European Parliament

Regulating disinformation with artificial intelligence

STUDY

Panel for the Future of Science and Technology

European Science-Media Hub

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 624.279 – March 2019

EN

Regulating disinformation with artificial intelligence

Effects of disinformation initiatives on freedom of expression and media pluralism

This study examines the consequences of the increasingly prevalent use of artificial intelligence (AI) disinformation initiatives upon freedom of expression, pluralism and the functioning of a democratic polity.

The study examines the trade-offs in using automated technology to limit the spread of disinformation online. It presents options (from self-regulatory to legislative) to regulate automated content recognition (ACR) technologies in this context. Special attention is paid to the opportunities for the European Union as a whole to take the lead in setting the framework for designing these technologies in a way that enhances accountability and transparency and respects free speech. The present project reviews some of the key academic and policy ideas on technology and disinformation and highlights their relevance to European policy.

Chapter 1 introduces the background to the study and presents the definitions used. Chapter 2 scopes the policy boundaries of disinformation from economic, societal and technological perspectives, focusing on the media context, behavioural economics and technological regulation. Chapter 3 maps and evaluates existing regulatory and technological responses to disinformation. In Chapter 4, policy options are presented, paying particular attention to interactions between technological solutions, freedom of expression and media pluralism.

EPRS | European Parliamentary Research Service

AUTHORS

This study was financed by the European Science-Media Hub (ESMH) budget. It was written by Prof. Chris MARSDEN, University of Sussex, and Dr Trisha MEYER, Vrije Universiteit Brussel (with an independent review by Dr Ian BROWN), at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

ADMINISTRATOR RESPONSIBLE

Mihalis KRITIKOS, Scientific Foresight Unit

To contact the publisher, please e-mail esmh@ep.europa.eu

LINGUISTIC VERSION

Original: EN

Manuscript completed in March 2019.

DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its authors and any opinions expressed herein should not be taken to represent an official position of the Parliament.

PE 624.279
ISBN: 978-92-846-3947-2
doi: 10.2861/003689
QA-01-19-194-EN-N

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (ESMH website)
https://sciencemediahub.eu/ (EPRS website)
http://epthinktank.eu (blog)

Executive summary

The European Parliament elections of May 2019 provide an impetus for European-level actions to tackle disinformation. This interdisciplinary study analyses the implications of artificial intelligence (AI) disinformation initiatives on freedom of expression, media pluralism and democracy. The authors were tasked to formulate policy options, based on the literature, expert interviews and mapping that form the analysis undertaken. The authors warn against technocentric optimism as a solution to disinformation online, that proposes use of automated detection, (de)prioritisation, blocking and removal by online intermediaries without human intervention. When AI is used, it is argued that far more independent, transparent and effective appeal and oversight mechanisms are necessary in order to minimise inevitable inaccuracies.

This study defines disinformation as 'false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit' in line with the European Commission High Level Expert Group report use of the term. The study distinguishes disinformation from misinformation, which refers to unintentionally false or inaccurate information.

Within machine learning techniques that are advancing towards AI, automated content recognition (ACR) technologies are textual and audio-visual analysis programmes that are algorithmically trained to identify potential 'bot' accounts and unusual potential disinformation material. In this study, ACR refers to both the use of automated techniques in the recognition **and** the moderation of content and accounts to assist human judgement. Moderating content at larger scale requires ACR as a supplement to human moderation (editing). Using ACR to detect disinformation is however prone to false negatives/positives due to the difficulty of parsing multiple, complex, and possibly conflicting meanings emerging from text. If inadequate for natural language processing and audiovisual material including 'deep fakes' (fraudulent representation of individuals in video), ACR does have more reported success in identifying 'bot' accounts. In the remainder of this report, the shorthand 'AI' is used to refer to these ACR technologies.

Disinformation has been a rapidly moving target in the period of research, with new reports and policy options presented by learned experts on a daily basis throughout the third quarter of 2018. The authors analysed these reports and note the strengths and weaknesses of those most closely related to the study's focus on the impact of AI disinformation solutions on the exercise of freedom of expression, media pluralism and democracy. The authors agree with other experts that evidence of harm is still inconclusive, though abuses resulting from the 2016 US presidential election and UK referendum on leaving the European Union ('Brexit') have recently been uncovered by respectively the United States (US) Department of Justice and United Kingdom (UK) Digital, Culture, Media and Sport Parliamentary Committee.

Restrictions to freedom of expression must be provided by law, legitimate and proven necessary, and as the least restrictive means to pursue the aim. The illegality of disinformation should be proven before filtering or blocking is deemed suitable. AI is not a 'silver bullet'. Automated technologies are limited in their accuracy, especially for expression where cultural or contextual cues are necessary. Legislators should not push this difficult judgement exercise in disinformation onto online intermediaries. While many previous media law techniques are inappropriate in online social media platforms, and some of these measures were abused by governments against the spirit of media pluralism, it is imperative that legislators consider which of these measures may provide a bulwark against disinformation without the need to introduce AI-generated censorship of European citizens.

Different aspects of the disinformation problem merit different types of regulation. We note that all proposed policy solutions stress the importance of literacy and cybersecurity. Holistic approaches point to challenges within the changing media ecosystem and stress the need to address media pluralism as well. Further, in light of the European elections in May 2019, attention has focused on strategic communication and political advertising practices. The options laid out in this study are

specifically targeted at the regulation of AI to combat disinformation, but should be considered within this wider context.

In this study, Chapter 2 scopes the problem of disinformation, focusing in particular on the societal, economic and technological aspects of disinformation. It explains how disinformation differs on the internet compared to other forms of media, focusing on (a) the changing media context and (b) the economics underlying disinformation online. The chapter also explores (c) the limits to AI technologies, and introduces us to (d) a typology of self- and co-regulatory solutions, that will be used in Chapter 4 to explore policy options. Chapter 2 provides insight into the societal, economic and technological origins of disinformation. Disinformation is a complex problem. Identifying components of the problem helps in the identification of the various components of the solution. We explain how disinformation differs on the internet compared to other state and self-regulated forms of media, examining both regulatory and media contexts. In addition, to comprehend the uses of disinformation, it is necessary to understand user behaviour. The behavioural economics of disinformation are examined, notably 'filter bubbles', and nudge regulation.

Chapter 3 reviews policy and technology initiatives relevant to disinformation and illegal content online with the aim of understanding: (a) how they recommend to use technology as a solution to curb certain types of content online; and (b) what they identify as necessary safeguards to limit the impact on freedom of expression and media pluralism. At the end of the chapter, the study (c) maps the existing initiatives onto the typology of self- and co-regulatory solutions. We analyse commitments and recommendations made towards transparency in technical interventions aimed at decreasing the prevalence of disinformation. Other EU initiatives also call for the pro-active measures by intermediaries through use of AI to aid removal of illegal content. The recently proposed EU regulation on the prevention of dissemination of terrorist content online targets rapid removal terrorist content by online intermediaries. Iterations of Article 13 of the proposed copyright in the digital single market directive suggest changing intermediary liability protections with a requirement to use filtering technologies. These policy developments fit in a context, where social media platforms and search engines are increasingly scrutinised on competition grounds and are called to shoulder their responsibility in the online ecosystem.

Chapter 4 presents policy options, paying particular attention to interactions between technological solutions, freedom of expression and media pluralism. The opportunities and drawbacks of various self-regulatory to legislative options are explored.

We conclude that legislation to protect freedom of expression may be premature and potentially hazardous with regard to fundamental rights: collaboration between different stakeholder groups with public scrutiny is preferable, where effectiveness can be independently demonstrated. Most importantly, options are interdependent – where regulation is proposed, it sits atop a pyramid of activities including co-regulation, self-regulation, technical standards and individual company/NGO/academic initiatives. There is no single option to solve the problem of disinformation.

With regard to the policy options, the authors would like to make the following comments:

1 We emphasise that disinformation is best tackled through **media pluralism and literacy** initiatives, as these allow diversity of expression and choice. **Source transparency indicators** are preferable over (de)prioritisation of disinformation, and users need to be given the opportunity to understand how their search results or social media feeds are built and edit their search results/feeds where desirable.

2 We advise against regulatory action that would encourage increased use of AI for content moderation purposes, without **strong human review and appeal processes**.

3 We argue that **independent appeal and audit** of platforms' regulation of their users be introduced as soon as feasible. When technical intermediaries need to moderate

content and accounts, detailed and transparent policies, notice and appeal procedures, and regular reports are crucial. We believe this is also valid for automated removals.

4 There is scope for standardising (the basics of) notice and appeal procedures and reporting, and creating a **self- or co-regulatory multistakeholder body**, such as the UN Special Rapporteur's suggested 'social media council'. As the Special Rapporteur recommends, this multistakeholder body could, on the one hand, have competence to deal with industry-wide appeals and, on the other hand, work towards a better understanding and minimisation of the effects of AI on freedom of expression and media pluralism.

5 Lack of independent evidence or **detailed research** in this policy area means the risk of harm remains far too high for any degree of policy or regulatory certainty. **Greater transparency must be introduced** into the variety of AI and disinformation reduction techniques used by online platforms and content providers.

Table of contents

Table of figures

Table of tables

List of acronyms

ACR	Automated content recognition
Ad	Advertisement/advertising
AFP	Agence France Presse
AI	Artificial intelligence
API	Applications programme interface
BEUC	Bureau Européen des Unions de Consommateurs (European Consumer Organisation)
CJEU	Court of Justice of the European Union
DLT	Distributed ledger technology
DNS	Domain name system
DWeb	Decentralised web (technology)
EBU	European Broadcasting Union
EC	European Commission
EDRi	European digital rights
EEAS	European External Action Service
EFJ	European Federation of Journalists
EMEA	Europe, Middle East and Africa
EP	European Parliament
EPRS	European Parliament Research Service
EU	European Union
EURID	Registry manager of the .eu and .ею
GDPR	General Data Protection Regulation (EU)
GEN	Global Editors Network
GNI	Global Network Initiative
ICT	Information communication technology
IRA	Russian Internet Research Agency
ISSP	Information Society Service Provider
HADOPI	Haute Autorité pour la Diffusions des Œuvres et la Protection des Droits sur Internet (French High Authority on the Distribution of Works and the Protection of Rights on the Internet)
HLEG	High Level Expert Group (in this context: on Disinformation)
IFCN	International Fact-Checking Network
KPI	Key performance indicator
NetzDG	Netzwerkdurchsetzungsgesetz (German Network Enforcement Act)
NGO	Non-governmental organisation
OBA	Online behavioural advertising
PEGI	Pan European Game Information
P2P	Peer-to-peer
TFEU	Treaty on the Functioning of the European Union
UK	United Kingdom
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organisation
US	United States
VPN	Virtual private network

1. Introduction

There is a desire and momentum within the European Commission to tackle illegal (and undesirable) content through online intermediaries. The European Parliament elections of May 2019 give immediate impetus to European-level actions to tackle disinformation. This interdisciplinary study analyses the implications of artificial intelligence (AI) disinformation initiatives on freedom of expression, media pluralism and democracy.

The authors were tasked to formulate policy options, based on the literature, expert interviews and mapping that form the analysis undertaken.[1] The authors warn against policy optimism that proposes use of automated detection, (de)prioritisation and removal by online intermediaries without human intervention as a solution to disinformation online. When automated technologies are used, it is argued that far greater appeal and oversight mechanisms are necessary in order to minimise the negative impact of evitable inacurracies.

This study defines disinformation as 'false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit' in line with the European Commission High Level Expert Group (HLEG) report's use of the term[2]. The study distinguishes disinformation from misinformation, which refers to unintentionally false or inaccurate information,[3] Further, in parts of the study, a distinction is made between public, private, electoral, and foreign disinformation, as this is helpful to understand differences in available regulatory approaches depending on the destination and origin of the disinformation.[4]

Within machine learning techniques that are advancing towards AI, automated content recognition (ACR) technologies are textual and audio-visual analysis programmes that are trained to identify

[1] This study consists of a literature review, expert interviews and mapping of policy and technology initiatives on disinformation in the European Union. This report has been written by internet regulatory experts: a socio-legal scholar with a background in law and economics of mass communications; a media scholar with a background in internet policy processes and copyright reform; and reviewed by a computer scientist with a background in internet regulation and fundamental human rights. We suggest that this is the bare minimum of interdisciplinary expertise required to study the regulation of disinformation on social media with an adequate degree of competence.

The authors conducted ten expert interviews in the context of the study. The details can be found in Annex 1 to the study. In addition to the desk research and expert interviews, the researchers took part in several expert seminars, including: the Annenberg-Oxford Media Policy Summer Institute, Jesus College, Oxford, 7 August 2018; the Google-Oxford Internet Leadership Academy at the Oxford Internet Institute, 5 September 2018; Gikii'18 at the University of Vienna, Austria, 13-14 September 2018; the Microsoft Cloud Computing Research Consortium annual workshop, St John's College, Cambridge, 17-18 September 2018. The authors thank all interview respondent and participants for the enlightening disinformation discussions; all errors remain our own.

[2] High Level Expert Group on Fake News and Online Disinformation (2018) *Report to the European Commission on A Multi-Dimensional Approach to Disinformation*, https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation, p. 10

[3] Wardle, C. and Derakhstan, H. (2017) *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making (DGI(2017)09),* Shorenstein Center on Media, Politics and Public Policy at Harvard Kennedy School for the Council of Europe, https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking

The EU's interinstitutional terminology database IATE (Inter-Active Terminology for Europe) specifically notes that disinformation should not be confused with misinformation, defined in IATE as 'information which is wrong or misleading but not deliberately so'. See Bentzen, N. (2015) *Understanding Propaganda and Disinformation,* European Parliament Research Service At a Glance, http://www.europarl.europa.eu/RegData/etudes/ATAG/2015/571332/EPRS_ATA(2015)571332_EN.pdf

We discussed different forms of disinformation in an expert interview with the author of the aforementioned *At a Glance,* Naja Bentzen (Policy Analyst in External Policies Unit at European Parliament Research Service, 13 September 2018).

[4] Public/private explains differences between exchanges that are posted in public or private fora. Electoral/foreign are both strategic forms of political influence. We view the former as originating primarily from domestic political actors, while the latter is foreign political influence, whether government or private

potential 'bot' accounts and unusual potential disinformation material[5]. In this study, ACR refers to both the use of automated techniques in the recognition **and** the moderation of content and accounts to assist human judgement,[6] Where necessary, which of the two (recognition or moderation) is implied is specified. Moderating content at larger scale requires ACR as a supplement to human moderation (editing).[7] The shorthand 'AI' to refer to these technologies is used in the remainder of the report.

AI to detect disinformation is however prone to Type I-II errors (false negatives/positives) due to the 'difficulty of parsing multiple, complex, and possibly conflicting meanings emerging from text'.[8] If inadequate for natural language processing and audiovisual material including 'deep fakes' (fraudulent representation of individuals in video), AI does have more reported success in identifying 'bot' accounts: 'Such 'bots' foment political strife, skew online discourse, and manipulate the marketplace'.[9] Note that disinformation has been a rapidly moving target in the period of research, with new reports and policy options presented by learned experts on a daily basis throughout the third quarter of 2018. The authors analysed these reports up to 26 October 2018,[10] and the following chapters note the strengths and weaknesses of those most closely related to the study's focus on the impact of AI disinformation solutions on the exercise of freedom of expression, media pluralism and democracy.

[5] Artificial Intelligence refers to advanced forms of machine learning, generally classified as algorithmic processes powered by advanced computing techniques such as neural networks and including in particular Deep Learning. The technical literature is vast, but of relevance to this report, see Klinger, J., Mateos-Garcia, J.C., and Stathoulopoulos, K. (2018) *Deep Learning, Deep Change? Mapping the Development of the Artificial Intelligence General Purpose Technology,* DOI: http://dx.doi.org/10.2139/ssrn.3233463. See also Zuckerberg, M. (15 Nov 2018) 'A Blueprint for Content Governance and Enforcement', *Facebook Notes*, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/ stating: 'Some categories of harmful content are easier for AI to identify, and in others it takes more time to train our systems. For example, visual problems, like identifying nudity, are often easier than nuanced linguistic challenges, like hate speech'.

