Spencer, "Coronavirus: 'Children may have been radicalised in lockdown'", *BBC News*, 30 June 2020, https://www.bbc.com/news/uk-wales-53082476.

18 A good example of this is the discussion regarding Apple's announcement that it would scan all shared photos that are uploaded to the Apple Cloud in case they include child exploitation images and, should that occur, Apple would inform the authorities. Jack Nicas, "Are Apple's Tools Against Child Abuse Bad for Your Privacy?", *The New York Times*, 18 August 2021, https://www.nytimes.com/2021/08/18/technology/apple-child-abuse-tech-privacy.html.

19 Barrie Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation", *Fordham International Law Journal* 43 (2019): 939, p.6.

post reactive content moderation refers to content flagged as possibly violating Terms of Use and therefore necessitating a review. [20]

The role of social heuristics is vital in understanding the allure of social media and technology and how the internet facilitates radicalisation because research shows individuals are less likely to move beyond an immediate circle as they look for affirmation and support for their beliefs, values and interests. [21] Social media and technology enable homophilic communities to remain because they sustain the 'information loop'. [22] In other words, people tend to trust people in their circle and will share information with the people in that circle, helping to reinforce assumptions, beliefs, outlooks and worldviews. The danger of information loops and homophilic communities has become more evident as people argue over facts in their search for 'truth'. Notable, and something that is significant to this research, is the verification of the old adage that falsehood tends to spread faster than truthful information. [23] This reality raises a unique challenge for those looking to prevent and counter radicalisation. If nefarious actors understand how algorithms operate and how they can be used to push content, these actors can 'game the system' and get social media platforms to push their toxic content to vulnerable individuals. This demands more thought into how platforms deal with the social system that they create, as opposed to simply relying on algorithms to remove material.

Over the last few years, research into online extremism and radicalisation has intensified as it has become clear that extremists are conscious of counter- and anti-terrorism measures. [24] It is evident that they adapt their messages to avoid detection, including moving to platforms they know will enable them to disseminate their messages. [25] They also know that much of the content moderation undertaken is technology-based, leading them to adapt the language, image and message. Moustafa Ayad provides an example of such adaption, noting that Egypt's President al-Sisi has attracted a neo-Nazi following. Members of this group avoid detection by mainstream social media platforms by ensuring that their content does not breach user guidelines – what they include on their pages is a link to Telegram Channels where the racist or extremist content is revealed. Ayad identifies a Facebook page with an image of the Egyptian god Horus, its wings wrapping a Swastika with a black banner running through it, added in order to avoid detection. The page includes a link to a Telegram channel holding 275-gigabytes of Nazi propaganda materials. [26]

There are indications that the COVID-19 pandemic and the resulting lockdowns have propelled people to spend more time online, exposing users, particularly young people, to more content,

---

20 See Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review* 131, no. 1598 (April 10, 2018); Spandana Singh, "Everything in Moderation", *New America*, July 2019, https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/introduction.

21 Andreas Schäfer, "Digital Heuristics: How Parties Strategize Political Communication in Hybrid Media Environments", *New Media & Society* (2021): 14614448211012101; Dongfang Gaozhao, "Flagging Fake News on Social Media: An Experimental Study of Media Consumers' Identification of Fake News", *Government Information Quarterly* (2021): 101591; Rune Karlsen and Toril Aalberg, "Social Media and Trust in News: An Experimental Study of the Effect of Facebook on News Story Credibility", *Digital Journalism* (2021): 1-17.

22 Brian Stelter, "Trump wants his base to watch Fox News or OANN on Mueller report day", *CNN News*, 18 April 2019, https://edition.cnn.com/2019/04/18/media/fox-oann-trump-tweet/index.html; Matthew Gertz, "I've Studied the Trump-Fox Feedback Loop for Months. It's Crazier Than You Think", *Politico*, 5 January 2018, https://www.politico.com/magazine/story/2018/01/05/trump-media-feedback-loop-216248/.

23 Vosoughi, Roy and Aral, "The Spread of True and False News Online", 1146-1151.

24 Counter-measures are primarily offensive approaches aimed at preventing, deterring or responding to an event or an action. The intention is to prevent, neutralise or mitigate the action. Anti-measures are defensive in orientation, with the goal being to understand the pull and push factors so as to take action that would reduce the vulnerability.

25 Clifford, "'Trucks, Knives, Bombs, Whatever", 23-29; Mia Bloom, Hicham Tiflati and John Horgan, "Navigating ISIS's Preferred Platform: Telegram1", *Terrorism and Political Violence* 31, no. 6 (2019): 1242-1254.

26 Moustafa Ayad, "Trump's Favorite Dictator Fueling New Pro-Hitler Movement", *The Daily Beast*, 11 August 2021, https://www.thedailybeast.com/al-sisi-trumps-favourite-dictator-fuelling-new-pro-hitler-movement-on-facebook-and-telegram.

some of which is toxic and dangerous. One study found a 70 per cent increase in hate speech between children and teens during online chats. It also found a 40 per cent increase in toxicity among young gamers communicating using gaming chat. [27] Obtaining information from a single source, and being exposed to only one set of 'facts', individuals end up constructing a unique vision of the world, one that is often masked by cognitive dissonance.

The social media ecosystem is a self-perpetuating data-generating cycle. In 2018, the total amount of data created, captured, copied and consumed globally was 33 zettabytes (ZB). Two years later, the amount rose to 59ZB. By 2025, it is expected that we will generate 175ZB of data. [28] The conventional profit-seeking business model monetises data. This can only be attained through interaction. [29]

A short primer on online radicalisation

Over the last two decades, research into radicalisation and online radicalisation has become sophisticated, with some scholars looking at the process while others have studied the pathways or triggers. Scholars have also sought to look at the causes of radicalisation through psychological assessments and the exploration of both socio-cultural and socio-economic factors. [30]

The studies generally indicate that contemporary radicalisation is a slow process, undertaken by actors who want to recruit members to their way of thinking. It involves social, political, economic, cultural, religious and psychological elements that lead an individual to adopt a set of ideas, values and desires that look to undermine mainstream society. The process does not occur in a vacuum, but requires a community of like-minded individuals who reinforce the individual's views. [31]

Online radicalisation needs to be seen as a two-step process. It begins through the dissemination of content on open social media platforms such as Facebook, Twitter and YouTube. The second stage involves the recruiter engaging with the recruit through private messages and/or encrypted platforms, allowing the recruiter to connect with the individual on a more personal level. Violent extremists watch the social media world and employ a host of strategies to recruit members and sympathisers.

