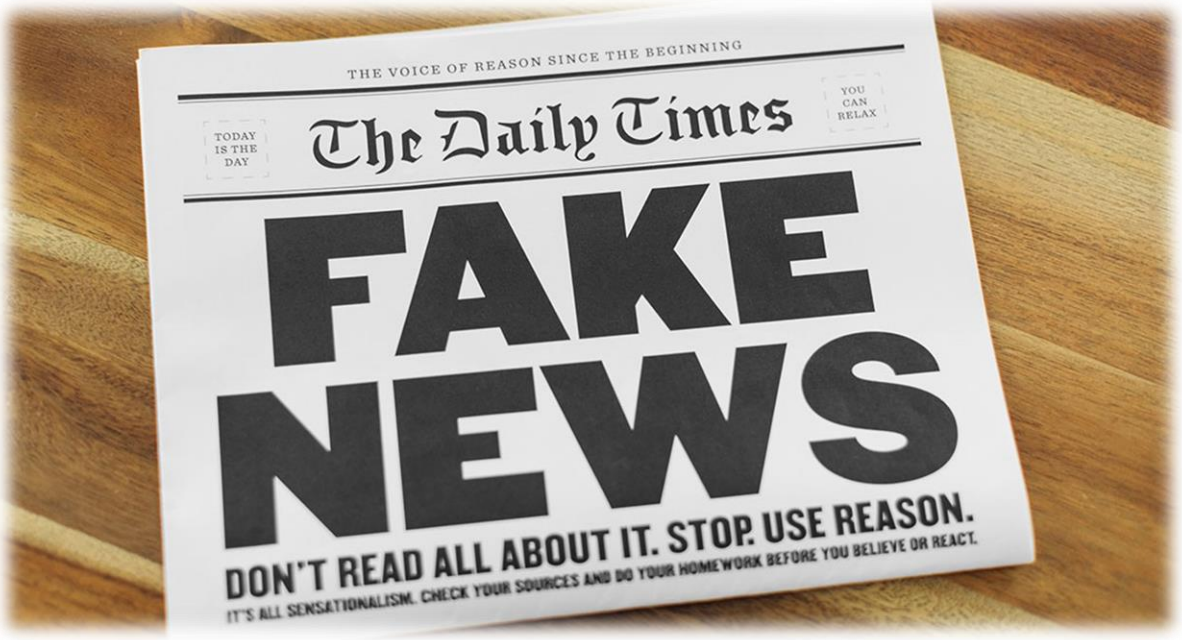
THE WILBERFORCE SOCIETY



Tackling Intentional Disinformation in the Social Media Era

Editors: Bryan Fong and Jemima Baar

Writers: Bryan Fong, Jemima Baar, Eugenio Nanni, Oliver Guest, Kristen Chin Yan Han, Tim Lee, and Dhruv Kanabar

The Wilberforce Society

Cambridge, UK

www.thewilberforcesociety.co.uk

February 2021

ABSTRACT

This paper examines the impact of intentional disinformation – the dissemination of false information with the deliberate intent to deceive or mislead – in the digital age. It proposes that social media has accelerated the impact of modern intentional disinformation, resulting in intentional disinformation becoming an insidious and persistent force in the current climate. The paper outlines the scale and scope of this problem, including an examination of how social media platforms provide malign actors with enhanced means of fabricating credibility and enable these actors to remain anonymous, as well as exacerbate extant psychological vulnerabilities of the masses to disinformation, and details the effects of increasingly widespread reduction in the perceived credibility of information. The paper then turns to proposing a range of policies to mitigate these issues. It maintains that a nuanced approach, targeted directly towards the roots of the issue on an international scale, is required to effectively combat international disinformation in the social media era. This comprises fostering greater transparency of social media platforms and their business models, implementing an effective universal fact-checking platform, founding a unified intentional anti-disinformation agency and providing education programmes to foster digital literacy.

EXECUTIVE SUMMARY

The paper distinguishes between two forms of social media. First, public forum social media platforms (e.g. Facebook and Twitter), upon which users are able to share posts, links, pictures, and other multimedia with other users, gated by varying privacy settings and degrees of connection between them. Second, private messaging social media platforms (e.g. WhatsApp and WeChat), which focus on direct messaging between users, or groups of users. It is noted that many public forum social media platforms also incorporate a private messaging platform attached to their services.

Although differing in their functions, it is concluded that both platform types are commonly used to facilitate intentional disinformation and have both significantly contributed in accelerating the impact of modern intentional disinformation. A combination of technology and social media platforms have enabled disinformation to spread more easily and efficiently, and aggravated peoples' susceptibilities to it for the following reasons:

I. Enhanced Methods of Fabricating Credibility

- Social media platforms have eliminated the barriers to entry to publishing information associated with traditional forms of media. This has:
    - Improved public discourse to some extent: citizen journalists can now share information that traditional journalists were not there to witness, or that a state does not want its public to see; but,
    - Enabled unvetted malign information to spread more easily.
- There are political and financial motivations for spreading intentional disinformation.
- Furthermore, the financial benefits of intentional disinformation for the publishers of disinformation websites, and the social media companies themselves, make the problem relatively intractable.

II. Anonymity of Perpetrators

- There are different forms of anonymity:
    - The perpetrators may be simply unidentifiable;
    - The perpetrators may be known but hidden, especially after content has been shared multiple times;
    - The perpetrators anonymous by using a false front or impersonation; or,
    - Even if the content's producer is known, their true affiliations may not be.

- Anonymity emboldens perpetrators to disseminate disinformation and thereby enables it to spread quickly, makes it almost impossible to hold perpetrators to account, and enables large disinformation campaigns to be easily concealed.

III. Vulnerability of the Masses

- People are susceptible to believing intentional disinformation, due to the following psychological phenomena:
    - Motivated reasoning;
    - Social proof; and
    - 'The Truth Effect'.
- These extant susceptibilities are exacerbated by developments in technology. People are, therefore:
    - More likely to believe disinformation; and,
    - Less likely to believe attempts to debunk intentional disinformation.

It is therefore concluded that technology and social media platforms make intentional disinformation more viral, more difficult to trace, and compound people's extant vulnerabilities to it. It is likely that these issues will worsen with improvements in technological capabilities.

Furthermore, it is concluded that, while social media companies' recent efforts to tackle intentional disinformation are encouraging, these efforts are in their early stages, and the scale of the problem is immense, which makes it unclear whether the companies will be able to overcome it on their own. More profoundly, due to its virality and engagement potential, intentional disinformation positively benefits the social media companies' business model, which relies on engagement and advertising. Even if the companies themselves do not condone disinformation, given their obligations to their shareholders and the financial impact such a comprehensive change of their business model would have, there are valid concerns that social media companies have insufficient impetus to eradicate disinformation completely. Policies proposed by governments and governing institutions are in similarly rudimentary stages.

Thus, to mitigate these issues, the following policies are proposed:

I. Social Media Transparency Models

- Two potential models for promoting transparency in information sources for social media posts and articles are discussed, with the paper ultimately proposing the positive reinforcement variant:
    - Negative reinforcement variant: a punitive model based on proactively scanning and visually negatively marking social media posts making factual claims (potentially limited only to those related to certain key disinformation-heavy topics) for questionable or insufficient information sources
    - Positive reinforcement variant: a reward-based model based on independent user submissions of social media posts for review of sufficient and credible information sources for factual claims – in which successfully reviewed posts would be positively marked accordingly
- An ad database model to promote transparency in broad trends regarding intentional disinformation is also proposed, centred around consistently structured and easily accessible databases on ads and ad-related information on social media platforms

II. Universal Fact-Checking

- A universal fact-checking service is discussed and proposed to provide a universal, consistent standard for fact-checking on a credible platform that is easily accessible by the masses – with special consideration given to elements such as:
    - The fact-checking and review processes
    - Use of technology
    - Concerns regarding freedom of speech
    - Jurisdiction and responsibility

III. A Unified Anti-Disinformation Agency

- To effectively implement the previous two policies, a UN-based unified anti-disinformation agency is discussed and proposed (the UN Debunking and OnLine Fake Information Neutralisation, or DOLFIN, Agency), to ensure effective and consistent implementation and a platform for unassailable credibility and objectivity – with special consideration given to elements such as:
    - Distribution of responsibilities
    - Structure
    - Funding

- Protection from abuse

IV. Fostering Widespread Digital Literacy

- A media, information, and digital literacy (MIDL) curriculum-addendum is discussed and proposed to provide a longer-term measure to combat the vulnerability of the masses to intentional disinformation
- Consideration is also given to more immediate digital literacy solutions to help minimise vulnerability of the masses to intentional disinformation, as a supplementary shorter-term measure

Overall, departing from traditional measures to tackle intentional disinformation (e.g. blanket legal measures and heavy-handed controls on social media platforms), this paper aims to take a more nuanced approach in creating a tailored suite of initial policies for tackling intentional disinformation. Specifically, this paper focuses on tackling the source issues directly instead of utilising sweeping measures, with an emphasis on solutions which combat intentional disinformation at the same level that it originates from – enhancing individual abilities to accurately distinguish veritable and trustworthy information, while reducing the protection afforded to potential perpetrators, on a widespread level.

TABLE OF CONTENTS

I. INTRODUCTION AND CONTEXT

Intentional disinformation is defined as the dissemination of false information with the deliberate intent to deceive or mislead.[1] While instances of disinformation have been present throughout recorded history,[2] the advent of social media has transformed intentional disinformation, turning it into a far more potent and impactful force in today's world.

With a myriad of applications, social media can be broadly split into two different classifications, public forum social media platforms, and private messaging social media platforms. Public forum social media sites (e.g. Facebook and Twitter) generally act as platforms where users are able to share posts, links, pictures, and other multimedia with other users – gated by varying privacy settings and degrees of connection between them. In contrast, private messaging social media platforms (e.g. WhatsApp and WeChat) focus on direct messaging between users, or groups of users – and many public forum social media platforms also incorporate a private messaging platform attached to their services as well.

However, though differing in their functions, both platform types are commonly used to facilitate intentional disinformation, and have both significantly contributed in accelerating the impact of modern intentional disinformation. Overall, the social media era has enabled intentional disinformation to be spread rapidly, at low costs, anonymously, and by anyone with access to the internet. As a result, intentional disinformation has become a far more insidious and persistent force in the current climate, requiring a vastly different set of approaches to fully tackle. Analysing the core issues that fuel intentional disinformation in the social media era, this paper therefore aims to provide a structured framework of potential solutions, tailored specifically towards mitigating and preventing modern intentional disinformation on an international scale.

[1] Daniel Chandler and Rod Munday, 'Disinformation', *A Dictionary of Media and Communication* (Oxford: Oxford University Press, 2016)
[2] J.M. Burkhardt, 'Combatting Fake News in the Digital Age' (2017) Library Technology Reports 53:8.

I.I. THE DANGERS OF MODERN INTENTIONAL DISINFORMATION

Intentional disinformation has harmful tangible impacts. Recent examples include the conspiracy theory that 5G cellular technology causes Covid-19, which incited groups to set 5G towers on fire,[3] and Covid-19 patients refusing treatment after reading false information online about the severity of the illness and the risks of going to hospital.[4] It is more difficult to measure the extent to which intentional disinformation has impacted political events given the high levels of privacy and anonymity that surround a person's vote. This means that analysis of the alleged impact of disinformation on major world events, such as the 2016 United Kingdom European Union membership referendum and the 2016 United States presidential election is circumspect or inconclusive.

Nevertheless, there are cases for which there is greater available data, which makes it possible to draw correlations between disinformation campaigns and political activity. For example, a Senate Intelligence Committee report found that disinformation campaigns linked to the Russian 'Internet Research Agency' that targeted black Americans and exploited racial tensions were a factor behind the low voting turnout of the black demographic in the 2016 Presidential Election.[5] Intentional disinformation also affected the 2018 Brazilian elections. To vote in Brazil, one must punch in a number for a candidate or party in an electronic voting machine. Disinformation spread over WhatsApp gave the wrong number for a particular politician, which caused some people to unintentionally vote for a different party.[6]

The prevalence of disinformation (either perceived or real) has also led to more intangible, but perhaps more worrying effects. In particular, it has fuelled suspicion towards the credibility of all forms of information, especially that produced by traditional media outlets.[7] As James Kulinski and Paul Quirk argue, in order for a democracy to function well, it is imperative that its populace

[3] Isobel Asher Hamilton, 'Here's what we know about the bizarre coronavirus 5G conspiracy theory that is leading people to set cellphone masts on fire', (*Business Insider*, 6 May 2020) <https://www.businessinsider.com/coronavirus-conspiracy-5g-masts-fire-2020-4?r=US&IR=T> accessed 14 February 2021.

[4] Chris Baynes, 'Coronavirus: Patients refusing treatment because of fake news on social media, NHS staff warn', *The Independent* (5 June 2020) <https://www.independent.co.uk/news/uk/home-news/coronavirus-fake-news-conspiracy-theories-antivax-5g-facebook-twitter-a9549831.html> accessed 14 February 2021.

[5] Renee DiResta and others, 'The Tactics & Tropes of the Internet Research Agency', (New Knowledge, 2019) p.85-89.

[6] Mike Isaac and Kevin Roose, 'Disinformation Spreads on WhatsApp Ahead of Brazilian Election', *The New York Times* (19 October 2018) <https://www.nytimes.com/2018/10/19/technology/whatsapp-brazil-presidential-election.html> accessed 14 February 2021.

[7] Janna Anderson and Lee Raine, 'The Future Of Truth And Misinformation Online' (Pew Research 2017)

is educated and well-informed.[8] The increasing levels of scepticism towards objective, *bona fide* facts, especially those that may challenge preconceptions, negatively impacts the quality of liberal democratic discourse and may increase political polarisation.

More broadly speaking, this diminishing of credibility of general information could perhaps pose the greatest danger from intentional disinformation in the social media era. Damaging the general populace's ability to effectively distinguish truth, this issue remains far more insidious and slow-evolving compared to many of the headline political and social issues generated by fake news regarding current events. However, with such a phenomenon beginning to manifest in the increased persistence of beliefs in fake news, and difficulties in establishing any line of commonly accepted facts amongst general populations, it is clear that this particular aspect is inexorably growing into a dire issue which governments and international bodies are struggling to face.

[8] James H. Kuklinski and Paul J. Quirk, 'Conceptual Foundations Of Citizen Competence' (2001) 23 Political Behavior 285.

I.II. THE SCALE AND SCOPE OF THE PROBLEM

Analysing the root causes of this issue, this paper argues that advances in technology, such as AI and social media, have significantly increased the prevalence and exacerbated the effects of intentional disinformation. In its exposition of the scale and scope of online disinformation, this paper considers the motivations behind spreading disinformation, how disinformation is spread on social media, and the extant psychological issues that make people susceptible to disinformation. It concludes that technology and social media platforms have enabled disinformation to spread more easily and efficiently, and aggravated peoples' susceptibilities to it. These themes are examined through a combination of case studies and analysis in the sections:

- II.I. Enhanced Methods of Fabricating Credibility
- II.II. Anonymity of Perpetrators
- II.III. Vulnerability of the Masses

II.I. Enhanced Methods of Fabricating Credibility examines how social media platforms have eliminated the barriers to entry to publishing information associated with traditional forms of media. This has improved public discourse to some extent – citizen journalists can now share information that traditional journalists were not there to witness, or that a state does not want its public to see – but it has also allowed unvetted malign information to spread more easily. The section also considers the political and financial motivations for spreading intentional disinformation. It is suggested that the financial benefits of intentional disinformation for the publishers of disinformation websites, and the social media companies themselves, make the problem relatively intractable.

II.II. Anonymity of Perpetrators examines how anonymity on social media platforms facilitates the spread of disinformation. Different forms of anonymity are considered: (i) the perpetrators may be simply unidentifiable, (ii) they may be known but hidden, especially after content has been shared multiple times, (iii) they may remain anonymous by using a false front or impersonation, or (iv) even if the content's producer is known, their true affiliations may not be. It is argued that anonymity emboldens perpetrators to disseminate disinformation and thereby enables it to spread quickly, makes it almost impossible to hold perpetrators to account, and enables large disinformation campaigns to be easily concealed.

II.III. Vulnerability of the Masses considers people's extant susceptibility to disinformation within the framework of how these existing issues may be exacerbated by developments in

technology. Motivated reasoning, social proof, and the truth effect are examined in relation to people's susceptibility to believing disinformation and the efficacy of debunking disinformation. It is suggested that these extant issues are likely to worsen as technology improves.

It is therefore concluded that technology and social media platforms make intentional disinformation more viral, more difficult to trace, and compound people's extant vulnerabilities to it. It is likely that these issues will worsen with improvements in technological capabilities.

I.III. EXISTING MEASURES

Before discussing additional measures for tackling intentional disinformation, it is important to understand the broad range of existing solutions that have been attempted. In recent years, there have been more concerted approaches both from social media companies themselves and governments to tackle intentional disinformation – albeit with distinct differences in their methods and scope.

i. Social Media Companies' Policies

The following section provides an overview of three of the largest social media companies' approaches to the issue:

Twitter

Twitter's policies on disinformation are among the most stringent. The aim of 'The Twitter Rules' is to "ensure all people can participate in the public conversation freely and safely".[9] To these ends, content containing abuse and inciting violence is strictly policed. Regarding disinformation specifically, Twitter's guidelines, quoted from their website, are as follows:

- Platform manipulation and spam: You may not use Twitter's services in a manner intended to artificially amplify or suppress information or engage in behaviour that manipulates or disrupts people's experience on Twitter.
- Civic Integrity: You may not use Twitter's services for the purpose of manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process.
- Impersonation: You may not impersonate individuals, groups, or organizations in a manner that is intended to or does mislead, confuse, or deceive others.
- Synthetic and manipulated media: You may not deceptively share synthetic or manipulated media that are likely to cause harm. In addition, we may label Tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context.[10]

[9] 'The Twitter Rules' (*Twitter*) < https://help.twitter.com/en/rules-and-policies/twitter-rules> accessed 14 February 2021.
[10] Ibid..

Content that is not in violation of Twitter's authenticity guidelines includes: inaccurate statements about parties, officials, or candidates, and polarising, biased, or controversial opinions.

During the Covid-19 pandemic, Twitter updated its regulations regarding misleading content.[11] This included:

- Putting warnings on tweets that include misleading or disputed claims about Covid-19;
- Using AI technology to flag the most urgent content to be reviewed on the basis of misleading claims, allowing Twitter to review potentially problematic tweets before they are reported by users;
- Reviewing manually (i.e. with human teams) any content that might require additional context; and
- Removing content including, but not limited to: denial of health authority recommendations; denial of established scientific facts; propagation of false or misleading information around Covid-19 diagnostic criteria or procedures; claims that specific groups are more or less susceptible to Covid-19.

One notable example of Twitter's more stringent guidelines in practice during the Covid-19 pandemic was the temporary ban imposed on Donald Trump Jr.'s account for posting a video discussing the benefits of hydroxychloroquinine, a substance controversially touted as a potential treatment for Covid-19.[12]

Facebook

Facebook's statement about their role in combatting disinformation is, "We cannot become arbiters of truth ourselves — it's not feasible given our scale, and it's not our role".[13] Thus, their policies are more directed towards reducing the spread of harmful content, rather than its removal. Facebook's website details the following policies regarding intentional disinformation:

[11] 'An update on our continuity strategy during COVID-19' (*Twitter*). <https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html> accessed 14 February 2021.
[12] 'Facebook And Twitter Restrict Trump Accounts Over 'Harmful' Virus Claim' (*BBC News*, 6 August 2020) <https://www.bbc.com/news/election-us-2020-53673797> accessed 21 August 2020.
[13] 'Working to Stop Misinformation and False News' (*Facebook*) <https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news> accessed 14 February 2021.

- "Disrupting economic incentives because most false news is financially motivated". This is primarily targeted at pages that post hoaxes or 'clickbait' images with the aim of enticing viewers to click on their websites, which contain many advertisements.
- "Building new products to curb the spread of false news". In other words, Facebook tries to build and maintain the site in such a way that disinformation does not become widespread. This includes making it easier to report or 'flag' posts that users perceive to be harmful or to contain disinformation. Facebook also collaborates with third-party fact-checkers to conduct more thorough research on individual pieces of content and publish reports explaining why content might have been flagged. However, they give little detail about removing content outright; rather, their policies aim to reduce the spread of problematic content and make publishing it less lucrative.
- "Helping people make more informed decisions when they encounter false news". This involves initiatives such as the Facebook Journalism Project, which is a collaboration with the News Literacy Project to produce PSAs for users to improve their ability to detect disinformation and respond appropriately, and the News Integrity Initiative, which is a collaboration with academic institutions and non-profit organisations to promote media literacy skills.[14]

YouTube

YouTube[15] outlines categories of content that may be removed, including:

- Misleading thumbnails;
- Technically manipulated videos that deliberately mislead;
- Scams;
- Videos that encourage voter suppression and incite interference with democratic processes.

As with all content that violates YouTube's community guidelines, most penalties are in the form of a strikes system. The first time a channel posts content in breach of any of YouTube's guidelines, the video is removed and the channel is given a warning. Each further offence incurs a strike to the channel. After three strikes, the channel itself is removed. Channels that attempt

[14] Ibid.

[15] 'Spam, deceptive practices & scams policies' (*YouTube*) < https://support.google.com/youtube/answer/2801973?hl=en&ref_topic=9282365> accessed 14 February 2021

to impersonate, misrepresent, or conceal affiliation with a government and those which artificially increase engagement (for example, through the use of fake accounts or 'bots') are removed.[16]

Evaluation

Social media companies' recent efforts to tackle intentional disinformation, and the variety of methods proposed – including long-term methods, such as Facebook's digital education schemes, as well as short-term methods, such as reviewing and flagging content – is encouraging. However, these efforts are in their early stages, and the scale of the problem is immense, which makes it unclear whether the companies will be able to overcome it on their own. More profoundly, due to its virality and engagement potential, intentional disinformation positively benefits the social media companies' business model, which relies on engagement and advertising. Even if the companies themselves do not condone disinformation, given their obligations to their shareholders and the financial impact such a comprehensive change of their business model would have, there are valid concerns that social media companies have insufficient impetus to eradicate disinformation completely.

[16] 'Spam, deceptive practices & scams policies' (*YouTube*) < https://support.google.com/youtube/answer/2801973?hl=en&ref_topic=9282365> accessed 14 February 2021.

ii. Government Responses

Governments and governmental institutions have also started to consider and implement regulations that tackle online intentional disinformation:

Singapore

Singapore has adopted one of the most stringent approaches to combatting intentional disinformation. The Protection from Online Falsehoods and Manipulation Bill, passed in October 2019, gives government ministers the power to require warnings to be placed on content that the government deems to be disinformation, and in extreme cases, the power to order its removal.

There are three "tiers" of sanctions that can be imposed on offenders under this law:

- A fine of up to $37,000 or five years in prison for sharing false information;
- A fine of up to $74,000 and a 10-year jail term for using fake accounts or bots;
- A fine of up to $740,000 and 10 years in prison for tech platforms that do not remove such content.

The extent of this governmental involvement raises normative concerns. Journalists, social media companies, and academics have expressed significant apprehensions about the prospect of the bill giving the Singaporean government excessive powers to censor free expression that is not actually intentional disinformation.[17]

Germany and France

Germany and France have also started regulating content on social media platforms, but in the sphere of hate speech rather than intentional disinformation. In January 2018, the German government passed the Network Enforcement Act (commonly known as NetzDG), which stipulates that social media companies must remove "obviously illegal" hate speech and other postings within 24 hours of receiving a notification, or face a €50m fine.[18] The NetzDG is effective, and has drawn praise from the U.K. Parliament's Digital, Culture, Media and Sport Committee, which stated: "As a result of this law, one in six of Facebook's moderators now works in Germany, which is practical evidence that legislation can work".[19] In France, judges are able to

[17] "Chilling': Singapore's 'fake news' law comes into effect' (*The Guardian*, 2 October 2019) < https://www.theguardian.com/world/2019/oct/02/chilling-singapores-fake-news-law-comes-into-effect> accessed 14 February 2021.

[18] Heidi Tworek and Paddy Leersen, 'An Analysis of Germany's NetzDG Law' *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* (2019).

[19] Digital, Culture, Media and Sport Committee, *Disinformation and 'fake news': Final Report* (2019, HC 1791).

order the removal of content deemed to be hate speech. Regarding disinformation, French law stipulates that social media platforms must disclose details of any funding received for promotion, broadcasting agencies must suspend foreign-influenced TV channels disseminating disinformation that may affect democratic processes, and failure to comply can lead to a one year prison sentence and €75,000 fine.[20]

United Kingdom

By contrast, enforceable regulation in the U.K. is in rudimentary stages of development. There have been proposals outlined by the U.K. Parliament's Digital, Culture, Media and Sport Committee,[21] which include:

- Encouraging greater transparency from publishers, advertisers, content creators, and intermediaries about sponsorship, algorithms, country of origin, and targeting;
- Increasing powers of regulators to impose appropriate penalties for non-compliance;
- Improving digital literacy to reduce penetration of disinformation.
- A proposal to create a new category of tech company to tighten legal liability because "Social media companies cannot hide behind the claim of being merely a 'platform' and maintain that they have no responsibility themselves in regulating the content of their sites";
- Creation of a new regulating authority with a Code of Ethics similar to Ofcom's Broadcasting Code as well as accessible channels for members of the public to make complaints and report potentially problematic content.