Bishop, C. (1995) *Neural Networks for Pattern Recognition*, Gloucestershire: Clarendon Press; Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. (2009) 'Reading Tea Leaves: How Humans Interpret Topic Models', in Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.) *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, pp. 288–96; Hillard, D., Purpura, S., and Wilkerson, J. (2008) 'Computer-Assisted Topic Classification for Mixed-Methods Social Science Research', *Journal of Information Technology & Politics 4(4)* 31–46; Monroe, B., Colaresi, M., and Quinn, K. (2008) 'Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict', *Political Analysis 16(4)* 372-403; Azevedo, L. (2018) 'Truth or Lie: Automatically Fact Checking News', in *Companion Proceedings of The Web Conference 2018 (WWW '18),* International World Wide Web Conferences Steering Committee, Geneva, Switzerland, pp. 807-811, DOI: https://doi.org/10.1145/3184558.3186567

[6] See Epstein, R. & Robertson, R.E. (2015) 'The Search Engine Manipulation Effect (SEME) and its Possible Impact on the Outcomes of Elections', *112 Proc Nat'l Acad. Sci.* E4512

[7] Klonick, K. (2018) 'Why The History Of Content Moderation Matters', *Content Moderation at Scale 2018 Essays: Techdirt*, https://www.techdirt.com/articles/20180129/21074939116/why-history-content-moderation-matters.shtml

[8] Doering, D. and Neff, G. (2018) 'Fake News as a Combative Frame: Results from a Qualitative Content Analysis of the Term's Definitions and Uses on Twitter', *4th Annual International Journal of Press/Politics Conference*, Oxford, 12 October. See also Bhaskaran, H., Harsh, M., and Pradeep, N. (2017) 'Contextualizing Fake News in Post-Truth Era: Journalism Education in India', *Asia Pacific Media Educator 27(1)* 41–50; Cohen, M. (2017) 'Fake News and Manipulated Data, the New GDPR, and the Future of Information'*, Business Information Review 34(2)* 81-85; Conroy, N, Rubin, V. and Chen, Y. (2015) 'Automatic Deception Detection: Methods for Finding Fake News', in *Proceedings of the Association for Information Science and Technology 52(1),* pp. 1–4; Rubin, Vi., Chen, Y., and Conroy, N.(2015) 'Deception Detection for News: Three Types of Fake News', in *Proceedings of the Association for Information Science and Technology*, St. Louis, MO: ASIST, pp. 1–4

[9] Lamo, M. and Calo, R. (2018) 'Regulating Bot Speech', *UCLA Law Review 2019*, http://dx.doi.org/10.2139/ssrn.3214572

[10] For October at EU level, see e.g. Christie, E.H. (2018) 'Political Subversion in the Age of Social Media', *CES Policy Brief, October*; Access Now, Civil Liberties Union For Europe, and European Digital Rights (2018) *Informing the 'Disinformation' Debate*, https://edri.org/files/online_disinformation.pdf. At national level, see e.g. UK House of Commons Select Committee on Media, Culture and Sport (2018) *infra n.35*

Evidence of disingenuous news is as old as the cuneiform tablets of Hammurabi.[11] The most recent iteration of the disinformation problem reached the European Union in the wake of claims of Russian interference in the 2016 US presidential election and the 2016 UK referendum on leaving the European Union. The problem of large-scale state-sponsored social media inaccuracy was first identified in Ukraine in 2011, when the Russian goverment was accused of deliberately faking news of political corruption.[12] Disinformation can also be economically profitable to economic actors who employ 'clickbait' tactics to lure users into reading/viewing false articles and advertisements.[13]

Not all bots are necessarily detrimental to media pluralism, with curated 'newsbots' serving a media literacy and education purpose[14]. However, curated news feeds from the major platforms, notably Google's YouTube, Facebook's social networks and the much smaller Twitter, have been criticized for continuing to promote 'popular' videos which propagate extremist views while denying wider distribution to independent content creators[15].

Telling falsities serves powerful interests, and citizens are at times unwilling or unable to discount proven untruths, due to confirmation bias, peer pressure and other media literacy factors[16]. In an expert interview, Prof. Milton Mueller expressed the following:

> 'Disinformation is a very longterm historical problem with human society. The fact that we can automate it and scale it up the way we can with social media is interesting, but I don't think it is qualitatively different from what we have seen. With the exception that it is more globalized, so foreign governments or foreign actors can partake and have access in ways that are both good and bad.'[17]

Many recent studies and reports have attempted to quantify the threat of disinformation.[18] We examined the very large amount of new evidence emerging in 2018, notably from HLEG members. HLEG Chair De Cock Buning has argued that at least in France and Italy in the period to 2018 'fake news is having a minimal direct impact. Its effect is limited mostly to groups of 'believers' seeking to reinforce their own opinions and prejudices'[19]. We agree that evidence of largr-scale harm is still

[11] Discussed in Enriques, L. (9 Oct 2017) 'Financial Supervisors and RegTech: Four Roles and Four Challenges', *Oxford University, Business Law Blog*, http://disq.us/t/2ucbsud

[12] See Sanovich, S. (2017) 'Computational Propaganda in Russia: The Origins of Digital Misinformation', *Oxford Computational Propaganda Research Project, Working Paper No. 2017(3),* http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Russia.pdf

[13] Lamb, K. (23 July 2018) 'I Felt Disgusted: Inside Indonesia's Fake Twitter Account Factories', *The Guardian,* https://www.theguardian.com/world/2018/jul/23/indonesias-fake-twitter-account-factories-jakarta-politic

[14] Harambam, J. & Helberger, N., and van Hoboken, J. (2018) 'Democratizing Algorithmic News Recommenders: How to Materialize Voice in a Technologically Saturated Media Ecosystem,' *Philosophical Transactions of The Royal Society A: Mathematical Physical and Engineering Sciences 376(2133)*, DOI 10.1098/rsta.2018.0088

[15] Rieder, B., Matamoros-Fernández, A., and Coromina, Ò. (2018) 'From Ranking Algorithms to 'Ranking Cultures': Investigating the Modulation of Visibility in YouTube Search Results', *Convergence 24(1)* 50–68, https://doi.org/10.1177/1354856517736982

[16] Fletcher, R., and Nielsen, R.K. (2018) 'Are People Incidentally Exposed to News on Social Media? A Comparative Analysis', *New Media & Society 20(7)* 2450–2468, https://doi.org/10.1177/1461444817724170

[17] Expert interview with Milton Mueller (Professor at Georgia Institute of Technology School of Public Policy; Director Internet Governance Project, 6 August 2018)

[18] Reports and studies by e.g. Pomerantsev/Weiss, Chatham House and the Legatum Institute (United Kingdom), Center for European Policy Analysis (CEPA), RAND Corporation (United States), StopFake, Words and Wars, A Guide to Russian propaganda (Ukraine), UCMC, Detector Media, Kremlin Influence Index, Kremlin Watch (Czech Republic), Deutsche Gesellschaft Auswertige Politik, Bundeszentrale fur Politische Bildung (Germany), Finnish Institute of International Affairs and StratCom Laughs (NATO).

See especially Benkler, Yochai, Robert Faris, and Hal Roberts (2018) *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*, Oxford: Oxford University Press.

[19] de Cock Buning, M. (10 Sept 2018) 'We Must Empower Citizens In The Battle Of Disinformation', *International Institute for Communications,* http://www.iicom.org/themes/governance/item/we-must-empower-citizens-in-the-battle-of-disinformation

inconclusive in Europe, though abuses resulting from the 2016 US Presidential election and UK referendum on leaving the European Union ('Brexit') have recently been uncovered by respectively the US Department of Justice and UKDigital, Culture, Media and Sport Parliamentary Committee. To show the power of targeted disinformation, consider the Centre for Media Pluralism and Media Freedom report from 2016 US elections:

> 'Facebook said in 2017 that 126 million users saw posts made by 80,000 Russian-backed accounts. On top of that, the company said at the same time that over 10 million Facebook users saw 3,000 ads, which cost about $100,000 to post. Twitter said around the same time that 36,746 inauthentic accounts automatically generated 1.4 million election-related tweets, reaching Twitter users about 288 million times. Google in 2017 said it found 18 YouTube channels associated with an influence campaign, which posted about 1,100 videos, seen more than 165,000 times.'[20]

The desire of governments and companies to filter or block content online is not new. Initiatives to tackle disinformation through technology follow preceding work seeking to counter phishing sites, malware, incitement to hatred, xenophobia, copyright infringement, child pornography, etc. The internet was built with open, unfettered communication in mind, providing exciting opportunities for freedom of expression and citizen engagement. However, not all uses are desirable in a democratic society subject to process of law.

In the European context, the EU-orchestrated Multistakeholder Forum industry self-regulatory **Code of Practice on Online Disinformation** is core to any reflection on technology-based solutions to disinformation, as it focuses on the actions of online intermediaries (social media platforms, search engines and online advertisers) to curb disinformation online.[21] It led from the HLEG report.[22] While the Code of Practice was criticised by its Sounding Board for not stipulating any measurable outcomes,[23] Professor Rasmus Kleis Nielsen argued the EU Code of Practice produced 'three potentially major accomplishments':[24]

1. Signatories including Facebook, Google, and Twitter commit to bot detection and identification by promising to 'establish clear marking systems and rules for bots to ensure their activities cannot be confused with human interactions'.
2. Submit their efforts to counter disinformation to external scrutiny by independent third party: 'an annual account of their work to counter Disinformation in the form of a publicly available report reviewable by a third party'.
3. A joint, collaborative effort based on shared commitments from relevant stakeholders including researchers, where signatories promise not to 'prohibit or discourage good faith research into Disinformation and political advertising on their platforms'.[25]

In Chapter 3 commitments and recommendations made towards transparency in technical interventions aimed at decreasing the prevalence of disinfomation are analysed. Other EU initiatives

[20] Hautala, L. (16 Oct 2018) 'Hackers, Trolls and the Fight over Your Vote in the 2018 Midterm Elections', *CNET*, https://www.cnet.com/news/hackers-trolls-and-the-fight-over-your-vote-in-the-2018-midterm-elections/

[21] EU Code of Practice on Disinformation (2018) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[22] High Level Expert Group on Fake News and Online Disinformation (2018) *Report to the European Commission on A Multi-Dimensional Approach to Disinformation*, https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

[23] European Commission (26 September 2018) *Code of Practice on Disinformation,* Press Release, https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[24] Nielsen, R.K. (24 Oct 2018) 'Misinformation: Public Perceptions and Practical Responses', *Misinfocon London*, hosted by the Mozilla Foundation and Hacks/Hackers, https://www.slideshare.net/RasmusKleisNielsen/misinformation-public-perceptions-and-practical-responses/1

[25] Nielsen, R.K. (26 Sept 2018) *Disinformation Twitter Thread*, https://twitter.com/rasmus_kleis/status/1045027450567217153

also call for the pro-active measures by intermediaries through use of AI to aid removal of illegal content. The recently proposed EU regulation on the prevention of dissemination of terrorist content online[26] targets rapid removal terrorist content by online intermediaries. Iterations of Article 13 of the proposed copyright in the digital single market directive[27] suggest changing intermediary liability protections with a requirement to use filtering technologies. These policy developments fit in a context, where social media platforms and search engines are increasingly scrutinised on competition grounds[28] and are called to shoulder their responsibility in the online ecosystem.[29]

This study analyses the causes of disinformation in an online context and the responses that have been formulated from a technology perspective. It examins the effects that AI-enhanced disinformation initiatives have on freedom of expression, media pluralism and the exercise of democracy, from the wider lens of tackling illegal content online and concerns to request proactive (automated) measures of online intermediaries,[30] thus enabling them to become censors of free expression. In line with the recommendations of the UN Special Rapporteur on Freedom of Opinion and Expression, the study calls for assessments of the impact of technology-based solutions on human rights in general, and freedom of expression and media pluralism in particular.[31]

Restrictions to freedom of expression must be provided by law, legitimate[32] and proven necessary and as the least restrictive means to pursue the aim.[33] The illegality of disinformation should be

[26] Proposed EU Regulation on Prevention of Dissemination of Terrorist Content Online (COM(2018) 640 final - 2018/0331 (COD)) https://ec.europa.eu/commission/sites/beta-political/files/soteu2018-preventing-terrorist-content-online-regulation-640_en.pdf

[27] Proposed EU Directive on Copyright in the Digital Single Market (COM(2016) 593 final – 2016/0280(COD)) https://ec.europa.eu/digital-single-market/en/news/proposal-directive-european-parliament-and-council-copyright-digital-single-market

[28] For a scholarly overview and discussion of ongoing platform and search engine competition cases, see Mandrescu, D. (2017) 'Applying EU Competition Law to Online Platforms: The Road Ahead – Part I', *Competition Law Review 38(8)* 353-365*;* Mandrescu, D. (2017) 'Applying EU Competition Law to Online Platforms: The Road Ahead – Part II', *competition Law Review 38(9)* 410-422. For an earlier call to co-regulation, see Marsden, C. (2012) 'internet Co-Regulation and Constitutionalism: Towards European Judicial Review' *International Review of Law, Computers & Technology 26(2*-3) 215-216

[29] Within the copyright context, see for instance Angelopoulos, C., and Quintais, J.P. (30 August 2018), 'Fixing Copyright Reforms: How to Address Online Infringement and Bridge the Value Gap', *Kluwer Copyright Blog,* http://copyrightblog.kluweriplaw.com/2018/08/30/fixing-copyright-reform-address-online-infringement-bridge-value-gap/; Stephens, H. (26 March 2018) *Internet Platforms: It's Time to Step Up and Accepty Your Responsibility (Or Be Held Accountable),* https://hughstephensblog.net/2018/03/26/internet-platforms-its-time-to-step-up-and-accept-your-responsibility-or-be-held-accountable/

[30] ACR techniques became newsworthy in 2016 with the development of eGLYPH for removal of terrorist content: see The Verge (2016) *Automated Systems Fight ISIS Propaganda, But At What Cost?,* https://www.theverge.com/2016/9/6/12811680/isis-propaganda-algorithm-facebook-twitter-google

[31] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) *Report to the United Nations Human Rights Council on A Human Rights Approach to Platform Content Regulation,* A/HRC/38/35, https://freedex.org/wp-content/blogs.dir/2015/files/2018/05/G1809672.pdf

See also UN Special Rapporteur on Freedom of Opinion and Expression et. al. (2017) *Joint Declaration on Freedom of Expression and 'Fake News,' Disinformation and Propaganda,* UN Document FOM.GAL/3/17, https://www.osce.org/fom/302796?download=true; Access Now, Civil Liberties Union for Europe, and European Digital Rights (EDRi, 2018) *Informing the 'Disinformation' Debate*, https://edri.org/files/online_disinformation.pdf

We discussed the implications of technology-driven solutions for freedom of expression in expert interviews with David Kaye (UN Special Rapporteur on Freedom of Opinion and Expression, 3 July 2018) and Joe McNamee (Executive Director at EDRi, 6 Sept 2018)

[32] Pursue one of the purposes set out in Article 19.3 International Covenant on Civil and Political Rights, i.e. to protect the rights or reputations of others; to protect national security, public order or public health or morals.

[33] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) *Report to the United Nations Human Rights Council on A Human Rights Approach to Platform Content Regulation,* A/HRC/38/35, https://freedex.org/wp-content/blogs.dir/2015/files/2018/05/G1809672.pdf, pars 44-48

proven before filtering or blocking is deemed suitable. AI is not a 'silver bullet'. Automated technologies are limited in their accuracy, especially for expression where cultural or contextual cues are necessary. The illegality of terrorist or child abuse content is far easier to determine than the boundaries of political speech or originality of derivative (copyrighted) works. We should not push this difficult judgement exercise in disinformation onto online intermediaries.

A final initial point needs making about media pluralism and fundamental rights. If the socio-technical balance is trending towards greater disinformation, a lack of policy intervention is not neutral, but erodes protection for fundamental rights to information and expression. While there remains insufficient research to authoritatively conclude that this is the case, it is notable that after previous democratic crises involving media pluralism and new technologies (radio, television, cable and satellite), parliaments passed legislation to increase media pluralism by for instance funding new sources of trusted local information (notably public service broadcasters), authorising new licencees to provide broader perspectives, abolishing mandatory licensing of newspapers or even granting postage tax relief for registered publishers, and introducing media ownership laws to prevent existing monopolists extending their reach into new media.[34] The UK House of Commons Select Committee on Media, Culture and Sport Interim Report on Disinformation and 'Fake News' states that '[i]n this rapidly changing digital world, our existing legal framework is no longer fit for purpose'.[35]

While many previous media law techniques are inappropriate in online social media platforms, and some of these measures were abused by governments against the spirit of media pluralism, it is imperative that legislators consider which of these measures may provide a bulwark against disinformation without the need to introduce AI-generated censorship of European citizens.

Different aspects of the disinformation problem merit different types of regulation. We note that all proposed policy solutions stress the importance of literacy and cybersecurity. Holistic approaches point to challenges within the changing media ecosystem and stress the need to address media pluralism as well. Further, in light of the European elections in May 2019, attention has turned to strategic communication and political advertising practices. The options laid out are specifically targeted at the regulation of AI to combat disinformation, but should be considered within this wider context.

In this study, Chapter 2 further scopes the problem of disinformation, focusing in particular on the societal, economic and technological aspects of disinformation. It explains how disinformation differs on the internet compared to other forms of media, focusing on (a) the changing media context and (b) the economics underlying disinformation online. The chapter also explores (c) the limits to AI technologies, and introduces (d) a typology of self- and co-regulatory solutions, that will be used in Chapter 4 to explore policy options.

Chapter 3 reviews policy and technology initiatives relevant to disinformation and illegal content online with the aim to understand (a) how they recommend to use technology as a solution to curb certain types of content online and (b) what they identify as necessary safeguards to limit the impact on freedom of expression and media pluralism. At the end of the chapter, (c) the existing initiatives are mapped onto the typology of self- and co-regulatory solutions.

Chapter 4 presents policy options, paying particular attention to interactions between technological solutions, freedom of expression and media pluralism. The opportunities and drawbacks of various self-regulatory to legislative options are explored.

[34] See e.g. C-288/89 (judgment of 25 July 1991, Stichting Collectieve Antennevoorziening Gouda and others [1991] ECR I-4007); Protocol on the System of Public Broadcasting in the Member States annexed to the EC Treaty; Council Directive 89/552/EEC on the Coordination of Certain Provisions Laid Down by Law, Regulation or Administrative Action in Member States concerning the Pursuit of Television Broadcasting Activities (particularly its seventeenth recital)

[35] UK House of Commons Select Committee on Media, Culture and Sport (2018) *Interim Report on Disinformation and 'Fake News'*, https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/36302.htm

The authors advise against regulatory action that would encourage increased use of AI for content moderation purposes, without strong human review and appeal processes. The study emphasises that disinformation is best tackled through media pluralism and literacy initiatives, as these allow diversity of expression and choice. Source transparency indicators are preferable over (de)prioritisation of disinformation, and users need to be given the opportunity to understand how their search results or social media feeds are built and make changes where desirable.

When technical intermediaries need to moderate content and accounts, detailed and transparent policies, notice and appeal procedures, and regular reports are crucial. The authors believe this is valid for automated removals as well. There is scope for standardising (the basics of) notice and appeal procedures and reporting, and creating a self-regulatory multistakeholder body that, on the one hand, has competence to deal with industry-wide appeals and, on the other hand, continually works towards a better understanding and minimisation of the effects of content moderation on freedom of expression and media pluralism.