Built into their recruitment strategies is a flexibility feature that allows them to post some information on a mainstream social media platform, but including a link to a third-party site carrying their toxic information. [32] For example, the Islamic State has a history of participating

---

27 Joanne Orlando, "Young people are exposed to more hate online during COVID. And it risks their health", *The Conversation*, 9 November 2020, https://theconversation.com/young-people-are-exposed-to-more-hate-online-during-covid-and-it-risks-their-health-148107.

28 Melvin M. Vopson, "The world's data explained: how much we're producing and where it's all stored", *The Conversation*, 4 May 2021, https://theconversation.com/the-worlds-data-explained-how-much-were-producing-and-where-its-all-stored-159964.

29 Steven Levy, "How 30 Random People in Knoxville May Change Your Facebook News Feed", *Backchannel*, 31 January 2015, https://medium.com/backchannel/revealed-facebooks-project-to-find-out-what-people-really-want-in-their-news-feed-799dbfb2e8b1.

30 For a review of some of the early radicalisation theory, see Alex P. Schmid, "Radicalisation, De-Radicalisation, Counter-Radicalisation: A Conceptual Discussion and Literature Review", *ICCT Research Paper* 97, no. 1 (2013): 22; Keiran Hardy, "Comparing Theories of Radicalisation with Countering Violent Extremism Policy", *Journal for Deradicalization* 15 (Summer 2018): 76-110.

31 See, for example, Peter R. Neumann, "The Trouble with Radicalization", *International Affairs* 89, no. 4 (2003): 873-893; Magnus Ranstorp, *Understanding Violent Radicalisation: Terrorist and Jihadist Movements in Europe* (London: Routledge, 2010); Mark Sedgwick, "The Concept of Radicalization as a Source of Confusion", *Terrorism and Political Violence* 22, no. 4 (2010): 479-494; Andrew Silke, "Holy Warriors: Exploring the Psychological Processes of Jihadi Radicalization", *European Journal of Criminology* 5, no. 1 (2008): 99-123.

32 Sajid Amit, Imran Rahman and Sadiat Mannan, "Social Media and Radicalisation of University Students in Bangladesh", *Journal of Policing, Intelligence and Counter Terrorism* 15, no. 3 (2020): 228-229.

in both mainstream and radical forums. Once they identify individuals or groups that espouse views critical of the mainstream, recruiters can then direct content towards them, hastening their radicalisation. [33]

This report offers an outline for studying the link between algorithms, online communities and toxic content that facilitates radicalisation. Toxic content comes in many forms, both in terms of the vector of delivery – words, images, videos and more – and in the content itself. It encapsulates such things as anti-social, discriminatory, offensive language, and content that presents a skewed vision of society driven by conspiratorial images and words that seek to entice a vulnerable person to join a group or movement that looks to change the established social, political or economic order, often through violence.[34] The report also highlights how algorithms keep track of a person's interests and interactions, pushing content that the programme believes the person wants to see and hear, thus leading the person to join a homophilic community where they can find fellow travellers. Algorithms are constructed by people to aid in decision-making. Drawing on copious amounts of data, they present options to the users. They are largely presented as neutral, free from human biases; however, they are open to abuse, manipulation and misuse if one knows how they operate. Therefore, if it is true that 70 per cent of the content watched on YouTube (over a billion hours a day) comes from recommendations, or algorithms, by understanding their role, we can begin to understand how nefarious actors can manipulate guidelines and content moderation to promote content aimed at undermining social cohesion.

[33] J.M. Berger, "Tailored Online Interventions: The Islamic State's Recruitment Strategy", *CTC Sentinel* 8, no.10 (2015): 19-23; J.M. Berger and Jonathon Morgan, *The ISIS Twitter Census* (Washington, D.C.: The Brookings Institution, 2015).

[34] Zeerak Waseem et al., "Understanding Abuse: A Typology of Abusive Language Detection Subtasks", *arXiv:1705.09899* (2017); Robert Faris, et al., 2016, "Understanding Harmful Speech Online", *Berkman Klein Center Research Publication* 2016-21 (2016).

Chapter 2: The Social Media Ecosystem

Writing over 2,300 years ago, Aristotle concluded that people search for happiness. He identified three types of happiness: pleasure, passion and purpose. In *Phaedrus*, Plato claimed that similarity leads to friendship; his student Aristotle developed this observation in *The Nicomachean Ethics*, asserting that people love those who are like them. These two ancient thinkers understood homophily, noting that humans look to form communities that reflect similar views, traits, ideas, norms and the like, and that these give people happiness. In other words, one would not want to live in a community with which one disagrees as that would not give happiness.

Historically speaking, geography, technology (or lack of) and social norms have hindered people's ability to form friendships. Social media and technology have, in some ways, removed the age-old restrictions, allowing people to connect with others without living in their local communities, and thus to form transnational homophilic communities that connect people on the grounds of religion, age, race, gender, culture, education, politics and the like. [35] Two other factors help explain the growth of online homophilic communities. Firstly, the use of algorithms that identify users' preferences. These algorithms are intended to make users' lives easier by tracking sites visited, content posted or material liked or disliked. Secondly, research indicates that once a homophilic community is formed, the members will draw on selective exposure to information, promoting the 'information loop'. In other words, people opt to interact with people and information that resonates with their outlooks, and consciously filter out information that challenges their perceptions. [36]

Centuries after Plato and Aristotle made their observations, the positive psychologist Mihaly Csikszentmihalyi developed the concept of "flow" to explore how individuals look for happiness, leading him to argue that happiness comes from transcendence, which comes from hyper-concentration which occurs when people are totally and utterly immersed in an activity. [37]

Social media relies on "flow". [38] It permits people to search for optimal happiness through engagement with information, images, words, videos and games, all of which facilitate transcendence as users can spend hours surfing, leading them to forget the physical world. Concomitantly, as social beings, humans search for communities, and social media facilitates the emergence of unique communities, allowing the user to live within a virtual bubble that they feel comfortable in because the members share their views. [39] This therefore means that social media enables people to search for happiness by encouraging them to take control over what information they acquire, leading them to look for other individuals who share their outlooks.

35 Pablo Barberá, "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data", *Political Analysis* 23, no. 1 (2015): 76-91. Miller McPherson, Lynn Smith-Lovin and James M. Cook, "Birds of a Feather: Homophily in Social Networks", *Annual Review of Sociology* 27, no. 1 (2001): 415-444.