European Union

In 2018, the European Commission published its 'Code of Practice on Disinformation',[22] which is the first set of worldwide standards to tackle intentional disinformation, and advocates greater collaboration between countries, companies, academics, and other stakeholders. Notable signatories include Facebook, Twitter, Google and Microsoft. The Code aims to tackle disinformation in the following ways:

- Disrupting the advertising revenues of perpetrators;

[20] 'Against information manipulation' (*French Government*, 2018) <https://www.gouvernement.fr/en/against-information-manipulation> accessed 14 February 2021.
[21] Digital, Culture, Media and Sport Committee, *Disinformation and 'fake news': Final Report* (2019, HC 1791).
[22] 'Tackling online disinformation' (*European Commission,* 2020) <https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation> accessed 14 February 2021.

- Improving the transparency of political advertisements;
- Tackling fake accounts and bots;
- Empowering reporting of disinformation and other poor sources of information, while also improving the findability of "authoritative content"; and
- Allowing researchers to monitor the issue using "privacy-compliant" access to data.

The Commission's broader action plan involves the following, quoted from their website,[23] with additional clarification from the writers of this paper in parentheses:

- Improving detection, analysis, and exposure of disinformation (e.g. through the use of A.I. technologies)
- Stronger cooperation and joint responses to threats (i.e. creating a global standard against which to design regulation and policy)
- Enhancing collaboration with online platforms and industry to tackle disinformation
- Raising awareness and improving societal resilience (media, information, and digital literacy)

Alongside the Code, the Commission established the European Digital Media Observatory. This is a hub for academics, fact-checkers, and other stakeholders to collaborate with media outlets and media literacy experts to support policymakers in tackling disinformation. They aim to achieve this through five pillars of operation:

- Mapping fact-checking organisations;
- Supporting and coordinating of research activities;
- Building a public portal;
- Ensuring secure and privacy-protected access to platforms' data; and
- Supporting public authorities.

[23] Ibid.

I.IV. ISSUES WITH EXISTING MEASURES AND COUNTER-PROPOSALS

While there has been a range of attempted solutions for intentional disinformation, it is clear that this still remains a rampant issue, which remains challenging in nature to combat effectively. In the case of government solutions, many of the existing attempted solutions have fallen short due to the constraints of using heavy-handed blanket solutions (i.e. broad legal measures) to tackle a nuanced and subjective issue. Too constricting of a blanket approach can result in unnecessary infringement upon personal liberties (generating subsequent backlash, which could render such solutions wholly counterproductive), and too lax of a blanket approach can end up being entirely ineffectual. In the case of social media companies' solutions, many of the existing solutions – though promising in many cases – fall short in scope and coverage. After all, social media companies only have jurisdiction over their platform to a certain degree – and have even less leeway when it comes to issues of infringement upon personal liberties than governments do.

As a result, this paper maintains that a nuanced, targeted approach on an international scale is required to effectively combat international disinformation in the social media era. Intentional disinformation in the social media era remains as insidious and as powerful as it is, precisely because it is an issue fostered at the individual level, over a vast and widespread scale – which also contributes to its resistance to many of the broad blanket measures that have been attempted in response to it. Targeting the root issues directly instead, this paper therefore attempts to prevent and address intentional disinformation at the same grassroots level – but on a scale large enough to avoid the issues of segregated jurisdiction and responsibility and claims of regional biases that plague many solutions attempted at smaller scales. Detailed further in Section III., this paper therefore ultimately aims to provide a framework of nuanced solutions, able to neutralise intentional disinformation at the exact same scope and level of granularity that it arises in.

II. KEY ISSUES

II.I. ENHANCED METHODS OF FABRICATING CREDIBILITY

The internet, and social media in particular, have dramatically lowered the barriers to entry of launching a credible disinformation campaign. At the same time, new technologies provide strong financial incentives to spread and facilitate disinformation. Given these raised incentives and lowered costs, it is unsurprising that disinformation has become prevalent in recent years.

i. Lowered Barriers to Entry

Technology has made it significantly easier for individuals or small groups to publish messages. Traditional forms of media, such as newspapers, are capital-intensive and logistically complex, and so require large institutions to run them. By contrast, newer forms of media, such as blogs and social networks, require little to no money or technical expertise. Almost anyone with an internet connection can therefore use them to publish. If social media users think that a particular message is important, they can share it to their own network, potentially bringing the message to the attention of a large audience, or 'going viral'.

The ensuing decrease in the power of gatekeeping institutions like newspapers has in many ways improved public discourse: several important stories that traditional journalists were unable or unwilling to tell have caught the public's attention in recent years. In the U.S., for instance, numerous videos of police brutality against people of colour have been shared, allowing these events to be documented, even if a traditional journalist was not present at the time.[24] Another example is the use of social media by activists during the Arab Spring to report on protests that state-controlled media was not covering.[25]

Broader access to public attention, and the conception of journalism as no more than a small number of specific outlets, have, by the same turn, made it easier for malign actors to spread disinformation, however. Moreover, disinformation on social media can seem similarly credible to any other message. Social networks do have features to show that a specific account is not an

[24] Sarah Almukhtar and others, 'Black Lives Upended by Policing: The Raw Videos Sparking Outrage' (*The New York Times,* 30 July 2015) <https://www.nytimes.com/interactive/2017/08/19/us/police-videos-race.html, https://www.nytimes.com/interactive/2017/08/19/us/police-videos-race.html> accessed 7 March 2020.
[25] Peter Beaumont, 'The Truth about Twitter, Facebook and the Uprisings in the Arab World' (*The Guardian,* 25 February 2011) <https://www.theguardian.com/world/2011/feb/25/twitter-facebook-uprisings-arab-libya> accessed 7 March 2020.

impersonation, such as Twitter's blue check mark.[26] However, since one would not necessarily expect legitimate citizen journalists to have this mark, its absence cannot necessarily be used to judge that an account is dubious. The problem is even more pronounced in the case of disinformation websites: it is easy to create a website that looks like that of a trustworthy outlet, and one would not expect to see a check mark or similar.[27] Moreover, in some cases, people are "source agnostic"; they are simply not concerned about whether an a source of information is credible.[28]

[26] Nellie Bowles, 'Twitter, Facing Another Uproar, Pauses Its Verification Process' (*The New York Times*, 9 November 2017) <https://www.nytimes.com/2017/11/09/technology/jason-kessler-twitter-verification.html> accessed 7 March 2020.

[27] Alice Marwick and Rebecca Lewis, 'Media Manipulation and Disinformation Online' (2017) Data & Society Research Institute 55 <https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf> accessed 7 March 2020. and Craig Silverman, Jane Lytvynenko and William Kung, 'Disinformation For Hire: How A New Breed Of PR Firms Is Selling Lies Online' (*BuzzFeed News*, 6 January 2020) <https://www.buzzfeednews.com/article/craigsilverman/disinformation-for-hire-black-pr-firms> accessed 7 March 2020.

[28] Santanu Chakrabarti, Lucile Stengel and Sapna Solanki, 'Duty, Identity, Credibility: "Fake News" and the Ordinary Citizen in India' (BBC 2018) 34 <http://downloads.bbc.co.uk/mediacentre/duty-identity-credibility.pdf> accessed 14 February 2021.

ii. An Uneven Playing Field

Although the superficial similarities between legitimate and illegitimate sources of information might imply a level playing field where the truth will be able to win out, several factors seem to make disinformation disproportionately likely to spread.

Novelty and emotion

Widely cited research from the M.I.T. Media Lab found that falsehoods are 70% more likely than accurate news to be retweeted on Twitter.[29] Since an individual user's retweeting affects what other users see, this leads to disinformation tweets reaching disproportionately many people and spreading disproportionately quickly.

The researchers hypothesised that two trends were responsible for the high virality of disinformation. First, disinformation is generally more novel. Second, it often evokes more emotion, with disinformation tweets tending to elicit words associated with surprise and disgust.[30] Although little follow-up research has been done, the findings of this research are concerning if correct: since disinformation is not constrained by reality, it can more easily be made novel and emotive. This may make it more likely to spread.

Traditional media's reporting on disinformation

Since disinformation's virality tends to be confused for validity, mainstream outlets sometimes report on it, amplifying it further. For instance, in response to what were hoax tweets, several major outlets published incorrect articles about a technical error having caused CNN to broadcast pornography.[31] Even when outlets correctly note that a claim is false, and frame the story as being about disinformation, mainstream coverage can still amplify disinformation.

One case study is the 'Pizzagate' conspiracy theory, which alleges that a paedophile ring comprising senior members of the American Democratic Party operates out of a Washington D.C. pizzeria. In the run up to the 2016 U.S. Presidential Election, the theory was conceived on 4chan, an anonymous online forum. It then spread to Reddit, then to Twitter, and was promoted

---

[29] Paul Dizikes, 'Study: On Twitter, false news travels faster than true stories', *Massachusetts Institute of Technology*, 2018 <https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308> [accessed 11 September 2019].

[30] Soroush Vosoughi, Deb Roy and Sinan Aral, 'The Spread of True and False News Online' (2018) 359 Science 1146. and Robinson Meyer, 'The Grim Conclusions of the Largest-Ever Study of Fake News' (*The Atlantic*, 9 March 2018) <https://www.theatlantic.com/technology/archive/2018/03/largest-study-ever-fake-news-mit-twitter/555104/> accessed 7 March 2020.

[31] Loren Grush, 'The CNN Porn Scare Is How Fake News Spreads' (*The Verge*, 25 November 2016) <https://www.theverge.com/2016/11/25/13748226/cnn-accidentally-airs-porn-fake-news-boston> accessed 8 March 2020.

by profit-motivated fake news sites.[32] Suspicions were fuelled when Reddit deleted the 'Pizzagate' subreddit, and tenuous links were drawn between Democrats and paedophiles, such as Bill Clinton using alleged sex trafficker Jeffery Epstein's private plane.[33] It came to a head when an individual fired shots in the pizzeria as part of a "self-investigation".[34] Although there does not seem to any quantitative analysis, some reports suggest that the ensuing news coverage on the New York Times and Fox News spread the theory further: although the mainstream coverage framed the claim as untrue, it brought the claim to the attention of more people, some of whom saw it as credible.[35] Indeed, in 2017, some people protested for a more serious investigation into the (false) matter.[36]

[32] 'Pizzagate': The Fake Story That Shows How Conspiracy Theories Spread' (*BBC News*, 2016) <https://www.bbc.com/news/blogs-trending-38156985> accessed 21 August 2020.
[33] 'Who Was Jeffrey Epstein?' (*BBC News*, 16 November 2019) <https://www.bbc.co.uk/news/world-us-canada-48913377> accessed 21 August 2020.
[34] 'Pizzagate': The Fake Story That Shows How Conspiracy Theories Spread' (*BBC News*, 2 December 2016) <https://www.bbc.com/news/blogs-trending-38156985> accessed 21 August 2020.
[35] Marwick and Lewis (n 4) 55–56; 'The Saga of "Pizzagate": The Fake Story That Shows How Conspiracy Theories Spread' (*BBC News*, 2 December 2016) <https://www.bbc.com/news/blogs-trending-38156985> accessed 8 March 2020.
[36] Michael Miller, 'Protesters Outside White House Demand 'Pizzagate' Investigation' (*The Washington Post*, 25 March 2017) <https://www.washingtonpost.com/news/local/wp/2017/03/25/protesters-outside-white-house-demand-pizzagate-investigation/> accessed 21 August 2020.

iii. Financial Incentives

As identified in the introduction and elsewhere in this paper, there may be strong political incentives to spreading disinformation: disinformation could potentially alter the outcomes of political processes by influencing who votes, or whom they support. Financial incentives created by social media and the internet also motivate some of the production of disinformation. Moreover, social media companies often have similar financial incentives to disinformation producers. Although they are not the perpetrators themselves, these incentives may reduce the likelihood of them enacting strong measures against the phenomenon.

Publishers of disinformation

During the 2016 U.S. Presidential Election, BuzzFeed News identified more than 100 pro-Trump websites being run from Veles, a small town in North Macedonia. The sites appeared to be news outlets but were actually publishing disinformation. Common false claims included that the Pope had endorsed Donald Trump or that Hilary Clinton was going to be indicted. The sites would generally plagiarise a story from American fringe political sites, write a compelling headline and then post it on social media. Many of the stories went viral, resulting in large numbers of people clicking through to the websites, generating substantial advertisement revenue.[37][38]

The people running these sites did not seem to have political objectives; their claims were that they were not interested in American politics and that the only reason why their disinformation supported Trump instead of other candidates was that this was more lucrative were compelling. Furthermore, prior to the election, some of them had applied a similar business model to health advice.[39]

Despite being run by apolitical actors, the potential political impact of such sites is clear, particularly when one considers that disinformation made up a greater share of the 20 most widely shared election stories on Facebook in the three months prior to the election.[40]

---

[37] Simon Oxenham, 'I Was a Macedonian Fake News Writer' (*BBC*, 29 May 2019) <https://www.bbc.com/future/article/20190528-i-was-a-macedonian-fake-news-writer> accessed 8 March 2020.

[38] Craig Silverman and Lawrence Alexander, 'How Teens In The Balkans Are Duping Trump Supporters With Fake News' (*BuzzFeed News*, 3 November 2016) <https://www.buzzfeednews.com/article/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo> accessed 8 March 2020.

[39] Samanth Subramanian, 'Inside the Macedonian Fake-News Complex' (*Wired,* 15 February 2017) <https://www.wired.com/2017/02/veles-macedonia-fake-news/> accessed 8 March 2020.

[40] Craig Silverman, 'This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook' (*BuzzFeed News*, 16 November 2016) <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook> accessed 9 September 2019.

Social media companies

Although they do not condone disinformation, social media companies benefit financially from disinformation being spread on their platforms. Many sites, such as Facebook, Twitter, YouTube and Instagram, are free to use, so rely on paid advertisements for revenue. Since disinformation is disproportionately engaging, it is likely to increase the amount of time that users spend on a social network. This gives more opportunities for the social media company to show advertisements and thereby make money. There may, therefore, be a conflict of interest for the companies: they may be incentivised to accept payment for advertisements that present false information from buyers with nefarious motives, and since these sites rely on maximum engagement from their users, they may be incentivised to allow misinformative posts to be made if those posts attract attention, despite the potentially harmful effects of this content.

Stuart Russell has pushed this analysis further, noting that social media companies particularly want users to click on the ads that they see, since advertisers pay more for this. This creates an incentive to turn people into "predictable clickers", because this allows the companies to better match advertisement and user. Since people seem to be more predictable when they are at an ideological extreme, the companies may benefit from people moving towards ideological extremes. Disinformation, with its novelty and emotiveness – as well as its often-extreme claims – is effective in this regard.[41]

Although the business model of social media companies is benefitted by disinformation, the companies also have incentives that could push them to tackle the phenomenon. They might want to appear proactive so as to avoid consumer backlash or overbearing regulation, for instance. They are also coming under increasing pressure from other companies to remove problematic content. In 2020, the 'Stop Hate for Profit' movement began in response to Facebook's refusal to remove a post containing Donald Trump's "When the looting starts, the shooting starts" comment. The aim of the movement is to encourage Facebook to be stricter about removing hate speech by withdrawing their advertisements from the platform. Over 1,000 companies joined these boycotts, including Ben & Jerry's, Pfizer, Puma, The North Face, Sony Interactive Entertainment, Vans, and Verizon.[42] It marks an interesting development whereby

[41] 'Human Compatible: Artificial Intelligence and the Problem of Control with Stuart Russell' (*Future of Life,* 8 October 2019) <https://futureoflife.org/2019/10/08/ai-alignment-podcast-human-compatible-artificial-intelligence-and-the-problem-of-control-with-stuart-russell/> accessed 14 February 2021.

[42] 'Stop Hate for Profit', <https://www.stophateforprofit.org> accessed 14 February 2021.

other companies have taken action to hold social media platforms accountable for the content that they allow to be published on their sites.

However, this movement is in early stages and, moreover, focuses on hate speech rather than intentional disinformation. Furthermore, given the potential impact on profitability, social media companies' resistance to take comprehensive action against the spread of disinformation should not be surprising.

iv. The Role of Videos in Disinformation

Although comparatively understudied, videos seem to present a particularly enhanced method of credibility for disinformation.[43] One reason for this is that videos are disproportionately likely to be consumed on social media. This results from people having a tendency to share videos more than other types of content with their network, increasing viewership.[44] Additionally, Facebook has designed its ranking algorithm to favour video, apparently in a bid to compete with YouTube.[45]

Videos may also be effective for disinformation because of the 'Realism Heuristic': researchers have suggested that because videos have a greater resemblance to reality than text or speech, people assume that they actually match reality more closely.[46] This subsection therefore initially considers deepfakes, which have the potential to resemble reality particularly strongly. Nevertheless, as will be seen, so-called 'cheap fakes' are also very concerning.

Deepfakes

Deepfakes refer to videos where machine learning software has been used to map one face onto another. Although the technology was originally used to create pornography that appears to feature celebrities, later videos have shown public figures saying messages that they never actually said.[47] In one notable example, a fairly convincing Barack Obama warns of the implications of deepfakes for public discourse.[48] Although initially the domain of computer scientists, software that produces deepfakes has become widely accessible and comparatively easy to use.[49] This has

[43] Joshua Tucker and others, 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature' 95, 47.
[44] Tucker and others (n 13).
[45] David Ingram, 'Facebook to Use Its News Feed to Push More Videos to Users' (*Reuters,* 14 December 2017) <https://www.reuters.com/article/us-facebook-video-idUSKBN1E8300> accessed 8 March 2020.
[46] Tucker and others (n 13) 48.
[47] Cade Metz, 'Internet Companies Prepare to Fight the "Deepfake" Future' (*The New York Times* , 24 November 2019) <https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html> accessed 8 March 2020.
[48] James Vincent, 'Watch Jordan Peele Use AI to Make Barack Obama Deliver a PSA about Fake News' (*The Verge*, 17 April 2018) <https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed> accessed 8 March 2020.
[49] Britt Paris and Joan Donovan, 'Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence' (Data & Society Research Institute, 18 September 2019) 13–15 <https://datasociety.net/library/deepfakes-and-cheap-fakes/> accessed 8 March 2020.

caused numerous outlets to warn of video's claim to truth being undermined, with profound implications for democracy.[50][51][52]

Cheap fakes

As Paris and Donovan note, however, the truth value of video had already been problematised by much less technologically sophisticated hoaxes, so called 'cheap fakes'.[53] One widely-shared example was a clip that appeared to show U.S. House Speaker Nancy Pelosi slurring her words, but had actually been slowed to 75% of its proper speed.[54] In the U.K., the Conservative Party edited two moments of an interview together to give the false impression that the then shadow Brexit Minister (now, Leader of the Opposition) Keir Starmer was unable to answer a question.[55] An even simpler strategy is claiming that a video of one event actually shows another. For instance, a video that supposedly shows a kidnapping, and that has been linked to numerous reprisals in India, was in fact part of an anti-kidnapping awareness campaign in Pakistan.[56]

It is therefore easy to produce not only text-based, but also video-based disinformation. As deepfake technology continues to improve, the quality of faked videos that can be easily obtained is likely to increase. Given that capability does not pose much of a hurdle, one would therefore want strong deterrents in place to prevent people from spreading disinformation. Regrettably, however, as the next section considers, anonymity often means that are few deterrents in place.

---

[50] Franklin Foer, 'The Era of Fake Video Begins' (*The Atlantic,* 15 May 2018) <https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/> accessed 8 March 2020.
[51] Joshua Rothman, 'In the Age of A.I., Is Seeing Still Believing? (*The New Yorker* , 12 November 2018) <https://www.newyorker.com/magazine/2018/11/12/in-the-age-of-ai-is-seeing-still-believing> accessed 8 March 2020.
[52] Jennifer Finney Boylan, 'Will Deep-Fake Technology Destroy Democracy?' (*The New York Times* (17 October 2018) <https://www.nytimes.com/2018/10/17/opinion/deep-fake-technology-democracy.html> accessed 8 March 2020.
[53] Paris and Donovan (n 19) 5–9.
[54] Dan Evan, 'Does This Video Show Nancy Pelosi Drunk and Slurring Her Speech?' (*Snopes*) <https://www.snopes.com/fact-check/nancy-pelosi-slurring-speech/> accessed 8 March 2020.
[55] Rachael Krishna, 'This Footage of Keir Starmer Being Interviewed on Good Morning Britain Was Edited before Being Posted by the Conservative Party' (*Full Fact*) <https://fullfact.org/news/keir-starmer-gmb/> accessed 8 March 2020.
[56] Timothy Mcloughlin, 'How WhatsApp Fuels Fake News and Violence in India' [2018] *Wired* <https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/> accessed 8 March 2020.

II.II. ANONYMITY OF PERPETRATORS

One of the underlying problems in relation to disinformation is the anonymity of perpetrators. An important caveat, however, is that even without anonymity, disinformation can flourish. In 2019, Donald Trump tweeted a video of Nancy Pelosi, the Speaker of the House of Representatives, edited to make her appear to slur her speech.[57] Similarly, tweets from Russia Today that violated advertising policies garnered approximately 50 million impressions in the run-up to the 2016 US Election.[58] In the last year, however, social media companies have more stringently policed disinformation disseminated by non-anonymous perpetrators. As cited in the introduction, in August 2020, Facebook and Twitter removed posts deemed to contain "harmful" claims about coronavirus on Donald Trump's re-election campaign's accounts. In July 2020, Twitter temporarily suspended the President's son, Donald Trump Jr., for sharing a clip that it claimed promoted "misinformation" about coronavirus and hydroxychloroquine.[59]

Yet, the companies' crackdown is in its rudimentary stages, so the impact of non-anonymous perpetrators is still not negligible. However, the following analysis will focus on how anonymity in particular can contribute to, and exacerbate the problem of, the dissemination of disinformation.

[57] 'Trump Retweets Doctored Video Of Pelosi To Portray Her As Having 'Lost It' (*Reuters*, 25 May 2019) <https://www.reuters.com/article/us-usa-trump-pelosi/trump-retweets-doctored-video-of-pelosi-to-portray-her-as-having-lost-it-idUSKCN1SU2CB> accessed 21 August 2020.

[58] Alexandre Alaphilippe and others, *Automated Tackling of Disinformation: Major Challenges Ahead* (2019) <http://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf> accessed 14 February 2021, p.15.

[59] 'Facebook And Twitter Restrict Trump Accounts Over 'Harmful' Virus Claim' (*BBC News*, 6 August 2020) <https://www.bbc.com/news/election-us-2020-53673797> accessed 21 August 2020.

i. The Different Forms of Anonymity

Several forms of anonymity can be identified. The perpetrators may be simply unidentifiable, or they may be known but hidden, especially after content has been shared multiple times. Perpetrators may also remain anonymous by using a false front or impersonation. Finally, even if the content's producer is known, their true affiliations may not be. The different forms of anonymity certainly overlap, but they are distinguished here for clarity, and to allow for an exploration of some of the nuances attached to each of them.

Unidentifiable perpetrators

The first and most obvious form of anonymity is where the source of a post, tweet, video, or other content is unknown. Certain social media sites such as 4chan, an anonymous message board website, do not require users to identify themselves.[60] Under the cloak of such anonymity, social media users may be inclined to post or share disinformation. As long as users remain unidentifiable, they can remain immune to the consequences of their actions.[61]

The 'Pizzagate' conspiracy theory, discussed in II.I., illustrates this problem well.[62] Further examples of the same issue were two inflammatory stories that appeared in the run up to the Italian general election in November 2017. The first alleged that a 9-year-old Muslim girl had been sexually assaulted by her 35-year-old husband and required hospitalisation. The second alleged that lawmaker Maria Elena Boschi, a member of former Prime Minister Matteo Renzi's party, was photographed mourning the death of mafia boss Salvatore Riina. Both were fabricated, although it remains unclear by whom.[63] When internet users are able to remain anonymous or untraceable, not only may they become emboldened, but it becomes difficult, if not impossible, to take punitive measures against them.