2. Scoping the online disinformation problem

Chapter 2 provides insight into the societal, economic and technological origins of disinformation. Disinformation is a complex problem. Identifying components of the problem helps in the identification of the various components of the solution.

Section 2.1 explains how disinformation differs on the internet compared to other state and self-regulated forms of media, examining both regulatory and media contexts.

To comprehend the uses of disinformation, it is necessary to understand user behaviour. The behavioural economics of disinformation are briefly examined in Section 2.2, notably filter bubbles, and nudge regulation.

This leads to brief discussion of AI technology and disinformation in Section 2.3.

To contextualise the discussion within regulatory policy, the chapter concludes in Section 2.4 by classifying self- and co-regulatory solutions, as a prelude to examining the forms of disinformation regulation in Chapter 3.

2.1. Freedom of expression, media pluralism and disinformation

There has been little academic legal research into the challenges to media pluralism from disinformation online[36], though significant research into the effect of increased legislative provisions for website blocking on freedom of expression exists[37]. 'Fake news' falls within a grey area of political expression because it encompasses both mis- and disinformation. The term is abused to label news or opinion disfavorable to one's position as 'fake'. At the same time online intermediaries (are pressured to) moderate content and accounts[38], which, considering the current lack of strong safeguards, can constitute private censorship. It is challenging to tackle disinformation, while protecting fundamental rights including media pluralism, data protection and freedom of expression[39].

There is nothing novel about the phenomenon, except that the power of the internet as a medium of reproduction of conspiracy theory lends itself to amplification of dis/misinformation (scale and scope effects)[40]. Hillary Clinton expressed forthright views about disinformation by her opponents

[36] Some literature addresses disinformation within the context of platform regulation. Syed, N. (2017) 'Real Talk About Fake News: Towards a Better Theory for Platform Governance', *Yale Law Journal 127(Forum)* 337-357. See also Georgetown Technology Law Review 2 (2) Special issue on Platform Regulation, at https://www.georgetownlawtechreview.org/issues/volume-2-issue-2/ which has 12 articles analyzing fake news/disinformation. See also Sanovich, Sergey (2017) 'Computational Propaganda in Russia: the Origins of Digital Misinformation' Woolley, S., and Howard, P.N. (Eds) Working Paper No.2017.3, *University of Oxford, UK: Project on Computational Propaganda*, http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Russia.pdf; Burshtein, S. (2017) 'The True Story on Fake News', *Intellectual Property Journal 29(3)*. See also Hurst, A. (2017) 'Fake News: Striking a Balance between Regulation and Responsibility,' *Society of Computers & Law, August/September* https://www.scl.org/articles/8961-fake-news-striking-a-balance-between-regulation-and-responsibility

[37] O'Leary, S. (2018) 'Balancing Rights in a Digital Age,' *Irish Jurist 59,* 59; Geiger, C. & Izyumenko, E. (2016) 'The Role of Human Rights in Copyright Enforcement Online: Elaborating a Legal Framework for Website Blocking', *American University International Law Review 32(1),* 43; Mac Síthigh, D. (2008) 'The Mass Age of Internet Law', *Information & Communications Technology Law 17(2),* 79-94.

[38] Such as deprioritising, blocking, removing/suspending content and accounts.

[39] Brown, I. (2013) *Online Freedom of Expression, Association, Assembly and the Media in Europe*, Council of Europe MCM(2013)007, Strasbourg: Council of Europe; Brown, I. (2013) *Transparency to Protect Internet Freedom: a Shared Commitment*, Strasbourg: Council of Europe; Korff, D. with Brown, I. (2013) *The Use of the Internet & Related Services, Private Life & Data Protection: Trends & Technologies, Threats & Implications,* Council of Europe T-PD(2013)07, Strasbourg: Council of Europe.

[40] While examples of such conspiracy include the violent extremist misogyny of the 'Gamergate trolls' and ISIS/Daesh devotees, the ideological organisation of such groups that lead to terror may be better indicated by mental illness and social marginalization than online discussion itself.

in her failed 2016 Presidential campaign, stating: 'The [I]nternet has given them a great advantage because they are able to sow discontent, divisiveness and false information incredibly easily'[41].

The UN Special Rapporteur on Freedom of Opinion and Expression, David Kaye, expressed concerns that governments – for instance Malaysia or Bangladesh – have passed laws that hamper democratic and legitimate free speech, associating opposition speech with disinformation[42]. The illegality of such content often is not clear cut[43]. Countering such hate speech online is a continued task of cybercrime policing, and is not the focus of this report.

Media pluralism policy has permitted differential regulation of media in democratic politics in previous eras. Broadcast regulation, advertising self-regulation, data protection via independent agency, and newspaper self-regulation, have provided a combined regulatory safety net using different techniques as part of a pyramid of regulation. Social media joins a long list of mass communications media that can serve the public sphere in allowing European citizens to communicate, share and participate in society, economy and politics. During our interview, Head of Public Policy for Belgium at Twitter, Stephen Turner, argued that Twitter aims to 'encourage[e] more open, civil dialogue and discourse on the platform, making room for multiple voices, including around political conversations.' He added that ' [t]he #blacklivesmatter movement is an example of how the platform can effectively elevate voices that may not otherwise have been heard through traditional media channels.'[44] There is a long tradition of literature on the public service value of media, some of which has assessed the threat to the late twentieth century tradition from the new media online[45].

Pluralism was an important European discussion in media ownership in the mid-1990s, with the European Commission Green Paper on the topic of 1995 leading to extensive debate at national level and the development of the Media Pluralism Monitor. The co-existence of online platforms of various sizes, including blogging, and traditional media, such as newspapers, means a more pluralistic media environment has to some extent come to pass, with new news platforms emerging. However, a 1999 Council of Europe report on media pluralism warned that the role of trusted information needed strengthening in digital media via some form of due prominence on internet news portals[46].

[41] Wylie, K. (9 October 2018) 'Hillary Clinton Sttacks Putin over Brexit as She Claims Democracy is 'Under Siege'', *The Independent,* https://www.independent.co.uk/news/world/americas/hillary-clinton-vladimir-putin-brexit-democracy-under-siege-a8575001.html

[42] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) *Malaysia Reply of 18 June,* https://www.ohchr.org/Documents/Issues/Opinion/Legislation/ReplyMalaysiaOL.pdf

We discussed Malaysia's disinformation measures in an expert interview with David Kaye (UN Special Rapporteur on Freedom of Opinion and Expression, 3 July 2018).

[43] Brown, I. and Korff, D. (2012) *Digital Freedoms in International Law*, Global Network Initiative.

[44] Expert interview with Stephen Turner (Head of Public Policy for Belgium at Twitter, 25 September 2018)

[45] Akdeniz, Y. (2011) '*Freedom of Expression on the Internet: Study of Legal Provisions and Practices related to Freedom of Expression, the Free Flow of Information and Media Pluralism on the Internet in OSCE Participating States,* Vienna: Office of the Representative on Freedom of the Media, Organisation for Security and Co-operation in Europe, http://www.osce.org/fom/80723. Gibbons, T. (2000) 'Pluralism, Guidance and the New Media', in C. Marsden (Ed.) *Regulating the Global Information Society*, Abingdon: Routledge, pp. 304-315. Marsden, C. (2000) 'Not So Special? Merging Media Pluralism with Competition and Industrial Policy', *Info 2(1)* 9-15.

[46] Marsden, C. (1999) *Pluralism in Multi-Channel Digital Communications: Suggestions for Regulatory Scrutiny,* MM-S-PL 1999-12, Study prepared on behalf of the Committee of Specialists on Media Pluralism, Directorate of Human Rights Strasbourg: Council of Europe, def 2, at section 5. A European Parliament report in 2001 explained that the Western intelligence agencies were developing surveillance capabilities of the internet as well as traditional telecommunications, a warning borne out by the Snowden revelations in 2013: European Parliament (2001) *Final Report on the Existence of a Global System for the Interception of Private and Commercial Communications (ECHELON interception system)*, Temporary Committee on the ECHELON.

As news organisations struggle to make the transition to the online environment, paid newspaper readership has drastically decreased (by 65% in the United Kingdom for instance[47]) and working conditions across the media sector have deteriorated. Director of the European Federation of Journalists (EJF), Renate Schroeder noted:

> *'There is a decline in good-working conditions, an increase in outsourcing of real core journalistic work, and there have been dismissals. Newsrooms have been squeezed, newsrooms have been merged. The impact of media concentration has been huge.'* [48]

Major contributory elements are the almost exclusive reliance on advertising for online business models and the disconnection between publishing and advertising led by major online intermediaries.[49] Online advertising encourages use of 'clickbait' techniques, which consequently has led to fears that disinformation can trade on 'clickbait' popularity amongst casual readers (this will be addressed in the next Section 2.2)[50]. Newspapers' declining readership, along with reliance on advertising from dominant social media platforms, advertising platforms and search engines, has led to fears that the eroding profitability of print newspapers will not be replaced by online revenues. As Independent Tech Policy and Digital Rights Reporter, Jennifer Baker explained to us:

> *'You can't imagine any other industry where they would say 'this is our product, news is our product, and we will give it away for free'. Retroactively some are going behind paywalls. But the truth is that they became hooked on advertising. And what gets advertising? Stories that are clickbait. If you create stories to make more clicks, you get more advertising. It's a bit like shutting the door after the horse has bolted. This data-driven advertising is an intrinsic part of the problem.'*[51]

The loss of journalists has led to problems for traditional news organisations in fact-checking both mis- and disinformation. The Atlantic Council's Digital Forensic Research Lab explains that: '[a]ccusing an outlet of deliberately presenting false facts is a serious act; doing so irresponsibly, without due levels of evidence, does a grave disservice both to the target of the accusation and to the broader concepts of accuracy and empirical evidence'[52]. This is a problem that confronts many news organisations, as '32 of 33 major American news outlets published stories with a tweet embedded from an [Russian government aponsored] Internet Research Agency troll account' and these accounts are becoming more sophisticated and harder to track with AI or human agency: 'the goal of these efforts isn't necessarily to deceive via a single post or tweet but to create the illusion of a groundswell of sentiment'[53]. The context is vital to understand that platforms and news organisations can make better efforts to control obvious disinformation, but that it is not a problem that can be eliminated entirely.

[47] Tobitt, C. (13 September 2018) 'National Newspaper ABCs: Free Evening Standard and Metro Only UK Papers to See Circulation Growth in August', *Press Gazette,* https://www.pressgazette.co.uk/national-abcs-free-evening-standard-only-uk-paper-to-see-circulation-growth-in-august/

[48] Expert interview with Renate Schroeder (Director at European Federation of Journalists – EFJ, 7 Sept 2018)

[49] It should be noted that some examples of successful online paywalls exist, such as The Wall Street Journal and The Financial Times. The Guardian is also pioneering a membership-support model.

[50] The infamous example is the Macedonian 'clickbait factory' serving disinformation in the 2016 US Presidential election: Osnos, E., Remnick, D., and Yaffa, J. (6 March 2017) 'Trump, Putin, and the New Cold War: What Lay Behind Russia's Interference in the 2016 Election—And What Lies Ahead?', *The New Yorker*, https://www.newyorker.com/magazine/2017/03/06/trump-putin-and-the-new-cold-war

[51] Expert interview with Jennifer Baker (Independent Tech Policy and Digital Rights Reporter, 13 September 2018)

[52] DFRLab (2018) 'Fake News: Defining and Defeating Real Techniques for Identifying Fake News and Disinformation', *Medium,* https://medium.com/dfrlab/fake-news-defining-and-defeating-43830a2ab0af?_branch_match_id=553166123622124243

[53] Glaser, A. (23 Oct 2018) 'Facebook Will Never Run Out of Moles to Whack: The Latest Disclosure of Russian Election Meddling Reveals the Limits of Social Media's New Dedication to Fighting False News', *Slate,* https://slate.com/technology/2018/10/project-lakhta-facebook-russia-election-meddling-midterms.html

Context helps us understand the growth and extent of disinformation[54]. Disinformation in industrial societies has been spread by both low tech and high tech means, and trust in media varies widely by medium. Recent Pew research research demonstrated what Professor Eli Noam wrote in 2001:

> *'It is easy to romanticize the past of democracy as Athenian debates in front of an involved citizenry, and to believe that its return by electronic means is nigh. A quick look to in the rear-view mirror, to radio and then TV, is sobering.Here, too, the then new media were heralded as harbingers of a new and improved political dialogue.But the reality of those media has been is one of cacophony, fragmentation, increasing cost, and declining value of 'hard' information.'*

He explained that: 'precisely because the internet is powerful and revolutionary, it also affects, and even destroys, all traditional institutions - including democracy. To deny this potential is to invite a backlash when the ignored problems eventually emerge'[55]. In 2018, it is arguably the time for a backlash against disinformation, but that policy response must be tempered by study of the history of mass communications regulation across Western Europe, public service television news is considered most trustworthy (although it should be noted that trust in media is nationally divergent across Europe). Citizens with populist views are less likely to trust public news organisations.[56] Disinformation fits within this wider context of a changing media landscape, business models and journalistic practices.

2.2. Behavioural economics, filter bubbles and nudges

2.2.1. Behavioural economics and disinformation

Behavioural or 'nudge' regulation has become a favoured 'light touch' regulatory technique in the last decade. The use of behavioural psychology insights to observe changes in the 'bounded rational' choices of consumers is commonplace in the online environment. Nudging was so familiar to internet regulatory scholars in the late 1990s that it came to be termed the leading example of the 'new Chicago School'[57], recognising imperfect information, bounded rationality and thus less than optimal user responses to competition remedies, driven by insights from the internet's architecture. Internet usage allows us to investigate the impact of extremely large nudges that operate in real time – what Yeung has described as the 'hyper nudge'[58]. The mass adoption by users of new products and services online (for instance the Google Chrome browser) can shape regulatory outcomes and private enforcement very substantially – whether that be by adopting new platforms, sharing data (including disinformation) in new ways, or blocking adverts (and reporting bot accounts) to make surfing the internet easier.

Motivations for the production and distribution of disinformation are an important and often overlooked feature of research into the phenomenon. Just as the internet enables, but did not create memes, so it also cultivates community/tribal understandings of the world, through inventions such

[54] For a graphical account, see Knight Foundation (2018) *Misinformation in Graphics*, https://www.knightfoundation.org/features/misinfo/

[55] Noam, E. (2001) *Will the Internet Be Bad for Democracy?*, Columbia Institute for Tele Information, New York, http://www.citi.columbia.edu/elinoam/articles/int_bad_dem.htm

[56] Matsa, K.E. (8 June 2018) 'Across Western Europe, Public News Media are Widely Used and Trusted Sources of News', Pew Research Centre, http://www.pewresearch.org/fact-tank/2018/06/08/western-europe-public-news-media-widely-used-and-trusted/; See also Pew Research Centre (27 Sept 2018) *Europe News Platforms, Topline Questionnaire*, Press Release, http://www.pewresearch.org/wp-content/uploads/2018/09/FT_18.09.27_EuropeNewsPlatforms_Topline.pdf

[57] Richardson, M. and Hadfield, G. (1999) *The Second Wave of Law and Economics*, Sydney: Federation Press; Lessig, L. (1998) 'The New Chicago School', *the Journal of Legal Studies* 27(2) 661-691. For comment, see Tushnet, M. (1998) 'Everything Old is New Again: Early Reflections on the New Chicago School', *Wisconsin Law Review 579*.

[58] Yeung, K. (2017) 'Hypernudge': Big Data as a Mode of Regulation by Design' *Information, Communication & Society* 20(1) pp.118-136

as creation myths. These are not new, and there has been a great deal of research into these examples of disinformation and how they are influenced by different media of communication[59]. *Which heuristics tend to lead to the spread of inaccurate and sensational news? What are the incentives of different actors in the media ecosystem to promote more or less accurate information?*[60]

Attempts to analyse disinformation which assume that all citizens are searching for an objective truth online will provide a partial picture obscured by confirmation bias and group membership signalling. While there is a great deal of controversy surrounding 'filter bubbles' and the extent to which individual news consumption is influenced by partisan politics[61], a 2017 Harvard study provides solid evidence that the highly populist and often disinformative source Breitbart was disproportionately popular on news consumption for Trump-leaning voters[62]. Marwick states 'sharers are principally concerned with signaling membership in particular communities rather than with the truth or falsity of the items they choose to share'[63].

Understanding online information practices is critical before creating law and policy[64]. As interviewee Jennifer Baker states, 'filter bubbles and echo chambers have always existed. What we see is at the moment is that the vectors, the volumes and the velocity with which new ideas come to you are hugely increased by social media use'.[65] The 'truth may be out there', but users have not all wanted only to share it, hence the multiplicity of political, religious and social viewpoints that exist in each European society. This was true in 'cheap talk' in drinking groups in pubs, canteen culture in the police, housekeepers' gossip, gentlemen's dining clubs, and workers' groups throughout history. Any attempt to signal towards 'truth' in news will have only limited effects in society[66].

2.2.2. Online behavioural advertising, elections and disinformation

Evidence has remained a problem for independent researchers trying to examine the AI practices of the platforms responsible for social media and political messaging, and the disinformation sources themselves such as the Russian Internet Research Agency (IRA) or the Leave.EU campaign[67]. Amongst platforms, Facebook responded to the Cambridge Analytica scandal in part by blocking much of what remained of third-party researcher access to their processes and their other platforms Instagram and WhatsApp. This made it even more difficult to gather independent evidence, with only around 30 approved academics allowed access to their Applications Programme Interface (API)

[59] Russell, N.W. (2016) *The Digital Difference: Media Technology and the Theory of Communication Effects,* Cambridge, MA: Harvard University Press. See also Periñán, B., 'The Origin of Privacy as a Legal Value: a Reflection on Roman and English Law', *American Journal of Legal History 52(1)*

[60] See e.g. Davis, E. (2017) *Why We Have Reached Peak Bullshit and What We Can Do About It,* Little: Brown.