36 Silvia Knobloch-Westerwick and Jingbo Meng, "Looking the other way: Selective Exposure to Attitude-Consistent and Counterattitudinal Political Information", *Communication Research* 36, no. 3 (2009): 426-448; R.K. Garrett, "Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate", *Journal of Communication* 59, no. 4 (2009): 676–699.

37 Mihaly Csikszentmihalyi, *Flow: The Psychology of Optimal Experience* (New York: Harper & Row, 1990).

38 Bandopadhyaya Suvojit, "Transcendence Through Social Media", *Journal of Media and Communication Studies* 8, no. 3 (2016): 25-30; Sophie H. Janicke-Bowles, et al., "Exploring the Spirit in US Audiences: The Role of the Virtue of Transcendence in Inspiring Media Consumption", *Journalism & Mass Communication Quarterly* 98, no. 2 (2021): 428-450; Stephen Graham, "Beyond the 'dazzling light': From Dreams of Transcendence to the 'remediation' of Urban Life: A Research Manifesto", *New Media & Society* 6, no.1 (2004): 16-25.

39 The terrorist Anders Breivik could live in a cosmopolitan city such as Oslo but be cut off from it completely in favour of a virtual existence where he spent hours, days, weeks and months playing games and interacting with people online. Ian Buruma, "One of Us: The Story of Anders Breivik and the Massacre in Norway by Åsne Seierstad – review", *The Guardian*, 26 February 2015, https://www.theguardian.com/books/2015/feb/26/one-of-us-the-story-of-anders-breivik-massacre-norway-asne-seierstad-review.

Key to the process of creating homophilic communities are data and algorithms – particularly ones that drive search engines. These enable people to look for information and connect with people who share their views, sentiments, values and hopes for the future, often in an uninhibited, unfiltered manner (there is always a platform or a service for what one needs or wants).[40] These algorithms look to provide the user with search results that are based on their past interests. In other words, they search for what the user has asked for and present the information in the manner and order that the algorithm has determined the user would want.

The social media ecosystem, algorithms and homophilic communities

The digital ecosystem exists on many levels and in many forms. There is the general mainstream ecosystem composed of leading social media and technology companies that provide many services to individual users and public bodies. Included in this part are such things as domain names, various protocols to manage and govern the World Wide Web, broadband and the like. Within this larger system, one identifies different operators. Some platforms and companies have chosen or are only able to provide specific services, be it in the form of messaging, storage, software or hardware. Other platforms and companies have managed to establish an ecosystem within the larger ecosystem as seen with Facebook, which has expanded beyond the initial platform and now offers such services as instant messaging, economic engagement, games, political interaction, cloud storage, currency and more. Google, Amazon and Microsoft have created similar ecosystems for their users.[41]

> YouTube is designed to catch you and not let go. Sometimes I access the site just to watch one thing, but then one of the related videos catches my eye. I watch that one, then another and another, and soon a five-minute visit has stretched far longer.[42]

The above quote from Jay David Bolter emphasises the core purpose of social media – getting the person to engage because it provides pleasure and happiness. However, to help the person reach happiness, the system requires data which is obtained through the interaction as the platform collects data. The process is driven by search engines and specific algorithms, which have given rise to algorithmic culture: the gradual abandonment of culture's publicness and the emergence of a strange new breed of elite culture purporting to be its opposite.[43] Simply, the sorting, classifying and hierarchy-ising of people, places, objects and ideas has been transferred to machines built to identify patterns in behaviour.

Surprisingly, the term 'algorithm' has proven rather difficult to define. Etymologically, the word algorithm has an interesting foundation. Its roots lie in the Greek word for number, arithmós (αριθμός), from which the English form arithmetic is drawn (it first appears in Chaucer's *Canterbury Tales* as *augrim*). In the contemporary world, an algorithm is understood to be a formal process, often expressed mathematically, that lays out a set of step-by-step procedures aimed at exposing a truth.[44]

---

[40] For example, people who do not want content moderation have flocked to forums known as chan boards or image boards where there is little oversight over content and one can post things anonymously. When 8chan was taken offline by Cloudflare and it failed to get other providers to work with the site, it reappeared as 8kun. Oscar Gonzalez, "8chan, 8kun, 4chan, Endchan: What you Need to Know", *CNET*, 7 November 2019, https://www.cnet.com/news/8chan-8kun-4chan-endchan-what-you-need-to-know-internet-forums/.

[41] Juan Carlos Miguel and Miguel Ángel Casado, "GAFAnomy (Google, Amazon, Facebook and Apple): The Big Four and the b-Ecosystem" in *Dynamics of Big Internet Industry Groups and Future Trends* (Springer: Cham, 2016): 127-148; David B. Nieborg and Anne Helmond, "The Political Economy of Facebook's Platformization in the Mobile Ecosystem: Facebook Messenger as a Platform Instance", *Media, Culture & Society* 41, no. 2 (2019): 196-218.

[42] Jay David Bolter, "Social Media Are Ruining Political Discourse", *The Atlantic*, 19 May 2019, https://www.theatlantic.com/technology/archive/2019/05/why-social-media-ruining-political-discourse/589108/.

[43] Ted Striphas, "Algorithmic Culture", *European Journal of Cultural Studies* 18, no. 4-5 (2015): 395-412.

[44] Ibid., 403-405.

At their most basic, algorithms are composed of problem-solving technologies shaped by a social process. [45] They help people make choices about food, health, education and so on. This reality stems from the fact that so much of our life occurs within the digital world where every movement and action is recorded, studied and catalogued, giving rise to surveillance capitalism. [46] Put simply, as we surf the internet, whether for pleasure or work, we leave a data trail that is collected by computer programmes that, over time, establish an ontological process based on interests, passions and desires, which allows them to anticipate what we want, like or desire. [47]

Algorithms structure possibilities and rank them. They determine which information is to be shown to a user, often basing it on their past interactions, allowing information to be ranked. Concomitantly, algorithms envision, plan for and execute with supposed detachment, objectivity and certainty. There is an expectation that they operate as filters, drawing on computational logic. [48] An integral feature built into algorithms is not only that they track one's movements but also the people that one interacts with. For example, when one posts a story on Facebook, the algorithm that powers the platform records people's reactions to the post. The increased reliance on algorithms has meant they shape the social world, which explains why they have attracted interest.

Social media relies on recommender algorithms and collaborative filters. These programmes and codes operate on two levels. Firstly, they track, catalogue and assess your online interactions – who you engage with, what you read, what you buy and so on. Secondly, by recording past engagement, including what people in the user's circle are doing, they can recommend similar products, engagements, information and so on. [49] If we are to understand that social media looks to create and promote homophilic communities, then the role of recommender algorithms and collaborative filters becomes very important – as they direct a specific type of content that reinforces narratives, assumptions, views and ideas.