Hidden perpetrators

A second and less obvious form of anonymity occurs when the source of a post, tweet, video, or other content is traceable, but hidden and obscure. This may occur when a false story has been shared or reposted numerous times, such that the majority of its viewers encounter it as content

[60] '4Chan' (*4chan.org*, 2020) <http://www.4chan.org/> accessed 21 August 2020.
[61] Janna Anderson and Lee Raine, 'The Future Of Truth And Misinformation Online' (*Pew Research*, 2017) <https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/> accessed 21 August 2020.
[62] Alice Marwick and Rebecca Lewis, 'Media Manipulation And Disinformation Online' [2020] Data & Society <https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf> accessed 21 August 2020, p.55-56.
[63] Yasmeen Serhan, 'Italy Scrambles To Fight Misinformation Ahead Of Its Elections' (*The Atlantic*, 2018) <https://www.theatlantic.com/international/archive/2018/02/europe-fake-news/551972/> accessed 21 August 2020.

shared by friends, prominent figures, or reputable news outlets. In these situations, most viewers of such content are unlikely to be aware of the origins of the content that they are viewing, even if they could trace it with sufficient persistence. The original perpetrator may even remove the original post after a period of time.

Such forms of anonymity are particularly relevant to this discussion, since disinformation tends to go viral easily. It has been noted that disinformation gets a far longer chain of retweets and reposts than real news because it is more likely to be surprising and emotive.[64][65] As such, it is easy for the true origins of these stories to be obscured amidst a chain of retweets and reposts.

Furthermore, viral content may be more easily taken at face value when viewers are engaging with reposted or retweeted content. Social media platforms have transformed their users from audience members to co-producers of content, including news. A UNESCO report noted that this facilitates the spread of content via 'trust networks' of family and friends, wherein inaccurate, false and malicious content can find increased traction.[66]

This effect may be particularly pronounced on private messaging platforms such as WhatsApp, as illustrated in India. Rumours of kidnappings were spread over WhatsApp, and substantiated with images and video taken from different contexts.[67] Lynch mobs were formed in response, which attacked passing travellers who were mistakenly accused of being child abductors.[68] More recently, WhatsApp was used in India to disseminate disinformation ahead of the 2019 elections. For instance, photographs were circulated purporting to show that Indian airstrikes on Pakistani territory had been successful, but these were in fact old images from other events.[69]

Politically-motivated disinformation was also spread via WhatsApp in Brazil. When Brazilians vote, they punch in a number for a candidate or party in an electronic voting machine. Disinformation spread over WhatsApp in 2018 gave the wrong number for a particular politician, which may have caused some people to unintentionally vote for a different party. This kind of

[64] Joshua Tucker and others, 'Social Media, Political Polarization
[65] Political Disinformation: A Review Of The Scientific Literature' [2018] SSRN Electronic Journal, p.37.
[66] Julie Posetti, 'News Industry Transformation: Digital Technology, Social Platforms And The Spread Of Misinformation And Disinformation' [2018] Handbook for Journalism Education and Training, UNESCO, p 59-60.
[67] Timothy McLaughlin, 'How Whatsapp Fuels Fake News And Violence In India' (*Wired*, 2018) <https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/> accessed 21 August 2020.
[68] 'How Whatsapp Helped Turn An Indian Village Into A Lynch Mob' (*BBC News*, 2018) <https://www.bbc.com/news/world-asia-india-44856910> accessed 21 August 2020.
[69] 'Whatsapp: The 'Black Hole' Of Fake News In India's Election' (*BBC News*, 2019) <https://www.bbc.com/news/world-asia-india-47797151> accessed 21 August 2020.

disinformation is particularly difficult to trace and therefore debunk, since WhatsApp uses end-to-end encryption.[70]

WhatsApp has taken measures such as limiting users to forwarding a message only five times.[71] This places a cap on the unfettered dissemination of viral messages, which may otherwise spread so far that their original source is obscured. Nevertheless, at the time of writing, some issues remain, such as the fact that forwarded messages do not indicate the original sender. This may reflect valid privacy considerations, but nonetheless contributes to the problem of anonymity, which can exacerbate the spread of disinformation via private chat groups and trust networks.

Impersonation

Impersonation is another form of identity concealment and deception. In the United States, Democratic primary candidates have been impersonated on Twitter, a problem not helped by the fact that the social media platform grants verified accounts to congressional candidates only after they win their primaries.[72]

Such impersonation attempts can be carried out by both humans and bots (i.e. software that automates the process of posting and sharing content), and may even form part of a concerted campaign of disinformation. The Internet Research Agency in Russia attempts to change the majority position in debates within and without Russia, or at least muddy the waters, through paying full-time staff to express opinions through fake accounts.[73] China's '50 Cent Army' plays a similar role.[74] A report by the European Parliament estimates that more than half of the accounts following Donald Trump on Twitter may be fake.[75] As will be discussed in the following section, this may contribute to the 'social proof' effect, whereby content is considered more valid when it has a larger following.[76] These accounts also post and repost news stories, and share opinions and comments. This may not always be disinformation. Nevertheless, the fundamental point remains:

---

[70] Fernando Hadad, 'Disinformation Spreads On Whatsapp Ahead Of Brazilian Election' (*New York Times*, 19 October 2018) <https://www.nytimes.com/2018/10/19/technology/whatsapp-brazil-presidential-election.html?module=inline> accessed 21 August 2020.

[71] Jacob Kastrenakes, 'Whatsapp Limits Message Forwarding In Fight Against Misinformation' (*The Verge*, 21 January 2019) <https://www.theverge.com/2019/1/21/18191455/whatsapp-forwarding-limit-five-messages-misinformation-battle> accessed 21 August 2020.

[72] Maegan Vazquez and Donie O'Sullivan, 'Twitter Tells New Congressional Candidates They'll Have To Win Their Primaries To Get Verified' (*CNN*, 6 August 2019) <https://edition.cnn.com/2019/08/06/politics/twitter-primaries-verification/index.html> accessed 21 August 2020.

[73] Adrian Chen, 'The Agency' (*New York Times*, 7 June 2015) <https://www.nytimes.com/2015/06/07/magazine/the-agency.html> accessed 21 August 2020.

[74] Damian Tambini, 'Fake News: Public Policy Responses' (*LSE Media Policy Project*, 2017).

[75] Alexandre Alaphilippe and others, *Automated Tackling of Disinformation: Major Challenges Ahead*, p.23

[76] Stephen Harkins, Kipling Williams, and Jerry Burger, *The Oxford Handbook of Social Influence*, Oxford Handbooks (Oxford University Press, 2017) 108 <doi.org/10.1093/oxfordhb/9780199859870.013.4>.

impersonation is a tool that can be used by malicious actors as part of campaigns to mislead and sow discord.[77]

The particular threat posed by troll and bot accounts may only worsen as technology improves, as it will likely result in more sophisticated impersonations of human users.[78] Some measures have been taken already. Twitter has improved their automated systems for identifying and challenging suspected spam or bot accounts. Such accounts are barred from engaging with other users or tweeting until they pass a 'challenge', such as confirming a phone number.[78] Nonetheless, researchers have suggested that Twitter is not flagging up content and taking action against these accounts quickly enough, and that improved machine learning algorithms may be needed to do so more swiftly.[79]

Opaque motivation or funding

A fourth form of anonymity occurs where the producer of content is known, but their financial, political, or other motivations are not publicly known. This has been highlighted as the key factor that sets today's politically-targeted content apart from conventional political advertising.[80] Users are generally oblivious to the partisan affiliations of what they view online.[81] There has been a dearth of regulations that compel political parties to disclose their sponsorship of online, micro-targeted content. This form of anonymity has allowed political actors to avoid being held accountable for the content they spread.[82]

An appropriate example is Russia's state-sponsored channels on YouTube, until now a relatively less scrutinised social media platform. Until recently, YouTube had failed to label news outlets such as NTW and Russia-24 as state-sponsored, even though they disseminated false reports such as a purported cover-up by a U.S. politician of an organ harvesting ring.[83] While YouTube has now labelled the channels appropriately, their response – as is often the case with problematic

[77] Nicholas Thompson and Issie Lapowsky, 'How Russian Trolls Used Meme Warfare To Divide America' (*Wired*, 2018) <https://www.wired.com/story/russia-ira-propaganda-senate-report/> accessed 21 August 2020.
[78] Yoel Roth and Del Harvey, 'How Twitter Is Fighting Spam And Malicious Automation' (*Twitter*, 2018) <https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html> accessed 21 August 2020.
[79] Andy Greenberg, 'Twitter Still Can't Keep Up With Its Flood Of Junk Accounts, Study Finds' (*Wired*, 2019) <https://www.wired.com/story/twitter-abusive-apps-machine-learning/> accessed 21 August 2020.
[80] Arjun Bisen, 'Disinformation Is Drowning Democracy' (*Foreign Policy*, 2019) <https://foreignpolicy.com/2019/04/24/disinformation-is-drowning-democracy/> accessed 21 August 2020.
[81] Ibid.
[82] Ibid.
[83] Paresh Dave and Christopher Bing, 'Russian Disinformation On Youtube Draws Ads, Lacks Warning Labels: Researchers' (*Reuters*, 2019) <https://www.reuters.com/article/us-alphabet-google-youtube-russia/russian-disinformation-on-youtube-draws-ads-lacks-warning-labels-researchers-idUSKCN1T80JP> accessed 21 August 2020.

accounts – seems reactive rather than preventive. More proactively, YouTube has invested $25 million in journalism on its platform to promote content from vetted sources.[84] This came in response to disinformation about mass shootings, such as videos claiming that the Las Vegas shooting was a hoax.[85] In general, however, YouTube has been inclined to merely demote content rather than take content down entirely.[86] Some argue that YouTube should remove such content altogether, although there are, of course, normative dimensions to this debate regarding freedom of speech.[87] There is certainly an impetus for exploring means of regulating content on YouTube given the rapid development of technologies that can create deepfakes and false videos that are increasingly realistic.

Encouragingly, Google, Facebook, and Twitter have started making data about who is paying for political advertising publicly available.[88] However, actors have responded by going to greater lengths to conceal their identities, through the use of virtual private networks (VPNs), internet phone services, and third parties that run ads on their behalf.[89] Indeed, Nathaniel Gleicher, the head of cyber security policy at Facebook, has noted that, unlike during the 2016 election campaign, the company is no longer able to attribute disinformation campaigns to Russia's Internet Research Agency, despite evidence of several continuities.[90] Similarly, the Swedish security service has been unable to attribute foreign influence campaigns to Russia, despite the resemblance they bear to those in the United States and Europe.[91]

[84] Alex Hern, 'Youtube To Crack Down On Fake News, Backing 'Authoritative' Sources' (*The Guardian*, 9 July 2018) <https://www.theguardian.com/technology/2018/jul/09/youtube-fake-news-changes> accessed 21 August 2020.
[85] Ibid.
[86] Samuel Stolton, 'In The Fight Against Fake News, Youtube Has A 'Bias Toward Keeping Content Up' (*Euractiv*, 2 May 2019) <https://www.euractiv.com/section/digital/news/in-the-fight-against-fake-news-youtube-has-a-bias-toward-keeping-content-up/> accessed 21 August 2020.

[87] Michael Posner, 'Dealing With Disinformation: Facebook And Youtube Need To Take Down Provably False "News"' (*Forbes*, 14 March 2019) <https://www.forbes.com/sites/michaelposner/2019/03/14/dealing-with-disinformation-facebook-and-youtube-need-to-take-down-provably-false-news/#f487e0e19e79> accessed 21 August 2020.
[88] Alexandre Alaphilippe and others, *Automated Tackling of Disinformation: Major Challenges Ahead*, p.16
[89] Facebook Uncovers Disinformation Campaign To Influence US Midterms' (*The Financial Times,* 1 August 2018) <https://www.ft.com/content/7af02014-94e1-11e8-b67b-b8205561c3fe> accessed 21 August 2020.
[90] Ibid.
[91] Kristine Berzina, 'Sweden — Preparing For The Wolf, Not Crying Wolf: Anticipating And Tracking Influence Operations In Advance Of Sweden's 2018 General Elections' (*The German Marshall Fund of the United States*, 7 September 2018) <http://www.gmfus.org/blog/2018/09/07/sweden-preparing-wolf-not-crying-wolf-anticipating-and-tracking-influence> accessed 21 August 2020.

ii. How Technology Enables Anonymity

As has been discussed above, technology has played a significant role in enabling the anonymous dissemination of information. The fundamental structure of the internet is built around the principles of openness and interoperability; it was not designed with a view to identifying end users or end user locations.[92] Furthermore, there is an unavoidable gap between technical attribution and human attribution. Even if one can perfectly identify the IP addresses from which a malicious campaign is conducted, one may not be able to connect the machine to the person behind the attack.[93]

Technological progress has also enabled internet users to adopt techniques to obscure their identity. This includes spoofing, which was, for instance, employed by Russian 'sock-puppet' accounts to spread ideologically-laden and divisive disinformation in the aftermath of terrorist attacks.[94] These accounts posed as users along all ends of the political spectrum, using the incident as a springboard to spread specific rhetoric, such as anti-Muslim rumours through a false US conservative Texan persona.[95] The technology of social media platforms has enabled content to go viral, which can obscure its true origins, as illustrated by the spread of disinformation in private WhatsApp chats, as discussed earlier.

However, it is important to not overplay the role of technology. As will be discussed in the next section, end users are, due to a range of psychological factors, vulnerable to disinformation in all forms. These susceptibilities are exacerbated by social media: users are not sufficiently cognizant of the methods and tricks that may be used by perpetrators to obscure their identities, and are thus vulnerable to taking information at face value. Therefore, while regulation of platforms and content is important, engendering greater awareness of the possible methods of deception amongst the population would make it much harder for malicious actors to succeed in their aims.

[92] Jason Jolley, *Attribution, State Responsibility, And The Duty To Prevent Malicious Cyber-Attacks In International Law* (University of Glasgow 2017) <http://theses.gla.ac.uk/8452/1/2017JolleyPhD.pdf> accessed 21 August 2020, 144-148
[93] Ibid.
[94] Martin Innes, Diyana Dobreva and Helen Innes, 'Disinformation And Digital Influencing After Terrorism: Spoofing, Truthing And Social Proofing' [2019] Contemporary Social Science 7-8
[95] Ibid.

II.III. VULNERABILITY OF THE MASSES

People seem to naturally be vulnerable to believing disinformation. On the most basic level, this is because different forms of disinformation often look similar to their genuine equivalents, making it difficult to distinguish between the two. BuzzFeed News superimposed the word "FALSE" onto a screenshot of a fake news article, for example, suggesting that its misleading nature might not otherwise be clear to readers.[96] Similarly, Russian channels Russia Today and Sputnik have the appearance of proper news programmes, despite serving propagandist purposes for the Russian state.[97] Moreover, fake videos of politicians talking are somewhat convincing, though not yet completely indistinguishable from real ones.[98]

Certain psychological phenomena appear to make people particularly vulnerable to believing at least some types of disinformation, and actors seeking to spread disinformation seem to deliberately exploit these phenomena. Furthermore, increasingly advanced AI could exacerbate these issues in the future.

The Malicious Use of Artificial Intelligence report, written by leading AI governance groups, sets out a helpful framework for considering such risks. First, AI could expand existing threats: its efficiency and scalability could allow actors to carry out far more attacks than they do currently, or to start carrying out attacks where they did not previously have the resources. Second, AI could introduce new threats that humans would not otherwise be able to carry out. This point is not considered here, since the attacks addressed all already exist in some form. Third, AI could alter the character of the threats: efficiency and scalability would lessen the need for actors to choose between scale and effectiveness of attacks, permitting numerous, potentially highly effective attacks.[99]

[96] Craig Silverman, 'This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook' (*BuzzFeed News*, 16 November 2016) <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook> accessed 9 September 2019.
[97] Cristopher Paul and Miriam Mathews, 'The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It' (2016) Perspective 5 <https://doi.org/10.7249/PE198>.
[98] David Mack, 'This PSA About Fake News from Barack Obama Is Not What It Appears' (*BuzzFeed News*, 17 April 2018) <https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed> accessed 14 February 2021.
[99] Miles Brundage and others, 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', (2018) *ArXiv:1802.07228 [Cs]* 18–22 <http://arxiv.org/abs/1802.07228> accessed 9 September 2019.

i. Psychological Vulnerabilities to Disinformation

Motivated reasoning

One way in which people are vulnerable to believing disinformation is motivated reasoning, a central theoretical concept in Psychology and Political Science.[100] Motivated reasoning is people's tendency to discount or argue against facts and arguments that conflict with their prior beliefs but much more readily accept facts and arguments that fit their beliefs.[101] In a commonly cited study of the phenomenon, participants were shown research that suggested either that capital punishment has a deterrent effect or that it does not. Proponents and opponents of capital punishment both rated the research that supported their pre-existing beliefs to be more convincing.[102]

To some extent, motivated reasoning protects people from disinformation. If a piece of disinformation does not support an individual's prior belief, they are likely to be sceptical of it. However, if the disinformation supports the individual's prior belief, they will be less sceptical. In this case, the disinformation may strengthen their initial belief, even if this belief is incorrect.

The precise population targeting made possible by social media advertising could help malign actors show specific pieces of disinformation only to people who already have the particular worldview required to find it convincing. As a paper for the European Parliament argues, this has the strategic advantage of keeping these "dark ads" hidden from other social media users. Other users cannot attempt to correct the disinformation; they may not even know that the actor was spreading disinformation.[103] Thus, disinformation campaigns can be disseminated to their target audiences effectively, while remaining undetected by opponents. The researchers argue that VoteLeave used this strategy during the Brexit campaign. The disinformation in the examples of dark ads given – about the costs of Britain's EU membership, for example – were,

[100] Thomas J. Leeper and Kevin J. Mullinix, 'Motivated Reasoning' (Oxford University Press, 2018) <https://doi.org/10.1093/obo/9780199756223-0237>.
[101] Milton Lodge and Charles S Taber, *The Rationalizing Voter* (Cambridge University Press, 2013) 149–50 <https://www.cambridge.org/core/books/rationalizing-voter/9E4E27965B612D5DDB1091BD907DF492> accessed 9 September 2019.
[102] Charles G. Lord, Lee Ross, and Mark R. Lepper, 'Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence.' (1979) 37.11 *Journal of Personality and Social Psychology* 2098–2109 <https://doi.org/10.1037/0022-3514.37.11.2098>.
[103] Alexandre Alaphilippe and others, *Automated Tackling of Disinformation: Major Challenges Ahead* (European Parliament 2019) 18. <http://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf> accessed 9 September 2019.

however, also repeatedly used by VoteLeave in public.[104][105] This suggests that VoteLeave's targeting of disinformation had a different strategic purpose. One possible reason is that it would have allowed the group's advertising budget to be spent primarily on advertising to those who were most likely to be influenced by the disinformation. This targeted and efficient spending may have led to a higher rate of persuasion amongst those who were more susceptible to the advertisements in the first place.

Even without buying advertisements, it is possible to 'organically' direct specific pieces of disinformation towards people that are, owing to motivated reasoning, more likely to be vulnerable to them. Certain hashtags on Twitter are more likely to be used and read by people with a particular political perspective. One could, for instance, target disinformation aiming to increase support for Brexit towards people who were already Eurosceptic, and were therefore more susceptible to this sort of disinformation, by using the "no2eu" hashtag. Unlike "dark ads", this tactic does not have the aforementioned advantages of concealing disinformation from those who are unlikely to believe it. Anyone can search for and read through tweets with a particular hashtag. It is unclear, however, how many Remainers (to continue the example) would look through a pro-Brexit hashtag.

Using the Malicious Use of AI framework, it is possible to see how more advanced AI could increase the likelihood of people's motivated reasoning being exploited. One potential application of AI is analysis of vast amounts of data to enable more precise targeting of individuals. Social media platforms already offer targeting of specific user groups for all ads purchased. Therefore, social media companies are incentivised to improve their means of targeting specific groups because it gives them a competitive advantage in the advertising market: improved targeting incentivises legitimate as well as illegitimate advertisers to spend their advertising budget on social media ads over other forms. More sophisticated AI technology would increase the effectiveness of the targeting offered to advertisers and therefore would be adopted readily by social media companies. This would benefit legitimate advertisers, but it would also allow malign actors to focus their advertising budget on users who are particularly likely to be vulnerable to them.

[104] Ibid, p. 19.
[105] Anthony Reuben, 'Are We Giving £350m a Week to Brussels?' (*BBC News*, 22 April 2016) <https://www.bbc.com/news/uk-politics-eu-referendum-36110822> accessed 9 September 2019.

AI could also increase the ability of malign actors to target users in organic ways (i.e. without buying advertising). It could analyse large amounts of data to find numerous hashtags (including those less obvious than the likes of "#no2eu") that are frequented by users who are vulnerable to motivated reasoning for that specific case. It seems plausible that sophisticated data analysis could reveal hashtags which tend to be used by people who are leaning in a political direction, but are not yet set upon it. Whereas "#no2eu" conveys a well-stated policy position, other hashtags may be used by people who are only somewhat Eurosceptic. For malign actors, the motivated reasoning of these individuals is most desirable to exploit: these individuals already have a prior belief that can be magnified, but this belief is not yet strong enough that they would necessarily vote without manipulation in the way the malign actor wishes. Malign actors could therefore post disinformation to these hashtags discovered by AI, broadening their reach.

Furthermore, natural language processing, which is the ability for software to use human language, could lower the costs of spreading disinformation.[106] Having software rather than humans write messages in order to flood specific hashtags with disinformation would eliminate labour costs. It would also lower the barriers to entry in other ways, such as by removing the need for the actor spreading disinformation to have a good knowledge of the language of the target population. Numerous, previously less powerful actors would be able to publish disinformation that attempts to exploit individuals' motivated reasoning. This could generate a far greater amount of disinformation, as well as increased difficulties in attributing it.

Social proof

A second way in which people are vulnerable to believing disinformation is through social proof. Social proof refers to people's tendency to observe the behaviour of others to guide them in how to behave in an ambiguous situation. It is commonly used in advertising, where companies emphasise the existing popularity of a product in order to persuade people who are unsure of whether to buy it into doing so.[107] Similarly, a well-cited study found that hotel guests were much more likely to comply with requests to reuse their towels if they were told that the majority of

[106] 'Natural-Language Processing', in Andrew Butterfield, Gerard Ekembe Ngondi, and Anne Kerr (eds), *A Dictionary of Computer Science* (Oxford University Press 2016) <https://www.oxfordreference.com/view/10.1093/acref/9780199688975.001.0001/acref-9780199688975-e-6410> accessed 10 September 2019.

[107] Stephen Harkins, Kipling Williams, and Jerry Burger, *The Oxford Handbook of Social Influence* (Oxford University Press 2017) 108 <doi.org/10.1093/oxfordhb/9780199859870.013.4>.

guests do so than if they were told about the environmental benefits.[108] Social proof can also have an effect on less conscious decisions; hearing others laugh generally makes people find a TV programme funnier.[109]

Relying on the wisdom of the crowds when it comes to judging the credibility of information could make people vulnerable to believing disinformation, particularly when using social media. This is because it is relatively easy for malign actors to manipulate who and how many appear to be in the crowd of social media users supporting a particular cause. For example, 'anti-vaxxers' have used co-ordination and a good understanding of Twitter to amplify their otherwise relatively small presence on the site.[110]

Whereas anti-vaxxers seem to generally use their genuine accounts, other exploitation of social proof uses fake accounts. 'Troll' accounts, for instance, appear to represent specific individuals, but are in fact controlled in bulk by human handlers. Bots are similar, though their activity is managed by software. These fake accounts can be used for 'astroturfing', where the appearance of a large grassroots movement and consensus in favour of a particular cause is suggested.[111]

As an example, Presidents Obama and Trump both seem to have had large numbers of fake followers on Twitter at one point.[112] This may have been in order to exploit social proof: the politicians appearing more popular might have made them become more popular. It is important to note that attribution is difficult here. Although it is possible that the presidents or their teams ordered this astroturfing, it could also have been carried out by an unconnected sympathiser, or even an opponent who wanted to make it appear that the presidents were astroturfing in order to discredit them.

A further example of astroturfing is that around 500 automated Twitter accounts appear to have been used in the 2016 American election to spread criticism of Ted Cruz. Researchers have

[108] Noah J. Goldstein, Robert B. Cialdini, and Vladas Griskevicius, 'A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels' (2008) 35.3 *Journal of Consumer Research* 472–82 <https://doi.org/10.1086/586910>.

[109] Michael J. Platow and others, '"It's Not Funny If They're Laughing": Self-Categorization, Social Influence, and Responses to Canned Laughter', (2005) 41.5 *Journal of Experimental Social Psychology* 542, 542–43 <https://doi.org/10.1016/j.jesp.2004.09.005>.

[110] Renee DiResta Lotan Gilad, 'Anti-Vaxxers Are Using Twitter to Manipulate a Vaccine Bill' (*Wired*, 8 June 2015) <https://www.wired.com/2015/06/antivaxxers-influencing-legislation/> accessed 9 September 2019.

[111] Samuel Woolley and Philip N. Howard, 'Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media' (Oxford University Press 2019) 3–5.

[112] Philip Bump, 'Trump Is Mad about the Size of His Crowd on Twitter' (*Washington Post*, 23 April 2019) <https://www.washingtonpost.com/politics/2019/04/23/trump-is-mad-about-size-his-crowd-twitter/> accessed 9 September 2019.

suggested that seeing (fake) people criticising the politician could have caused real people to view him more negatively, and therefore be less likely to vote for him.[113] This astroturfing may have drowned out any positive messages about Cruz, preventing people from being influenced through social proof into supporting him. Making criticism of Cruz artificially prominent could also have had the strategic advantage of raising awareness about any of his genuine drawbacks. This could cause people to choose, on a rational basis, not to vote for him.