[61] Pariser E, (2011) *The Filter Bubble: What the Internet is Hiding From You*, London: Penguin Press.

[62] Benkler, Y. et al (2017) *Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election,* Harvard Berkman-Klein Center, https://cyber.harvard.edu/publications/2017/08/mediacloud

[63] Marwick, A.E. (2018) 'Why Do People Share Fake News? A Sociotechnical Model of Media Effects', *Georgetown Technology Law Review 2(2)*, at 474, https://www.georgetownlawtechreview.org/issues/volume-2-issue-2/

[64] Narayanan, V. et al. (2018) *'Polarization, Partisanship and Junk News Consumption over Social Media in the US', COMPROP Data Memo 2018(1),* Computational Propaganda Project, http://comprop.oii.ox.ac.uk/wpcontent/uploads/sites/93/2018/02/Polarization-Partisanship-JunkNews.pdf

[65] Expert interview with Jennifer Baker (Independent Tech Policy and Digital Rights Reporter, 13 September 2018)

In the expert interviews, Jennifer Baker (Independent Tech Policy and Digital Rights Reporter, 13 September 2018) and Monique Goyens (Director-General at European Consumer Organisation – BEUC, 31 August 2018) raised tracking and micro-targeting practices as primary issues to tackle when considering disinformation.

[66] Marwick, A. & Lewis, R. (2017) 'Media Manipulation and Disinformation Online', *Data & Society*, https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformation

[67] For an example of the activities of the latter, see https://leave.eu/support-our-research-into-remainer-election-spending/

specifically to research disinformation[68]. Access to YouTube and other Alphabet/Google properties and their attempts at more ethical design[69], and to Twitter, is easier but remains challenging.

Disinformation is a new version of an existing regulated problem, with potentially dramatic negative effects on democracy and media pluralism. Media pluralism and literacy goes hand in hand with any technological intervention. In tackling disinformation[70], the effectiveness of the technological measures needs to be considered, alongside awareness raising of individual and social responsibility for the provision and appreciation of verifiable truthful content. This should be carried out by independent platforms rather than a single central authority, as examined in Chapter 3.

There is an overwhelming need for greater independent research into the processes that online platforms have put in place, including those to combat disinformation. Some greater transparency has been promised through the EU Code of Practice on Disinformation. Facebook refused UK parliamentary requests to allow CEO Mark Zuckerberg to give evidence to its fake news inquiry, and it was recently fined the maximum permitted by the UK data protection authority for its various infractions associated with the Cambridge Analytica investigation in the UK (Ireland's investigation is ongoing)[71]. Neither the Russian IRA nor the Leave.eu organisation cooperated with the UK parliamentary inquiry, with the latter found to have breached electoral law[72].

Disinformation is also part of the debate on Online Behavioural Advertising (OBA) more generally, and the specific regulatory tools provided in the General Data Protection Regulation (GDPR)[73] and the proposed Electronic Privacy Regulation[74]. As to the type of regulation required and its venue, there is a large gap between, for instance, time-limited electoral campaign regulation and advertising self-regulation.[75] To fill this gap requires a more holistic view of regulation, while at the same time researching the type of disinformation targeted. National parliament reports, civil society reports, Data Protection Regulator responses to the Cambridge Analytica scandal, and other expert reports have continually drawn attention to the fact that the use of micro-targeting is a particularly pernicious and effective form of digital advertising, and that the catalogued techniques used for disinformation are a politically sensitive example of the wider use of OBA. Whether one considers this 'digital canvassing'[76] or simply effective commercial communications, it is a problem that has dogged internet use since the first spam email was sent.

[68] Hill, R. (25 April 2018) 'Academics: Shutting down Facebook API Damages Research, Oversight, Competition. Open Letter Throws Heavy Shade on Social Network's Research Initiative', *The Register*, https://www.theregister.co.uk/2018/04/25/shutting_down_facebook_api_damages_research_oversight_competition_warn_academics/

[69] Article 19 (14 June 2018) *Google: New Guiding Principles on AI Show Progress But Still Fall Short on Human Rights Protections*, https://www.article19.org/resources/google-new-guiding-principles-on-ai-show-progress-but-still-fall-short-on-human-rights-protections/

[70] And other undesirable uses of online communication, as the history of electoral and defamation reform shows, Noam (2001) supra.

[71] UK Information Commissioner's Office (25 Oct 2018) *ICO Issues Maximum £500,000 Fine to Facebook for Failing to Protect Users' Personal Information,* News, https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2018/10/facebook-issued-with-maximum-500-000-fine/

[72] See UK House of Commons (2018) supra n.35

[73] EU Regulation 2016/679 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data, and repealing Directive 95/46/EC (General Data Protection Regulation) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

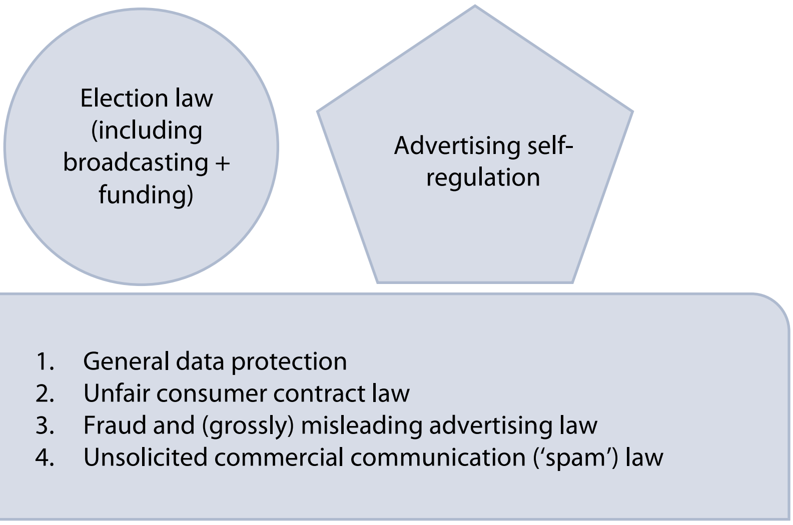
[74] Proposed EU Regulation concerning the Respect for Private Life and the Protection of Personal Data in Electronic Communications and Repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications) (COM(2017)10 final – 2017/0003(COD)) https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-privacy-and-electronic-communications

[75] McStay, Andrew (2011) *The Mood of Information: A Critique of Online Behavioural Advertising*, London: A&C Black.

[76] Individual activist to elector lobbying is the oldest established form of political advertising, dating to at least Anicent Athenian times, and discussed on numerous occasions in the Old Testament of the Bible, and the crucifixion story

The issue of OBA extends beyond disinformation on social media, to include the issue of whether online political advertising should be permitted at all, by domestic or foreign organisations. There is a vast gulf between two types of media regulation. Commercial speech in newspapers and advertising self-regulation in printed and online publications, are bound to traditional freedom of expression restrictions, but are primarily self-regulated[77]. Broadcast content, together with electoral spending by political parties, are rigidly regulated by both broadcast and electoral regulators. Disinformation thus exposes a challenging divergence of regulations in fairness requirements. The regulatory solutions can be mapped between the two extremes, which are represent in the figure below.

Figure 2.1 Disinformation solutions from electoral law to advertising self-regulation


Election law (including broadcasting + funding)
Advertising self-regulation
1. General data protection
2. Unfair consumer contract law
3. Fraud and (grossly) misleading advertising law
4. Unsolicited commercial communication ('spam') law

Generic regulation applies to varying degrees, especially in the EU context where privacy and consumer law is well developed, especially for online transactions with the e-Privacy and Distance Selling Directives.[78] EU *sui generis* privacy law also applies, notably the General Data Protection Regulation (GDPR). The recently constituted Expert Group to the EU Observatory on the Online Platform Economy may assist further research into the wider implications of OBA use by platforms[79].

Problems of jurisdiction also arise. In the Cambridge Analytica-Facebook Brexit referendum case, the UK Electoral Commission needed the cooperation of almost entirely non-domestic companies. It also dealt with targeted social media profiling rather than standard electoral expenditure in qualifying media such as outdoor posters or newspaper adverts. UK electoral law does not permit any type of television or radio advertising by political parties or other concerned parties, except for short explicitly allocated party political broadcast slots.[80]

concerns Barabbas supporters lobbying the crowd to support his release: AALEP (19 April 2011) *Biblical Accounts of Lobbying*, http://www.aalep.eu/biblical-accounts-lobbying

[77] Such as the protection of privacy and the prohibition to defamation, and incitement hatred or violence.

[78] Directive 2002/58/EC concerning the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector (Directive on Privacy and Electronic Communications) https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:HTML

Directive 97/7/EC on the Protection of Consumers in Respect of Distance Contracts, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A31997L0007

[79] European Commission (2018) *Expert Group to the EU Observatory on the Online Platform Economy, Team responsible E-Commerce and Platforms (Unit F.2)*, https://ec.europa.eu/digital-single-market/en/expert-group-eu-observatory-online-platform-economy

[80] Separately, the Information Commissioner examined whether Cambridge Analytica abused the consent of the Facebook and other identities which it targeted in its search for gullible undecided voters, and issued instructions to Facebook to cease any such cooperation with third parties.

There is an evident disinformation problem with OBA for political purposes, whether in elections or outside, and the apparently simple legislative reform (though complex in its implementation) of subjecting online advertising to the same electoral processes and laws as offline equivalents needs to be analysed, taking into account the explicitly and uniquely targeted nature of OBA. The UK Parliamentary Disinformation Committee's Interim Report noted that it 'might be difficult to advocate a total ban on micro-targeting political advertising online, a preferable alternative could be to limit the amount of "lookalike micro-targeting"' to stop similar messages to small groups of voters, as opposed to broadcast-type wide distribution messages[81]. The UK trade body for advertisers called for 'a total ban on micro-targeting political advertising online, with a minimum limit on the number of voters who are sent individual political messages'[82]. Facebook told the Committee in June 2018:

> *'We are heavily investing in advanced technologies and machine learning [AI] to better assess advertisements that fall into specific categories (like political and issues adverts) so we can identify and enforce policies and tools that may apply'[83].*

It is clear that measuring the use of OBA, including by use of AI, is important to scoping the disinformation problem. Technology has thus far been deployed to increase the micro-targeting of voters, with relatively little AI use to stop non-voters (i.e. foreign accounts and bot accounts) from influencing voters.

The table below maps some solutions in the case of disinformation in advertising.

Table 2.1 Regulatory options for tackling disinformation in advertising

Commercial (advertising) 0: internal organisational checks on origin/trolls 1: ban all (foreign) adverts on certain topics (e.g. Google/Facebook/Twitter in Ireland abortion referendum) 2: publish transparent adverts register (e.g. Twitter Ads Transparency Centre) 3: co-regulatory agreement on ad registry/origin (European Advertising Standards Alliance/National advertising associations) 4: agency to regulate advertising , including preventing non-consensual OBA (Consumer protection bodies/Election commissions/Data Protection authorities) 5: amend and strengthen existing legislation (e.g. on misleading advertising, political spending, election silence period, data protection, etc.)
Public accounts Ban all identified political bot accounts (note many millions of news/celebrity/marketing bot accounts) **Private (messaging)** Prevent private messages being made public; restrict number of recipients of messages

2.3. AI techniques and problems

Over time, AI solutions to detect and remove illegal/undesirable content have become more effective, but they also raise questions about who is the 'judge' in determining what is legal/illegal, and desirable/undesirable in society. Underlying AI use is a difficult choice between different

[81] UK House of Commons (2018*) Interim Report on Disinformation and 'Fake News'*, supra n.35.

[82] Institute of Practitioners in Advertising (IPA) (FKN0093) at p38 in UK House of Commons (2018) *Interim Report on Disinformation and 'Fake News'*, supra n.35

[83] Facebook letter from Rebecca Stimson to Damian Collins, 8 June 2018, at p.38 in UK House of Commons (2018) *Interim Report on Disinformation and 'Fake News'*, supra n.35

elements of law and technology, public and private solutions, with trade-offs between judicial decision-making, scalability, and impact on users' freedom of expression.

Professor Mireille Hildebrandt explains the scale and scope that can create disinformation problems in social media platforms:

> *'Due to their distributed, networked, and data-driven architecture, platforms enable the construction of invasive, over-complete, statistically inferred, profiles of individuals (exposure), the spreading of fake content and fake accounts, the intervention of botfarms and malware as well as persistent AB testing, targeted advertising, and automated, targeted recycling of fake content (manipulation).'*[84]

Hildebrandt warns that we must avoid the machine learning version of the Thomas self-fulfilling prophecy theorem – that 'if a machine interprets a situation as real, its consequences becomes real'[85]. This is reiterated in support of Langdon Winner's insights into biases of technology[86]. Hildebrandt explains that 'data-driven systems parasite on the expertise of domain experts to engage in what is essentially an imitation game. There is nothing wrong with that, unless we wrongly assume that the system can do without the acuity of human judgment, mistaking the imitation for what is imitated'[87]. Some of the claims that AI can 'solve' the problem of disinformation do just that. Limiting the automated execution of decisions on AI-discovered problems is essential in ensuring human agency and natural justice: the right to appeal. That does not prevent the suspension of bot accounts at scale, but ensures the correct auditing of the system processes deployed.

Public and private actors have suggested that AI could play a larger role in future identification of problematic content – but these systems have their own prejudices and biases. Neither law nor technology are neutral: they embody the values and priorities of those who have designed them ('garbage in, garbage out'). It leads us to ask:

> *Do AI systems, that use algorithmic processes to identify 'undesirable' content and nudge it out of consumers' view, provide a means for effective self-regulation by platforms?*

It is an open question, with no definitive answer at this stage. Technical research into disinformation has followed several tracks:

- identifying bots as distinct from human accounts[88];
- identifying the real world effects of internet communication on social networks[89],

[84] Hildebrandt, M. (2018) 'Primitives of Legal Protection in the Era of Data-Driven Platforms', *Georgetown Law Technology Review 2(2)* at p. 253 footnote 3

[85] Merton, R.K. (1948) 'The Self-Fulfilling Prophecy', *The Antioch Review 8(2),* 193-210.

[86] Winner, L. (1989) *The Whale And The Reactor: A Search For Limits In An Age of High Technolo*gy, Chicago: University of Chicago Press, p.29. See also Chander, A. and Vivek, K. (2018) 'The Myth Of Platform Neutrality', *Georgetown Law Technology Review 2(2)* 400-416.

[87] Hildebrandt, M. (2018) supra n.84, at p. 255. The imitation game is often known as the Turing test, after Turing, A.M. (1950) 'Computing Machinery and Intelligence', *Mind 49*, 433-460. See also Mitchell, T. (1997) *Machine Learning,* New York: McGraw-Hill Education, pp. 7–9.

[88] Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., and Crowcroft. J. (2017) 'Of Bots and Humans (on Twitter)', in *ASONAM '17 Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* pp. 349-354; Perez, B., Musolesi, M., and Stringhini, G. (2018) 'You are Your Metadata: Identification and Obfuscation of Social Media Users using Metadata Information', *ICWSM.*

[89] Including the 'Dunbar number' of friends that can be maintained, which has not measurably increased with the Internet: Dunbar, R. I. M. (2016) 'Do Online Social Media Cut Through the Constraints that Limit the Size of Offline Social Networks?', *Royal Society Open Science 2016(3)*, DOI: 10.1098/rsos.150292. Quercia, D., Lambiotte, R., Stillwell, D. Kosinski, M., and Crowcroft, J. (2012) 'The Personality of Popular Facebook Users', in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12),* pp. 955-964, https://doi.org/10.1145/2145204.2145346

- assessing the impact of disinformation via media consumption and electoral outcomes[90];
- researching security threats from disinformation[91];
- researching discrimination and bias in the algorithms used to both propagate and increasingly to identify and/or disable disinformation[92].

The UK Parliament AI Committee reported on some of these issues in 2017[93]. There are an enormous number of false positives in taking material down. It is very difficult for AI to tell the difference between a picture of fried chicken and a Labrapoodle dog, simply because of the limited number of the attempts by algorithms to match these things[94]. Human intervention is necessary to analyse these false positives that could lead to over-censorship of legitimate content that is machine-labelled incorrectly as disinformation.

Online disinformation consumption includes that of video news and newspapers, whose readerships have largely migrated online[95], but also images and amateur montages of video ('deep fakes') that are far harder to detect as disinformation. Textual analysis of Twitter or news sites can only explore the tip of the iceberg of disinformation, as video and images are much more difficult to examine comprehensively. Only a partial view of AI effectiveness exists outside corporate walls:

> *'Facebook says its AI tools—many of which are trained with data from its human moderation team—detect nearly 100 percent of spam, and that 99.5 percent of terrorist-related removals, 98.5 percent of fake accounts, 96 percent of adult nudity and sexual activity, and 86 percent of graphic violence-related removals are detected by AI, not users.'*[96]

This level of AI removals sounds impressive, though these are unaudited company claims, but Facebook's AI detects: 'just 38 percent of the hate speech-related posts it ultimately removes, and at the moment it doesn't have enough training data for the AI to be very effective outside of English and Portuguese'[97]. In 2018, researchers have claimed that trained algorithmic detection of fact verification may never be as effective as human intervention, with serious caveats (each has accuracy of only 76%): 'future work might want to explore how hybrid decision models consisting

[90] Zannettou, S. et al. (2018) *Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web*, arXiv:1801.09288v1; Nilizadeh, S. et al. (2017) 'POISED: Spotting Twitter Spam Off the Beaten Paths', *CCS*. Chatzakou, D. et al. (2017), 'Mean Birds: Detecting Aggression and Bullying on Twitter', *WebSci*. Hine, G. et al. (2017) 'Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and its Effects on the Web', *ICWSM*.