A short overview of Salafi-jihadis pushing content on social media: the traditional model

When Salafi-jihadis initially entered the World Wide Web, the content they created was heavy on text and unprofessional, focusing mostly on discussion forums. This was primarily down to a lack of digital literacy and technological expertise. These Salafi-jihadis relied on traditional media to transmit their messages, which limited their ability to communicate freely with the world, as decisions about what was transmitted ultimately rested with influential news editors.

Technological innovation, particularly digital cameras and software editing programmes, coupled with a greater understanding of strategic communication, have allowed jihadis to adapt their messaging campaign, making it more widely available. [50] With technology and

45 Ulrike Klinger and Jakob Svensson, "The End of Media Logics? On Algorithms and Agency", *New Media & Society* 20, no. 12 (2018): 4655.

46 Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Nature at the New Frontier of Power* (London: Profile Books, 2019).

47 Kelley Cotter, "Playing the Visibility Game: How Digital Influencers and Algorithms Negotiate Influence on Instagram", *New Media & Society* 21, no. 4 (2019): 895-913.

48 Mike Ananny, "Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness", *Science, Technology, & Human Values* 41, no. 1 (2016): 93-117.

49 Badreesh Shetty, "An In-Depth Guide to How Recommender Systems Work", *Bulletin Beta*, 12 November 2021, https://builtin.com/data-science/recommender-systems.

50 In the early 2000s, online Salafi-jihadi content was mainly in Arabic and the websites themselves were rather unstable, with the URL changing regularly in part because the creators were trying to avoid detection. Maura Conway and Lisa McInerney, "Jihadi Video and Auto-radicalisation: Evidence from an Exploratory YouTube Study", *European Conference on Intelligence and Security Informatics*. Springer, Berlin, Heidelberg, 2008. http://doras.dcu.ie/2253/2/youtube_2008.pdf; Marcelo Royo-Vela and Katherine A. McBee, "Is IS Online Chatter Just Noise?: An Analysis of the Islamic State Strategic Communications", *International Journal of Strategic Communication* 14, no. 3 (2020): 179-202; Jason Burke, "The Age of Selfie Jihad: How Evolving Media Technology is Changing Terrorism", *CTC Sentinel* 9, no. 11 (2016): 16-22; Berger, "Tailored Online Interventions", 19-23.

social media changing, Salafi-jihadis developed specific programmes to address recruitment, proselytisation, legitimacy, intimidation and so on. By the 2010s, with millennials joining Salafi-jihadi groups, the use of social media drastically changed as the groups drew on new technology to spread their message, which centred on highlighting the subjugation and exploitation of Muslims. This led them to refer to the injustices and point to abuses committed by Arab leaders and the legacy of Western colonialism. Their goal was to highlight and manipulate grievances. Using GoPro cameras, 4G technology and applications such as Periscope, Twitch and Facebook Live, they brought their campaign of terror into people's homes – even live-streaming their acts of violence.

Groups such as the Taliban and Islamic State have created social media accounts. Looking at their accounts highlights two different approaches. For the Taliban, social media, particularly YouTube, has been important because its sympathisers use it to project positive images of what the Taliban is doing. The Taliban relies on these sympathisers because it is officially banned from Facebook and YouTube. The Islamic State Amal News Agency provides a different example. The Agency is very good at using technology and social media platforms to generate video and audio news, as well as training guides, in a manner that appeals to a young audience. It is challenging to remove the content because Amal has developed a sophisticated distribution system that relies not only on official Islamic State channels and distributors, but also on sympathisers, private channels and game platforms. [51]

Down the rabbit hole... self-radicalisation through online engagement

The term self-radicalisation is a misnomer as no one can become radicalised without some support, and yet it is becoming increasingly obvious that over the last few years more people have adopted anti-social, anti-establishment views – with the common denominator being online engagement. Reading case histories of those who have gone down the QAnon rabbit hole since the outbreak of the pandemic tells researchers something about their online engagements – how one online story led to another and then another, leading the individual to construct a new reality, one centred on cognitive dissonance.

A former QAnon disciple has outlined the process, noting how the lockdown in her home state, her extrovert character and a need for emotional, psychological and spiritual space led her to watch a series on YouTube that a friend had sent; this compelled her to 'do more research' that eventually pulled her down into the rabbit hole and nearly claimed her relationship. [52] Notably, this is not a unique case, but just one example of the many people who are succumbing to these wild ideas that have led to deaths and destruction. The story of a 13-year-old boy is representative of this sort of transformation. Having experienced a traumatic event at school and wanting more information, the boy turned to Google. The search engine provided results that, in the words of one of his parents, "flooded his developing brain with endless bias-confirming 'proof' to back up whichever specious alt-right standard was being hoisted that week. Each set of results acted like fertilizer sprinkled on weeds: A forest of distortion flourished." [53]

---

[51] For example, around 2014, the Islamic State opened a Twitter account – @ISILCats – showing its fighters playing with cats. It also included cat memes and images of life in Raqqa. Supposedly, the purpose was to show how normal life is under the Islamic State in Raqqa.

[52] Anastasiia Carrier, "QAnon Almost Destroyed My Relationship. Then My Relationship Saved Me From QAnon", *Politico*, 13 August 2021, https://www.politico.com/news/magazine/2021/08/13/qanon-radicalization-bernie-sanders-supporter-503295; Gemma Conroy, "Believers in QAnon and other conspiracy theories reveal how they climbed out of the rabbit hole", *ABC News*, 23 May 2021, https://www.abc.net.au/news/science/2021-05-23/ex-conspiracy-theorists-reveal-how-they-got-out-qanon/100153732.

[53] Anonymous, "What Happened After My 13-Year-Old Son Joined the Alt-Right", *The Washingtonian*, 5 May 2019, https://www.washingtonian.com/2019/05/05/what-happened-after-my-13-year-old-son-joined-the-alt-right.

In summary, the need to recognise the role of algorithms in facilitating the development and growth of homophilic communities is significant as these seem to sustain polarisation in society. In the past, those who held anti-social ideas had to invest considerable time and energy in finding like-minded individuals. But now, with increased interconnectedness, individuals can search for and connect with those who share their sentiments and easily create a group that mutually reinforces their toxic ideas. The search and engagement which is undertaken online allows them to reach transcendence.