Advancements in AI could, to some extent, increase people's susceptibility to disinformation spread through social proof. Bot accounts are already employed to share existing pieces of social media content or to follow certain figures in order to generate social proof. It is therefore unclear that AI could make exploiting social proof any cheaper or more accessible. However, improved population targeting could make the automated exploitation of social proof more effective. Psychological research suggests that individuals are more likely to conform to the opinions of people that they perceive to be similar to themselves.[114] In one experiment, for instance, participants were more likely to comply with requests if they (falsely) believed that they shared a birthday with the requester.[115] Malign actors could exploit this vulnerability by creating fake accounts that they know will appear similar to an individual or group. This would increase the effect of social proof for the individual or group, and therefore, the effectiveness of disseminating disinformation.

More advanced natural language processing could also make the automated exploitation of social proof more effective. It could allow bots to express the desired viewpoint in new ways, rather than simply retweeting or copying an existing message. This might make astroturfing seem more credible, and so less likely to be detected and discounted.

The Truth Effect

A third way in which people are vulnerable to disinformation is through the truth effect, sometimes called the illusory truth effect. This refers to people's tendency to view a statement whose truth is ambiguous as more credible if they have previously encountered it. The effect is thought to occur due to repetition increasing fluency, which is the ease with which statements are processed. This occurs because fluency is used as a heuristic to judge accuracy. The truth effect occurs equally for true statements as for false ones with the only constraint seeming to be that

[113] Woolley and Howard, 190–94.
[114] Harkins, Williams, and Burger, 122.
[115] Jerry M. Burger and others, 'What a Coincidence! The Effects of Incidental Similarity on Compliance' (2004) 30.1 *Personality and Social Psychology Bulletin* 35–43 <https://doi.org/10.1177/0146167203258838>.

individuals have to be uncertain about the truthfulness of the statement. A 2010 meta-analysis of 51 earlier studies called the evidence supporting the existence of the truth effect "very robust".[116]

The truth effect could make people more vulnerable to disinformation because the spreader of disinformation only has to repeat the false claim in order to make it more believable, as long as the claim is not self-evidently false. Indeed, a recent study found that participants rated fake new headlines from the 2016 U.S. election as more credible if the researchers had earlier shown these headlines to them.[117] Furthermore, some researchers have suggested that one of the reasons why Russian propaganda in the West is so repetitive is that it makes use of this effect in order to seem more credible.[118]

However, there is comparatively little research on the effect of multiple repetitions on the truth effect.[119] The current understanding is that some repetition increases credibility, but that excessive repetition causes it to drop again, although psychologists do not yet have a precise understanding of what the optimal amount of repetition is for maximum credibility.[120] The extent to which Russian or other disinformation campaigns hit the optimum balance is therefore also unclear.

Improvements in AI's natural language processing could affect malign actors' exploitation of the truth effect in a similar way to the exploitation of motivated reasoning. Automating the production of disinformation would reduce labour costs, lowering the barriers to entry for disinformation campaigns that exploit people's vulnerability to the truth effect. Moreover, natural language processing would reduce the need for malign actors to be able to speak the language of the target population. Far more actors, potentially producing far more disinformation, could therefore seek to exploit the truth effect in order to influence people.

[116] Alice Dechêne and others, 'The Truth About the Truth: A Meta-Analytic Review of the Truth Effect' (2010) 14.2 *Personality and Social Psychology Review* 238, 239 <https://doi.org/10.1177/1088868309352251>.
[117] Gordon Pennycook, Tyrone D. Cannon, and David G. Rand, 'Prior Exposure Increases Perceived Accuracy of Fake News' (2018) 147.12 *Journal of Experimental Psychology: General* 1865–80 <https://doi.org/10.1037/xge0000465>.
[118] Paul and Mathews, 4.
[119] Dechêne and others, 254.
[120] Nicole Ernst, Rinaldo Kühne, and Werner Wirth, 'Effects of Message Repetition and Negativity on Credibility Judgments and Political Attitudes' (2017) 11 International Journal of Communication 3265 <https://doi.org/10.5167/uzh-139745>.

ii. Consequences of Psychological Vulnerabilities

Debunking and 'The Backfire Effect'

People's vulnerability to disinformation is particularly concerning given that correcting people's beliefs appears to be very difficult. Indeed, some research suggests that corrective messaging not only fails to correct false beliefs, but rather, causes people to believe the misinformation even more strongly. Nyhan and Reifler's 2010 paper, which has been widely cited, including by hundreds of news outlets, had participants read a mocked-up news report about the Iraq War.[121] The report falsely stated that there had been Weapons of Mass Destruction (WMDs) in the country. Participants who subsequently received a debunking of this piece of information were more likely to believe that there had been WMDs than those that did not receive a debunking.[122] A 2012 study involving the same researchers generated the same result among one specific sub-group of participants.[123]

People's vulnerability to this 'backfire' or 'boomerang' effect has been linked to two of the psychological phenomena set out earlier. Several researchers have argued that there is a stronger form of motivated reasoning in which people are not just highly sceptical of opposing arguments, but actively argue against them, further entrenching their prior beliefs. In addition, since debunking generally involves repeating the original misinformation in order to address it, it could contribute to the truth effect, where familiarity with a claim is used as a heuristic for its credibility.[124]

The impact of the backfire effect should not be overstated, however. As Full Fact, the factchecking organisation, notes, none of the five more recent studies in the area have found statistically significant evidence of backfiring.[125] Moreover, the studies that indicate a backfire effect have far fewer participants than the studies that give cause for scepticism.[126] This suggests

[121] Daniel Engber, 'We've Been Told We're Living in a Post-Truth Age. Don't Believe It.'(*Slate Magazine,* 3 January 2018) <https://slate.com/health-and-science/2018/01/weve-been-told-were-living-in-a-post-truth-age-dont-believe-it.html> accessed 11 September 2019.
[122] Brendan Nyhan and Jason Reifler, 'When Corrections Fail: The Persistence of Political Misperceptions' (2010) 32.2 *Political Behavior* 303–30 <https://doi.org/10.1007/s11109-010-9112-2>.
[123] Brendan Nyhan, Jason Reifler, and Peter A. Ubel, 'The Hazards of Correcting Myths About Health Care Reform' (2012) 51.2 *Medical Care* 127–32 <https://doi.org/10.1097/MLR.0b013e318279486b>.
[124] Stephan Lewandowsky and others, 'Misinformation and Its Correction: Continued Influence and Successful Debiasing' (2012) 13.3 *Psychological Science in the Public Interest* 106, 117–19) <https://doi.org/10.1177/1529100612451018>.
[125] Amy Sippitt, *The Backfire Effect: Does It Exist? And Does It Matter for Factcheckers?* (Full Fact 2019) 4 <https://fullfact.org/blog/2019/mar/does-backfire-effect-exist/> accessed 14 February 2021.
[126] Nyhan and Reifler, 312, 316; Nyhan, Reifler, and Ubel, 128; Thomas Wood and Ethan Porter, 'The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence' (2016) *SSRN Electronic Journal* <https://doi.org/10.2139/ssrn.2819073>.

that the initial results in favour of backfiring may have been due to random noise in the data or quirks in the participants in those studies rather than a fundamental aspect of human psychology. Indeed, one of the authors of the studies which found a backfire effect has since moderated his views, no longer believing, in the majority of cases, that debunking is unhelpful.[127]

Even if debunking is not actively unhelpful, it certainly appears not to be as effective as one might like, or as it would be if people were completely rational and objective. This can be seen in a 2012 review of several earlier studies, all employing the same methodology: participants read about a warehouse fire and learned that it was caused by gas cylinders that had been negligently left there. Some were then told that there were in fact no gas cylinders. In the subsequent memory test, references to the gas cylinders being the cause of the fire dropped by no more than half among the participants who received the correction. This suggests a limited effectiveness of corrective messaging.[128] A 2017 meta-analysis of 52 previous studies, including some which used authentic examples of misinformation, similarly found a "large" effect of people still somewhat believing misinformation after receiving a corrective message.[129]

Lewandowsky summarises the possible explanations for why debunking does not tend to be fully effective. One explanation, for instance, is that people build mental models to understand unfolding events. Individuals continue using their model, even if corrective messages subsequently suggest that part of it is false, as they do not have another way of explaining the events. In addition, as explained above, the notion of fluency suggests that people use their familiarity with a claim as a heuristic for judging its credibility. Since corrective messaging tends to involve repeating the false claim, it potentially increases the claim's fluency and perceived credibility. Lewandowsky uses this understanding of why debunking can fail to give recommendations to fact-checkers about how to minimise the risk of this. He suggests, for instance, that fact-checkers use explanations of why a piece of disinformation was incorrect, in order to replace the gap in people's mental models. Furthermore, fact checkers should also avoid repeating the misinformation in their corrective messaging to avoid increasing fluency.[130]

---

[127] Brendan Nyhan, 'Fact-Checking Can Change Views? We Rate That as Mostly True' (*The New York Times*, 6 November 2016 <https://www.nytimes.com/2016/11/06/upshot/fact-checking-can-change-views-we-rate-that-as-mostly-true.html> accessed 11 September 2019.

[128] Lewandowsky and others.

[129] Man-pui Sally Chan and others, 'Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation' (2017) 28.11 *Psychological Science* 1531–46 <https://doi.org/10.1177/0956797617714579>.

[130] Lewandowsky and others, 113–17.

More advanced AI could plausibly make disinformation even harder to fully debunk. As it is currently easier to write false information than to generate a fake video, people are likely to be less sceptical of videos than written information. However, technology is rapidly improving. Indeed, as existing deepfakes suggest, AI could, in future, make it easy to generate fake videos. If people remain trusting of videos, they will be more vulnerable to initially believing disinformation. Furthermore, they will be less receptive to corrective messaging. This is because scepticism at the time of receiving false information is known to make subsequent debunking more effective.[131] Major outlets such as BuzzFeed News have already started alerting people to the capabilities of deepfakes, however.[132] This may cause people to become more sceptical of the videos they watch, and thereby reduce the risk of them being convinced.

[131] Chan and others, 1541; Lewandowsky and others, 116.
[132] Mack.

iii. Conclusions on Vulnerability of the Masses

Due to the aforementioned psychological vulnerabilities – motivated reasoning, social proof and 'The Truth Effect' – people are highly susceptible to being influenced by disinformation; indeed, these susceptibilities are already being exploited by malign actors. The prospect of more sophisticated technology, particularly in the field of AI, only serves to exacerbate these extant problems. While debunking disinformation may not be as ineffective as some may assume due to the reasonably tenuous studies on 'The Backfire Effect', debunking is still fraught with difficulties due to the impact of 'motivated reasoning' and 'The Truth Effect'. Thus, any significant policies directed towards debunking disinformation need to consider providing explanations for why a piece of misinformation was incorrect rather than simple statements that the information is false and avoiding repetitions of the disinformation in their corrective messaging.

Furthermore, the impact of strategies such as debunking, coupled with the growing publicity given to disinformation may have another psychological effect on people: widespread scepticism towards all sources of information. This problem will be addressed in the following section.

II.IV. REDUCED CREDIBILITY OF INFORMATION

i. Sources of Reduced Credibility of Information

The pervasiveness of disinformation, either perceived or real, may begin to erode public trust in all sources of information, even those that are factually accurate. This may occur organically, whereby individuals become confused between genuine information and disinformation that looks identical, to the extent that they treat all the information with suspicion.[133] This scepticism is also fuelled by politicians, most famously Donald Trump, who use the terms "disinformation" or "fake news" to discredit factually correct but critical information and news coverage. Donald Trump's tweets, such as, "Really disgusting that the failing New York Times allows dishonest writers to totally fabricate stories",[134] and, "The Huffington Post is a total joke & laughing stock of journalism, as is gross Arianna Huffington. They don't report the facts!",[135] feed into an extant distrust of all information sources. More perniciously, they also undermine the authenticity of traditional media outlets, which, due to the high barriers to entry, are considered to be more objective sources of information than internet sources, even if they may have some political leanings.

When authoritative figures publicly undermine trust in traditional media outlets, there seems to be a tendency amongst the electorate to shift away from objective information sources and gravitate towards more subjective information that appeals to their personal beliefs and emotions.[136] The so-called "post-truth era", therefore, causes a cycle of disinformation: the (perceived or real) prevalence of disinformation causes individuals to lose trust in objective sources of information, so that they turn towards more subjective information, which may be disinformation, and the disinformation multiplies to meet this need, and so on.[137]

Furthermore, it appears that the prevalence of disinformation has led to a general disregard for healthy criticism and traditional opposition based on *bona fide* facts. Any information that

[133] Janna Anderson and Lee Raine, 'The Future Of Truth And Misinformation Online' (*Pew Research*, 19 October 2017) <https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/> accessed 22 August 2020.
[134] Donald Trump, 2016, <https://twitter.com/realdonaldtrump/status/689458468835569664?lang=en> accessed 22 August 2020.
[135] Donald Trump, 2015, < https://twitter.com/realdonaldtrump/status/570238975157387264> accessed 22 August 2020.
[136] Janna Anderson and Lee Raine, 'The Future Of Truth And Misinformation Online'.
[137] Ibid.

opposes a prevailing view amongst certain individuals (usually fuelled by politicians) may be dismissed as "fake news".[138]

[138] 'How President Trump Took 'Fake News' Into The Mainstream' (*BBC*, 12 November 2018) < https://www.bbc.com/news/av/world-us-canada-46175024> accessed 22 August 2020.

ii. Consequences of Reduced Credibility of Information

James Kulinski and Paul Quirk argue that in order for a democracy to function well, it is imperative that the populace is educated and well-informed.[139] The increasing levels of scepticism towards objective facts, especially those that may challenge preconceptions, may therefore have a severe impact on liberal democracies. Having a discussion based upon concrete indisputable factual information is essential to stimulate meaningful debates. Yet, if individuals become unwilling to engage in constructive debates that may better inform their opinions, and further, become hardened to certain political positions through exposure to disinformation, which is compounded by motivated reasoning and the truth effect, as discussed in II.III, the fundamental tenets of liberal democracy will be at risk. If those who disseminate disinformation are foreign actors, as is alleged to have taken place during the 2016 U.S. Elections[140] and in Sweden,[141] the sovereignty of a country and its electorate is undermined as well.

The "post-truth era" is in early stages, so it is difficult to ascertain the current effect of intentional disinformation on the electorate. However, there already seems to be a trend of increased suspicion towards information, particularly that produced by traditional media outlets.[142] If such attitudes continue and are hardened by actors and politicians, this attitude may become endemic. Strategies to mitigate the impacts of intentional disinformation therefore need to address the shift in public mood away from traditional media outlets as trustworthy sources of information.

Indeed, the impact of a perception of a reduction in the credibility of information also has an impact on the effectiveness of the strategies that may be considered to mitigate the impact of intentional disinformation. As was discussed in II.III., debunking strategies may not be entirely effective due to the impact of motivated reasoning and the truth effect. If there is also a general distrust in all sources of information, and traditional sources of information in particular, a strategy that simply offers more information will be ineffective, and may even exacerbate the problem. It is therefore imperative that any debunking strategy relies not on an existing source of information (and particularly not on traditional media outlets), but rather comprises a separate, unassailably objective and trustworthy debunking source.

---

[139] James H. Kuklinski and Paul J. Quirk, 'Conceptual Foundations Of Citizen Competence' (2001) 23 Political Behavior 285.
[140] Jennifer Hansler, 'US Accuses Russia Of Conducting Sophisticated Disinformation And Propaganda Campaign' (*CNN*, 5 August 2020) <https://edition.cnn.com/2020/08/05/politics/state-department-russian-disinformation-report/index.html> accessed 22 August 2020.
[141] Martin Kragh and Sebastian Asberg, 'Russia's Strategy For Influence Through Public Diplomacy And Active Measures: The Swedish Case' (2017) 40 Journal of Strategic Studies 773.
[142] Janna Anderson and Lee Raine, 'The Future Of Truth And Misinformation Online' (*Pew Research* 2017)).

III. SOLUTIONS

As illustrated in the previous sections, intentional disinformation is a pervasive problem in the modern social media era, manifesting through several severe issues that destabilise many traditionally solid pillars of information verification and processing in society. Left unchecked, it is easy to see that these could spiral out of control and generate huge negative impacts across multiple arenas, and create significant opportunities for malicious intentional efforts to materialise. Indeed, over the past few years, it has grown increasingly common for instances of such intentional disinformation to appear on global news, serving as an alarmingly growing influence in an unsettling number of national and international issues.

However, having dissembled intentional disinformation into several of its key components in the prior section, it is now possible to tailor solutions to address each of these key aspects of this "fake news" phenomenon; and in doing so, tackle the issue of intentional disinformation overall.

In the following sections, this paper will therefore explore various transparency models, a universal fact-checking service, a unified international anti-disinformation agency, and efforts to foster widespread digital literacy, which each individually tackle different aspects of the intentional disinformation problem.

Many of these solutions are able to operate as standalone initiatives, and each would provide significant benefits even through just their standalone implementation. However, intentional disinformation is a multi-faceted and complex issue, and these solutions have each been identified as tackling only specific aspects of this greater problem. As a result, this paper would highly recommend implementing the entire suite of solutions, as proposed in the following sections, in order to most effectively combat, in its entirety, the issue of modern intentional disinformation in the social media era.

As mentioned previously, it is also important to note that these solutions proposed are not the only possible options. It is clear to see that there are a range of possible methods that could be (and have been) employed to curtail the effects of intentional disinformation – namely, far stricter methods including legal measures and heavier-handed restrictions and controls on social media platforms and their operations. However, as previously discussed, this paper takes the view that it is necessary to take a more nuanced approach, targeted more directly at the root causes and issues of modern intentional disinformation, to be effective. As a large part of the issue lies with individual people's perceptions of truth and credibility being warped, the set of solutions

proposed here takes the stance that providing individuals greater transparency, clarity, and self-ability to accurately distinguish the truth would be far more effective in combatting modern intentional disinformation than heavy-handed blanket measures (some of which have already been attempted to limited success).

III.I. SOCIAL MEDIA TRANSPARENCY MODELS

i. Overview

As discussed in the issues section of this paper, two key problems amplifying the effect of intentional disinformation in the social media era, are that enhanced methods of fabricating credibility in fake news and the greater anonymity afforded to perpetrators. These are particularly problematic. They allow more significant quantities of fake information to spread through social media and be taken as verified truth, while the perpetrators responsible risk minimal exposure or real consequences. These effects have contributed significantly to the sheer volume of intentional disinformation regularly present on social media, which has led to many of the dire real-world consequences seen today.

However, it is possible to respond to such issues related to disinformation, without undermining freedom of speech or risking the democratic harms associated with making governments or companies arbiters of truth. If individuals can reliably judge what information is accurate, they are unlikely to believe disinformation when they see it. Moreover, as explained in the 'Vulnerability of the Masses' section, more significant scepticism towards a piece of disinformation makes people more receptive to subsequent corrective messaging, even if they did initially believe the disinformation. By empowering users to assess the credibility of information, therefore, the effects of disinformation can be primarily neutralised, without needing to trust a government or company with the power to discredit or remove information from social networks.

Two policies can empower users to better make these judgments about specific claims: digital literacy training, which is addressed later in the paper, and transparency models promoting greater transparency related to factual claims and advertisements on social networks, which is addressed in this section.

To facilitate the more significant potential for self-verification, social networks should provide individual users with the information that they need to judge the claims that they see. Therefore, when showing content to users, social networks should be transparent with users about the information sources for claims, if the sources are credible, and provide further information linked to credible sources if the claim is related to current significant intentional disinformation trends.

In addition, social networks should be transparent about broad trends relating to advertising on their platforms. This means providing academics and journalists with access to the data required to research this area. This transparency would make it easier for experts to understand the spread of any remaining disinformation and design and assess retroactive measures to protect people.

Both models would also make it more likely that actors spreading disinformation will be exposed, providing a disincentive against making use of disinformation campaigns and diminishing the greater anonymity that modern social media platforms typically allow perpetrators.

It is important to note that these transparency models largely focus on public forum social media platforms and do not directly impact private messaging social media platforms. By nature, private messaging platforms cannot easily be subjected to such transparency models without breaching privacy laws. However, as much of the disinformation spread through private messaging platforms originates from public forum platforms – and as many users of social media operate across both – transparency models on public forum social media platforms would likely have a rollover effect into private messaging platforms. Indeed, with successful transparency models, the same network effect that makes fake news so pervasive through private messaging platforms could eventually work in the opposite direction, spreading information from posts verified as backed by credible sources and combatting intentional disinformation.

ii. Transparency about Information Sources

Overview

An important transparency model would be transparency about information sources, mainly targeting intentional disinformation spread through factual claims on public forum social media platforms. Without a culture of substantiating them with credible sources, such factual claims provide one of the major platforms for intentional disinformation to spread through social media. This is because, as referenced in the prior sections on 'enhanced methods of fabricating credibility' and 'anonymity of perpetrators', modern intentional disinformation is particularly problematic due to an increased ability to fabricate legitimacy in claims and articles, as well as a decreased risk in producing and sharing disinformation since modern social media tends to allow perpetrators to easily avoid being traced back to as a source of such claims.

As encouraging transparency around factual claims through providing information sources helps solve both of these issues, this is an important transparency model to implement to combat the spread of intentional disinformation over social media. This is because a model that encourages social media users to provide reputable information sources for their posts containing factual

claims, paired with a public visual indication on their posts verifying whether this has been achieved, significantly dismantles fake information's ability to fabricate credibility. The public visual indications also hold people more accountable for the credibility of the information they share, to ensure their information is from reputable sources – thus encouraging individuals to pre-emptively look deeper into the credibility of claims before sharing them. Additionally, by encouraging the provision of sources, this continuously links posts sharing intentional disinformation back to the original source, weakening the anonymity of perpetrators generating intentional disinformation (which traditionally relies on a chain of unreferenced sharing through social media, until the original perpetrators are long forgotten), and potentially disincentivising the production of intentional disinformation for social media. Overall, such a transparency model would help identify baseless claims on social media while fostering credibility for reputable claims, protecting social media users and fostering a greater online culture for requiring credible proof and sources before believing in anything online. In turn, fostering such a culture could also help diminish the issues discussed in the 'vulnerability of the masses' section, which leave much of the population susceptible to intentional disinformation on social media.

Though achieving similar goals, there are two main variants of this transparency model that could be implemented: the positive reinforcement variant, and the negative reinforcement variant.

Negative reinforcement variant

The negative reinforcement variant of the information source transparency model would seek to penalise posts on public forum social media, which make factual claims but do not provide links to reputable sources to substantiate such claims. This is the most direct and straightforward variant between the two, as it would directly place a negative visual marker on such posts. Such visual markers would distinguish offending posts as less credible due to their failure to provide reputable sources.

This would directly penalise factual posts without reputable sources provided, thus directly incentivising effective sourcing and evaluation of sources before making factual claims on social media and providing a clear indication to other social media users of questionable factual claims.

However, this could present several issues in implementation. First and foremost, the actual ability to screen for offending posts, which match the specific criteria for making factual claims and lacking substantiation with reputable sources, would be challenging to implement. This would necessitate sifting through the titanic quantities of posts on public forum social media platforms, which would be extremely expensive; not to mention impossible by hand and highly

prone to error if automated. This is because a large amount of public forum social media posts are not created with the intent to make factual claims, and the ability to distinguish these can require contextual and linguistic understanding which an algorithm might be hard-pressed to develop. Additionally, the ability to distinguish which factual claims are substantiated with reputable sources could similarly be challenging to manage. This vulnerability to error could then pose additional issues, as negative visual indications on posts immediately denote the post and poster as questionable in credibility and reputability. If erroneously attributed, this could generate backlash ranging from irritation at the platform to outright accusations of libel and slander.

This could be made more manageable by only screening posts related to key topics commonly fraught with disinformation. It would be much easier to implement an algorithm highlighting specific posts for screening based on certain keywords related to current disinformation trends. This would have to draw from a database of keywords that are kept up to date with the latest disinformation trends, which could be organised through cooperation with dedicated fact-checking sites like Snopes.com or the universal fact-checking service proposed in the next section. However, following this, the same issues would still arise in separating factual claims from regular posts and identifying substantiation with reputable sources.

Positive reinforcement variant

The positive reinforcement variant of the information source transparency model would seek to reward posts on public forum social media which provide links to reputable sources for their factual claims. This is probably the least invasive and easiest to implement variant between the two, as it simply involves implementing a positive visual marker on posts which make factual claims substantiated with reputable sources. Such visual markers would distinguish posts with greater credibility and reputability; and by contrast, would indicate unmarked posts as more questionable due to their lack of reputable sources to substantiate their claims.