[91] See the research group that evolved in part from FP7 European Internet Science, continued in part as the WebSci conference: e.g. Ibosiola, D. et al. (2018) 'Movie Pirates of the Caribbean: Exploring Illegal Streaming Cyberlockers', *ICWSM*; Zannettou, S. et al. (2018) 'Understanding Web Archiving Services and Their (Mis)Use on Social Media', *ICWSM*; Zannettou, S. et al. (2017) 'The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources', *IMC*.

[92] Alexander J., and Smith, J. (2011) 'Disinformation: A Taxonomy', *IEEE Security & Privacy 9(1)*, 58-63, doi: 10.1109/MSP.2010.141; Michael, K. (2017) 'Bots Trending Now: Disinformation and Calculated Manipulation of the Masses [Editorial]', *IEEE Technology and Society Magazine 36(2)*, 6-11, doi: 10.1109/MTS.2017.2697067

[93] UK House of Lords (2017) *AI Select Committee: AI Report Published* https://www.parliament.uk/business/committees/committees-a-z/lords-select/ai-committee/news-parliament-2017/ai-report-published/ (note the report is published in non-standard URL accessed from this link)

[94] Reddit poster (2017) Artificial Intelligence Can't Tell Fried Chicken from Labradoodles, https://www.reddit.com/r/funny/comments/6h47qr/artificial_intelligence_cant_tell_fried_chicken/

[95] Nielsen, R.K. and Ganter, S. (2017) 'Dealing with Digital Intermediaries: A Case Study of the Relations Between Publishers and Platforms', *New Media & Society 20(4)*, 1600-1617, doi: 10.1177/1461444817701318

[96] Koebler, J., and Cox, J. (23 Aug 2018) 'The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People', *Motherboard*, https://motherboard.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works

[97] Koebler and Cox (2018) supra n.96

of both fact verification and data-driven machine learning judgments can be integrated'[98]. This is a sensible approach where resources allow for such a wide spectrum of solutions.

AI cannot be the only way to regulate content in future[99]. 'Mechanical Turks' are people employed—subcontracted, typically—to carry out these activities[100], in parts of the world where their own cultural understanding of the content they are dealing with may not be ideal. One of the problems is that they are responding to a perceived need to remove more content, rather than addressing fair process and due process. Subcontracting to people on very low wages in locations other than Europe is a great deal cheaper than employing a lawyer to work out whether there should be an appeal to put content back online. The incentive structure will be for platforms to demonstrate how much content they have removed, when a very important factor may be examples of successful appeals to 'put back' legitimate content online[101]. Research director at the Tow Center for Digital Journalism, Jonathan Albright argues: 'Since the 2016 election, I've come to the realization — horribly, and it's very depressing — that nothing has gotten better, despite all the rhetoric, all of the money, all of the PR, all of the research.'[102] He argues that content moderation at scale still needs executive human intervention to support AI: 'And I don't mean, like, contract moderators from India — I mean high-level people. The companies need to invest in human capital as well as technological capital, but that doesn't align with their business model'.

Executive Director at European Digital Rights, Joe McNamee, calls for transparency and detail in reporting on providers' responses to requests of government bodies, in order to better understand the accuracy of the notices and the (legal) consequences of actions. He provides the example of incitement to violence:

> *'I want transparency on what everybody did. So if we imagine that 50% of Europol reports are wrong and we get a transparency report that says that member states carried out investigations in relation to 100% of them. Then it would come out in the wash that the member states would say that they didn't investigate half of them because they aren't illegal. I think transparency on like 'Person X uploaded incitement to violence. What happened? It was removed. The police was interested. The police wasn't interested. There was a prosecution. There wasn't a prosecution. There was a prosecution, but it wasn't incitement to violence. So clearly it wasn't illegal'. Okay, that's interesting. Now we've got transparency. Now we know what's happening. Now we know that there's incitement to violence that can't be prosecuted because of bad law, can't be investigated because of lack of police.'*[103]

Michael Veale, Reuben Binns and Max Van Kleek explain how to move beyond transparency and explicability to replicability: to be able to run the result and produce the answer that matches the answer they have[104]. The greater the transparency, the greater the amount of information you give to users, the less the degree to which that help is limited. Users are told that if they do not agree to the effectively unilateral Terms of Service, they can no longer use the service. Transparency and explanation is necessary, but it is a small first step towards better regulation[105]. A satisfactory

[98] Perez-Rosas, V., Kleinberg, B. Lefevre, A. and Mihalcea, R. (2018) *Automatic Detection of Fake News*, http://web.eecs.umich.edu/~mihalcea/papers/perezrosas.coling18.pdf

[99] Discussed by Marietje Schaake MEP in April: Schaake, M. (4 April 2018) 'Algorithms Have Become So Powerful We Need a Robust, Europe-Wide Response', *The Guardian* https://www.theguardian.com/commentisfree/2018/apr/04/algorithms-powerful-europe-response-social-media

[100] Kotaro, H. et.al (2017) 'A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk', in *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI'18)*, Paper no. 449.

[101] Google (2018) *YouTube Transparency Report*, https://transparencyreport.google.com/youtube-policy/overview

[103] Expert interview with Joe McNamee (Executive Director of European Digital Rights – EDRi, 6 Sept 2018)

[103] Expert interview with Joe McNamee (Executive Director of European Digital Rights – EDRi, 6 Sept 2018)

[104] Veale, M., Binns, R., and Van Kleek, M. (2018) 'The General Data Protection Regulation: An Opportunity for the CHI Community? (CHI-GDPR 2018)', *Workshop at ACM CHI'18,* 22 April 2018, Montreal, arXiv:1803.06174

[105] Edwards, L. and Veale, M. (2017) *Slave to the Algorithm? Why a 'Right to Explanation' is Probably Not the Remedy You are Looking for*, https://ssrn.com/abstract=2972855. Erdos, D. (2016) 'European Data Protection Regulation and Online

solution to algorithmic transparency might be the ability to replicate the result that has been achieved by the company producing the algorithm. Algorithms change all the time: there are good reasons to keep them trade secrets. Replicability would be the ability to look at the algorithm in use at the time and, as an audit function, run it back through the data to produce the same result. It is used in medical trials as a basic principle of scientific inquiry. It would help to create more trust in what is otherwise a black box that users and regulators simply have accept. The European Commission has used the overarching phrase 'a fair deal for consumers'[106]. We return to these thoughts on transparency, oversight and appeal in use of AI to tackle disinformation in Chapter 4.

2.4. Policy responses to internet speech problems

2.4.1. Safeguarding human rights in internet regulation

Disinformation online needs to be viewed within the broader context of how policymakers have adapted media law to the internet. The baseline response was described by Pamela Samuelson in 1999 as five key policy challenges:

> *'1. whether they can apply or adapt existing laws and policies to the regulation of internet activities, or whether new laws or policies are needed to regulate internet conduct;*
>
> *2. how to formulate a reasonable and proportional response when new regulation is needed;*
>
> *3. how to craft laws that will be flexible enough to adapt to rapidly changing circumstances;*
>
> *4. how to preserve fundamental human values in the face of economic or technological pressures tending to undermine them; and*
>
> *5. how to coordinate with other nations in internet law and policy making so that there is a consistent legal environment on a global basis'*[107].

These lessons of adaption or new laws, proportionality, flexibility, and international coordination remain sound. Protecting and preserving fundamental rights are more important than individual sectoral innovations (to 'move fast and break things'). Techno-economic progress must respect fundamental rights responsibilities, as stated in constitutions throughout the Industrial Revolution in advanced nations, from the US Bill of Rights 1789 and French 'Declaration of the Rights of Man and of the Citizen' 1789, to the Convention for the Protection of Human Rights and Fundamental Freedoms of the Council of Europe, and the EU Charter of Fundamental Rights.

The internet is a powerful technology for society and individuals to express their rights, as well as an environment in which such rights can be abused and curtailed due to legal, economic, technological, security and other incentives for powerful actors. We identify disinformation as a problem set in which human rights are primarily regulated by the Terms of Service of the platforms to which they subscribe, though there are bright lines of human rights protection against unlawful censorship without appeal or redress. Privacy and freedom of expression are in constant tension on the internet, whether that is debated in copyright enforcement, network neutrality and the open internet, or the specific freedom of expression problem of disinformation.

New Media: Mind the Enforcement Gap', *Journal of Law and Society 43(4)* 534-564, http://dx.doi.org/10.1111/jols.12002

[106] Vestager, M. (2018) 'Competition and A Fair Deal for Consumers Online', *Netherlands Authority for Consumers and Markets Fifth Anniversary Conference,* 26 April 2018, The Hague, https://ec.europa.eu/commission/commissioners/2014-2019/vestager/announcements/competition-and-fair-deal-consumers-online_en

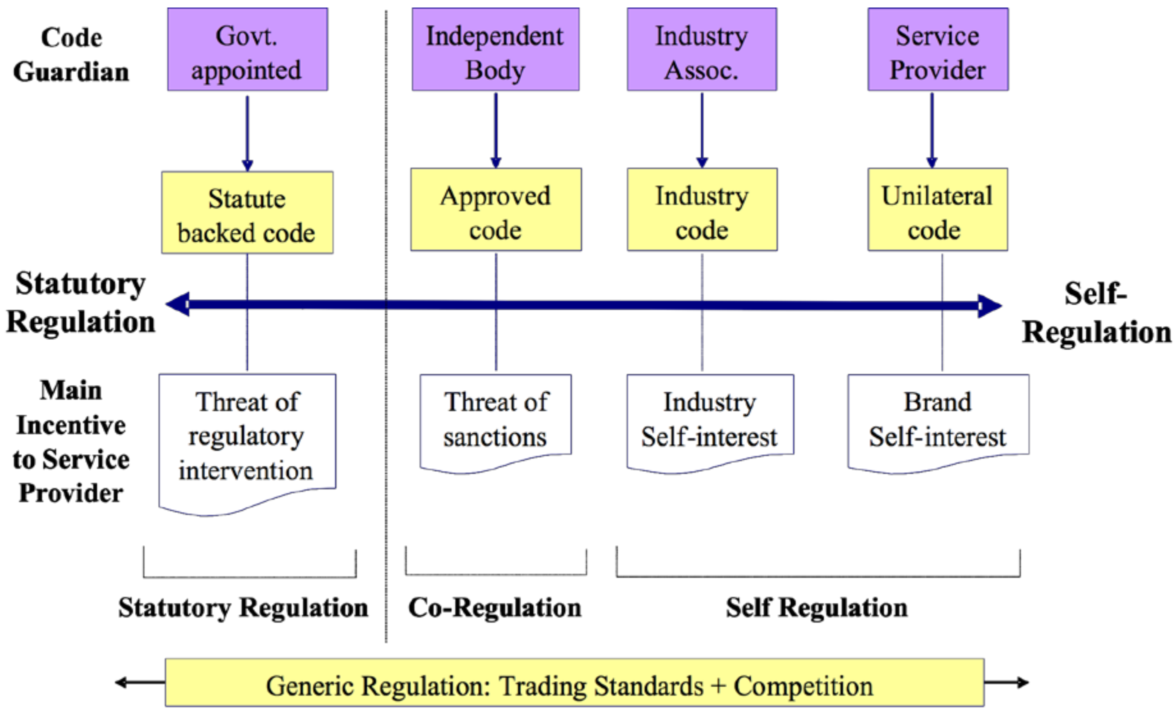
[107] Samuelson, P. (1999) 'A New Kind of Privacy? Regulating Uses of Personal Data in the Global Information Economy', *California Law Review 87,* 751-778.

2.4.2. Classifying self- and co-regulatory solutions

This chapter has explained the genesis of the disinformation problem, and some aspects that need analysis in understanding its context. Technological self-regulation does not exist in a regulatory vacuum, and disinformation also needs placing into the regulatory discussion of how to provide solutions.

The diagram below demonstrates the types of media consumption online and the types of regulation that may be available.

Figure 2.2 Millwood-Hargrave Diagram of SROs, regulatory type and incentive structure[108]



Most of the technical actions observed will fall into the right-hand columns as unilateral or industry coordinated actions, such as the combined July 2018 operations of Google, Facebook and Twitter to disable nearly a thousand accounts alleged to be operated by agencies of Iranian and Russian intelligence services, or the removal of the Jones 'Infowars' accounts (though note Twitter did not respond alongside its peers[109]).

This introduction to the topic, its development, technological means of content recognition, and the classification of different systems by which to regulate content, leads us to Chapter 3 which deals with mapping of existing and proposed regulatory and technical solutions to disinformation online.

[108] Millwood-Hargrave, M. (2007) *Report for Working Group 3 of the Conference of Experts for European Media Policy, More Trust in Content – The Potential of Co- and Self-Regulation in Digital Media*, Leipzig: 9-11 May.

[109] YouTube, Facebook, Apple and Spotify removed the Infowars site in a coordinated action: see https://www.bbc.co.uk/news/world-us-canada-45442417

3. Mapping and evaluating the existing solutions

Chapter 3 reviews policy and technology initiatives relevant to disinformation and illegal content online with the aim of understanding (a) how they recommend to use technology as a solution to curb certain types of content online and (b) what they identify as necessary safeguards to limit the impact on freedom of expression and media pluralism. At the end of the chapter (c) the existing initiatives are mapped onto the typology of self- and co-regulatory solutions.

Section 3.1 maps policy initiatives relevant to disinformation and illegal content online. We pay attention to the overall set of solutions they identify, as well as their recommendations specific to technology-based initiatives. We also compare the safeguards these policy initiatives emphasise in four key action areas of online intermediaries (advertising, automated processes, content/account moderation, and fact-checking/trustworthiness). We will identify the limits of technology in regulating content online, as seen by the intermediaries.

Section 3.2 explores AI-enabled content/account moderation (removal/filtering, blocking and (de)prioritisation of content and advertising, as well as disabling/suspension of accounts). We map the actions of three major intermediaries (Facebook, Google and Twitter) on content/account moderation and evaluate their safeguards to minimise impact on freedom of expression and media pluralism.

Section 3.3 maps the existing initiatives onto the spectrum of solutions to regulating content online introduced in Chapter 2. The aim of this comparison is to explain the regulatory focus.

3.1. Policy initiatives

Within Europe, online disinformation is currently tackled by regulators from a variety of regulatory angles. It can be limited through stipulations and actions against defamation, incitement to hatred and violence[110] or the ban on certain misleading advertising techniques[111]. Within the context of electoral campaigns, the problem can be tackled by regulating spending and transparency of political campaigns, enforcing data protection rules and bolstering against cyberattacks.[112] Figure 3.1 provides a timeline of relevant disinformation policy initiatives at a European level.

[110] Such as found, for instance, in the UN International Covenant on Civil and Political Rights (1966), Articles 19 and 20; the EU Multistakeholder Code of Conduct on Countering Illegal Hate Speech Online (May 2016); and Proposed EU Regulation on Prevention of Dissemination of Terrorist Content Online (COM(2018) 640 final - 2018/0331 (COD))

[111] E.g. EU Directive 2006/114/EC concerning Misleading and Comparative Advertising; EU Directive 2010/13/EU on Audiovisual Media Services (the Coordination of Certain Provisions Laid Down by Law, Regulation or Administrative Action in Member States concerning the Provision of Audiovisual Media Services)

[112] Such as recently addressed, for instance, in the EC Elections Package, which contains (amongst others) a Proposed EU Revised Regulation on the Statute and Funding of European Political Parties and European Political Foundations (COM/2017/0481 final - 2017/0219 (COD)); EC Guidance on the Application of Union Data Protection Law in the Electoral Context (COM(2018) 638 final); and a EC Recommendation on Election Cooperation Networks, Online Transparency, Protection against Cybersecurity Incidents and Fighting Disinformation Campaigns in the Context of Elections to the European Parliament (C(2018) 5949 final)

Figure 3.1 Recent EU disinformation initiatives

2015

- European Council Conclusions calling for Strategic Communications Action Plan (March 2015)
- EEAS East StratCom Task Force (March 2015)
- EC Communication on European Agenda on Security (April 2015)
- Europol Internet Referral Unit (July 2015)
- EU Internet Forum on Terrorist Content Online (Dec 2015)

2016

- EU General Data Protection Regulation (April 2016, entry into force May 2018)
- EU Multistakeholder Code of Conduct on Countering Illegal Hate Speech Online (May 2016)
- EP Resolution on EU Strategic Communication to Counteract Propaganda (Nov 2016)

2017

- EU Proposed Revised Regulation on E-Privacy (Jan 2017)
- EU Directive on Combatting Terrorism (Article 21, March 2017)
- European Council Conclusions on Internal Security and the Fight against Terrorism (June 2017)
- EC Communication on Tackling Illegal Content Online (Sept 2017)
- EU Proposed Revised Regulation on the Statute and Funding of European Political Parties and European Political Foundations (Sept 2017)

2018

- EC Recommendation on Measures to Effectively Tackle Illegal Content Online (March 2018)
- EC High Level Expert Group Report on Fake News and Online Disinformation (March 2018)
- EDPS Opinion on Online Manipulation and Personal Data (March 2018)
- EC Communication on a European Approach to Tackling Online Disinformation (April 2018)
- EC and HR Joint Communication on Increasing Resilience and Bolstering Capabilities to Address Hybrid Threats (June 2018)
- European Council Conclusions calling for Disinformation Action Plan (June 2018)
- Proposal for EU Regulation on Prevention of Dissemination of Terrorist Content Online (Sept 2018)
- EC Recommendation on Election Cooperation Networks, Online Transparency, Protection against Cybersecurity Incidents and Fighting Disinformation Campaigns in the Context of Elections to the European Parliament (Sept 2018)
- EC Guidance on the Application of Union Data Protection Law in the Electoral Context (Sept 2018)
- EC Regulation proposal establishing the European Cybersecurity Industrial, Technology and Research Competence Centre and the Network of National Coordination Centres (Sept 2018)
- EU Multistakeholder Code of Practice on Disinformation (Sept 2018)
- Facebook, Google, Twitter, Mozilla Roadmaps for implementing the Code of Practice (Oct 2018)

More broadly, institutional support can be provided to safeguard media pluralism[113], encourage fact-checking[114] and enhance media literacy[115].