Extremists have always relied on their own skills and knowledge to push their content, leading the technology and social media community to develop counter-extremist programmes that rely on technological solutions aimed at blocking and/or removing the content. However, with extremists adapting, including taking a more extradiegetic approach to developing content, the prospect of them 'gaming the system' increases.

Chapter 3: 'Gaming the system' and the prospect of exploitation

Attempts at exploiting the online world emerged as soon as we moved from Web 1.0 to Web 2.0.[54] Technological innovation made the Web more organic and responsive to specific social trends, and designers and programmers looked to create pages that resonated with search engines and etiquette guides.[55]

An integral feature of Web 2.0 is the e-commerce economy: the ability to monetise content and interaction. However, because there is so much content out there – with much of it free to access – it is not always easy or simple to attract users and to make money. This has led to some trying to 'game the system' or 'play the invisibility game'.

The phrase 'gaming the system' refers to users of data within the social media world who act on their knowledge of the rules governing the behaviour of systems to get to a specific outcome. This term can be contrasted with the more benign phrase 'playing the invisibility game'. The former refers to attempts to undermine the integrity of the system by looking to use automation to attract users to the content they are promoting whereas the latter reflects working with and within the system.[56]

Those wanting to exploit algorithms understand that search engines are not neutral engines but rather mechanisms powered by algorithms that could be exploited.[57] Gillespie points out that what drives the desire to 'game the system' is the need for user interaction, in the form of requests to link back to, like, retweet and spread the content, in the hope that if enough people do it, the search engines will pick up their content and position it high enough in the listing. This in turn creates a unique loop as the more users referred to the content, the higher it appears on the search engine, and the more people are likely to see it.[58]

The value of looking at attempts to 'game the system' when exploring online radicalisation is because extremists increasingly recognise that platforms and public bodies are more aware of their efforts to disseminate their toxic content. This recognition has led them to develop strategies that allow their content to remain on mainstream platforms as they tailor it to meet Terms of Use.[59] This adaption is a form of 'gaming the system', although another iteration of 'gaming the system' will refer to ways in which extremists look to mechanisms within the platforms that allow them to promote their specific content as they search for sympathisers and potential recruits.

From a commercial perspective, a person 'gaming the system' is undermining the integrity of a system's outcomes as algorithms interpret a user's behaviour based on underlying assumptions about how users will behave and what that behaviour signifies.[60] The reality, however, is that everyone seeks to maximise use and in doing so make money. Consumers want to use the

54 The term Web 2.0 was popularised by Tim O'Reilly as he identified a major change in how users came to interact with the World Wide Web, as technology allowed them to connect and share information and experiences in more meaningful ways. Tim O'Reilly, *What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software*, https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html.

55 Tarleton Gillespie, "Algorithmically Recognizable: Santorum's Google Problem, and Google's Santorum Problem", *Information, Communication & Society* 20, no. 1 (2017): 63-80; James Grimmelmann, "The Google Dilemma", *New York Law School Law Review* 53, no. 4 (2008): 939–950.

56 Cotter, "Playing the Visibility Game", 896-897.

57 Gillespie, "Algorithmically Recognizable", 63-80; Grimmelmann, "The Google Dilemma", 939–950.

58 Gillespie, "Algorithmically Recognizable", 64.

59 Jordan Eschler and Amanda Menking, "'No Prejudice Here': Examining Social Identity Work in Starter Pack Memes", *Social Media & Society* 4, No. 2 (2018): 1-13.

60 Cotter, "Playing the Visibility Game", 895-913.

online space to acquire goods and services at the best price, without exerting too much effort (there are tools aimed at finding the 'best deal' if one provides a certain amount of personal data). For sellers, having the 'right data' means that in the cut-throat world of commerce, they can have an edge on the competition. [61]

In looking to understand how one can 'game the system', it is useful to look at the role of digital influencers, as these micro- and macro-celebrities use social media for financial rewards. The influencer develops a virtual relationship with their followers who regard the person as someone they can trust and therefore they will follow their recommendations. [62] For the influencer, success is measured by the number of followers they have, as the larger the following, the more likely producers and advertisers are to pay them to promote products or services. [63]

Several factors help determine the ability of digital influencers to attract a large audience. Firstly, the digital influencer must understand how they can use the programme that manages the platform to maximise their exposure and reach more people. [64] This form of algorithmic knowledge or literacy is particularly valuable as it means a user can increase their exposure without too much in the way of effort or resources. Acquiring this knowledge can be challenging because of the proprietary nature of the algorithm. Nevertheless, qualitative research among social media users indicates patterns of experiential learning whereby the users interact with algorithms, reflect on their observations, and adopt measures to test their hypotheses. [65] Along with a better understanding of how the platforms work, digital influencers recognise that what would help promote their message is genuineness and authenticity.

A separate tactic in the 'gaming the system' arsenal is using content – whether words, images or memes – that is not outwardly violent and does not look to inspire violence. Maura Conway highlights this in her study of memes, pointing out that in 2015, more than 50 per cent of the Islamic State's online content was not aimed at offending – it was images of well-stocked markets, well-equipped hospitals, children's playgrounds and the like. [66] Such content could evade the attention of the content moderators as it was anodyne and therefore less likely to be flagged.

Bots: another way to 'game the system'

Content creators have become more innovative in how they search for new users and interact with them. Interactive software known as a bot can be used to encourage users to click on the content. [67] This use of bots goes back to the early days of the World Wide Web when they were used in gaming and chat rooms, giving rise to specific bots known as chatbots

61 Joseph Turow, Lee McGuigan and Elena R. Maris, "Making data mining a natural part of life: Physical retailing, customer surveillance and the 21st century social imaginary", *European Journal of Cultural Studies* 18, no. 4-5 (2015): 464-478.

62 Studies indicate consumers are 70 per cent more likely to purchase a product if it is recommended by someone they know personally and therefore trust, which explains why over 90 per cent of marketers have used an influencer. Paul Harrigan, et al., "Identifying influencers on social media", *International Journal of Information Management* 56 (2021).

63 Kelley Cotter, "Playing the Visibility Game", 895-913.

64 Ibid., 896-897.

65 Michael A. DeVito, et al., "How People Form Folk Theories of Social Media Feeds and What it Means for how we Study Self-Presentation", *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, https://dl.acm.org/doi/10.1145/3173574.3173694; Erin Klawitter and Eszter Hargittai, "'It's like learning a whole other language': The role of algorithmic skills in the curation of creative goods", *International Journal of Communication* 12 (2018): 3490-3510.