This variant would be far easier and less problematic to implement, as it would not necessarily need to implement a screening process. This is because, while the negative reinforcement variant needs to penalise all posts that do not meet the information source requirements (and thus needs to proactively screen for such offending posts), the positive reinforcement variant would simply seek to provide an additional mark of credibility to posts with factual claims that meet the information source requirements. As most individuals posting factual claims seek to convince other users of their claims, most such individuals would seek to gain the additional credibility from the positive visual marker (and distance themselves from the potential questionability that

the lack of such a marker implies about their factual claim). Therefore, a screening process could be avoided in favour of a self-nomination process, wherein individuals posting factual claims can request an evaluation of their post for the information source requirements in order to obtain a positive visual marker (if successful). This would allow the circumvention of the screening and erroneous selection issues found in the negative reinforcement method, as this could rely on self-motivated individuals to submit their factual claims for evaluation. Following the self-nomination of a post for evaluation, the post could be checked for citing reputable sources through automation or manual review, as the self-nomination function would drastically narrow down the number and type of post that is submitted for review – thereby allowing the actual evaluation to either be more easily automated or conducted manually.

However, comparatively, since this variant does not involve proactive screening, this variant would likely be less comprehensive in marking all of the factual claims that link to reputable sources (since not all of the posts which meet the criteria will necessarily submit themselves for evaluation). It is likely though, that a significant enough number will due to the motivations of individuals in posting factual claims in the first place, and the added questionability that their posts would have by contrast without the positive visual marker from a successful evaluation. As such, this variant would, at most, risk having some factual claims with reputable sources not explicitly classified as more credible (through inaction on the part of the poster), but would still ensure that all factual claims without reputable sources would remain in the more questionable category. This could be viewed as a weaker solution to the issue of enhanced methods of fabricating credibility since it only marks factual claims without reputable sources as potentially more questionable (since a factual claim without a positive visual marker could be either a factual claim citing reputable sources that did not elect to be evaluated, or a factual claim lacking reputable sources). However, it is this softer classification that would allow the positive reinforcement method to avoid complications like accusations of libel and slander from posters, while still providing significant penalty to factual claims lacking reputable sources from the contrasted questionability that lacking a positive visual marker connotes.

This could even be taken a step further, rewarding individuals who consistently post factual claims verified to cite reputable sources, with a similar positive visual marker on the individual's profile itself. Similar to Twitter and Instagram's verified status for the official accounts of public figures, and Facebook's 'Top Fan' mechanic for regular contributors to groups and pages, this would help identify an individual on a social media platform that consistently provides reputable

sourced posts with factual claims. This could be easily implemented for individuals that post more than a certain number of posts with factual claims over a rolling time period, which are all consecutively verified to have cited reputable sources – thus adding additional credibility to individuals who consistently post claims citing reputable sources. This would reinforce a community culture of consistently citing reputable sources in factual claims, providing greater protection for other users from intentional disinformation, and further weakening the fabricated credibility of intentional disinformation (as factual claims citing fake news and disreputable sources would not have a positive visual marker on the post or poster).

Implementation, expansions, and concerns

After evaluating both variants, this paper supports the implementation of the positive reinforcement variant of the information source transparency model for public forum social media platforms. This variant presents as more feasible in terms of cost, sustainability, effective implementation, and error avoidance. Additionally, providing sufficient deterrence and penalty for factual claims without reputable sources by simply publicly and visually rewarding factual claims citing reputable sources, this method weakens intentional disinformation spread across public forum social media platforms while simultaneously avoiding accusations of libel, slander, and suppressing freedom of speech.

Furthermore, though the negative reinforcement variant could be challenging to implement, the screening process touched upon in that variant, using keywords from a database related to current disinformation trends, could be used in some capacity as an expansion to supplement the positive reinforcement variant. Specifically, this could be used to provide a non-invasive pop-up link to objective information on the disinformation topic, on social media posts related to the topic. This is similar to some of the strategies implemented by YouTube and Google to deal with topics that attract significant rumours or conspiracy theories. For instance, in March 2019, YouTube trialled out a fact-check surfacing information panel in the Indian version of their webpage. This added an additional pop-up containing fact-checks from verified organisations whenever a user searched for a topic prone to misinformation.[143] Google's approach is similar, with an add-on comment box or 'Knowledge Panel' pointing out relevant fact-checker information during a search.[144]

[143] Daniel Funke, 'Youtube Is Now Surfacing Fact Checks In Search. Here's How It Works.' (Poynter, 2019) <https://www.poynter.org/fact-checking/2019/youtube-is-now-surfacing-fact-checks-in-search-heres-how-it-works/> accessed 25 August 2020.
[144] Ibid.

This adds additional incentive for the keyword database to be organised in cooperation with fact-checking sites like Snopes.com and the universal fact-checking service proposed in the next section; since not only are they the organisations best suited to maintain a keyword database of current disinformation trends, but they would also have readily accessible objective information on such disinformation trends that could be linked to. The objectivity of such fact-checking services would also be an asset, as previous trials of such pop-up fact-checking services have been vulnerable to accusations of bias by linking to organisations with arguable political biases. For instance, in January 2019 Google's tool was accused of supporting liberal biases by appending fact-checks from the Washington Post to the republican-supporting Daily Caller, which was an influential contributor to Google's decision to suspend the tool pending further improvement.[145]

This would effectively expand upon and supplement the positive reinforcement variant while remaining relatively non-invasive and avoiding censorship. This is because this method more directly combats intentional disinformation, yet merely provides the same consistent objective information on all posts related to a disinformation trend as an optional read via a pop-up link which makes no direct judgment on the post itself. This would likely support the credibility of factual claims with reputable sources, while weakening the credibility of factual claims spreading intentional disinformation, and while also remaining immune to accusations of libel, slander, or censorship; thereby further weakening the enhanced methods of fabricating credibility of modern intentional disinformation on social media.

However, one large issue remains (which would exist in implementing either variant), which is how to determine the basis of 'reputable' sources. It is simple enough to evaluate a factual claim for the presence of cited sources vs. no citations; however, the evaluation of reputable vs disreputable sources is a potentially contentious process that could be prone to accusations of bias. The most efficient solution to this would be to regulate this with an existing organisation, external to the social media platforms, that is unassailably objective and can evaluate the reputability of news sources (or of the factual information in the linked sources, but this is a much more intensive process). This therefore links closely to fact-checking so, while this could be regulated by the social media companies themselves, or by standalone organisations (i.e. NGOs or government-based ones), these could face significant issues. These include the low cost-effectiveness of such organisations needing to build new infrastructure for such regulation, and

[145] Daniel Funke, 'Blame Bugs, Not Partisanship, For Google Wrongly Appending A Fact Check To The Daily Caller' (Poynter, 2018) <https://www.poynter.org/fact-checking/2018/blame-bugs-not-partisanship-for-google-wrongly-appending-a-fact-check-to-the-daily-caller/> accessed 25 August 2020.

the potential accusations of ulterior motives of such organisations weakening the effectiveness of the transparency model as a whole (particularly if organisations related to political or regional interests in any way are used to regulate these criteria across global social media platforms). As a result, this paper supports using fact-checking organisations to regulate the determination of 'reputable' sources. In particular, given similar requirements of unassailable objectivity and fact-checking, this paper supports the regulation of these criteria through the same organisation overseeing the universal fact-checking service proposed later in this paper. As detailed further in Section III.III., providing both potential solutions through the same organisation would help in cost-efficiency and streamlining similar resources and expertise under the same organisation, in a structure specifically built to enable as much objectivity as possible. Further details on how such an organisation would interact with the public forum social media platforms to regulate this are expanded on in Section III.III..

It is important to note though, that this transparency model does not directly target intentional disinformation spread in private messaging social media platforms, which remains a large contributor of intentional disinformation. However, as many users of private messaging social media platforms also frequently use, post, and share information from public forum social media platforms, this transparency model would still have an indirect spill over effect weakening the credibility of intentional disinformation spread through private messaging social media platforms and encouraging a greater culture of requiring reputable sources and citations before believing anything spread through social media.

iii. Transparency about Broad Trends

The Ad Database Model

Although social networks need to be transparent with individual users about the context needed to make decisions about specific pieces of content, social networks should also be transparent with society about broad trends relating to disinformation. This means providing researchers with the data needed to answer questions relating to the scale, nature and effects of disinformation campaigns. This would better allow policymakers to assess existing measures to tackle disinformation and retroactively design new ones. In addition, an increased understanding of the current disinformation situation would make it easier to hold producers of disinformation accountable, creating a stronger deterrent against the practice.

Although this data would be most useful if it included content from users as well as from advertisers, it may be prudent to restrict the scope to advertising only. Making the content of

individual users easily accessible for research could be considered an invasion of their privacy. Given that advertising's role is to diffuse a message, it is unclear that advertisers could legitimately have the same concern. Moreover, given that advertising, unlike much of the content that individual users might post, is intended to influence the public, it seems appropriate to subject advertisers to a higher degree of public scrutiny. Indeed, for political advertising at least, this is a norm in many forms of traditional media. In the UK, for instance, printed election material has to include an imprint detailing who is responsible for it.[146]

Transparency about broad trends could therefore take the form of databases of ads that have appeared on a particular social network. In order for researchers to gain a fuller understanding of the disinformation landscape, each ad should be accompanied by additional data associated with it. The Mozilla Foundation suggests, for instance, that the precise targeting criteria for the ad, the cost of the ad, and precise details about who saw and engaged with the ad should be available. A machine-readable version of graphics should also be provided to allow for automated analysis of the large amount of data.[147]

The advantages of ad databases

Ad databases would likely help reduce two of the previously identified issues relating to disinformation. Moreover, they could be useful for tackling issues in advertising other than disinformation.

As noted in the 'Vulnerability of the Masses' section, human psychology appears to contribute in several ways to susceptibility to believing disinformation. Analysis of ads known to be disinformation could reveal which of these biases are mainly exploited by spreaders of disinformation. The designers of digital literacy programmes could use this information to design interventions that aim to address the most exploited biases and make their programmes more effective at combatting disinformation. If it emerges that social proof is exploited far more than the other phenomena, for instance, teaching people about this way in which they are vulnerable should be a key focus of these programmes. Moreover, analysis of which demographics are

[146] Helen Warrell, 'Efforts to Prevent Foreign Manipulation of UK Election Flounder' *Financial Times* (London, 10 December 2019) <https://www.ft.com/content/19daf806-1a98-11ea-97df-cc63de1d73f4> accessed 23 January 2020.

[147] 'Facebook and Google: This Is What an Effective Ad Archive API Looks Like' (*The Mozilla Blog*, 27 March 2019) <https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like> accessed 15 October 2019.

particularly targeted by false claims could help policymakers make informed decisions about which groups should be prioritised to receive digital literacy training.

Ad databases would further reduce the vulnerability of the masses by allowing fact-checkers to understand which groups were targeted with particular false claims in advertising. They may also be able to use data about how shared or otherwise engaged with the ad as a proxy for which groups were particularly likely to be convinced by an ad's false claim. Fact-checkers could therefore target their corrective messaging at this group specifically. Furthermore, data about how many people saw or engaged with particular pieces of disinformation could allow fact-checkers to 'triage' disinformation campaigns, allowing them to focus their finite resources on tackling the false claims which are the most influential.

One potential risk of this functionality is that it could be used by intentional disinformation campaigns to identify particularly vulnerable demographics to target. The impact of this would likely be comparatively small, however. Since social networks already provide advertisers with campaign analytics, propagators of intentional disinformation can already get a sense for which demographics they are successfully influencing. Therefore, only new propagators, who do not already have access to these analytics, would significantly benefit from these databases for targeting disinformation. Even this risk could be reduced by implementing a process to verify if a user of a database is an academic or journalist, and therefore more likely to have a legitimate interest. This verification process could plausibly be used to prevent accountability from legitimate researchers, however, so may create more risk than it mitigates.

An additional way in which ad databases could help tackle disinformation is through increasing the potential costs of disinformation campaigns. As identified in the 'Anonymity of Perpetrators' section, there is currently relatively little risk associated with spreading disinformation. This is because it is often difficult to link a false claim to a particular actor, meaning that there is only a low chance of the perpetrator suffering legal or reputational harm. Making it easier to research the disinformation landscape would make being caught more likely and provide a more substantial disincentive against spreading disinformation. Politicians' electoral campaigns may be deterred from smearing their opponents in social media ads, for example, by the risk that journalists could use the ad databases to find and document this behaviour, making the politician appear less trustworthy to potential voters. Similarly, lobby groups may be deterred from astroturfing by the threat of having regulators crack down on them, should they be caught.

Data about who was targeted with the disinformation may also may the motives of the perpetrator clearer. It would be possible to see whether disinformation aiming to suppress voter turnout, such as by publicising a false election date, is targeted at a particular minority, for instance. This would imply a desire to reduce specifically this group's electoral power. Knowing this motive could make attribution of the disinformation campaign easier, in cases where the advertiser has obscured their identity.

Transparency in trends relating to advertising may additionally bring benefits other than tackling disinformation. A complete analysis of this potential impact is outside the scope of this paper. It is conceivable, however, that this policy could, for instance, facilitate research into which companies disseminate ads that make use of gendered stereotypes. The backlash from this could discourage businesses from using this form of advertising in future. Similarly, companies may be deterred by the fear of backlash from engaging in what might be seen as exploitative advertising, such as the promotion of junk food to children.

Facebook's ad database

With its 'Ad Library', Facebook, the biggest social media company, has already taken steps towards the ad database model. The firm's offering is currently far too limited to serve the role envisioned here, however. The Ad Library allows anyone to search by keyword through all ads currently running on Facebook or Instagram. In addition, for political ads, there is an archive going back seven years. Information is given about the rough number of people who saw the ad, the rough amount spent on it, and the percentage of people by age and gender who saw it. It is also possible to see individual advertisers' total spending and number of ads.[148] Facebook additionally provides an API for people who have their identity confirmed by the company. With this tool, it is possible to request and receive, in bulk, specific pieces of data from the Ad Library.[149]

Facebook's Ad Library API could facilitate large-scale analysis of broad trends in advertising on Facebook or Instagram. However, several researchers have reported that the tool is "so plagued by bugs and technical constraints that it is effectively useless as a way to comprehensively track

[148] 'What Is the Facebook Ad Library and How Do I Search It?' (*Facebook Help Centre*) <https://www.facebook.com/help/259468828226154> accessed 16 October 2019.
[149] David Blood, 'How Facebook Shares Its Advertising Data' *Financial Times* (London, 1 June 2019) <https://www.ft.com/content/7e7f0564-82e9-11e9-b592-5fe435b57a3b>.

political advertising".[150] It appears to have numerous software glitches, which can cause any requests for data to fail or be answered incorrectly. It is also designed in a way which makes it difficult to access a very large amount of data, such as all the ads in a particular country. Like the non-API Ad Library, it is also missing key data, such as which demographics the advertiser wanted to see the ad, or detailed information about who saw it.[151][152][153][154]

Prior to Facebook releasing its API, several groups developed tools to independently gather data about political advertising on Facebook. ProPublica details, for instance, how volunteers can install its 'Political Ad Collector' web extension. This captures the ads seen by the volunteers, as well as the justification given by Facebook to the user for showing that particular ad. Since this justification references very precise demographics, ProPublica is able to use it to gather nuanced data about who was being targeted with what kind of advertising. The Political Ad Collector could therefore be a good alternative or complement to Facebook's own flawed tool. ProPublica notes, however, that Facebook has repeatedly made small changes to the site which prevent the web extension from working properly. At one point, for instance, the Political Ad Collector identified ads by searching for the word 'Sponsored', since this has to appear in all ads on Facebook. Facebook changed the site's code so that humans would still see 'Sponsored', but the tool would read this as "SpSonSsoSredS", so not capture the ad.[155][156] Therefore, there does not appear to yet be a particularly effective way of gauging the advertising landscape on Facebook.

Twitter's ad database

As of 22nd November 2019, political advertising is not allowed on Twitter.[157] Therefore, it may seem as if disinformation that aims to influence political outcomes can no longer be spread

[150] Matthew Rosenberg, 'Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised' *The New York Times* (New York, 25 July 2019) <https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html> accessed 17 October 2019.
[151] Blood (n 10).
[152] 'Data Collection Log – EU Ad Transparency Report' (*Mozilla Ad Transparency*) <https://adtransparency.mozilla.org/eu/log/> accessed 17 October 2019.
[153] 'Facebook's Ad Archive API Is Inadequate' (*The Mozilla Blog*, 29 April 2019) <https://blog.mozilla.org/blog/2019/04/29/facebooks-ad-archive-api-is-inadequate> accessed 17 October 2019.
[154] 'Facebook Ads Library Assessment' (*Ambassador for Digital Affairs*) <https://disinfo.quaidorsay.fr/en/facebook-ads-library-assessment> accessed 17 October 2019.
[155] Jeremy Merrill and Ariana Tobin, 'Facebook Moves to Block Ad Transparency Tools – Including Ours' (*ProPublica*, 28 January 2019) <https://www.propublica.org/article/facebook-blocks-ad-transparency-tools> accessed 18 October 2019.
[156] Jim Waterson, 'Facebook Restricts Campaigners' Ability to Check Ads for Political Transparency' *The Guardian* (27 January 2019) <https://www.theguardian.com/technology/2019/jan/27/facebook-restricts-campaigners-ability-to-check-ads-for-political-transparency> accessed 18 October 2019.
[157] Kate Conger, 'Twitter Will Ban All Political Ads, C.E.O. Jack Dorsey Says' *The New York Times* (30 October 2019) <https://www.nytimes.com/2019/10/30/technology/twitter-political-ads-ban.html> accessed 10 January 2020.

through ads on the platform. This would remove the need for a database to increase transparency around political advertising on Twitter to reduce disinformation.

However, the complexity of defining what is political means that ads which many people would consider political are likely to remain on Twitter. The company's definition is content that "references a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome". Additionally, no ads are allowed from candidates, political parties or government officials, or, in the US, some additional types of organisations.[158] Ads relating to broad political topics, such as animal rights or feminism, are therefore allowed, as long as these ads do not reference specific measures.[159]

It seems plausible, however, that promoting broad messages about political issues could influence people's decisions about specific measures linked to these issues. Therefore, even with the new restrictions, advertisers on Twitter could still use disinformation to influence political outcomes. One journalist suggested, for example, that employers could take out ads to falsely claim that a town was bankrupted by its decision to raise the minimum wage.[160] This disinformation could perhaps unduly influence people into voting against real proposals to raise the minimum wage.

Twitter has taken steps to reduce the potential impact of such instances of disinformation by reducing the access to targeting on ads related to broad political issues: advertisers cannot filter so narrowly be geography, for instance, nor target users by political keyword such as "conservative". This could be a useful step, since, as noted in the 'Vulnerability of the Masses' section, precise population targeting is likely to increase the effectiveness of disinformation campaigns. Targeting could lead to instances of disinformation receiving less criticism than they would otherwise, for instance, and allows spreaders of disinformation to spend their advertising budget more efficiently.

However, given the potential for disinformation, it still seems like an effective ad database would be useful. Regrettably, Twitter's current offering seems comparatively limited. The site's 'Ad Transparency Center' seems only to allow users to search by the advertiser, for instance.[161] This

[158] 'Political Content' (*Twitter*) <https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html> accessed 10 January 2020.
[159] Makena Kelly, 'Twitter Will Ban Candidate Ads and Limit Issue Ads' (*The Verge*, 15 November 2019) <https://www.theverge.com/2019/11/15/20966854/twitter-super-pacs-political-ads-facebook-climate-change-abortion-paid-promotion> accessed 10 January 2020.
[160] Emily Stewart, 'Twitter Is Walking into a Minefield with Its Political Ads Ban' (*Vox*, 15 November 2019) <https://www.vox.com/recode/2019/11/15/20966908/twitter-political-ad-ban-policies-issue-ads-jack-dorsey> accessed 10 January 2020.
[161] 'Ads Transparency Center' (*Twitter*) <https://ads.twitter.com/transparency> accessed 10 January 2020.

could make it difficult for researchers or journalists to find political advertisers they do not already know. Additionally, there is no easy way to download the information in the database, making it difficult for researchers to conduct quantitative studies of Twitter ads.[162] Furthermore, the database is not comprehensive: it only includes certain ad types on the platform and does not include ads that ran but were subsequently deleted by the advertiser.[163] All these factors are likely to severely limit the usefulness of the tool to actors that want to scrutinise political advertising on Twitter.

The role of regulators

Regulators would likely be helpful in pushing social networks to build ad databases that are useful to researchers. An initial role of regulators might be to establish clear frameworks, detailing what types of data should be made available to researchers and in what ways. However, as set out in prior subsections, these regulators may need to take on more of an enforcement role, should social networks not voluntarily conform to this framework.

To some extent, the European Commission has tried to establish this framework. In 2018, the Commission convened a multi-stakeholder forum on online disinformation.[164] This led to a voluntary Code of Practice, which has been signed by Facebook and Twitter, among others.[165] With point 13 of the code, signatories commit to "not to prohibit or discourage good-faith research into disinformation and political advertising on their platforms".[166] The code lacks any detail on how social networks should facilitate research, however. It does not specify what types of content, or associated metadata, should be made available, for example, nor how to provide this data to researchers. As a result, companies are not committed to specific standards, so they may find it easier to only enact half-measures.

Despite these shortcomings, the European Commission's forum provides a good starting point. A regulatory body should be set up to help set out clear guidelines for ad databases for social media companies and enforce their implementation. This should provide clear instructions on

---

[162] 'Twitter Ads Transparency Center Assessment' (*French Ambassador for Digital Affairs*) <https://disinfo.quaidorsay.fr/en/twitter-ads-transparency-center-assessment> accessed 10 January 2020.
[163] 'Ads Transparency Center FAQs' (*Twitter*) <https://business.twitter.com/en/help/ads-policies/ads-transparency-center-faqs.html> accessed 10 January 2020.
[164] Mariya Gabriel, 'Statement by Commissioner Gabriel on the Code of Practice on Online Disinformation' (*European Commission Press Release Database*, 26 September 2018) <https://europa.eu/rapid/press-release_STATEMENT-18-5914_en.htm> accessed 18 October 2019.
[165] 'Code of Practice on Disinformation' (*Digital Single Market - European Commission*, 17 June 2019) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation> accessed 18 October 2019.
[166] 'EU Code of Practice on Disinformation' 8 <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation>.

the type of content to provide (and associated metadata), the format and level of detail they should be provided in, and measures to ensure ease of access. A model ad database, or direct reviews and feedback on individual platforms' databases, could additionally be provided by regulators to help ensure that all platforms provide sufficient data at a consistently high quality.

While this responsibility could be enforced regionally through national boards or NGOs, the international nature of the social media platforms in question would benefit from a singular international regulator to enable consistency and ease of compliance. Similar to the 'Information Source' transparency model discussed earlier in this section, this paper supports the regulatory and enforcement responsibility for such an ad database model being handled by the same international organisation handling the 'Information Source' transparency model and the universal fact-checking service. As described in Section III.III., such a service would benefit from the global scale and overlap across multiple areas related to intentional disinformation, to effectively enforce the ad database model cost-effectively.

III.II. UNIVERSAL FACT-CHECKING

i. Overview

Another key issue surrounding intentional disinformation is the reduced general credibility of information that fake news has engendered. Since people are aware that disinformation exists, it is all too easy for a voter or a politician, or anyone for that matter, to discredit information that does conform to their world view by simply labelling it as "fake news". Consider its frequent use in recent times, increasing 356% from 2016 to 2017 alone[167]. Indeed, the UK government no longer uses the term "fake news" and has instead adopted the use of "misinformation" and "disinformation" in favour to avoid confusion[168].

Such a phenomenon has severe ramifications and recursively hinders attempts to dispel intentional disinformation, as attempts at debunking by traditional media outlets or government sources are all too frequently dismissed as fake news.

This paper proposes a potential solution in the form of a universal fact-checking service, which would actively monitor current trends and claims that could potentially include intentional disinformation, and publish definitive research and articles analysing such claims and their veracity.

Such a service would aim to provide a universal, consistent standard for fact-checking by disseminating accurate information widely accepted as reliable and easily accessible to anyone. In doing so, this would provide a platform wherein credibility and objectivity could be reliably defended, combatting the reduced credibility of information that other platforms are assailed by.

ii. Existing Fact-Checking Services

There have already been several attempts at services with similar aims. Below are brief discussions of a few notable examples.

Snopes.com

Arguably the first such attempt on a meaningful scale was Snopes.com, a website founded by David and Barbara Mikkelson in 1994 as a way of debunking urban legends and conspiracy theories. It has since evolved to cover a much broader range of topics, including science and

[167] Bente Kalsnes, 'Fake News' [2018] Oxford Research Encyclopedia of Communication
[168] House of Commons: Digital, Culture, Media, and Sports Committee, 'Disinformation And 'Fake News': Final Report' (House of Commons 2019)

politics, and a broader range of types of content, including images, articles, and social media posts.