We review a sample of these policy initiatives dealing with illegal content and disinformation online in the paragraphs below.[116] These serve as a necessary background for the comparative analysis provided in Section 3.1.4.

3.1.1. Illegal content online

Of crucial importance to this study on technology-based solutions to disinformation, Articles 12-15 of the **E-Commerce Directive** (2000/31/EC)[117] set out the limits of liability of internet intermediary service providers for illegal activity and content on their networks. 'Information Society Service Providers' (ISSPs or intermediaries) are not subject to liability for their European customers' content so long as they have no actual or constructive knowledge of that content: if they 'hear no evil, see no evil and speak no evil'. Article 12 protects the ISSP where it provides 'mere conduit'' with no knowledge of, nor editorial control over, content or receiver ('does not initiate [or] select the receiver'). However, liability increases as the intermediary's editorial control increases. Where intermediaries provide hosting services, they are protected from liability, in two ways:

- the provider does not have actual knowledge of illegal activity or information and is not aware of facts or circumstances from which the illegal activity is apparent; or
- the provider, upon obtaining such knowledge or awareness, acts expeditiously to remove or to disrupt access of the information.

As mere ciphers for content, ISSPs have a safe harbour from liability; should they engage in any filtering of content, they become potentially liable. They have to take action when they are notified of illegal activity or content on their networks. Intermediaries have been acting as the fabled 'three wise monkeys' in relation to internet content liability since the dawn of the commercial internet,

[113] The EU's most impactful tools are its competition rules on abuse of a dominant position, state aid and merger control (Articles 101 and 102 TFEU). It also takes a monitoring and capacity building approach, as illustrated amongst others through its support of the Media Pluralism Monitor and capacity building and training of journalists. See for instance European Commission (2018) *Media Pluralism Monitor,* https://ec.europa.eu/digital-single-market/en/media-pluralism-monitor-mpm and European Commission (18 Sept 2018) *The European Union Strengthens its Support to Media Freedom and Young Journalists in the Western Balkans*, Press Release (IP/18/5789), http://europa.eu/rapid/press-release_IP-18-5789_en.htm

[114] The EEAS East StratCom Task Force is the EU's most ambitious internal debunking effort. It was set up after the European Council mandated the High Representative and the member states to develop an action plan on strategic communications in its March 2015 Conclusions. The Task Force's mandate pertains to addressing Russia's ongoing disinformation campaigns through strategic communications and research. This can consist amongst others of better explaining EU policies and strengthening the media in the Eastern Partnership region; explaining, correcting and raising awareness of disinformation narratives through amongst others the Disinformation Review, http://www.EUvsDisinfo.eu, @EUvsDisinfo; and analyzing and reporting on disinformation trends. See EEAS (2017) *Questions and Answers about the East StratCom Task Force,* https://eeas.europa.eu/headquarters/headquarters-homepage/2116/-questions-and-answers-about-the-east-stratcom-task-force_en

[115] The European Commission plays a supporting role in digital and media literacy through programmes, prizes, coordination and sharing of best practices among member states. For instance, one of the actions has been to develop a Digital Competence Framework for Citizens. The European Commission also has supported worldwide safer internet efforts for over twenty years through the safer internet action plan and successor programmes. For instance, in 2018 they launched a series of #SaferInternet4EU initiatives in coordination with Safer Internet centres across Europe. See European Commission (6 Feb 2018) *Launch of the #SaferInternet4EU Initiatives on Safer Internet Day,* Press Release, https://ec.europa.eu/digital-single-market/en/news/launch-saferinternet4eu-initiatives-safer-internet-day

[116] The criterion for selecting these initiatives to the exclusion of many others is their level of detail in providing recommendations on how to approach technological regulation.

[117] EU Directive 2000/31/EC on Certain Legal Aspects of Information Society Services, in particular Electronic Commerce, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031

reflected in pioneering case law[118]. Thus 'masterly inactivity', except when prompted by law enforcement is the economically most advantageous policy open to them. Articles 12-15 of the E-Commerce Directive have not been amended since being implemented in national law in 2002, but have been subject to extensive judicial interpretation across Europe.[119]

The dangers of a regime that incentivises companies to take down content on notice, but not proactively search for illegal and otherwise harmful content has been recognised officially. The European Council in 2014 declared they will: '[r]aise awareness among judges, law enforcement officials, staff of human rights commissions and policymakers around the world of the need to promote international standards, including standards protecting intermediaries from the obligation of blocking internet content without prior due process.'[120] European Commission consultations on online platforms, assessment of the formally self-regulatory Code of Conduct fighting hate speech, and its overall Digital Single Market strategy all employ the 'bully pulpit' to argue for greater responsibility by online platforms[121]. This will lead to more private enforcement of censorship, of media freedom concern given the lack of appeal guarantees[122].

In response to pressure from those affected negatively by the legal 'masterly inactivity' approach of intermediaries[123], the European Commission explained in its September 2017 **Communication on Tackling Illegal Content Online** what further action online platform intermediaries should be required to consider, summarised in the table below[124].

[118] Marsden, C. (2012) 'Internet Co-Regulation and Constitutionalism: Towards European Judicial Review', *International Review of Law, Computers & Technology 26(2-3)* 215-216. For UK law, see *Shetland Times Ltd v Jonathan Wills and Another*, 1997 FSR (Ct Sess. OH), 24 October 1996; *Godfrey v Demon Internet Service* [2001] QB 201. For US law, see *Cubby v CompuServe* (1991) 766 F Supp 135, *Playboy Enterprises, Inc. v. Frena*, 839 F. Supp. 1552 (M.D. Fla. 1993), *Stratton Oakmont Inc v. Prodigy* 1995 NY Misc. 23 Media L. Rep. 1794, *American Civil Liberties Union v Reno* (1997) 21 US 844 of 27 June No 96–511, and Digital Millennium Copyright Act 1998, s 512(k)(1)(A–B).

[119] For comprehensive academic commentary, see for instance Husovec, M. (2017) *Injunctions Against Intermediaries in the European Union. Accountable But Not Liable?,* Cambridge, UK: Cambridge University Press; Rosati, E. (2019) *Copyright and the Court of Justice of the European Union,* Oxford: Oxford University Press; Sartor, G. (2017) *Providers Liability: From the eCommerce Directive to the Future,* In-Depth Analysis for EP IMCO Committee, IP/A/IMCO/2017-07, Brussels: European Parliament.

[120] EU Human Rights Guidelines on Freedom of Expression Online and Offline (2014) https://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/EN/foraff/142549.pdf, Paragraph 34.

[121] EC Communication on A Digital Single Market Strategy for Europe (COM (2015) 192 final), para. 3.3; European Commission (2015) *Public Consultation on the Regulatory Environment for Platforms, Online Intermediaries, Data and Cloud Computing and the Collaborative Economy* https://ec.europa.eu/digital-single-market/news/public-consultation-regulatory-environment-platforms-online-intermediaries-data-and-cloud; European Commission (6 Dec 2016) *Fighting Illegal Online Hate Speech: First Assessment of the New Code of Conduct*, Press Release, http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50840; EC Communication on Online Platforms and the Digital Single Market: Opportunities and Challenges for Europe (COM (2016) 288), p. 9

[122] See Frosio, G. (2017) 'From Horizontal to Vertical: An Intermediary Liability Earthquake in Europe, *Journal of Intellectual Property Law and Practice 12(7)* 565-575. See European Commission (2016) *Full Report on the Results of the Public Consultation on the Regulatory Environment for Platforms, Online Intermediaries and the Collaborative Economy,* https://ec.europa.eu/digital-single-market/en/news/full-report-results-public-consultation-regulatory-environment-platforms-online-intermediaries

[123] Lemley, M. (2006) 'Terms of Use', *Minnesota Law Review 91(2)* 459-483. See generally on intermediary and access provider liability, Marsden, C. (2018) 'Regulating Intermediary Liability and Network Neutrality', in I. Walden (Ed.) *Telecommunications Law and Regulation*, 5th Edition, Oxford: Oxford University Press, pp. 733-788. See specifically on intermediary liability in copyright, Meyer, T. (2017) *The Politics of Online Copyright Enforcement in the EU: Access and Control,* Cham: Palgrave Macmillan.

[124] EC Communication on Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms (COM(2017) 555 final) https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms

Table 3.1 Online platform action called for by European Commission (2017)

Section	**Online Platform action**
3.1	Take action on illegal content without a court order or administrative decision, especially where **notified by a law enforcement authority**
4.1	Prioritise removal in response to notices received from **law enforcement bodies and other public or private sector 'trusted flaggers'**
4.1	Fully automated removal should be applied where the circumstances leave little doubt about the illegality of the material (such as where **removal is notified by law enforcement** authorities)
3.2.1	In a limited number of cases platforms may remove content notified by **trusted flaggers without verifying legality** themselves
3.3.1	Platforms should not limit themselves to reacting to notices, but adopt **effective proactive measures** to detect and remove illegal content
5.1	Platforms should take measures (such as **account suspension or termination**) which dissuade users from repeatedly uploading illegal content of the same nature
5.2	Platforms are strongly encouraged to use fingerprinting tools to filter out content that has already been identified and assessed as illegal
4.1	Platforms should report to law enforcement authorities whenever they are made aware of or encounter evidence of criminal or other offences

In March 2018, the European Commission issued a follow-up **Recommendation on Measures to Effectively Tackle Illegal Online Content**[125]**.** The recommendation provides guidance and considerations, again primarily for hosting providers. Similar to the September 2017 Communication[126], it calls for pro-active measures, but also states that human verification should be included where possible. A distinction is made between 'all types of illegal content' and 'terrorist content', where the Commission specifically calls for use of automated means to remove, block or prevent re-upload of terrorist content. Informing the user whose content has been blocked or removed and providing the option of a counter-notice are emphasised, with the exception of manifestly illegal content that relates to criminal offences involving a threat to the life, or safety of persons. The Commission also encourages transparency through clear explanations and regular reports on content moderation policies, and cooperation with member states, trusted flaggers and among hosting providers is recommended.

In September 2018, the European Commission proposed a **Regulation on the Prevention of the Dissemination of Terrorist Content Online**.[127] Most remarkable/controversial in the proposal is the one hour rule to take terrorist content offline 'following a removal order from national

[125] EC Recommendation on Measures to Effectively Tackle Illegal Online Content (C(2018) 1177 final), https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online

[126] EC Communication on Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms (COM(2017) 555 final), https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms

[127] Proposed EU Regulation on Prevention of Dissemination of Terrorist Content Online (COM(2018) 640 final - 2018/0331 (COD)), https://ec.europa.eu/commission/sites/beta-political/files/soteu2018-preventing-terrorist-content-online-regulation-640_en.pdf

competent authorities'.[128] The proposal calls for a duty of care obligation (once again, advocating proactive measures by online service providers), increased cooperation and financial penalties for non-compliance, but also mentions safeguards through redress mechanisms, transparency and accountability through annual reporting. At a European level, the **Europol Internet Referral Unit**[129] is specifically tasked to analyze, detect, flag and request removal of terrorist and violent extremist content. On a self-regulatory level, the **EU Internet Forum on Terrorist Content Online**[130] precedes the proposed regulation.

Previously, in May 2016, online intermediaries had also agreed to an **EU Code of Conduct on Countering Illegal Hate Speech Online**.[131] They committed to 'review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary.'[132] The code of conduct also includes commitments on education/awareness of users, collaboration with civil society organisations and trusted reporters, clear community guidelines, clear procedures for flagging/reporting, best practice sharing and further discussions on transparency and alternative/counter narratives. Relevant to the topic of this study on freedom of expression, the code of conduct points to the chilling effect of illegal hate speech on democratic discourse:

> *'The spread of illegal hate speech online not only negatively affects the groups or individuals that it targets, it also negatively impacts those who speak out for freedom, tolerance and non-discrimination in our open societies and has a chilling effect on the democratic discourse on online platforms.'*[133]

In sum, the European Union has a well-established tradition on liability protections for digital intermediaries through the E-Commerce Directive and relevant case law. There is extensive use of notice and action processes to reduce the availability of illegal content online. As will be noted in Section 3.2, intermediaries also take voluntary action, often based on their terms of service.

At the same time, momentum has built within the EU to request hosting providers to take pro-active measures in tackling illegal content. This shift in European Commission thinking already became noticeable in the proposed Copyright in the Digital Single Market Directive[134] and in the proposed Prevention of the Terrorist Content Dissemination Online Regulation[135]. The Communication[136] and Recommendation[137] on tackling illegal content online confirm that the European Commission deems more stringent requirements for online platform intermediaries desirable beyond these policy areas of anti-terrorism and copyright.

[128] European Commission (12 September 2018) *State of the Union 2018: Commission Proposes New Rules to Get Terrorist Content Off the Web*, Press Release (IP/18/5561), http://europa.eu/rapid/press-release_IP-18-5561_en.htm

[129] Europol (22 July 2016) *Europol Internet Referral Unit One Year On*, Press Release, https://www.europol.europa.eu/newsroom/news/europol-internet-referral-unit-one-year

[130] EU Code of Conduct on Countering Illegal Hate Speech Online (2016), http://europa.eu/rapid/press-release_IP-15-6243_en.htm

[131] Initially Facebook, Microsoft, Twitter and YouTube signed the code of conduct. In the first half of 2018, Instagram, Google+, Snapchat and Dailymotion declared their intention to join as well. See European Commission (2018) *Countering Illegal Hate Speech Online #NoPlace4Hate*, Press Release, http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300

[132] EU Code of Conduct on Countering Illegal Hate Speech Online (2016), p. 2.

[133] Idem, p.1.

[134] Article 13 pertains to use of ACR technologies to prevent the availability of copyright content online. See Proposed EU Directive on Copyright in the Digital Single Market (COM(2016) 593 final – 2016/0280(COD))

[135] Proposed EU Regulation on Prevention of Dissemination of Terrorist Content Online (COM(2018) 640 final - 2018/0331 (COD))

[136] EC Communication on Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms (COM(2017) 555 final)

[137] EC Recommendation on Measures to Effectively Tackle Illegal Online Content (C(2018) 1177 final)

3.1.2. Disinformation online

In parallel to its longstanding policy activities on illegal content, the European Commission established a **High Level Expert Group (HLEG) on Fake News and Online Disinformation**[138] in early 2018. The group of thirty-eight experts published their Report in March 2018.[139]

The report reviewed current practices on disinformation, ranging from transparency and accountability-enhancing practices, trust-enhancing practices and algorithm changes, to media and information literacy. In recommending responses and actions, the High Level Expert Group focused primarily on 'improv[ing] the findability of, and access to, trustworthy content', as 'filtering out disinformation is difficult to achieve without hitting legitimate content, and is therefore problematic from a freedom of expression perspective'[140]:

1. enhancing transparency of digital ecosystem, in particular in terms of funding sources, online news sources and journalistic processes, and fact-checking practices;
2. promoting media and information literacy: reassessment and adjustment of educational policies, reaching out to citizens of all ages;
3. developing tools for empowering users and journalists; and
4. safeguarding the diversity and sustainability of the European news media ecosystem.[141]

In terms of process and evaluation, the group recommended 'a self-regulatory approach based on a clearly defined multi-stakeholder engagement process, framed within a binding roadmap for implementation, and focused on a set of short and medium-term actions'.[142] This includes a call for a European code of practices, key performance indicators (KPIs), timeframes, progress reports, and a permanent review mechanism. The HLEG also recommends the development of European Centres for Research on Disinformation.[143]

In April 2018, the European Commission responded to the HLEG report and published a **Communication on a European Approach to Tackling Online Disinformation**[144]. The Commission highlighted five priority areas for action. Similar to the HLEG, media literacy and pluralism are mentioned, but there is also attention paid to elections, strategic communication and the role of digital intermediaries:

1. a more transparent, trustworthy and accountable online ecosystem:
    a. online platforms to act swiftly and effectively to protect users from disinformation;
    b. strengthening fact-checking, collective knowledge, and monitoring capacity on disinformation;
    c. fostering online accountability; and

[138] European Commission (12 Jan 2018) *Experts Appointed to the High-Level Group on Fake News and Online Disinformation*, Press Release, https://ec.europa.eu/digital-single-market/en/news/experts-appointed-high-level-group-fake-news-and-online-disinformation

[139] High Level Expert Group (HLEG) on Fake News and Online Disinformation (2018) *A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation*, https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

Future policy options were discussed with Madeleine de Cock Buning (Chair of the High Level Expert Group on Fake News and Online Disinformation).

[140] Idem, p. 31.

[141] Idem, pp. 22-30.

[142] Idem, p. 35.

[143] Idem, pp. 31-33.

[144] EC Communication on Tackling Online Disinformation: a European Approach (COM(2018) 236 final)

   d. harnessing new technologies;

2. secure and resilient election processes;
3. fostering education and media literacy;
4. support for quality journalism as an essential element of a democratic society; and
5. countering internal and external disinformation threats through strategic communication.[145]

The Commission then convened a Multistakeholder Forum whose first output was an **EU Code of Practice on Disinformation**, published in September 2018. The emphasis lies here on commitments from online intermediaries, such as social media platforms, search engines and advertisers.[146] The Multistakeholder Forum is composed of a working group and a sounding board. In the code of practice, disinformation is defined as

> *'verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm.'*[147]

Noteworthy in this definition is the emphasis on 'verifiably' false or misleading information. The Code of Practice includes the following commitments:

1. scrutiny of ad placements, political and 'issue-based' advertising:
   a. disrupt advertising and monetisation incentives for relevant behaviours;
   b. ensure that advertisements are clearly distinguishable from editorial content;
   c. enable public disclosure of political advertising;
   d. use reasonable efforts towards devising approaches to publicly disclose 'issue-based advertising';
2. integrity of services:
   a. put in place clear policies regarding identity and the misuse of automated bots;
   b. put in place policies on what constitutes impermissible use of automated systems and to make this policy publicly available on the platform and accessible to EU users;
3. empowering users:
   a. help people make informed decisions when they encounter online news that may be false, including by supporting efforts to develop and implement effective indicators of trustworthiness in collaboration with the news ecosystem;
   b. invest in technological means to prioritise relevant, authentic and authoritative information;
   c. invest in features and tools to make it easier to find diverse perspectives;
   d. support efforts aimed at improving critical thinking and digital media literacy;
   e. encourage market uptake of tools that help consumers understand why they are seeing particular advertisements;
4. empowering the research community:

[145] Idem, pp. 6-16.