66 Maura Conway, "Routing the Extreme Right", *The RUSI Journal* 165, no. 1 (2020): 110; Charlie Winter, *Documenting the Virtual Caliphate* (London: Quilliam, 2015), pp. 30-37.

67 There are different types of bots – some are simple scripts of less than a page of code, whereas others are more experimental, aimed at showcasing the latest artificial intelligence techniques. Most bots are the simpler ones. As is often the case, when building a bot, it is generally best to employ the simplest approach that provides good real-world results. For a typology of bots, see William Marcellino, et al., "Counter-Radicalization Bot Research", (Santa Monica: RAND Corporation, 2020), https://www.rand.org/content/dam/rand/pubs/research_reports/RR2700/RR2705/RAND_RR2705.pdf.

whose purpose is to interact with humans and in doing so draw information from the human.[68] Increasingly, bots are used to spread misinformation, promote disinformation, connect people to a social network, disrupt a social network and obtain and exploit personal information to facilitate criminal activities.

Interest in social bots stems from the ease with which one can manufacture the programme – a simple bot is normally less than a page of code – and because of increased evidence of how bots can impact on electoral processes, social discourse, financial services and so on.[69] Social media encourages bot use because social media ecosystems depend on easily accessible application programming interfaces to allow the development of apps and interactive features; this programming can also be used for the development of bots.[70] And yet platforms also look to minimise the surreptitious use of bots through various mechanisms that look to detach and shut down bot activities.[71]

The most concerning bots are those that are programmed to harvest information. These bots explore social media platforms, sending requests which, if accepted, enable the bot to collect information on the user, such as what they post or receive on their news feed. The impact of bots became more obvious when the Islamic State, Russia, North Korea and other rogue actors launched successful bot attacks against their enemies. One study highlights how manipulative social bots are by looking at how their programmers designed them to initially appear as a benign, non-political engagement that over time became more overtly political – aimed at slowly persuading the exposed human to adapt their outlooks.[72]

In 2014, the Islamic State developed an application called the 'Dawn of Glad Tidings' which enabled it to promote thousands of tweets a day about life under IS until Twitter shut it down. Notably, the application, which many downloaded, demanded a lot of private information from users. J.M. Berger, who has studied the application, points out that the tweets it spread included links, hashtags and images. Moreover, the application also retweeted content from those who had signed up, thus increasing its amount of content, meaning it would appear higher in search engines. Significantly, the Islamic State designers were cognisant of Twitter's attempts to limit spam through its spam-detection algorithms, which is why the application included a mechanism that allowed tweets to be spaced out, creating the illusion that a human was doing the tweeting.[73]

This underlies a growing trend within social media to rely on automation to promote content; individuals and groups look for ways to use the tools that the platforms have used in their meteoric rise to advance their interests. The platforms encourage such a pursuit because they want users to engage as every click generates revenue.

---

68 Amit Kumar Tyagi and G. Aghila, "A Wide Scale Survey on Botnet", *International Journal of Computer Applications* 34, no. 9 (2011).

69 Emilio Ferrara, et al., "The Rise of Social Bots", *Communications of the ACM* 59, no. 7 (2016): 96-104; Taberez Ahmed Neyazi, "Digital Propaganda, Political Bots and Polarized Politics in India", *Asian Journal of Communication* 30, no. 1 (2020): 39-57.

70 Microsoft, for example, offers a preview version of its "Bot Framework"; its purpose is to provide a unified programming interface that will enable bots to interact with different services, including Skype, Slack, Facebook Messenger, Kik and Office 365 email. William Marcellino, et al., "Counter-Radicalisation Bot Research", p.6.

71 A University of British Columbia study of the Facebook Immune System revealed that the software only removed around 20 per cent of bot activity. Yazan Boshmaf, et al., "The Socialbot Network: When Bots Socialize for Fame and Money", Proceedings of the *Twenty-Seventh Annual Computer Security Applications Conference*, Association for Computing Machinery, 2011.

72 Andrew Weisburd, Clint Watts and J.M. Berger, "Trolling for Trump: How Russia Is Trying to Destroy Our Democracy", *War on the Rocks*, 6 November 2016, https://warontherocks.com/2016/11/trolling-for-trump-how-russia-is-trying-to-destroy-our-democracy/.

73 J.M Berger, "How ISIS Games Twitter", *The Atlantic*, 16 June 2014, https://www.theatlantic.com/international/archive/2014/06/isis-iraq-twitter-social-media-strategy/372856/.

Chapter 4: Platform Governance

Growing concern with toxic content has led governments and public bodies to encourage and compel technology and social media companies to introduce measures aimed at governing what is uploaded and kept. Platform governance has become central to understanding the online world as different parties – platform companies, users, advertisers, governments and other political actors – are all involved in regulating the online space. [74] Conversely, watching the leaders of Amazon, Microsoft, Google, Facebook and Twitter testify before Congress has helped to explain why platform governance has been piecemeal and disjointed as the interactions highlight the lack of understanding by policymakers of the technology. [75]

The academic and policy worlds have responded to the need to address online radicalisation through many studies and assessments, giving rise to a better understanding of how specific online characteristics encourage violent extremism, with the internet operating as an echo-chamber allowing the radicalised to search and find fellow travellers who reinforce their world view. [76] In this echo-chamber, and depending on the rules that govern the platform, extremists foster and more importantly normalise a social environment that reflects and magnifies their worldview. [77] Finally, platforms and technology allow the process to occur without physical contact, making countering the threat more challenging.

The pervasive use of social media, and the increasing use of these platforms by extremists, has required social media and technology companies to look for ways to counter, hinder and prevent extremists from using their products for anti-social purposes. One tool that has been used is content moderation, which is understood to be "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." [78] Such an expansive definition underlies the challenge of moderating content as it involves a myriad of people and machines. This is made all the more challenging by the fact that one is dealing with platforms that transcend geographical boundaries, which means abiding by different regulatory systems, cultures, ideas, histories and politics.

Recognising the power of social media, governments and the platforms themselves have sought to find ways to undermine online radicalisation efforts through initiatives such as de-platforming, suspending and blocking accounts, all of which indicate a commitment to govern platforms.

Tautologically, content moderation takes place either implicitly or explicitly. The latter refers to companies laying out what is permissible and what is not. Assessment is determined through community standards, terms of service and internal moderation guidelines that companies may keep private. Explicit moderation is easier to ascertain as the focus is on preventing hate speech, violence, breaches of copyright and illegal actions. Because the violations are clearer,

74 Robert Gorwa, "What is Platform Governance?", *Information, Communication & Society* 22, no. 6 (2019): 854-871.

75 Senator Orrin Hatch's question to Mark Zuckerberg as to how he can sustain a business model when users do not pay for the service highlights this lack of understanding, as does Senator Brian Schatz's question about whether Facebook can read emails that he sends on WhatsApp.