Their method broadly follows the steps outlined below, but due to the variety of the content they cover, this method is fairly flexible[169]:

- Topics to verify are selected based on search frequency, and questions asked on their site, as long as the topic is appropriate for their service (for example, questions they deem inappropriate listed on their site include "What is this celebrity's sexuality?" and "Is there a god?").
- One member of their editorial staff starts with preliminary research and the first draft of a report. They begin by contacting the original publisher of the content under investigation to acquire clarification or more robust referencing.
- Supporting research is then conducted by other staff, and all the research is collated and edited into a full report. All relevant sources are published with the report.
- Readers can suggest corrections through a contact form which are then reviewed by an editor and added to an updated report if appropriate.

Note that Snopes.com is transparent about its referencing and has a process in place for reviewing its published reports.

It is financed[170] largely through advertising and all single contributions either "exceeding $10,000 or comprising more than 5% of our total annual revenue" are published on the Disclosures page of the site. Notable such contributions include $506,000 across 2017 and 2018 from Facebook in a fact-checking partnership and $590,000 in shareholder financing across 2018 and 2019.

FactCheck.org

More focused on the problem of disinformation in political processes is FactCheck.org, a non-profit organisation run by the Annenberg Public Policy Centre of the University of Pennsylvania. It works explicitly on US politics and aims to cover Democrat- and Republican-related content equally. Their process[171] outlined below is quite similar to that of Snopes.com.

[169] 'Transparency' (Snopes.com) <https://www.snopes.com/transparency/> accessed 25 August 2020.
[170] 'Disclosures' (Snopes.com) <https://www.snopes.com/disclosures/> accessed 25 August 2020.
[171] 'Our Process - Factcheck.Org' (FactCheck.org) <https://www.factcheck.org/our-process/> accessed 25 August 2020.

- The content they cover can be from talk shows (such as on Fox, CNN, NBC, etc.), from television ads, from campaign rallies, or from floor debates such as on C-SPAN, among other types.
- They begin by contacting the original publisher if a statement is considered confusing, inaccurate, or misleading. If the publisher clarifies satisfactorily, they move on. If not, they conduct independent research.
- FactCheck.org attempts as much as possible to use primary sources in their research including SEC corporate records; the Bureau of Economic Analysis; the Congressional Budget Office; the Kaiser Family Foundation for healthcare data.
- Each report is edited after writing by a line editor who ensures that the report is comprehensible and has adequate context; a copy editor who ensures proper grammar, style, and tone; a separate fact-checker who reviews every statement in the report to verify factual accuracy; and finally by the director of the Annenberg Public Policy Centre. Each of the four editors must not have been involved with the writing team.
- Every report contains links to source materials used in writing and editing.
- Corrections can be suggested by email which will then be reviewed and, if appropriate, the report will be edited with a note explaining the nature and date of the change.

Note again the transparency of referencing and the existence of a review process.

FactCheck.org was financed primarily by the Annenberg Foundation but began accepting donations in 2010[172]. They do not accept funds from unions, advocacy groups, or corporations. The only exception to this policy is Facebook which started investing in 2017 to tackle the spread of disinformation on the site.

Facebook

Facebook uses third-party fact-checkers to monitor content on its site, including posts, links, photos, and videos[173][174]. These can be flagged actively by users, but Facebook also uses AI to monitor user patterns, such as comments expressing disbelief at certain content, and conducts investigations accordingly. If content is found to be false or even partly false, there are several

[172] 'Our Funding - Factcheck.Org' (FactCheck.org) <https://www.factcheck.org/our-funding/> accessed 25 August 2020.
[173] 'How Facebook's Fact-Checking Program Works' (Facebook, 2020) <https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works> accessed 25 August 2020.
[174] 'Fact Checking On Facebook' (Facebook) <https://www.facebook.com/business/help/182222309230722> accessed 25 August 2020.

actions that Facebook may take against it. This may include placing it lower on people's News Feed (or removing it from the Explore page and any hashtag pages if the content is on Instagram, which Facebook owns) and placing misinformation labels on it, so users are aware of its factual inaccuracy. People and pages that repeatedly offend risk seeing reduced distribution and losing their ability to monetise their content with advertisements in the future.

Facebook uses different fact-checking organisations to monitor content coming out of different countries. For example: in the US, there is FactCheck.org; in Spain, there is Newtral; France has 20 Minutes and Les Observateurs de France 24; AFP has branches in many countries including Canada, Spain, and Sri Lanka, among many others mostly in Africa, Asia, and South America; and the UK has Full Fact. (Most countries have several organisations each; above are just some notable examples.)

There appears to be less transparency about Facebook's process, but the individual organisations provide more detail on their websites.

iii. Universal Fact-Checking Model

While the services described above are noteworthy attempts at creating an independent fact-checking organisation to deal with the spread of disinformation, they are all unfortunately fairly small in scale. This creates two problems. (1) The scope of their activity is seriously limited as they cannot be expected to be able to cope with the volume of relevant content that they would need to investigate and (2) they therefore fail to create a consistent standard for fact-checking that can be referred to internationally.

Universal one needs to be large-scale and governed by a multi-partite system to avoid any accusations of vested interests and/or with a robust process that is unassailably objective (but requires more funding if not large-scale and multi-partite)

Fact-checking process

The initial process may go as follows:

- Reports must be published as quickly as possible (while maintaining the required high standard of research) as any delay may allow disinformation to become better embedded in people's minds. Therefore, it is advisable to use AI (discussed further below) to find content worth reporting on and to find relevant sources.
- Begin by contacting the original publisher of the content to request clarification and sourcing. If their response is satisfactory, move on, although it may be worth publishing

a report detailing this response for the benefit of those that otherwise would not see the clarification. (Indeed, it may be worth publishing a complimentary report on content with robust sources and sound reasoning to improve people's understanding of a topic and to provide them with an example of content worth engaging with.) If the response is not satisfactory, conduct further research and begin gathering evidence to put in a report.

- Given the importance of urgency, the writing team may need to proceed with their research without a response if the publisher does not clarify quickly. The report can be amended later should a satisfactory response come.
- Primary sources should be used where possible. The service may even try to build robust relationships or partnerships with outside experts such as research institutes, universities, and think tanks that can be consulted when writing reports.
- The research should be collated into an initial report by an editor working closely with the writers.
- After the report is nominally finished, it should be edited for: context and comprehensibility; style, tone, grammar, and formatting; and for factual accuracy by an editor who reviews all the sources used. Finally, it could be edited by a committee representative or other senior member of the service to ensure the report is in line with the service's values. Each category of editing should be done by a different individual editor not involved with the original writing team.
- Naturally, full transparency of sources and referencing would be essential given the service's aim to inspire faith in factual information. However, it may not be prudent to publish sources being used in ongoing, unfinished reports as this may encourage external interference by parties with interest in influencing or undermining the writing process.

Review and appeal process

To ensure fairness, any report should be open to review if a member of the public, an organisation, or a public office should disagree with it with some grounds for doing so. The review process may go as follows:

- All material should be subject to appeal. This could come from a claim made through the service's website if coming from a private entity (such as a person or a firm) or through some similar official channel if coming from a public office.
- The review team would need to be sufficiently separated from the original writing team. This is because there would be too much opportunity for coercion or bribery if the

original writing team is used. For this reason, an entirely separate team of researchers drawn from the general workforce of the service should be used. However, there should not be an entirely separate department specifically for reviews of already published reports as this would be too vulnerable to external influence. Instead, all researchers in the service should work on both original reports and reviews.

- The review team should conduct further research on the topic and amend the original report with full detail on the nature and date of any changes published with the updated version.
- Of course, knowledge changes over time as new information is brought to light[175] so reports should be amended accordingly and/or new reports on the topic could be published in which references are made to older reports highlighting the latest information.

The processes outlined here are inspired by those of existing fact-checking services. They are merely suggestions and should be adjusted where appropriate to make them more contextually effective, robust, and efficient.

Use of technology

It was mentioned above that technology will be instrumental in allowing the service to cope with the ever-increasing volume of relevant content (including but not limited to television interviews and debates, conventional media articles, social media posts, images and videos, and official statements) as entirely human teams simply would not have the capability to do so. Artificial intelligence will be needed to get through it all.

Full Fact (mentioned above - works with Facebook in the UK) uses AI for two of its tools. "Trends" logs repetitions of claims to identify what claims are being made and by whom. "Live" uses text from TV subtitles to cross-reference claims with reliable sources to determine whether statements made on TV are accurate. The use of AI here dramatically reduces the time requirement from Full Fact's human team.

Technology like that used by Full Fact's "Trends" tool could be useful in identifying unreliable sources (e.g. media outlets, politicians, political commentators) and technology used in their

[175] Consider the COVID-19 pandemic. Advice to the public changed (and continues to do so) as new information on the severity of the disease and infectivity of the virus was discovered.

"Live" tool could be extended to make it easier for the general public to know what claims are true or false without the need for extensive personal research.

Despite how promising this technology is, the scale of its implementation is far less than that which would be required for a service as described in this section. Therefore, significant investment would undoubtedly be necessary to achieve the desired expansion.

Concerns about freedom of speech

It would be reasonable at first to raise concerns about freedom of speech. In discussing censorship as a policy to combat disinformation, this would be a severe drawback to consider. This is because censorship aims to construct some "true narrative" of a sequence of events against which to compare published content and then any content which contradicts this narrative is taken down.

However, this fact-checking service would not involve any sort of censorship so would not have this aim. Instead, the goal would be to provide a consistent, international standard for fact-checking as described at the beginning of this section.

The service would focus on increasing faith in factual information by flagging content which is misleading or false and providing thorough reports explaining the decision to do so. In this way, it would not infringe on any freedom of speech regulation.

Jurisdiction and responsibility

A key question surrounding any potential universal fact-checking service would naturally be what organisation would hold jurisdiction and responsibility for its implementation.

Indeed, a common and easy way to implement this would be through individual organisations, operating at a regional or national level. However, this could easily run into issues of scale, funding, and regional jurisdiction; for instance, as most online disinformation is international, such organisations would naturally have to regularly try and distinguish whether or not an issue falls into their coverage zone, and if they would even be able to access the relevant information to verify truths in other countries. Individual international organisations could also potentially host such a service, but individual private or national organisations would still be vulnerable to potential accusations of single party bias or vested interests, which would severely hinder efforts at establishing the credibility and objectivity needed for an effective universal fact-checking service.

As a result, this paper supports implementation at a higher level to combat accusations of bias more effectively. In particular, implementation within an international organisation, with a credibly structured rotating multi-party system, would be suggested as, without this, it becomes significantly more challenging to overcome the issue of reduced credibility of information. Indeed, debunking is less successful nowadays, primarily because people doubt the truthfulness of the debunking sources themselves. Even with robust internal processes ensuring credibility and objectivity, it is still necessary to go higher and have a multi-partite system with equal input and vetting across many different parties, to dispel that potential lack of credibility. Creating an unassailably objective debunking source that cannot be affected by reduced credibility of information would allow for the universal fact-checking service to effectively combat intentional disinformation.

It would additionally make sense to merge this organisation with the organisation overseeing social media transparency model regulation. This is because the overlap in both solutions' relation to intentional disinformation, and requirements for multi-partite objectivity and vetting, makes it cost-effective and efficient to centralise both responsibilities together; as described in Section III.III..

III.III. A UNIFIED INTERNATIONAL ANTI-DISINFORMATION AGENCY

i. Overview

Though it is possible to conduct the transparency and universal fact-checking models as standalone, regional services, it is difficult to enact significant impacts on an issue as globally pervasive as fake news, without comparable scale and global coverage. Indeed, despite the impressive number of smaller projects of these types (particularly for fact-checking), they are primarily limited in their effectiveness and credibility by the small-scale nature of their work. More work is needed to create an internationally consistent service; particularly for universal fact-checking. The service relies upon unassailable objectivity and credibility to be effective and immune to accusations of bias – which is only possible with a service produced by international cooperation at the highest levels.

However, as such services increase in scale, so too do issues with efficiency and cost. Given that the transparency and universal fact-checking models have a high degree of overlap in coverage and resources employed, this paper proposes that both should be provided under the umbrella of a singular organisation, to maximise efficiency and reduce wastage.

Specifically, this paper proposes the foundation of the UN Debunking and OnLine Fake Information Neutralisation (DOLFIN) Agency. The service would be internationally run as a UN agency to minimise the effect of national biases; primarily publicly funded to achieve the required scale; technologically equipped to cope with the vast volumes of relevant content, and have consistent and transparent processes with frequent committee rotation to minimise the risk of any biases or external influences from special interest groups.

ii. Distribution of Responsibilities

As such an organisation would naturally work in close concert with several private organisations and national interest, it is crucial to clearly lay out the distribution of responsibilities and jurisdictions of such a centralised regulatory agency. The two main functions the organisation would be responsible for would be the social media transparency models, and universal fact-checking described previously in the paper.

Transparency Models

The transparency models are the functions of DOLFIN, which could present the most significant complications in distributing responsibilities. This is because this function is very closely related to, and dependent on, each social media platform's inherent digital infrastructure and

management. As a result, in overseeing these transparency models, DOLFIN naturally runs a thin line between enforcement and direct implementation; the latter of which could give rise to concerns of infringement on private company operations and data, and resource-efficiency issues and limitations for DOLFIN.

However, as the 'Information Sources' transparency model is a new model for all of these platforms that requires a certain standard of diligence in implementation to be effective, DOLFIN would be prudent to take on a direct role in reviewing submitted posts – at least initially. To minimise inefficiency, the social media platforms could accumulate the posts submitted for review into a database, then periodically send these over to DOLFIN for review. Following DOLFIN's review of these for referencing credible sources, they would send the results back for the social media platforms to add visual markers to the posts (or not) as appropriate. This allows control of each platform's digital processes and infrastructure to remain solely within their own hands, while only the review process is outsourced to DOLFIN – thus ensuring that the review processes are carried out correctly. DOLFIN would then be able to perform external, periodic reviews of each social media platforms' posts (using their own database of the posts they receive for review) to ensure that the platforms are actually putting positive visual markers on the ones DOLFIN has green-lit, and none others.

Eventually, as time passes, a clear precedent would be set by DOLFIN in managing this transparency model – augmenting the clear guidelines that DOLFIN would follow in implementing this model, that would be readily accessible to all. As such, the responsibility for implementing the review process could be slowly decentralised, pending evaluations, for the largest social media platforms. Allowing them to implement the entire review process in-house (though still sending the databases of submitted posts to DOLFIN), this would allow social media platforms to improve response times and the efficiency of the entire process. DOLFIN could then take an overseer/regulatory role for those platforms – performing periodic reviews and spot-checks instead of performing the entire review process themselves.

The informative pop-up aspect of the 'Information Sources' transparency model (wherein posts related to or referencing topics which are trending for disinformation have a non-invasive pop-up appended to them, which link to further information on the topic), would also require distributing responsibility between the social media companies and DOLFIN. Specifically, DOLFIN could provide all social media companies with a regularly updated database of keywords and topics related to the most common/pressing current disinformation trends; as well

as the actual links to further information on such topics through their universal fact-checking processes. Meanwhile, the social media companies would then be responsible for actually implementing the screening process (using the database of keywords and topics) and appending the pop-up links to such posts through their systems; thus allowing them to best implement these processes in a way that contextually fits their platforms. Similar to the 'Information Sources' model's review process, DOLFIN would then take on a supervisory role, performing periodic checks and assessments for each platform, ensuring that this system is well-implemented; and issuing advice, warnings, and reports if not.

The 'Broad Trends' transparency model (or 'Ad-Database' model), this is a far less bespoke process, as it merely involves the collation of information each platform already receives for its paid advertisements – with the bulk of the work involved in organising this data into an intuitive and accessible format for analysis. As a result, this requires far less direct involvement from DOLFIN, as it is not a process which requires much direct precedent or enforcement. As such, DOLFIN could simply provide the parameters of the databases that should be provided, and perform periodic reviews on each platform to enforce this model.

Universal Fact-Checking

Compared to the transparency models, the universal fact-checking function of DOLFIN is far less complicated to distribute responsibilities for, as it does not impact direct functions or processes of social media companies. As a result, the universal fact-checking process would be managed and implemented entirely by DOLFIN, and would likely form the bulk of their operations.

iii. Structure

International cooperation in the management of a conjoined transparency model and fact-checking service of the nature and scope described here is of paramount importance. Otherwise, the potential for national bias would be too significant – particularly in the case of the universal fact-checking service. For example, consider the validity of reports on the recent trade war between the US and China published by such a service based in either country. Naturally, each service might face incentives to unduly attack content coming from the other country while promoting content from its own country – meaning that some national bias would most likely affect the service in some way. Cultural differences may additionally contribute to this, and this might be more difficult to combat as there may not be any factual disagreement, but merely several different interpretations of the same news story. For example, consider the possibility of

disagreement between interpretations coming from liberal and conservative countries. In this way, cultural bias may affect the service without an obvious solution.

As such, this paper proposes that DOLFIN should be domiciled as a UN agency, as such services require international cooperation at the highest and most universal level. An entirely new organisation could logically work, but this would likely be simply too impractical compared to using a pre-existing organisation such as the UN.

The service should not be run by a smaller international organisation, such as the EU for example, as the regional boundaries of a non-universally international organisation could still allow for suspicion of regional biases. To understand why this is the case, consider the perception of a fact-checking service run by the EU in the lead-up to the Brexit referendum. Any reports debunking a pro-Brexit message, even if such a message did indeed include intentional disinformation, might appear biased just by nature of the potential regional biases inherent in the organisation – thus diminishing the inherent credibility of the organisation.

Additionally, given the global nature of the platforms regulated by the transparency models, having DOLFIN domiciled within the international organisation with the most significant global representation would allow DOLFIN to effectively enforce the transparency models proposed. Without the backing of the UN and its member states, an organisation like DOLFIN would struggle to implement any real ramifications behind non-compliance with its transparency models. Under the UN, however, failure to comply with DOLFIN's regulations could allow for a unilateral legal threat from the member nations (subject to agreement before DOLFIN's formation), that should provide sufficient motivation for social media platforms to comply; as well as real potential consequences if they do not.

Given the issues tackled by DOLFIN, the agency would most likely be run by the UN Economic and Social Council (ECOSOC). Although, given the relevance of political disinformation to matters of international security and the keeping of the peace, it may be advisable to reserve some (but less than a majority) of seats on the primary committee for representatives from states on the UN Security Council (UNSC). The central committee should liaise with other organs and subsidiaries of the UN to decide on policy extending from the investigations carried out by the service. For example, the committee may advise the UNSC or General Assembly if the service determines that a particular country or organisation is particularly problematic.

As with any UN subsidiary, ideally, all committee members should be experts in fields relevant to the service including but not limited to media, disinformation, computer science, politics, sociology, and economics.

iv. Funding

An international organisation responsible for enforcing transparency models and a universal fact-checking service – including a sizeable human team, a robust structure of committee rotation, and advanced technology – would undoubtedly come at great expense. As a result, it is necessary to discuss potential funding methods for such an organisation to ensure sustainable access to appropriate resources.

An organisation providing services of this nature would need to be non-profit, as otherwise there would be incentives to focus resources for financial gain rather than for the pursuit of factual accuracy and clarification. This would open the service up to influence from special interest groups, defeating the purpose of having it at all.

Most of the funding would likely come from the international organisation managing it, i.e. the UN as proposed by this paper. Government provision (with international collaboration) appears sensible as governments do not face the same need to seek profits that private corporations face. In this case, the funding would come, by extension, from the taxpayer. Of course, this is not perfect as the UN may also have its own interests to promote but it is a more certain way of ensuring the necessary scale.

For comparison, Snopes.com is mainly funded through advertising, so this may be an avenue worth exploring. However, this may undermine the gravitas of a publicly funded international operation of such significance.

Crowdfunding campaigns may be an additional option for funding, as it could allow for financial contributions from private individuals, while still implementing checks and balances that prevent financial contributions of a size large enough to promote significant external influences. It is unlikely that many individuals would be able to contribute sufficiently significant sums to gain a considerable influence; especially if the identities of donors are published (as is done by FactCheck.org with donations of $1000 or more). It is unlikely that this will have a material effect on such a large project, but if there are convincing (and genuine) guarantees that donations will be used only to fund the service's improved operation, this could be a valuable way of inspiring further faith in its aim.

DOLFIN could also collaborate with relevant corporations (such as Facebook or Google given the role of their services in the potential spread of disinformation), but this would allow a greater opportunity for external influence and conflicts of interest, as these are the same organisations that DOLFIN would be seeking to regulate through its transparency models.

Similar to political financing, the concern with private donations being used to promote external influences is significantly more pressing in the context of corporate investment than in the context of donations from private individuals, due to the average comparable size of such investments. As a result, funding should largely be drawn from the international umbrella organisation (i.e. the UN in this case), with private donations limited as much as possible to crowdfunding/grassroots financing schemes, or with hard caps on donations from private corporations, to minimise the risk of external pressures.

v. Protection from Abuse

In addition, it would be necessary to regularly rotate committee members, chair-people, or any senior members of the service to prevent groups or individuals from gaining excessive power and influence over DOLFIN. This is particularly important given the necessity of international cooperation in the organisation, as a senior member may face pressure to be biased in favour of their own country and allies. Therefore, to maintain the unassailable credibility that DOLFIN requires for its processes to be accepted and effective, it would be necessary to put in place specific measures to prevent abuse within the organisation. Below are three possible rotation structures to address this issue:

Regular short-term rotations

Very frequent rotation (i.e. annually or every two years) of senior members would be an effective way of preventing excessive influence from a particular group from taking hold.

However, such short terms may also impede progress on more significant issues over longer periods of time. With an organisation like a national government, long terms are necessary to decide on and implement policy to carry out effective change. Here though, this is arguably not so important as the service provided is intuitively not one that requires long committee terms. This is because the role of the committee only involves ensuring and facilitating the continuation of effective fact-checking. Therefore, short terms may not be problematic unless some sort of major restructuring is required. However, even such restructuring may not even require significant input from the DOLFIN committee, as it could just be planned and carried out by a higher UN authority, e.g. ECOSOC.

Longer term rotations

A longer term (e.g. 4-5 years per term) seems the obvious solution if a short term is considered insufficient to carry out significant change effectively. The obvious drawback is that longer terms allow greater opportunity for excessive influence by special interest groups as they have longer to establish themselves with coercion or bribery.

In addition, it would be advisable to avoid having the terms of committee members of the fact-checking service running in parallel with election cycles in major economies (such as the UK, the US, large EU member states, etc.). Otherwise, there may be split attention from voters and policymakers who are unable to devote due time to each cycle.

Overlapping rotation cycles

Since there are noteworthy issues with both long and short terms, the service could instead be structured with some sort of overlapping rotation cycle (examples of this include the US midterms and the UNSC). As an example, DOLFIN's committee could rotate every four years, while the leadership specifically responsible for the universal fact-checking service could rotate every two years; resulting in a midterm rotation similar to that of the US elections. A similar example would be to have the two committees each on four-year rotations, but staggered so that there is a two-year overlap across them.

The committee rotations could also be modelled on the system of the UNSC[176] instead. Consider a specific proposed example as follows: sixteen seats on the committee, each on two-year terms; six reserved for representatives from Security Council member states, the remaining ten reserved for non-members; rotate three UNSC committee members and five non-UNSC committee members each year. Some logistical planning would be required to maintain the balance of seats for representatives from UNSC and non-UNSC committee members in line with the rotation of the actual UNSC members, but this should be straightforward. It may be considered problematic for the rotations to run in parallel due to potential split attention for those involved in both the UNSC rotation and the fact-checking service's rotation. However, since the service would probably be a subsidiary of the ECOSOC and not directly related to the Security Council, this should not be of great concern.

[176] The UNSC's system works as follows: there are five permanent members (US, UK, China, Russia, France) and ten non-permanent members. The ten non-permanent members each serve two-year terms with five rotated out each year. Geographical regions are guaranteed a certain number of seats: three to Africa; three to Asia-Pacific; two to Eastern Europe; three to Latin America and the Caribbean; five to Western Europe and others.

Rotation structures of this kind would go some way to addressing the problems associated with the long and short terms described above, as the overlapping terms are long enough to allow for significant change if necessary, but the midterm rotations would provide an opportunity to curb detrimental influence. As a result, this paper supports an overlapping rotation cycle as an ideal structure for DOLFIN to follow to minimise potential abuse. However, again, it would be advisable to avoid having these terms running in parallel with those of election cycles in major countries especially since it is likely that much of the service's work will focus on content coming out of these countries.

Regardless of what type of rotation system is employed, note the following suggested details and clarifications. In order to avoid bribery, coercion, or the accumulation of excessive power by any country, no UN member state should be allowed to hold a seat for more than a certain number of terms consecutively. For the same reason, no individual person should be allowed to hold a seat for more than a certain number of terms in their lifetime and certainly not consecutively. The reason for the slightly different policies for countries and individuals is that people are naturally more vulnerable to coercion or bribery than a country particularly over long periods of time and to deny countries the possibility of holding a seat more than twice is very impractical.