[146] European Commission (26 September 2018) *Code of Practice on Disinformation,* Press Release, https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

We conducted expert interviews with the Directors of two organisations on the Sounding Board: Monique Goyens (Director-General at European Consumer Organisation – BEUC, 31 August 2018) and Renate Schroeder (Director at European Federation of Journalists – EFJ, 7 Sept 2018).

[147] EU Code of Practice on Disinformation (2018), p. 1.

a. support good faith independent efforts to track and research disinformation and political advertising, including the independent network of fact-checkers facilitated by the European Commission;

b. convene an annual event to foster discussions within academia, the fact-checking community and members of the value chain.[148]

The Sounding Board, composed of representatives of the media, civil society, fact-checkers and academia, was scathing in its opinion on the self-regulatory approach:

> *'the 'Code of practice' as presented by the working group contains no common approach, no clear and meaningful commitments, no measurable objectives or KPIs, hence no possibility to monitor process, and no compliance or enforcement tool: it is by no means self-regulation, and therefore the Platforms, despite their efforts, have not delivered a Code of Practice.'*[149]

In sum, the recommendations in the Report of the High Level Expert Group[150] focus primarily on the role that social media platforms can play in supporting the media ecosystem, fact-checking and literacy efforts. The EC Communication[151] includes recommendations on media literacy and pluralism as well, although in a significantly reduced form. It adds reflections on election processes and strategic communication, which were given further attention during Commission President Juncker's State of the Union address in September 2018.[152] Importantly the Communication picked up on the HLEG's reflections on a transparent digital ecosystem and set up the EU multistakeholder forum.

The resulting EU Code of Practice[153] focuses on (electoral) ads, includes a short section on automated bots, and addresses the platforms' role in supporting/enabling literacy, fact-checking and research. The Code of Practice mainly recaps existing measures and does not aim to provide industry standards. Note that media trust (source transparency) indicators are mentioned in each of these proposals. Discussions are ongoing at several levels, including in the multistakeholder forum.[154]

3.1.3. Freedom of expression/media pluralism in online content regulation

Three leading initiatives emphasise freedom of expression and media pluralism in online content regulation.

Firstly, the **UN Special Rapporteur on the Promotion and the Protection of the Right to Freedom of Opinion and Expression** in April 2018 published a Report on a Human Rights Approach to Platform Content Regulation.[155] Similar to the Special Rapporteurs' Joint Declaration on Freedom of Expression and 'Fake News,' Disinformation and Propaganda (published in March

[148] Idem, pp. 4-8.

[149] European Commission (26 September 2018) *Code of Practice on Disinformation,* Press Release, https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[150] HLEG on Fake News and Online Disinformation (2018) *A Multi-Dimensional Approach to Disinformation*

[151] EC Communication on Tackling Online Disinformation: a European Approach (COM(2018) 236 final)

[152] EC President Juncker announced an Elections Package during his 2018 State of the Union address. See European Commission (2018) *State of the Union 2018: European Commission Proposes Measures for Securing Free and Fair European Elections,* Press Release (IP/18/5681), http://europa.eu/rapid/press-release_IP-18-5681_en.htm

[153] EU Code of Practice on Disinformation (2018)

[154] See e.g. Reporters Without Borders (3 April 2018) *RSF and its Partners Unveil the Journalism Trust Initiative to Combat Disinformation,* https://rsf.org/en/news/rsf-and-its-partners-unveil-journalism-trust-initiative-combat-disinformation; and Santa Clara University Markkula Center for Applied Ethics (2017) *The Trust Project,* https://thetrustproject.org

Media trust indicators were also discussed during an expert interview with Renate Schroeder (Director at European Federation of Journalists – EFJ, 7 Sept 2018).

[155] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) *Report to the United Nations Human Rights Council on A Human Rights Approach to Platform Content Regulation,* (A/HRC/38/35), https://freedex.org/wp-content/blogs.dir/2015/files/2018/05/G1809672.pdf

2017)[156], the report points to the balancing test on restrictions to freedom of expression (legality, necessity and proportionality, and legitimacy) and liability protection for technical intermediaries from third party content. The Special Rapporteur raises concerns around content standards. These pertain to vague rules, hate, harassment and abuse, context, real-name requirements, and disinformation.

The report sets the bar high, laying out human rights principles for company content moderation. The UN Special Rapporteur's human rights principles for company content moderation are:

- human rights by default, legality, necessity and proportionality, and non-discrimination when dealing with content moderation;
- prevention and mitigation of human rights risks, transparency when responding to government requests;
- due diligence, public input and engagement, rule-making transparency when making rules and developing products;
- automation and human evaluation, notice and appeal, remedy, user autonomy when enforcing rules; and
- decisional transparency.[157]

The Special Rapporteur raises concern on 'the delegation of regulatory functions to private actors that lack basic tools of accountability', indicating that their 'current processes may be inconsistent with due process standards and whose motives are principally economic'.[158] The report also specifies that 'blunt forms of action, such as website blocking or specific removals, risk serious interference with freedom of expression'[159] and that technological measures that restrict news content

> *'may threaten independent and alternative news sources or satirical content. Government authorities have taken positions that may reflect outsized expectations about technology's power to solve such problems alone.'*[160]

An interesting proposal is to launch and empower an independent 'social media council', similar to press councils in the newspaper sector, to provide transparency in how technology companies interpret and implement their standards, allow for industry-wide complaints and inter-company cooperation to provide remedies.[161]

Secondly, the Santa Clara Principles on Transparency and Accountability in Content Moderation[162] are listed in full. The principles were developed in early 2018 by a group of US academics and digital rights advocates concerned with free speech in online content moderation. It is not regulatory in nature, but can contribute to the policy debate, both within the US and the EU. These principles are detailed and suggest standards for transparency reporting, notice and appeal mechanisms, and will

[156] UN Special Rapporteur on Freedom of Opinion and Expression et. al. (2017) *Joint Declaration on Freedom of Expression and 'Fake News,' Disinformation and Propaganda,* UN Document FOM.GAL/3/17, https://www.osce.org/fom/302796?download=true

[157] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) *Report to the United Nations Human Rights Council on A Human Rights Approach to Platform Content Regulation* (A/HRC/38/35), https://freedex.org/wp-content/blogs.dir/2015/files/2018/05/G1809672.pdf, section IV, pars 44-63

[158] Idem, par 17

[159] Ibidem

[160] Idem, par 31

[161] Idem, pars 58, 59, 63, 72. A similar idea is raised in Wardle, (2017) supra n5.. We discuss this pertinent report more fully in Section 4.2.

[162] ACLU Foundation of Northern California, Center of Democracy and Technology, Electronic Frontier Foundation, New America's Open Technology Institute et.al. (2018) *Santa Clara Principles on Transparency and Accountability in Content Moderation,* https://santaclaraprinciples.org/

be included in the analysis in Section 3.2 (as a comparative example from outside the international and EU policy circles).

NUMBERS

Companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines.

At a minimum, this information should be broken down along each of these dimensions:

- total number of discrete posts and accounts flagged.
- total number of discrete posts removed and accounts suspended.
- number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by category of rule violated.
- number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by format of content at issue (e.g., text, audio, image, video, live stream).
- number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by source of flag (e.g., governments, trusted flaggers, users, different types of automated detection).
- number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by locations of flaggers and impacted users (where apparent).

This data should be provided in a regular report, ideally quarterly, in an openly licensed, machine-readable format.

NOTICE

Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension.

In general, companies should provide detailed guidance to the community about what content is prohibited, including examples of permissible and impermissible content and the guidelines used by reviewers. Companies should also provide an explanation of how automated detection is used across each category of content. When providing a user with notice about why her post has been removed or an account has been suspended, a minimum level of detail for an adequate notice includes:

- URL, content excerpt, and/or other information sufficient to allow identification of the content removed.
- the specific clause of the guidelines that the content was found to violate.
- how the content was detected and removed (flagged by other users, governments, trusted flaggers, automated detection, or external legal or other complaint). The identity of individual flaggers should generally not be revealed, however, content flagged by government should be identified as such, unless prohibited by law.
- explanation of the process through which the user can appeal the decision.

Notices should be available in a durable form that is accessible even if a user's account is suspended or terminated. Users who flag content should also be presented with a log of content they have reported and the outcomes of moderation processes.

APPEAL

Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.

Minimum standards for a meaningful appeal include:

- human review by a person or panel of persons that was not involved in the initial decision.
- an opportunity to present additional information that will be considered in the review.
- notification of the results of the review, and a statement of the reasoning sufficient to allow the user to understand the decision.

In the long term, independent external review processes may also be an important component for users to be able to seek redress.

Thirdly, the **European Parliament** in May 2018 published a **Resolution on Media Pluralism and Media Freedom in the European Union**.[163] The resolution highlights freedom of expression and freedom of opinion as:

> *'indispensable conditions for the full development of individuals and their active participation in a democratic society, for the realisation of the principles of transparency and accountability and for the fulfilment of other human rights and fundamental freedoms'.*[164]

Media freedom, pluralism and independence are deemed 'crucial components of the right to freedom of expression', as 'the media play an essential role in democratic society, by acting as public watchdogs, while helping to inform and empower citizens, through widening their understanding of the current political and social landscape, and fostering their conscious participation in democratic life'.[165]

As regards AI, the European Parliament urges institutions to refrain from:

> *'technical control over digital technologies – i.e. blocking, filtering, jamming and closing down digital spaces – or the de facto privatization of control measures by pressuring intermediaries to take action to restrict or delete internet content'.*[166]

It calls for full transparency in use of algorithms, artificial intelligence and automated decision-making. Flagging and fact-checking by users and third party organisations are encouraged, as well as 'avoiding the spread of unverified or untrue information with the sole purpose of increasing online traffic through the use of, for example so-called clickbait'.[167] Its positive measures proposed include investment in media plurality and independence, media and digital literacy, and communication strategies 'in order to empower citizens and online users to recognise and be aware of dubious sources of information and to spot and expose deliberately false content and propaganda'.[168]

3.1.4. Comparative analysis of policy recommendations

The previous sections make it clear that the multi-faceted nature of disinformation calls for varied action. Indeed, the HLEG stressed that 'the best responses to disinformation are multi-dimensional, with stakeholders collaborating in a manner that protects and promotes freedom of expression, media freedom, and media pluralism'[169]. In this section, the focus is on four key areas (advertising, automated processes, content/account moderation, and fact-checking/trustworthiness) where transparency has frequently been requested.

Explanation, reporting, review and appeal are all components of transparency and are crucial to any technology-based approach that seeks to minimise harm on freedom of expression. Transparency allows for better understanding of the effects of technological responses. The following Tables 3.2

[163] European Parliament Resolution on Media Pluralism and Media Freedom in the European Union (P8_TA(2018)0204), http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2018-0204+0+DOC+PDF+V0//EN

[164] EP Resolution on Media Pluralism and Media Freedom in the EU (2018), par A

[165] Idem, par E

[166] Idem, par 24

[167] Idem, par 35

[168] Idem, par 36

[169] HLEG on Fake News and Online Disinformation (2018), p. 3

and 3.3 map and assess the provision on transparency as found in the key policy initiatives discussed above.

AI technologies are limited in their accuracy, especially when needing to assess contextual or cultural cues. In particular, mandating proactive measures to detect and remove illegal content endangers freedom of expression. This is especially the case when the proposal is broadly defined, as in the European Commission's Communication and Recommendation on Tackling Illegal Content Online.[170]

The Communication provides some backstops, calling for clear policies, notices and reversibility safeguards. These were not included in the Recommendation. There is acknowledgement in the proposals that human oversight is necessary 'where detailed assessment of the relevant context is required'[171]. Distinctions are thus made between types of content. The Communication considers that 'fully automated deletion or suspension of content is acceptable when its illegality has already been established'[172]. The Recommendation calls for automated prevention (filtering) of re-upload of terrorist content. The UN Special Rapporteur and the European Parliament prioritise freedom of expression and privacy over security, and call for human rights approaches/impact assessments in product and policy development.

In the tables below, the authors notice that recommendations on **advertising** are only included in more recent proposals dealing specifically with disinformation. It draws attention to the use of advertising for disinformation purposes, even though existing legislation[173] already provides protection and remedies for consumers. The proposals call for transparency on the source and placement of ads, as well as informing users why they are seeing certain ads. The EC Communication on Tackling Online Disinformation takes this approach one step further and recommends restricting targeting options for political advertising.[174] The HLEG report also highlights the 'follow-the-money' approach, which aims at restricting advertising and thus revenues of promoters of disinformation.[175]

Indeed, the broader question at hand when reflecting on economic drivers underlying disinformation are clickbait practices, which the policy initiatives additionally recommend tackling through trustworthiness or **source transparency** indicators, and prioritisation of 'relevant, authentic and authoritative'[176] and alternative content. As will be further explored in Section 3.2, source transparency is arguably preferable over prioritisation of trustworthy content or restricted advertising options from the perspective of plurality of voices and opinions.

Finally, it is striking that the recent proposals specific to disinformation do *not* address **content/account moderation**, despite its relevance in providing transparency, appeal and review in the removal of disinformation. Recommendations on content/account moderation can rather be found in the EC's Recommendation and Communication on Tackling Illegal Content Online[177], the

[170] EC Communication on Tackling Illegal Content Online (COM(2017) 555 final, p.12) calls to 'adopt effective proactive measures to detect and remove illegal content', the EC Recommendation on Effective Measures to Tackle Illegal Content Online (C(2018) 1177 final), preamble 24) echoes, mentioning 'proportionate and specific proactive measures taken voluntarily by hosting service providers, including by using automated means in certain cases'.

[171] EC Recommendation on Effective Measures to Tackle Illegal Content Online (C(2018) 1177 final), par. 20

[172] EC Communication on Tackling Illegal Content Online (COM(2017) 555 final), p.14

[173] For instance, EU Directive 2006/114/EC concerning Misleading and Comparative Advertising

[174] EC Communication on Tackling Online Disinformation: a European Approach (COM(2018) 236 final) p.7

[175] HLEG on Fake News and Online Disinformation (2018) p. 32

[176] EU Code of Practice on Disinformation (2018) commitment 8

[177] EC Recommendation on Measures to Effectively Tackle Illegal Online Content (C(2018) 1177 final); EC Communication on Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms (COM(2017) 555 final)

UN Special Rapporteur's Report on Platform Content Regulation[178] , and the Santa Clara Principles on Transparency and Accountability in Content Moderation[179]. In these latter proposals, emphasis is given to clear policies, notice/appeal opportunities and explanation for content removal or account suspension. The European Commission makes an exception where notification and appeal of removal is not deemed appropriate for criminal offences.[180]

Transparency reports to allow for monitoring of the effects of measures are also widely supported. The UN Special Rapporteur calls for the development of industry-wide accountability mechanisms on content moderation, and mentions the possible need for legislative and judicial intervention to ensure robust remediation.[181]

Table 3.2 Mapping technical tools against policy proposals (overview)

Transparency proposals per *technological concern*	EU Code of Practice on Disinformation (Sept 2018)	EC Communication on Tackling Online Disinformation (April 2018)	EU HLEG Report on Fake News and Online Disinformation (March 2018)	EC Recommendation Measures to Effectively Tackle Illegal Content Online (March 2018)	EC Communication on Tackling Illegal Content Online (Sept 2017)	UN Special Rapporteur Report on Platform Content Regulation (April 2018)	Santa Clara Principles on Transparency and Accountability in Content Moderation (May 2018)	EP Resolution on Media Pluralism / Freedom (May 2018)
(Political) advertising	**X**	**X**	**X**					
Automated processes (automated filtering, automated ranking, etc.)	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Content / account moderation				**X**	**X**	**X**	**X**	
Source transparency and fact-checking	**X**	**X**	**X**					**X**

[178] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) *Report to the United Nations Human Rights Council on A Human Rights Approach to Platform Content Regulation* (A/HRC/38/35)

[179] ACLU Foundation of Northern California, Center of Democracy and Technology, Electronic Frontier Foundation, New America's Open Technology Institute et.al. (2018) *Santa Clara Principles on Transparency and Accountability in Content Moderation,* https://santaclaraprinciples.org/

[180] EC Recommendation on Measures to Effectively Tackle Illegal Online Content (C(2018) 1177 final) par 10; EC Communication on Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms (COM(2017) 555 final) p.17

[181] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) pars 58-59

Table 3.3 Mapping technical tools against policy proposals (details)