76 See, for example, Gabriel Weimann and Katharina von Knop, "Applying the Notion of Noise to Countering Online Terrorism", *Studies in Conflict & Terrorism* 31, no. 10 (2008): 883-902; Robyn Torok, "Developing an explanatory model for the process of online radicalisation and terrorism", *Security Informatics* 2, no. 1 (2013): 1-10.

77 See, for example, Tim Stevens and Peter R. Neumann, "Countering Online Radicalisation: A Strategy for Action", *International Centre for the Study of Radicalisation and Political Violence*, 2009, https://icsr.info/wp-content/uploads/2010/03/ICSR-Report-The-Challenge-of-Online-Radicalisation-A-Strategy-for-Action.pdf; Ines von Behr, et al., "Radicalisation in the Digital Era: The Use of the Internet in 15 Cases of Terrorism and Extremism", (Santa Monica: RAND Corporation, 2013); https://www.rand.org/pubs/research_reports/RR453.html; Thomas Zeitzoff, "How Social Media is Changing Conflict", *Journal of Conflict Resolution* 61, no. 9 (2017): 1970–1991.

78 James Grimmelmann, "The Virtues of Moderation", *Yale Journal of Law and Technology* 17, no. 1 (2015): 42-109.

companies may take *ex-post* action, which means removing the content after it has been posted, or they may use *ex-ante* moderation, where content is not allowed onto the platform. [79]

The exponential growth of online content, and the complexity of the online world as it connects so many different platforms, means that platforms cannot regulate everything that is uploaded. Consequently, they have turned to machine learning, artificial learning and specific algorithms to help them identify and remove content deemed to breach community and user guidelines. These algorithmic moderation tools classify the content based on matching or prediction assessment; however, it is becoming clear that they struggle with extradiegetic content, which requires human involvement in the process. [80]

From the moment the World Wide Web changed from having a small number of writers creating static websites that allowed no real engagement by users to something that sought out user engagement, the seed for content moderation was laid. With Web 2.0, a myriad of technology and services emerged, all aimed at getting users to spend time, money and effort creating content mostly for monetised purposes. Put differently, Web 1.0 was highly limited in terms of content –something that is simply no longer the case. Instead, the challenge is the enormous amount of content that is uploaded, all of which needs to be sorted so that it can bring forth engagement through such tools as search engines, which sort out the upload chaos through a system that counts links and identifies interdependence. [81]

In Western society, content moderation is extremely challenging because of the centrality of free speech, which includes protecting an individual's ability to disseminate hate speech. In *Abrams v. The United States* (1919), the Supreme Court found that the most effective way to combat such speech is with speech that reveals falsity and challenges offensiveness. [82] The message from *Abrams* that one should rely on the marketplace of ideas to address toxic speech was arguably effective prior to the internet. However, the increased ability to connect with people and share falsehoods and offensive materials requires some form of governance to control what is uploaded and shared, leading regulators to look for ways to exert authority over transnational corporations. [83]

Over the last decade, and often in response to specific acts of violence, governments have shown their exasperation with social media and technology companies – encouraging them to take more proactive, regulatory measures to address toxic content. Hannah Bloch-Wehba correctly writes "As a general matter, efforts to enlist intermediaries to curb unwanted speech on a global basis are nothing new, although they are constantly evolving." [84]

The nature of the internet means that the ability of governments to regulate and moderate content is limited, as a government's power ends at the water's edge. To address the need for content moderation, the social media and technology companies have developed and adopted a host of measures to address the dissemination of toxic content. Content moderation refers to a set of practices aimed at ensuring that the users abide by guidelines in terms of how they use platforms.

---

79 Sander, "Freedom of Expression in the Age of Online Platforms", 947.

80 Robert Gorwa, Reuben Binns and Christian Katzenbach, "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance", *Big Data & Society* 7, no. 1 (2020).

81 James Grimmelmann, "The Google Dilemma", *New York Law School Law Review*, Vol 53 No. 4 (2008): pp.939-950.

82 *Abrams v. United States* 205 U.S. 616 (1919).

83 For example, in 2015, the French privacy regulator CNIL demanded that Google remove all links to pages containing false or damaging information about a person, should that person look to assert the 'right to be forgotten'. A year later, Google used a geo-blocking measure to prevent Europeans from seeing a delisted page, however, the company refused to apply the blocking measure globally, a position that the European Court of Justice upheld.

84 Hannah Bloch-Wehba, "Global Platform Governance: Private Power in the Shadow of the State", *Southern Methodist University Law Review* 72, no. 1 (2019): 40.

The process involves a large labour force and automated tools. It is largely led by the platforms but also involves law enforcement and specialist public bodies such as Internet Referral Units that must balance the right to free speech with the need to provide security, while working within the parameters of economic liberalism which call on the government to let the market decide. [85] The need to moderate content has led companies to adopt a two-prong strategy: automated and manual. [86] The former underlies the reality that humans cannot oversee, assess and review the copious amounts of data that are uploaded daily, necessitating the use of algorithms to detect, filter, flag, separate and remove content or accounts that violate the terms of use. Manual content moderation accepts that machines are fallible, shaped by the biases of their programmers, which is why platforms have hired, trained and deployed human moderators to address disputed content.

Government regulatory content moderation and the Online Safety Bill

Possibly because of frustration at the piecemeal way in which platforms have attempted to address online hate speech, radicalisation and violence, the UK Government – following in the footsteps of others – has drafted the Online Safety Bill. [87]

The draft legislation places a duty of care on social media platforms and some technology companies to protect people from the harm that could be caused by their products. This approach differs from the German Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG), [88] which penalises the platforms for not removing extremist content. Under the NetzDG, social media companies are fined after the fact for hosting illegal or harmful content. The legislation mandates that the platform has 24 hours to remove content that is "evidently unlawful"; failure to remove the content could lead to a fine of up to €50 million. In June 2020, the German Parliament added a provision requiring platforms to remove "illegal content". [89] Australia has taken a different approach through its Online Safety Act, which includes a complaint-based removal notice scheme for cyber abuse. The legislation also empowers the eSafety commissioner to block sites if they threaten the online safety of Australians. Linked to the measure is the Safety by Design initiative, led by the eCommissioner, which looks to educate designers and programmers to think about users' safety as they build their products, rather than only considering the desire to be first and to maximise profits.