It is important to note that specific details of optimal periods and restriction of this form would need to follow additional research and negotiation within the UN to be as contextually appropriate and palatable to all UN member states as possible. However, the points discussed thus far in this section should provide an ideal framework from which DOLFIN or a comparable organisation could be founded.

III.IV. FOSTERING WIDESPREAD DIGITAL LITERACY

i. Overview

Contrasting with the prior solutions discussed in this paper, fostering widespread digital literacy provides a longer-term and more indirect solution to intentional disinformation. While transparency models, the universal fact-checking service, and DOLFIN, all provide reactive solutions to the effects of intentional disinformation, digital literacy provides a more preventative solution tackling the root causes of fake news.

In particular, digital literacy combats the 'vulnerability of the masses' outlined in the issues section of this paper. Since the public may not be aware of the potential inaccuracy of the content they encounter (which could lead people to make decisions, such as voting choices in elections, against their interests), digital literacy could provide an effective solution to this by educating people to understand these issues better and make appropriate choices through their own agency.

The solution to this problem proposed in this section is media, information, and digital literacy training (henceforth, MIDL training)[177].

The goal of such a policy would be to equip all members of society with the skills necessary to understand how and why nefarious influences might work when spreading disinformation. This would be a pre-emptive policy, in that it is not used to combat any particular piece of misinformative content. Instead, the idea would be that people would be able to apply these skills to critically analyse, and avoid being unduly influenced by, misinformative content as they encounter it. Ideally, all members of the population would develop strong critical thinking skills and knowledge of the issue of disinformation, which would require formal education to be totally effective. Naturally, the most meaningful benefits of MIDL training, as it will be described in this section, would only truly be reaped in the long term, as a generation of appropriately educated students reaches adulthood.

There have been few significant attempts at implementing MIDL training in school education as proposed in this section; however, where it has been implemented, there seems to have been noticeable success. In summary, the proposal outlined in this section is to introduce content to school curricula worldwide to develop students' critical thinking, digital competence, and responsibility when encountering information, particularly online. This is especially important

[177] This acronym is used in this paper to encompass all relevant literacy training. Since they often have different names, "MIDL" seeks to cover all those names.

given the increasingly important place in society of such technologies. The curriculum should be built into existing subjects to allow students to learn the applications of these powerful resources in traditional learning contexts and it should be regulated and frequently updated to ensure consistent and effective results. Ideally, there should also be grants made available to as many students as is practical, to ensure that financial backgrounds will not interfere with their learning. In the short term, condensed curricula could also be distributed to adults who are already out of school to provide them with similar skill sets with incentives given to those participating to ensure adequate engagement.

ii. Notable Efforts to Foster Digital Literacy

Some countries have already begun implementing MIDL in their school curricula. Below are outlines of this type of policy in Sweden and Singapore.

Sweden

The Swedish government tasked the Swedish Media Council with reducing the country's vulnerability to cyber threats. The Education, Audiovisual and Culture Executive Agency (EACEA) of the EU summarised Sweden's policy with particular emphasis on the changes to school education[178].

The key parts of the national strategy are broken down as follows:

- Headmasters need to be equipped with the skills to "strategically lead digital development work".
- Staff are to be trained to identify and use appropriate resources available to them in and out of the classroom.
- Students must have access to such resources.

The summary splits the learning policies into formal and informal learning.

Informal learning most notably the tool "MIL for me" by the Swedish Media Council, which is available at Betterinternetforkids.eu and on their website. It provides training materials designed to strengthen MIDL skills, particularly concerning anti-democratic messages. "MIL for me" won Insafe's prize for the best educational tool in MIDL for children and young people[179].

[178] European Commission/EACEA/Eurydice, 'Digital Education At School In Europe. Eurydice Report.' (Education, Audiovisual and Culture Executive Agency (EACEA, Education and Youth Policy Analysis) 2019).
[179] Insafe works to improve online safety for young people across Europe.

In addition, the Swedish Media Council is running the No Hate Speech Movement in Sweden. This campaign aims to prevent intolerance in the form of sexism and racism (among other forms) to "shield democracy from violent extremism". This is done by promoting source evaluation and critical thinking, particularly online given the use by extremist groups of the internet and social media to "distribute propaganda and other materials that glorify and reinforce norms relating to masculinity and violence".

In short, the changes to formal learning involve the introduction of programming in mathematics and technology lessons; the expansion of problem-solving and critical thinking skills into contexts involving the use of technology; and the use and understanding of digital tools and their impact on society.

A paper by Heintz et al. (2017)[180] provides far greater detail on the specific changes to the school curriculum. In March 2017, the Swedish government accepted Skolverket's (Swedish National Agency for Education) proposal for a new curriculum to be implemented by Autumn 2018. Most changes are to the specifications for mathematics, technology, and social studies courses with the introduction of basic computer science and programming across several disciplines being most noteworthy. Heintz *et al.* (2017) discusses K-12 education (Grade 1 to Grade 12 i.e. age 7-19) but focuses on K-9 (Grade 1 to Grade 9, i.e. age 7-16). These changes are summarised below.

In maths, students will now be taught:

- Step-wise functions
- The design and use of algorithms in various environments
- Programming for mathematical problem-solving

The subject, technology, was introduced into Swedish teaching in 1994 to adapt students' learning to the modernising world. In the new curriculum, students learn about:

- Components and uses of computers
- Technical solutions using electronics and programming
- The use of programming and digital tools for drawing, modelling, etc.
- Safety when using technology, e.g. use of electricity, sharing information digitally, and data protection

[180] Fredrik Heintz and others, 'Introducing Programming And Digital Competence In Swedish K-9 Education' [2017] Informatics in Schools: Focus on Learning Programming.

- The use of technology in society and the workplace
- Benefits, risks, and limitations associated with the internet and other global systems

Lastly, in social studies, the curriculum is expanded to include:

- The social, ethical, and legal aspects of digital and other media
- Responsible use of digital and other media
- The influence of personal worldviews on the news
- Source evaluation of digital content
- The portrayal of individuals and groups as affected by factors such as ethnicity and gender
- The use of hidden programming to control information on digital media
- Opportunities and risks associated with the internet and other technology

For further detail, including a breakdown of when in a child's school career they may expect to encounter each part of the content outlined in brief above, see Section 5 of Heintz *et al.* (2017).

As evidence of the effectiveness of these policies, note that the Open Society Institute based in Sofia, Georgia created the Media Literacy Index, which ranks 35 European countries on various MIDL skills. Sweden ranked 4th in both 2018[181] and 2019[182].

However, it has been raised that the nature of state education in Sweden leads to inconsistent results[183]. More specifically, since state education is decentralised, local authorities, not central government, are responsible for funding schools. This causes severe disparities in the quality of education reforms, including in the provision of high-speed internet, qualified teachers, and digital hardware such as computers for use by students and staff.

In summary, Sweden is expanding its school curriculum to include critical thinking skills and personal safety and responsibility applied to digital contexts and digital and technological skills applied to traditional learning contexts (such as mathematics). In addition, it is providing informal

[181] Marin Lessenski, 'COMMON SENSE WANTED RESILIENCE TO 'POST-TRUTH' AND ITS PREDICTORS IN THE NEW MEDIA LITERACY INDEX 2018' (European Policies Initiative (EuPI) of the Open Society Institute - Sofia 2018).
[182] Marin Lessenski, 'Just Think About It. Findings Of The Media Literacy Index 2019' (European Policies Initiative (EuPI) of the Open Society Institute – Sofia 2019).
[183] Mark Scott, 'Sweden Tries To Make Digital Lightning Strike Twice' (POLITICO, 2018) <https://www.politico.eu/article/sweden-education-system-digital-revamp-coding-stockholm-school/> accessed 25 July 2020.

learning tools for students and is promoting similar skills to the general population (rather than just to school children) in the context of preventing intolerance and violence.

Singapore

The Ministry of Education is the Singaporean organisation coordinating the changes to the curriculum[184].

They are beginning to introduce a framework for children to learn and use in their learning and everyday life:

- Find ("Gather and evaluate information, and use digital resources in a safe and responsible manner.");
- Think ("Interpret and analyse data, and solve problems.");
- Apply ("Use software and devices to facilitate the use of knowledge and skills in different ways.");
- Create ("Produce digital products, and collaborate online.").

The goal is to equip their students with the skills to improve online safety and MIDL.

The Ministry of Education website provides further detail concerning what is introduced at each stage of education.

In primary education:

- Cyber wellness will see more focus in CCE (Character and Citizenship Education)
- Students will be taught coding as part of the "Code for Fun" programme

In secondary education and junior college:

- There will be more emphasis on the application of mathematics in computation
- The Ministry of Education aims to have all students in Secondary 1 (age 12-13) and above using a Personal Learning Device (PLD) (i.e. a laptop or tablet) by 2024
- $200 (expanded from $150) will be made available to all eligible students to fund the purchase of a PLD with further grants available for lower-income households

[184] 'Strengthening Digital Literacy - Committee Of Supply Debate 2020' (Moe.gov.sg, 2020) <https://www.moe.gov.sg/microsites/cos2020/refreshing-our-curriculum/strengthen-digital-literacy.html> accessed 25 August 2020.

- More schools are offering O-Level and A-Level Computing (these qualifications form the GCE, General Certificate of Education, which is done at the end of secondary school)

In higher education:

- There will be greater emphasis on baseline digital competencies including quantitative reasoning, computational thinking, and cyber wellness
- There will be more coverage of AI for students pursuing careers that may involve it (such as finance and cyber-security)

Note the similarities with Sweden's policies. There is an emphasis placed on the combined use of traditional and digital learning techniques and cyber wellness.

Furthermore, Facebook is implementing its We Think Digital initiative in Asia-Pacific[185], including in Singapore[186]. This programme provides resources to equip people with the skills necessary to become responsible "digital citizens".

iii. Fostering Long-Term Digital Literacy

The reforms in Sweden and Singapore described above are unequivocally impressive. Given the ever-growing role of technology and digital media in modern life, similar reforms must be introduced globally. Below is a series of guidelines for any government or other appropriate subsidiary organisation seeking to improve MIDL in the population under its jurisdiction.

The overall plan should equip every person (and crucially, every voter) with the skills necessary to become less vulnerable to disinformation. These skills include, but are not limited to:

- Independent, as well as collaborative, critical thinking and problem solving
- Source evaluation (such as motivations for publishing content, whether true or disinformation)
- Understanding of technology and social media
- Rudimentary understanding of natural and social sciences as well as trust in science and the scientific community
- Appreciation for personal and social responsibility when using social and other media.

[185] 'Partners - We Think Digital' (We Think Digital) <https://wethinkdigital.fb.com/partners/> accessed 25 August 2020.
[186] 'Singapore First Port Of Call In Asia Pacific For Facebook'S New Digital Literacy Initiative' (CNA, 2019) <https://www.channelnewsasia.com/news/technology/singapore-first-country-asia-pacific-facebook-digital-literacy-11312828> accessed 25 July 2020.

To foster long-term digital literacy, the goal should be to educate all people, formally and from childhood, to develop familiarity and comfort with technology and media. In order to achieve this, MIDL education must be implemented in every stage of school education.

It should be implemented into several existing subjects, rather than being introduced as its own entirely separate subject, as it is easier to build MIDL education into the curriculum in several small sections. With this strategy, students will learn to solve more complex problems in familiar contexts (rather than having entirely standalone content) while also understanding the motivation for using digital tools.

Moreover, emphasis must be placed, not only on the risks and costs associated with technology and the internet, but also on the opportunities and benefits of using such powerful tools. In this way, students will be taught to appreciate the value of the resources available to them while also cultivating a healthy respect for the responsibility required to use them to benefit society. Put frankly, this technology is here to stay, so it would be far more constructive to teach safe and responsible use for work and social life.

It may be wise to have a separate subject or series of classes for "critical thinking in the digital age" as it may be awkward to fit this into existing subjects. While some aspects of this could be worked into the study of modern languages and the humanities (including social sciences), education authorities could also create a small curriculum to have it taught separately to emphasise its importance. It may be practical to work this into the ICT (information and communications technology) curriculum (or international equivalent). Since most of the content necessary to develop adequate MIDL skills is being implemented into other existing subjects, "critical thinking in the digital age" is likely to be relatively small so should fairly easily fit into ICT or its own short curriculum.

Ideally, governments would provide grants to students, particularly to those from disadvantaged backgrounds, to aid funding of a personal learning device such as a laptop or tablet, as is done in Singapore.

An additional longer-term aim may be to create a centralised board that rates each country's MIDL curriculum. This could be similar to the Media Literacy Index by the Open Society Institute described above. With such a board, international consistency could be achieved to prevent significant disparities in vulnerability to disinformation and intolerance. The board could rate MIDL curricula on such factors as effectiveness, comprehensiveness, ease of understanding,

and synergy with the existing curriculum, among others. This, unlike the implementation of MIDL in national education, is not immediately essential so will not be discussed further; however, it is advised that it be considered in future as such a board would undoubtedly be effective in promoting MIDL globally. Indeed, given the international scope of this initiative, and relevance to intentional disinformation overall, this could even fall under the responsibilities of the UN DOLFIN Agency (described earlier in this paper), to foster greater synergistic efficiency in holistically tackling international intentional disinformation.

However, overall, each country should be left to determine its own MIDL curriculum's particulars to make it consistent with their respective social and cultural norms. The creation of a centralised board would be instrumental here as it will prevent such norms from encroaching unduly on the curriculum's effectiveness.

Potential issues

Naturally, there could be problems with such a policy which would need to be addressed.

Firstly, building the curriculum into existing subjects would require all teachers to be trained accordingly, which may be costly. However, this is arguably easier than training a few dedicated specialists to create a new department at each school, and may be justifiable as a long-term initiative for modernising education along with the developing times.

Additionally, though the centralised board monitoring the development of MIDL curricula is not immediately essential, its absence from the picture does make the coordination of a consistent level of familiarity with and understanding of MIDL more difficult. As a result, in the initial stages, without a centralised board, it is essential to note that there could be significant volatility in the efficiency of fostering digital literacy between countries.

Furthermore, when Sweden's policy was described above, it was mentioned that the decentralised funding of their education system was at least partly responsible for intranational disparity in the quality of education, and particularly in the effectiveness of the new reforms. In addition, as with any part of school education, some students (even in the same year group or class at the same school) will find MIDL education more difficult than others. As a result, it is of paramount importance to prevent significant differences in outcomes by region, social class, ethnicity, gender, or any other such factor. To this end, there will need to be significant proactive regulation. In the UK, this could come from Ofsted (Office for Standards in Education) for instance, but every country should implement regulation from an equivalent organisation.

Also, given that technology is developing so quickly, every MIDL curriculum would need to be updated frequently to prevent it from going out of date. This raises some conflict between whether schools should focus on traditional learning which will remain somewhat constant over time (e.g. mathematics, science, humanities), or move more proactively with the times by changing the curriculum periodically to cover new technology and norms. A healthy balance will need to be struck, one which of course will take time to find and which may even change. Cooperation, both intra- and internationally, will be essential to determine what works and what does not.

iv. Fostering Short-Term Digital Literacy

In the short term, teaching MIDL to adults who have already left school could be constructive to encourage responsible use of tech and media. This could help bridge the gap between the longer-term education provided to students (which would only realise its full impact many years down the line), with the immediate victims and proponents of intentional disinformation in the current global adult population. This could allow for a more short-term immediate impact through digital literacy efforts and potentially greater effectiveness in the long-term digital literacy strategy, as it would prevent significant disparities or conflicts between students and their parents or adult associates in what they respectively understand regarding digital literacy.

This could involve distributing abridged versions of the school curriculum as pamphlets or online courses:

- Pamphlets could include brief introductions to MIDL skills and the motivations to learn them at all. (The latter may involve a short description of the problems associated with disinformation as described in the Issues section of this paper.) They could also provide exercises and examples for people to use to develop their skills and directions to find additional resources such as government websites or online courses.
- The online courses would provide further detail on everything from the problems of disinformation to ways of improving MIDL skills. There should be example exercises and activities available to test and develop understanding. Content could be broken down by age group (or general proficiency). This way, students could be afforded a continuation of the content covered in school, younger adults already well-versed in the use of technology could be given more advanced content, and the elderly who may be less comfortable with technology could be given simpler content. This tailoring of content will allow for ease of understanding for all people, regardless of prior experience, while also

maintaining effectiveness. More specifically, things like critical thinking and source evaluation (for bias, reliability, etc.) should exist in some form in every course for every age/proficiency group. In contrast, things like programming and applications to mathematical modelling should be reserved for those more comfortable with the basics.

It may be worth providing some incentives for adults to engage with these resources. For example, there could be some small financial incentive such as vouchers for retail or a free gift for completing courses. Employers could also be provided financial incentives, such as tax breaks, for dedicating a short period of time every month/6 months/year for their employees to complete these courses or for distributing pamphlets.

One of the main benefits of including this short-term strategy, as well as the long-term strategy, is that results can be achieved quite quickly. Even before a full school curriculum is laid out, basic information can be distributed relatively easily. This can be done in stages starting with media outlets, public sector offices, and other places requiring more immediate attention with further distribution done over time for logistical ease. This would act as a first step while the other long-term policies proposed in this paper are trialled and developed. In this way, fully digitally educated and critically thinking voters are brought into an adult population which is also well-versed in the same matters, albeit to a lesser extent.

The primary concern with this short-term strategy should be the potential for backlash or resistance. People may see it as a waste of time and political biases may make people, and by extension politicians, unreceptive to the idea. It is for this reason that incentives may be necessary.

It is important to note though, the short-term strategy is not the primary aim of MIDL training. Combatting the short-term effects of intentional disinformation requires reactive processes, which other potential solutions in this paper already provide. This short-term aspect of MIDL training and fostering digital literacy forms an additional constructive way of stimulating positive change with more immediate effect; but it is crucial to keep in mind that such policies' primary focus would be to enact long-term sustained benefits in enduring widespread population digital literacy. Thereby safeguarding future populations against intentional disinformation and issues of vulnerability of the masses.

IV. OVERALL STRATEGY AND CONCLUDING REMARKS

As discussed throughout this paper, modern intentional disinformation remains a pervasive and deeply nuanced issue, and one that continues to grow alarmingly worldwide. Weakening individuals' abilities to accurately determine truth and credibility, while providing greater protection and lower barrier to entry for perpetrators, this issue has remained largely unresolved by most traditional methods of rectification (e.g. blanket legal measures and heavy-handed controls on social media platforms). In large part, this is due to accusations of censorship, and diminishing trust in the unbiased objectivity of the very institutions that traditionally have served as arbiters of truth in the media, that typically accompany such measures.

As a result, this paper has taken a more nuanced approach to resolving the issue of modern intentional disinformation. Opting to tackle the source issue directly instead of through broad, sweeping measures, this paper has targeted potential solutions which combat intentional disinformation at the same level that it originates from – enhancing individual abilities to accurately distinguish veritable and trustworthy information, while reducing the protection afforded to potential perpetrators, on a widespread level.

Specifically, these solutions compose of two core segments, targeting differing time horizons. Providing a more immediate, reactionary measure to modern intentional disinformation, this paper proposes the foundation of the UN Debunking and OnLine Fake Information Neutralisation (DOLFIN) Agency. Providing a longer-term, deep-rooted measure, this paper proposes an addition to national curriculums, fostering digital literacy.

The first measure, DOLFIN, would act as a multi-partite organisation with global coverage, mandate, and contribution, to implement two key elements of the solutions suite proposed: the social media transparency models, and the universal fact-checking service.

The former comprises of two main transparency models aimed at increasing transparency in information on public forum social media platforms. The proposed 'Information Source' transparency model operates through a self-opted verification process, where individuals making a post including factual claims can request for their post to be reviewed for the appropriate inclusion of citations from reputable sources, for the claims made. Upon successful review, this would then show up as a positive visual indicator on the post, indicating comparable credibility – and by extension, conveying increased questionability and cause for scepticism on posts lacking such an indicator. The proposed 'Broad Trends' transparency model, or 'Ad Database' model,

operates through a standardised collation of data on key statistics surrounding advertisements on public forum social media platforms; providing greater ease in analysis and exposure of trends and potential perpetrators of intentional disinformation. As both models require some degree of cooperation with social media platforms on a global scale (and as the 'Information Source' transparency model particularly would benefit from an unassailably objective authority determining the basis of reputable sources in their review processes), both models would benefit strongly in implementation under the responsibilities of DOLFIN.

The other solution under DOLFIN's responsibility, the universal fact-checking service, would provide a service for investigating and publishing unbiased articles or reports on various intentional disinformation trends, with global coverage. Providing analysis of such trends, with unbiased coverage of all key viewpoints, such a service would provide a much-needed source for objective analysis and confirmation (or debunking) of factual claims. Given the issue of reduced credibility of information that modern intentional disinformation has generated (eroding trust in the ability of many institutions, including long-established arbiters of truth, to remain objective, accurate, and credible), such a service would undoubtedly require the global multi-partite structure and accountability of DOLFIN to be effective.

Operating with a longer-term focus, the second segment of the proposed solution suite involves measures for fostering widespread digital literacy. This is mainly concentrated on long-term measures, centred around proposed additions to national curricula; namely, through introducing core courses on media, information, and digital literacy (MIDL) competency, and integrating these with existing 'traditional' subjects and learning styles. Constructing a solid groundwork for future sustainable resilience to intentional disinformation in the next generation and beyond, this solution forms a necessary counterpart to the short-term reactionary measures implemented through the UN DOLFIN Agency.

Given the rapidly moving, technologically advanced, and continuously evolving, nature of modern intentional disinformation, it is necessary that any solution set proposed is able to comprehensively provide coverage for the immediate and pressing known risks of fake news – as well as effective preparation for the unknown, yet inevitable, future risks that modern intentional disinformation will shift to. As such, this paper strongly supports the dual-pronged approach proposed through DOLFIN and fostering widespread digital literacy. Such an approach provides society with immediate and long-term enduring protection from fake news – through tools that allow individuals to naturally protect themselves from modern intentional

disinformation, and effectively educate themselves to remain resilient against any future forms of intentional disinformation. Therefore, while it is certainly true that there are alternate routes to combatting modern intentional disinformation (e.g. through channels of law or matters of direct corporate governance), this paper believes that these should be implemented (if and when necessary) on-top of the solutions proposed in this paper, to manage the contextual, geographical, and cultural idiosyncrasies of modern intentional disinformation's global impact.

Serving as a baseline solution suite, this paper seeks to provide an effective and consistent global groundwork for tackling modern intentional disinformation; ultimately providing just an indicative framework, or step in the right direction. As the situation progresses and the nuances and circumstances surrounding fake news grow clearer (or evolve further), further research and policies should certainly be conducted, to ensure that any policies implemented remain as contextually appropriate and robust as possible to deal with modern intentional disinformation in the social media era.

V. BIBLIOGRAPHY

'4Chan' (4chan.org, 2020) <http://www.4chan.org/> accessed 21 August 2020.

'Ads Transparency Center' (Twitter) <https://ads.twitter.com/transparency> accessed 10 January 2020.

'Ads Transparency Center FAQs' (Twitter) <https://business.twitter.com/en/help/ads-policies/ads-transparency-center-faqs.html> accessed 10 January 2020.

'Against information manipulation', French Government, 2018, <https://www.gouvernement.fr/en/against-information-manipulation>

Alaphilippe, A, et al. 'Automated Tackling of Disinformation: Major Challenges Ahead, 2019', <http://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf> [accessed 9 September 2019].

Almukhtar, S, et al. 'Black Lives Upended by Policing: The Raw Videos Sparking Outrage' The New York Times (30 July 2015) <https://www.nytimes.com/interactive/2017/08/19/us/police-videos-race.html, https://www.nytimes.com/interactive/2017/08/19/us/police-videos-race.html> accessed 7 March 2020.

'An update on our continuity strategy during COVID-19', Twitter, <https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html>

Anderson, J, and Raine, L. 'The Future Of Truth And Misinformation Online' (Pew Research 2017).

Baynes, C. 'Coronavirus: Patients refusing treatment because of fake news on social media, NHS staff warn', The Independent, 2020 <https://www.independent.co.uk/news/uk/home-news/coronavirus-fake-news-conspiracy-theories-antivax-5g-facebook-twitter-a9549831.html>

BBC, 'How President Trump Took 'Fake News' Into The Mainstream' (2018) < https://www.bbc.com/news/av/world-us-canada-46175024> accessed 22 August 2020.

Beaumont, P. 'The Truth about Twitter, Facebook and the Uprisings in the Arab World' The Guardian (25 February 2011) <https://www.theguardian.com/world/2011/feb/25/twitter-facebook-uprisings-arab-libya> accessed 7 March 2020.

Berzina, K. 'Sweden — Preparing For The Wolf, Not Crying Wolf: Anticipating And Tracking Influence Operations In Advance Of Sweden'S 2018 General Elections' (The German Marshall Fund of the United States, 2018) <http://www.gmfus.org/blog/2018/09/07/sweden-preparing-wolf-not-crying-wolf-anticipating-and-tracking-influence> accessed 21 August 2020.