The approach by the British Government places a positive obligation on platforms to protect people from the harm that they could suffer from using social media by looking at the services and products offered to users.

In summary, the platforms' attempts at governance to date have proven rather limited when it comes to countering extremist content because the dominance of economic liberalism has made whole-of-society content moderation governance challenging because there is a demand that private enterprise introduces measures of social responsibility to their business model, but at the same time adheres to neoliberal economic ideals that call for limited regulation and respect for basic human rights such as freedom of speech. Attempts to develop a single

85 Bharath Ganesh and Jonathan Bright, "Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation", *Policy & Internet* (2020): 6-19.

86 Singh, "Everything in Moderation".

87 Alex Hern, "Algorithms on social media need regulation, says UK's AI adviser", *The Guardian*, 4 February 2020, https://www.theguardian.com/media/2020/feb/04/algorithms-social-media-regulation-uk-ai-adviser-facebook.

88 Netzdurchsetzunggesetz [NetzDG] [Network Enforcement Act], 1 September 2017, https://germanlawarchive.iuscomp.org/?p=1245.

89 The law has attracted many criticisms, particularly from human rights advocates such as Human Rights Watch which describes the law as vague and overreaching. Moreover, it was claimed that the legislation turned the platforms into censors which would err on the side of caution in removing content to avoid the fines. "Germany: Flawed Social Media Law", *Human Rights Watch*, 14 February 2018, https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

regulatory regime are torpedoed because governments refuse to relinquish jurisdictional control, which benefits social media and technology companies which do not want to spend the resources to address the ills of the social world.

Chapter 5: Conclusion and Policy Recommendations

This exploratory study sought to explore the ways nefarious actors look to 'game the system' as they adapt to the regulatory regime imposed by public bodies and platforms to curtail the spread of toxic information. The intention was to emphasise that a siloed approach, one that relies on platforms or states taking individualistic action, is counterproductive, allowing extremists to exploit these gaps in the regulatory regime.

There is a need for a concrete, homogenous, regional and global response to the way platforms operate, which begins by reassessing their place in the social world. Simply put, seeing them as neutral service providers who play no role in how information is presented is simply untrue as information is filtered to users through algorithms whether operating as search engines or as content moderators. Although it is true that most users do not understand the way content is presented to them, an increasing community of individuals, as demonstrated by the digital influencers, have learned how to 'game the system' as they look to monetise their online presence. If these digital entrepreneurs can figure out how to increase their exposure and use algorithms to promote their content, so can nefarious actors.

Education, education, education

A study published by the American Psychological Association found that when users search the internet for information, they feel smarter. [90] The Yale University researchers conducted a series of experiments that revealed that after searching the internet for information, the user had an inflated sense of their own knowledge, even if they did not find the information. Moreover, the users believed the search made their brains more active. Such studies highlight the need for better education when it comes to the internet and the online ecosystem. Users must be trained to recognise that searching for information and seeing the information does not mean one has found 'the truth'.

Users need to learn that search engines are driven by algorithms aimed at assessing our requests and fulfilling them, which means answers tend to reflect what the algorithm thinks we want, which requires an education programme.

Algorithmic resilience as a prevention measure

There has been a big push towards developing resilience among internet users to countenance the spread of extremism. The concept, originally found in engineering, has been adapted to encourage a notion of bouncing back and returning to a state of equilibrium following adversity.[91] In social science and public policy, a resilience framework takes two forms. Firstly, it refers to a training mechanism that allows an individual to resist a process or an event. Secondly, it is about building up an ability to respond to an attack. Resilience, in other words, looks to enable individuals to resist something nefarious or dangerous by building up their character and values, while also giving them the equipment, i.e., training, to prevent the occurrence.

In the sphere of online radicalisation, resilience has appeared in two forms: specific training programmes aimed at building resilience among vulnerable communities and general initiatives looking at more societal solutions. [92]

---

[90] Matthew Fisher, Mariel K. Goddu and Frank C. Keil, "Searching for Explanations: How the Internet Inflates Estimates of Internal Knowledge", *Journal of Experimental Psychology: General* 144, no. 3 (2015): 674-687.

[91] David E. Alexander, "Resilience and Disaster Risk Reduction: An Etymological Journey", *Natural Hazards and Earth System Sciences* 13, no. 11 (2013): 2707-2716.

[92] William Stephens and Stijn Sieckelinck, "Being Resilient to Radicalisation in PVE Policy: A Critical Examination", *Critical Studies on Terrorism* 1, no. 1 (2020): 147-148.

Public–private partnership in developing public governance

Writing after the 2015 Paris and Brussels attacks, Europol, Europe's law enforcement agency, pointed out that terrorists use social media and the internet to spread propaganda, recruit and fundraise. [93] Four years earlier, Professor Gary Marchant pointed to the gap between technology and policy, underlining that policy simply cannot keep up with the rate of scientific advances. Merchant added that what complicates the situation is the public's appetite for more technology. [94] With this in mind, it is unsurprising that the regulatory system ends up being out-of-date and out-of-sync with the technology, which also grants the platforms more leeway in formulating a regulatory regime that is siloed. Therefore, it is important to get industry and policymakers to interact more, ideally through track 1.5 dialogues which encourage joint problem-sharing in a supportive environment. The Australian model of Safety by Design should become an industry model in that it looks to educated designers and programmers and most importantly venture capitalism to think about the social implications of their research, as the intention is to encourage in-built safety measures. One can only wonder how the online world would have appeared today had some thought been given to how nefarious actors could exploit social media and technology.

In summary, unless a whole-of-society governance regime is developed in which industry and government work together we will continue to put band-aids on gashing wounds.

93 Europol, "European Union Terrorism Situation and Trend Report", 2016, https://www.europol.europa.eu/sites/default/files/documentseuropol tesat_2016.pdf [https://perma.cc/DSZ6-UAPR].

94 Gary E. Marchant, "The Growing Gap between Emerging Technologies and the Law", in Gary E. Marchant, Braden Allenby and Joseph Herkert (eds.), *The Growing Gap between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* (Dordrecht, Germany: Springer, 2011): 19-32.

Title: “ONLINE RADICALISATION AND ALGORITHMS”
By Isaac Kfir

The Henry Jackson Society
Millbank Tower, 21-24 Millbank
London SW1P 4QP, UK

www.henryjacksonsociety.org

HJS
Henry
Jackson
Society

DEMOCRACY | FREEDOM | HUMAN RIGHTS

CENTRE ON RADICALISATION & TERRORISM

December 2021