Bisen, A. 'Disinformation Is Drowning Democracy' (Foreign Policy, 2019) <https://foreignpolicy.com/2019/04/24/disinformation-is-drowning-democracy/> accessed 21 August 2020.

Blood, D. 'How Facebook Shares Its Advertising Data' Financial Times (London, 1 June 2019) <https://www.ft.com/content/7e7f0564-82e9-11e9-b592-5fe435b57a3b>.

Bowles, N. 'Twitter, Facing Another Uproar, Pauses Its Verification Process' The New York Times (9 November 2017) <https://www.nytimes.com/2017/11/09/technology/jason-kessler-twitter-verification.html> accessed 7 March 2020.

Boylan, J. F., 'Will Deep-Fake Technology Destroy Democracy?' The New York Times (17 October 2018) <https://www.nytimes.com/2018/10/17/opinion/deep-fake-technology-democracy.html> accessed 8 March 2020.

Brundage, M, et al. 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', ArXiv:1802.07228 [Cs], 2018, <http://arxiv.org/abs/1802.07228> [accessed 9 September 2019].

Bump, P. 'Trump Is Mad about the Size of His Crowd on Twitter', Washington Post, 2019 <https://www.washingtonpost.com/politics/2019/04/23/trump-is-mad-about-size-his-crowd-twitter/> [accessed 9 September 2019].

Burger, J. M., et al. 'What a Coincidence! The Effects of Incidental Similarity on Compliance', Personality and Social Psychology Bulletin, 30.1 (2004), <https://doi.org/10.1177/0146167203258838>.

Burkhardt, J. M., 'Combatting Fake News in the Digital Age', Library Technology Reports, 53:8.

Chakrabarti, S, et al. 'Duty, Identity, Credibility: "Fake News" and the Ordinary Citizen in India' (BBC 2018) 34 <http://downloads.bbc.co.uk/mediacentre/duty-identity-credibility.pdf>.

Chan, M. S., et al. 'Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation', Psychological Science, 28.11 (2017), <https://doi.org/10.1177/0956797617714579>.

Chandler, D, and Munday, R. 'Disinformation', A Dictionary of Media and Communication (Oxford: Oxford University Press, 2016).

Chen, A. 'The Agency' (2015) <https://www.nytimes.com/2015/06/07/magazine/the-agency.html> accessed 21 August 2020.

"Chilling': Singapore's 'fake news' law comes into effect', The Guardian, 2019 < https://www.theguardian.com/world/2019/oct/02/chilling-singapores-fake-news-law-comes-into-effect>

'Code of Practice on Disinformation' (Digital Single Market - European Commission, 17 June 2019) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation> accessed 18 October 2019.

Conger, K. 'Twitter Will Ban All Political Ads, C.E.O. Jack Dorsey Says' The New York Times (30 October 2019) <https://www.nytimes.com/2019/10/30/technology/twitter-political-ads-ban.html> accessed 10 January 2020.

'Data Collection Log — EU Ad Transparency Report' (Mozilla Ad Transparency) <https://adtransparency.mozilla.org/eu/log/> accessed 17 October 2019.

Dave, P, and Bing, C. 'Russian Disinformation On Youtube Draws Ads, Lacks Warning Labels: Researchers' (Reuters, 2019) <https://www.reuters.com/article/us-alphabet-google-youtube-russia/russian-disinformation-on-youtube-draws-ads-lacks-warning-labels-researchers-idUSKCN1T80JP> accessed 21 August 2020.

Dechêne, A, et al. 'The Truth About the Truth: A Meta-Analytic Review of the Truth Effect', Personality and Social Psychology Review, 14.2 (2010), <https://doi.org/10.1177/1088868309352251>.

DiResta, R, and Gilad, L. 'Anti-Vaxxers Are Using Twitter to Manipulate a Vaccine Bill', Wired, 2015 <https://www.wired.com/2015/06/antivaxxers-influencing-legislation/> [accessed 9 September 2019].

DiResta, R, et al. 'The Tactics & Tropes of the Internet Research Agency', (New Knowledge, 2019) p.85-89

'Disclosures' (Snopes.com) <https://www.snopes.com/disclosures/> accessed 25 August 2020.

'Disinformation and 'fake news': Final Report', Digital, Culture, Media and Sport Committee, U.K. Parliament, 2019

Dizikes, P., 'Study: On Twitter, false news travels faster than true stories', *Massachusetts Institute of Technology*, 2018 <https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308> [accessed 11 September 2019].

Engber, D. 'We've Been Told We're Living in a Post-Truth Age. Don't Believe It.', Slate Magazine, 2018 <https://slate.com/health-and-science/2018/01/weve-been-told-were-living-in-a-post-truth-age-dont-believe-it.html> [accessed 11 September 2019].

Ernst, N, et al. 'Effects of Message Repetition and Negativity on Credibility Judgments and Political Attitudes', 2017 <https://doi.org/10.5167/uzh-139745>.

'EU Code of Practice on Disinformation' <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation>.

European Commission/EACEA/Eurydice, 'Digital Education At School In Europe. Eurydice Report.' (Education, Audiovisual and Culture Executive Agency (EACEA, Education and Youth Policy Analysis) 2019).

Evan, D. 'Does This Video Show Nancy Pelosi Drunk and Slurring Her Speech?' (Snopes) <https://www.snopes.com/fact-check/nancy-pelosi-slurring-speech/> accessed 8 March 2020.

'Facebook Ads Library Assessment' (Ambassador for Digital Affairs) <https://disinfo.quaidorsay.fr/en/facebook-ads-library-assessment> accessed 17 October 2019.

'Facebook and Google: This Is What an Effective Ad Archive API Looks Like' (The Mozilla Blog, 27 March 2019) <https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like> accessed 15 October 2019.

'Facebook And Twitter Restrict Trump Accounts Over 'Harmful' Virus Claim' (BBC News, 2020) <https://www.bbc.com/news/election-us-2020-53673797> accessed 21 August 2020.

Facebook Uncovers Disinformation Campaign To Influence US Midterms' (The Financial Times, 2018) <https://www.ft.com/content/7af02014-94e1-11e8-b67b-b8205561c3fe> accessed 21 August 2020.

'Facebook's Ad Archive API Is Inadequate' (The Mozilla Blog, 29 April 2019) <https://blog.mozilla.org/blog/2019/04/29/facebooks-ad-archive-api-is-inadequate> accessed 17 October 2019.

'Fact Checking On Facebook' (Facebook) <https://www.facebook.com/business/help/182222309230722> accessed 25 August 2020.

Foer, F. 'The Era of Fake Video Begins' [2018] The Atlantic <https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/> accessed 8 March 2020.

Funke, D. 'Blame Bugs, Not Partisanship, For Google Wrongly Appending A Fact Check To The Daily Caller' (Poynter, 2018) <https://www.poynter.org/fact-checking/2018/blame-bugs-not-partisanship-for-google-wrongly-appending-a-fact-check-to-the-daily-caller/> accessed 25 August 2020.

Funke, D. 'Youtube Is Now Surfacing Fact Checks In Search. Here's How It Works.' (Poynter, 2019) <https://www.poynter.org/fact-checking/2019/youtube-is-now-surfacing-fact-checks-in-search-heres-how-it-works/> accessed 25 August 2020.

Gabriel, M. 'Statement by Commissioner Gabriel on the Code of Practice on Online Disinformation' (European Commission Press Release Database, 26 September 2018) <https://europa.eu/rapid/press-release_STATEMENT-18-5914_en.htm> accessed 18 October 2019.

Goldstein, N. J., et al. 'A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels', Journal of Consumer Research, 35.3 (2008), 472–82 <https://doi.org/10.1086/586910>.

Greenberg, A. 'Twitter Still Can't Keep Up With Its Flood Of Junk Accounts, Study Finds' (Wired, 2019) <https://www.wired.com/story/twitter-abusive-apps-machine-learning/> accessed 21 August 2020.

Grush, L. 'The CNN Porn Scare Is How Fake News Spreads' (The Verge, 25 November 2016) <https://www.theverge.com/2016/11/25/13748226/cnn-accidentally-airs-porn-fake-news-boston> accessed 8 March 2020.

Hadad, F. 'Disinformation Spreads On Whatsapp Ahead Of Brazilian Election' (New York Times, 2018) <https://www.nytimes.com/2018/10/19/technology/whatsapp-brazil-presidential-election.html?module=inline> accessed 21 August 2020.

Hamilton, I. A., 'Here's what we know about the bizarre coronavirus 5G conspiracy theory that is leading people to set cellphone masts on fire', Business Insider, 2020 <https://www.businessinsider.com/coronavirus-conspiracy-5g-masts-fire-2020-4?r=US&IR=T>.

Hansler, J. 'US Accuses Russia Of Conducting Sophisticated Disinformation And Propaganda Campaign' (CNN, 2020) <https://edition.cnn.com/2020/08/05/politics/state-department-russian-disinformation-report/index.html> accessed 22 August 2020.

Harkins, S, et al. 'The Oxford Handbook of Social Influence, Oxford Handbooks' (Oxford: Oxford University Press, 2017), <doi.org/10.1093/oxfordhb/9780199859870.013.4>.

Heintz, F, et al. 'Introducing Programming And Digital Competence In Swedish K-9 Education' [2017] Informatics in Schools: Focus on Learning Programming.

Hern, A. 'Youtube To Crack Down On Fake News, Backing 'Authoritative' Sources' (The Guardian, 2018) <https://www.theguardian.com/technology/2018/jul/09/youtube-fake-news-changes> accessed 21 August 2020.

House of Commons: Digital, Culture, Media, and Sports Committee, 'Disinformation And 'Fake News': Final Report' (House of Commons 2019).

'How Facebook's Fact-Checking Program Works' (Facebook, 2020) <https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works> accessed 25 August 2020.

'How Whatsapp Helped Turn An Indian Village Into A Lynch Mob' (BBC News, 2018) <https://www.bbc.com/news/world-asia-india-44856910> accessed 21 August 2020.

'Human Compatible: Artificial Intelligence and the Problem of Control with Stuart Russell' <https://futureoflife.org/2019/10/08/ai-alignment-podcast-human-compatible-artificial-intelligence-and-the-problem-of-control-with-stuart-russell/>.

Ingram, D. 'Facebook to Use Its News Feed to Push More Videos to Users' Reuters (14 December 2017) <https://www.reuters.com/article/us-facebook-video-idUSKBN1E8300> accessed 8 March 2020.

Innes, M, et al. 'Disinformation And Digital Influencing After Terrorism: Spoofing, Truthing And Social Proofing' [2019] Contemporary Social Science.

Isaac, M, and Roose, K. 'Disinformation Spreads on WhatsApp Ahead of Brazilian Election', The New York Times, 2018, <https://www.nytimes.com/2018/10/19/technology/whatsapp-brazil-presidential-election.html>

Jolley, J. 'Attribution, State Responsibility, And The Duty To Prevent Malicious Cyber-Attacks In International Law' (University of Glasgow 2017) <http://theses.gla.ac.uk/8452/1/2017JolleyPhD.pdf> accessed 21 August 2020.

Kalsnes, B. 'Fake News' [2018] Oxford Research Encyclopedia of Communication.

Kastrenakes, J. 'Whatsapp Limits Message Forwarding In Fight Against Misinformation' (The Verge, 2019) <https://www.theverge.com/2019/1/21/18191455/whatsapp-forwarding-limit-five-messages-misinformation-battle> accessed 21 August 2020.

Kelly, M. 'Twitter Will Ban Candidate Ads and Limit Issue Ads' (The Verge, 15 November 2019) <https://www.theverge.com/2019/11/15/20966854/twitter-super-pacs-political-ads-facebook-climate-change-abortion-paid-promotion> accessed 10 January 2020.

Kragh, M, and Asberg, S. 'Russia's Strategy For Influence Through Public Diplomacy And Active Measures: The Swedish Case' (Journal of Strategic Studies, 2017).

Krishna, R. 'This Footage of Keir Starmer Being Interviewed on Good Morning Britain Was Edited before Being Posted by the Conservative Party' (Full Fact) <https://fullfact.org/news/keir-starmer-gmb/> accessed 8 March 2020.

Kuklinski, J. H., and Quirk, P. J., 'Conceptual Foundations Of Citizen Competence' (2001) 23 Political Behavior.

Leeper, T. J., and Mullinix, K. J. 'Motivated Reasoning' (Oxford University Press, 2018) <https://doi.org/10.1093/obo/9780199756223-0237>.

Lessenski, M. 'Common Sense Wanted Resilience To 'Post-Truth' And Its Predictors In The New Media Literacy Index 2018' (European Policies Initiative (EuPI) of the Open Society Institute - Sofia 2018).

Lessenski, M. 'Just Think About It. Findings Of The Media Literacy Index 2019' (European Policies Initiative (EuPI) of the Open Society Institute – Sofia 2019).

Lewandowsky, S, et al. 'Misinformation and Its Correction: Continued Influence and Successful Debiasing', Psychological Science in the Public Interest, 13.3 (2012), <https://doi.org/10.1177/1529100612451018>.

Lodge, M, and Taber, C. S. The Rationalizing Voter (Cambridge: Cambridge University Press, 2013), <https://www.cambridge.org/core/books/rationalizing-voter/9E4E27965B612D5DDB1091BD907DF492> [accessed 9 September 2019].

Lord, C. G., et al. 'Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence.', Journal of Personality and Social Psychology, 37.11 (1979), <https://doi.org/10.1037/0022-3514.37.11.2098>.

Mack, D. 'This PSA About Fake News from Barack Obama Is Not What It Appears', BuzzFeed News, 2018 <https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed>.

Marwick, A, and Lewis, R. 'Media Manipulation and Disinformation Online' (Data & Society Research Institute 2017) <https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf>, accessed 21 August 2020.

McLaughlin, T. 'How Whatsapp Fuels Fake News And Violence In India' (Wired, 2018) <https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/> accessed 21 August 2020.

Merrill, J, and Tobin, A. 'Facebook Moves to Block Ad Transparency Tools — Including Ours' (ProPublica, 28 January 2019) <https://www.propublica.org/article/facebook-blocks-ad-transparency-tools> accessed 18 October 2019.

Metz, C. 'Internet Companies Prepare to Fight the "Deepfake" Future' The New York Times (24 November 2019) <https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html> accessed 8 March 2020.

Meyer, R. 'The Grim Conclusions of the Largest-Ever Study of Fake News' [2018] The Atlantic <https://www.theatlantic.com/technology/archive/2018/03/largest-study-ever-fake-news-mit-twitter/555104/> accessed 7 March 2020.

Miller, M. 'Protesters Outside White House Demand 'Pizzagate' Investigation' The Washington Post (2020) <https://www.washingtonpost.com/news/local/wp/2017/03/25/protesters-outside-white-house-demand-pizzagate-investigation/> accessed 21 August 2020.

'Natural-Language Processing', in A Dictionary of Computer Science, ed. by Butterfield, A, et al. (Oxford University Press, 2016) <https://www.oxfordreference.com/view/10.1093/acref/9780199688975.001.0001/acref-9780199688975-e-6410> [accessed 10 September 2019].

Nyhan, B, and Reifler, J. 'When Corrections Fail: The Persistence of Political Misperceptions', Political Behavior, 32.2 (2010), <https://doi.org/10.1007/s11109-010-9112-2>.

Nyhan, B, et al. 'The Hazards of Correcting Myths About Health Care Reform', Medical Care, 51.2 (2012), 127–32 <https://doi.org/10.1097/MLR.0b013e318279486b>.

Nyhan, B. 'Fact-Checking Can Change Views? We Rate That as Mostly True', The New York Times, 2016, section The Upshot <https://www.nytimes.com/2016/11/06/upshot/fact-checking-can-change-views-we-rate-that-as-mostly-true.html> [accessed 11 September 2019].

'Our Funding - Factcheck.Org' (FactCheck.org) <https://www.factcheck.org/our-funding/> accessed 25 August 2020.

'Our Process - Factcheck.Org' (FactCheck.org) <https://www.factcheck.org/our-process/> accessed 25 August 2020.

Oxenham, S. '"I Was a Macedonian Fake News Writer"' BBC (London, 29 May 2019) <https://www.bbc.com/future/article/20190528-i-was-a-macedonian-fake-news-writer> accessed 8 March 2020.

Paris, B, and Donovan, J. 'Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence' (Data & Society Research Institute 2019) 13–15 <https://datasociety.net/library/deepfakes-and-cheap-fakes/>.

'Partners - We Think Digital' (We Think Digital) <https://wethinkdigital.fb.com/partners/> accessed 25 August 2020.

Paul, C, and Mathews, M. 'The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It', Perspective, 2016, <https://doi.org/10.7249/PE198>.

Pennycook, G, et al. 'Prior Exposure Increases Perceived Accuracy of Fake News', Journal of Experimental Psychology: General, 147.12 (2018), 1865–80 <https://doi.org/10.1037/xge0000465>.

'Pizzagate': The Fake Story That Shows How Conspiracy Theories Spread' (BBC News, 2016) <https://www.bbc.com/news/blogs-trending-38156985> accessed 21 August 2020.

Platow, M. J., et al. '"It's Not Funny If They're Laughing": Self-Categorization, Social Influence, and Responses to Canned Laughter', Journal of Experimental Social Psychology, 41.5 (2005), <https://doi.org/10.1016/j.jesp.2004.09.005>.

'Political Content' (Twitter) <https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html> accessed 10 January 2020.

Posetti, J. 'News Industry Transformation: Digital Technology, Social Platforms And The Spread Of Misinformation And Disinformation' [2018] Handbook for Journalism Education and Training, UNESCO.

Posner, M. 'Dealing With Disinformation: Facebook And Youtube Need To Take Down Provably False "News"' (Forbes, 2019) <https://www.forbes.com/sites/michaelposner/2019/03/14/dealing-with-disinformation-

facebook-and-youtube-need-to-take-down-provably-false-news/#f487e0e19e79> accessed 21 August 2020.

Reuben, A. 'Are We Giving £350m a Week to Brussels?', BBC News, 2016, section EU Referendum <https://www.bbc.com/news/uk-politics-eu-referendum-36110822> [accessed 9 September 2019].

Rosenberg, M. 'Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised' The New York Times (New York, 25 July 2019) <https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html> accessed 17 October 2019.

Roth, Y, and Harvey, D. 'How Twitter Is Fighting Spam And Malicious Automation' (Twitter, 2018) <https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html> accessed 21 August 2020.

Rothman, J. 'In the Age of A.I., Is Seeing Still Believing? | The New Yorker' [2018] The New Yorker <https://www.newyorker.com/magazine/2018/11/12/in-the-age-of-ai-is-seeing-still-believing> accessed 8 March 2020.

Scott, M. 'Sweden Tries To Make Digital Lightning Strike Twice' (POLITICO, 2018) <https://www.politico.eu/article/sweden-education-system-digital-revamp-coding-stockholm-school/> accessed 25 July 2020.

Serhan, Y. 'Italy Scrambles To Fight Misinformation Ahead Of Its Elections' (The Atlantic, 2018) <https://www.theatlantic.com/international/archive/2018/02/europe-fake-news/551972/> accessed 21 August 2020.

Silverman, C, and Alexander, L. 'How Teens In The Balkans Are Duping Trump Supporters With Fake News' BuzzFeed News (3 November 2016) <https://www.buzzfeednews.com/article/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo> accessed 8 March 2020.

Silverman, C, et al. 'Disinformation For Hire: How A New Breed Of PR Firms Is Selling Lies Online' BuzzFeed News (6 January 2020) <https://www.buzzfeednews.com/article/craigsilverman/disinformation-for-hire-black-pr-firms> accessed 7 March 2020.

Silverman, C. 'This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook' (BuzzFeed News, 16 November 2016) <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook> accessed 9 September 2019.

'Singapore First Port Of Call In Asia Pacific For Facebook'S New Digital Literacy Initiative' (CNA, 2019) <https://www.channelnewsasia.com/news/technology/singapore-first-country-asia-pacific-facebook-digital-literacy-11312828> accessed 25 July 2020.

Sippitt, A. 'The Backfire Effect: Does It Exist? And Does It Matter for Factcheckers?' (London: Full Fact, 2019), <https://fullfact.org/blog/2019/mar/does-backfire-effect-exist/>.

'Spam, deceptive practices & scams policies', YouTube, <https://support.google.com/youtube/answer/2801973?hl=en&ref_topic=9282365>

Stewart, E. 'Twitter Is Walking into a Minefield with Its Political Ads Ban' (Vox, 15 November 2019) <https://www.vox.com/recode/2019/11/15/20966908/twitter-political-ad-ban-policies-issue-ads-jack-dorsey> accessed 10 January 2020.

Stolton, S. 'In The Fight Against Fake News, Youtube Has A 'Bias Toward Keeping Content Up' (Euractiv, 2019) <https://www.euractiv.com/section/digital/news/in-the-fight-against-fake-news-youtube-has-a-bias-toward-keeping-content-up/> accessed 21 August 2020.

'Stop Hate for Profit', <https://www.stophateforprofit.org>

'Strengthening Digital Literacy - Committee Of Supply Debate 2020' (Moe.gov.sg, 2020) <https://www.moe.gov.sg/microsites/cos2020/refreshing-our-curriculum/strengthen-digital-literacy.html> accessed 25 August 2020.

Subramanian, S. 'Inside the Macedonian Fake-News Complex | WIRED' [2017] Wired <https://www.wired.com/2017/02/veles-macedonia-fake-news/> accessed 8 March 2020.

'Tackling online disinformation', European Commission, 2020 <https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation>

Tambini, D. 'Fake News: Public Policy Responses' (LSE Media Policy Project, 2017).

'The Twitter Rules', Twitter, < https://help.twitter.com/en/rules-and-policies/twitter-rules>

Thompson, N, and Lapowsky, I. 'How Russian Trolls Used Meme Warfare To Divide America' (Wired, 2018) <https://www.wired.com/story/russia-ira-propaganda-senate-report/> accessed 21 August 2020.

'Transparency' (Snopes.com) <https://www.snopes.com/transparency/> accessed 25 August 2020.

'Trump Retweets Doctored Video Of Pelosi To Portray Her As Having 'Lost It' (Reuters, 2019) <https://www.reuters.com/article/us-usa-trump-pelosi/trump-retweets-doctored-video-of-pelosi-to-portray-her-as-having-lost-it-idUSKCN1SU2CB> accessed 21 August 2020.

Trump, D. 2015, < https://twitter.com/realdonaldtrump/status/570238975157387264> accessed 22 August 2020.

Trump, D. 2016, <https://twitter.com/realdonaldtrump/status/689458468835569664?lang=en> accessed 22 August 2020.

Tucker, J, et al. 'Social Media, Political Polarization, And Political Disinformation: A Review Of The Scientific Literature' [2018] SSRN Electronic Journal.

'Twitter Ads Transparency Center Assessment' (French Ambassador for Digital Affairs) <https://disinfo.quaidorsay.fr/en/twitter-ads-transparency-center-assessment> accessed 10 January 2020.

Tworek, H, and Leersen, P. 'An Analysis of Germany's NetzDG Law', Transatlantic Working Group, 2019.

Vazquez, M, and O'Sullivan, D. 'Twitter Tells New Congressional Candidates They'll Have To Win Their Primaries To Get Verified' (CNN, 2019) <https://edition.cnn.com/2019/08/06/politics/twitter-primaries-verification/index.html> accessed 21 August 2020.

Vincent, J. 'Watch Jordan Peele Use AI to Make Barack Obama Deliver a PSA about Fake News' (The Verge, 17 April 2018) <https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed> accessed 8 March 2020.

Vosoughi, S, et al. 'The Spread of True and False News Online' (2018) 359 Science 1146.

Warrell, H. 'Efforts to Prevent Foreign Manipulation of UK Election Flounder' Financial Times (London, 10 December 2019) <https://www.ft.com/content/19daf806-1a98-11ea-97df-cc63de1d73f4> accessed 23 January 2020.

Waterson, J. 'Facebook Restricts Campaigners' Ability to Check Ads for Political Transparency' The Guardian (27 January 2019) <https://www.theguardian.com/technology/2019/jan/27/facebook-restricts-campaigners-ability-to-check-ads-for-political-transparency> accessed 18 October 2019.

'What Is the Facebook Ad Library and How Do I Search It?' (Facebook Help Centre) <https://www.facebook.com/help/259468828226154> accessed 16 October 2019.

'Whatsapp: The 'Black Hole' Of Fake News In India's Election' (BBC News, 2019) <https://www.bbc.com/news/world-asia-india-47797151> accessed 21 August 2020.

'Who Was Jeffrey Epstein?' (BBC News, 2019) <https://www.bbc.co.uk/news/world-us-canada-48913377> accessed 21 August 2020

Wood, T, and Porter, E. 'The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence', SSRN Electronic Journal, 2016 <https://doi.org/10.2139/ssrn.2819073>.

Woolley, S, and Howard, P. 'Computational Propaganda: Political Parties, Politicians, And Political Manipulation On Social Media' (S Woolley and P Howard, Oxford University Press 2019).

'Working to Stop Misinformation and False News', Facebook, <https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news>