Transparency proposals per *technological concern*	
(Political) advertising	**EU Code of Practice on Disinformation (Sept 2018)** Scrutiny of ad placement through e.g. brand safety, third party verification, additional assessment opportunities for advertisers Separation of editorial content and advertising Transparency on why people are seeing certain ads Transparency on source and spending of political ads *In the annex on best practices: (political) advertising policies* **EC Communication on Tackling Online Disinformation (April 2018)** Restrictions and transparency on sponsored content, especially political ads Transparency on why people are seeing certain ads **EU HLEG Report on Fake News and Online Disinformation (March 2018)** Transparency on funding source Transparency on ad placement Distinction between sponsored content, including political advertising, and other content Discourage dissemination and amplification of disinformation for profit, based on clear, transparent, and non-discriminatory criteria
Automated processes (automated filtering, automated ranking, etc.)	**EU Code of Practice on Disinformation (Sept 2018)** • Clear policies regarding identity and misuse of automated bots • *In the annex on best practices: service integrity policies, such as spam and misrepresentation policies* **EC Communication on Tackling Online Disinformation (April 2018)** • Clear marking systems and rules for bots • Detailed information on algorithms that prioritise display of content, as well as development of testing methodologies • Voluntary use of trustworthy electronic identification and authentication systems (to enable traceability of disinformation source) **EU HLEG Report on Fake News and Online Disinformation (March 2018)** • Transparent and relevant information on functioning of algorithms, without prejudice to IPRs **EC Recommendation Measures to Effectively Tackle Illegal Content Online (March 2018)** • Human oversight and verifications where appropriate and, in any event, where detailed assessment of relevant context is required **EC Communication on Tackling Illegal Content Online (Sept 2017)** • Human-in-the-loop principle in determining the illegality of content in areas where error rates are high or contextualisation is necessary • Reversibility safeguard for (erroneous) automated re-upload filters • Mention use and performance of automated re-upload filters in terms of service • Counter notices and notification of decision also for automated content removal **UN Special Rapporteur Report on Platform Content Regulation (April 2018)**

	• Rules rooted in rights, rigorous human rights impact assessments for product and policy development, operations with ongoing assessment, reassessment and meaningful public and civil society consultation (adoption of Guiding Principles on Business and Human Rights) • Recognition of significant limitations of automation. At a minimum, rigorous audit and development of technology with broad user and civil society input **Santa Clara Principles on Transparency and Accountability in Content Moderation (May 2018)** • Explanation on how automated content detection is used **EP Resolution on Media Pluralism / Freedom (May 2018)** • Full transparency in use of algorithms, artificial intelligence and automated decision-making • Human rights-based approach with remedies and safeguards for any EU digital policy and strategy
Content/account moderation	**EC Recommendation Measures to Effectively Tackle Illegal Content Online (March 2018)** • Easy to access, user-friendly mechanisms for submission of notices, with confirmation of receipt and follow-up on decision (where contact details are known) • Notification of removal / disabling, opportunity for counter notices and notification of decision (where contact details are known, with exception of criminal offences) • Out-of-court dispute settlement on removal / disabling • Clear, easily understandable and sufficiently detailed explanation of content policy, with detail on content considered illegal • Transparency reports on number of notices and counter-notices received and the time needed for taking action, amount and type of content removed (focus on content considered illegal) • In case of terrorist content, confirmation of receipt and follow-up on decision with competent authority or Europol • Member States and hosting service providers to submit reports to Commission to allow for monitoring of effects **EC Communication on Tackling Illegal Content Online (Sept 2017)** • Easily accessible and user-friendly mechanism for submission of notices, with confirmation of receipt and follow-up on decision • Opportunity for counter notices (with exception of criminal offences) and notification of decision (incl. reasons in case of negative review) • Clear, easily understandable and sufficiently detailed explanation of content policy in terms of service, with detail on safeguards to prevent over-removal • Transparency reports on number and type of notices received and actions taken, the time taken for processing, the source of the notification, and counter notices (possibility of standardisation mentioned) **UN Special Rapporteur Report on Platform Content Regulation (April 2018)** • Radically different approaches to transparency at all stages of operations, from rule-making to implementation and development of 'case law' framing the interpretation of private rules • Development of industry-wide accountability mechanisms, such as a social media council to hear complaints from individual users and gather public feedback on recurrent content moderation problems, such as over-censorship (possible need for legislative and judicial intervention mentioned) **Santa Clara Principles on Transparency and Accountability in Content Moderation (May 2018)** • Detailed explanation on content removal or account suspension • Meaningful opportunity to appeal content removal or account suspension

	• Notification of results of review, incl. reasons for decision • Transparency reports on the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines
Source transparency and fact-checking	**EU Code of Practice on Disinformation (Sept 2018)** • Invest in products, technologies and programs to help people make informed decisions, e.g. trustworthiness indicators • Prioritise 'relevant, authentic and authoritative information' • *In the annex on best practices: advice, guides, tools on fact-checking* **EC Communication on Tackling Online Disinformation (April 2018)** • Develop fair, objective, and reliable indicators for source transparency • Facilitate (alternative) content discovery • Develop European network of fact-checkers **EU HLEG Report on Fake News and Online Disinformation (March 2018)** • Develop source transparency indicators, focusing on source, ownership, and adherence to ethical and journalistic codes • Test integration of source transparency indicators in ranking algorithms • Dilute disinformation with quality information • Develop plug-ins / apps to provide better information access control, e.g. content displayed according to quality signals, alternative content recommendations, reporting tools • Increase transparency and efficiency of fact-checking practices through data sharing and collaborative research • Publish reports and appeals processes on flagging and fact-checking **EP Resolution on Media Pluralism / Freedom (May 2018)** • Independent and impartial certified third-party fact-checking organisations • Obligations and instruments in relation to source verification • Enable users to report and flag potential disinformation • Display and label disinformation revealed as such to stimulate public debate and prevent re-emergence

3.2. Technological initiatives

In the following section, the existing AI-enabled actions taken by technical intermediaries are mapped with a special focus on filtering/removal, blocking and (de)prioritisation of content and advertising, as well as disabling/suspension of accounts. They pertain to content/account moderation, but are relevant to the specific case of disinformation as well. The actions of three major intermediaries are also mapped against this typology and scrutinised compared to the key technological concerns raised in the previous section.

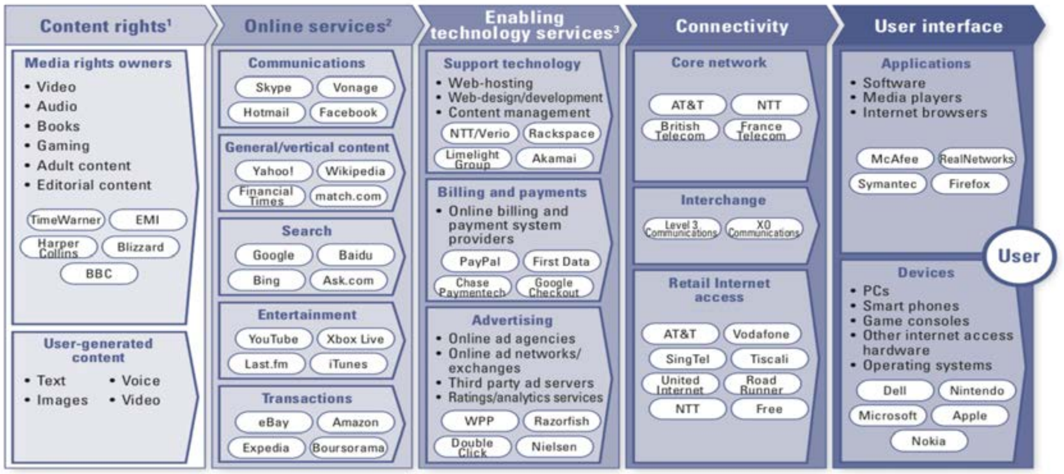
3.2.1. Filtering, blocking and (de)prioritisation

As the policy proposals above indicate, social media platforms and search engines are the target of requests to respond to disinformation online. However, it is necessary to recognise that they only constitute one set of technological players within a much larger internet ecosystem. Within A.T.

Kearney's[182] overview of the internet value chain, they would be identified as being part of online services (see Figure 3.2). When considering technological initiatives (especially within the European Parliament's meaning in its May 2018 Resolution on Media Freedom and Pluralism – 'blocking, filtering, jamming and closing down digital spaces'[183]), we need a broader scope and also examine the actions undertaken by e.g. email providers, operating systems, cloud providers, network providers and internet access providers.

In the categorisation below, different methods of content/account moderation are first highlighted. Then a brief comparison of the actions taken by three prominent platforms: Facebook, Google and Twitter is provided.

Figure 3.2 A.T. Kearney's internet value chain[184]



Notes: ¹Content rights abbreviated to CR in subsequent value chains
²See online services categories list in methodology for details
³Enabling technology/services abbreviated to ETS in subsequent value chains

Source: A.T. Kearney analysis

Filtering or removal of content (including advertisements) is the most effective, but also the most invasive, method of tackling disinformation. We understand filtering of content to be an *ex ante* measure that technical providers undertake to prevent the upload/posting of content, while removal of content is *ex post* and upon awareness, request or order[185].

The closer filtering or removal takes place to the source, the more accurate it will usually be. Thus it is preferable for the content creator to take this action; but media platforms, websites and apps might also intervene and remove content of their users. So for instance, online media platforms will prevent the upload of copyrighted content by users. YouTube Content ID is a widely deployed example of ex-ante filtering. With YouTube Content ID, uploaded files are scanned against

[182] A.T. Kearney (2010) *internet Value Chain Economics. Gaining a Deeper Understanding of the internet Economy,* https://www.atkearney.com/documents/20152/434237/internet-value-chain-economics.pdf/285d1a4d-a49c-43d1-5966-9fcca69aa55a

[183] European Parliament Resolution on Media Pluralism and Media Freedom in the European Union (P8_TA(2018)0204) http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2018-0204+0+DOC+PDF+V0//EN

[184] A.T. Kearney (2010) *internet value chain economics. Gaining a deeper understanding of the internet economy,* https://www.atkearney.com/documents/20152/434237/internet-value-chain-economics.pdf/285d1a4d-a49c-43d1-5966-9fcca69aa55a

[185] Barnes, R., Cooper, A., Kolkman, O. Thaler, D., and Nordmark, E. (2016) *RFC 7754 – Technical Considerations for internet Service Blocking and Filtering, March 2016*, https://tools.ietf.org/html/rfc7754#page-27

databases of works provided by copyright owners. If a match is found, copyright owners can decide to either block, monetise or track a video containing their work.[186]

Notice and Action procedures are also widely used for the removal of content (this is also used for blocking of content). With the exception of manifestly illegal (such as child abuse images or terrorist content) or disruptive (such as spam, viruses) content, *ex post* take down is preferable, as appeal of the measure is usually possible. It is worth noting that prevention of uploading content can also de facto happen at a network level, if network and internet access providers (de)prioritise or block certain protocols.

Table 3.4 Internet filtering, implications and examples

Filtering	**Implication**	**Example**
Ex ante	Content scanned prior to upload/post	Checking content based on available databases, e.g. YouTube Content ID
Ex post	Take down of content	Notice and Action procedures based on request/order

Blocking of content (including advertising) is a widely available technological solution, from users, email providers, search engines and social media platforms to network and internet access providers, as access to the original content is not required (see Table 3.5). For instance, Cleanfeed is a technology deployed by British Telecom in 2004 to ensure users would not access previously identified alleged child pornography content. It differs from ex-ante filtering in two respects: the content is not removed, but rather the user's access is blocked; there should in theory be an opportunity to reverse-engineer the filter and discover which content has been blocked. Appeals against blocking using this system are possible in some derivatives of the system[187]. All blocking can usually be circumvented by the use of a Virtual Private Network (VPN) to encrypt the user's traffic and obfuscate their location[188].

We adopt a broad definition of content: it can include (user-generated) content, advertising, spam, viruses, websites, etc. Similar to filtering, blocking can occur ex-ante and thus voluntary, or ex-post and upon awareness, request or order, with ex-post blocking taking preference from the perspective of users' ability to review and appeal. In cases where content is permissible under the provider's terms of service, but is culturally or legally unacceptable, blocking will occur to ensure that the content is blocked in specific locations, but available in others.

As an example, search engine filters can prevent access to search for the content, as with Google's rules against hate speech, or to prevent access to specific copyright-infringement material following a court order. Table 3.5 below shows the technologies, examples and their usage.

[186] Google (2018) *How Content ID Works* https://support.google.com/youtube/answer/2797370?hl=en

[187] See Marsden, C. (2011) *Internet Co-Regulation*, Cambridge: Cambridge University Press, at 166-178.

[188] Clayton, R. (2005) *Anonymity and Traceability in Cyberspace*, Cambridge Computer Lab Technical Report 653, http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-653.pdf

Table 3.5 Blocking/filtering technologies, use and examples

Blocking	**Usage**	**Examples**
Browser-based	Implemented direct or via a third party browser extension	Advert blocking or cookie blocking e.g. Privacy Badger
E-mail and messaging	Filter headers, such as sender and subject, to accept or reject messages	Most email systems use Bayesian filters to stop spam
Client-side	Installed as software on each computer	Every common anti-virus defence
Search engine	Can prevent access to search for content	Google's rules against hate speech, or to prevent access to specific copyright-infringement material following a court order
DNS-based, implemented at the DNS layer	Prevent lookups for domains that do not fit within a set of policies	Not Safe For Work rules; OpenDNS
Content-limited access providers	Can offer access to 'safe' internet content on an opt-in or mandatory basis	Implement parental, government or regulatory control over use
Network-based	Implemented at network or application layer as a web proxy	Cleanfeed child abuse filters

(De)prioritisation of content (including advertising) occurs at many levels as well, from user options to see less content related to particular persons or subjects (e.g. Facebook's options to hide certain ads) to automated ranking through algorithms and network-based solutions(e.g. 'shadow banning')[189]. We opt for a broad definition of (de)prioritisation: it can also include (de)prioritisation of advertising (thus implying demonetisation) or certain internet protocols (e.g. P2P).

In the context of disinformation, deprioritisation might entail content given less prominence in users' feeds. Prioritisation occurs when content from certain providers, such as media organisations or fact checkers, is given preference over or shown side-by-side with false information. At the level of network and internet access providers, paid prioritisation of particular content or services has been contested within net neutrality debates, as it gives preference/voice to big commercial players[190]. Google was also found to have abused its dominant position in search to prioritise its own comparison shopping services.[191] Algorithmic (de)prioritisation is key to the user experiences that social media platforms and search engines offer, yet is not straightforward in its implications for freedom of expression and media pluralism.

Disabling and suspension of accounts (including advertising accounts) are temporary and permanent solutions for technology providers to deal with their users' abuse of their terms of service and/or legislation. Important to note is that this action implies a client-service relationship, although the service can be offered free of charge to the client. It is available to a wide range of actors, such as email providers, social media platforms, cloud services, network providers and internet access

[189] Wagner, K. (14 April 2018) *Here's How to See, Edit and Delete the Topics that Facebook Advertisers Use to Target You,* Recode, https://www.recode.net/2018/4/14/17236072/facebook-mark-zuckerberg-ad-advertising-pixel-data

[190] Marsden, C. (2017) *Network Neutrality: From Policy to Law to Regulation,* Manchester: Manchester University Press.

[191] European Commission (27 June 2017) Antitrust: Commission Fines Google €2.42 Billion for Abusing Dominance as Search Engine by Giving Illegal Advantage to Own Comparison Shopping Service, Factsheet, http://europa.eu/rapid/press-release_MEMO-17-1785_en.htm

providers. Disabling and suspension of accounts often occurs at scale. As the severity and/or frequency of the violation increases, the action becomes more punitive and permanent. Take for instance the HADOPI graduated response mechanism for copyright regime in France. When a violation was noticed, an internet user first received two warnings (aimed at education and prevention) before the case was transferred to the judiciary.[192]

Finally, in this context, the implications of innovative deployment of technologies, such a decentralised web (Dweb[193]) technologies are not yet clear. DWeb has been encouraged by funding in Horizon Europe and other EU programmes. Specific funding for (AI- and blockchain-based) technologies to prevent disinformation was included in ICT funding under H2020 in 2018[194]. In October 2018, the European Parliament passed a **Resolution on Distributed Ledger Technologies and Blockchains**, in which Recital 54 '[c]alls on the Commission to evaluate the safety and efficiency of electronic voting systems, including those that employ DLTs [Distributed Ledger Technologies], for both private and public sectors'[195]. The extent to which the funding for, and incentives to deploy, such technologies is encouraged by the different disinformation options considered in Chapter 4 would need careful impact assessment[196]. It can be expected that corporate developments based in the United States will continue to fund research, such as the MIT-Harvard Ethics and the Harvard Berkman-Klein Center for Internet and Society Governance of AI Initiative[197], even in the absence of European support. Similarly, they will continue to research filtering technologies.

[192] Sanchez, L. (14 Aug 2018) *Hadopi: Beaucoup d'Avertissements Mais Peu de Condamnations*, LeMonde.fr, https://www.lemonde.fr/les-decodeurs/article/2018/08/14/hadopi-beaucoup-d-avertissements-mais-peu-de-condamnations_5342325_4355770.html. Also see Meyer, T. (2012) Graduated Response In France: The Clash of Copyright and the Internet, *Journal of Information Policy 2,* 107-127, DOI: 10.5325/jinfopoli.2.2012.0107 for a critical review of the French initiative.

[193] Benet, J. (2014) *IPFS - Content Addressed, Versioned, P2P File System (DRAFT 3)*, https://ipfs.io/ipfs/QmR7GSQM93Cx5eAg6a6yRzNde1FQv7uL6X1o4k7zrJa3LX/ipfs.draft3.pdf. See more generally, Ayala, D. (31 July 2018) *Introducing the Dweb*, Mozilla Blog, https://hacks.mozilla.org/2018/07/introducing-the-d-web/

[194] COM(2018) 236 final supra Chapter 3.1.4 at p.11, stating specifically:

'Artificial intelligence, subject to appropriate human oversight, will be crucial for verifying, identifying and tagging disinformation;

– Technologies for media to enable customizable and interactive online experiences can help citizens discover content and identify disinformation;

– Innovative technologies, such as blockchain, can help preserve the integrity of content, validate the reliability of information and/or its sources, enable transparency and traceability, and promote trust in news displayed on the internet. This could be combined with the use of trustworthy electronic identification, authentication and verified pseudonyms; and

– Cognitive algorithms that handle contextually-relevant information, including the accuracy and the quality of data sources, will improve the relevance and reliability of search results.'

This led to the allocation of €21m funds in HORIZON2020 Work Programme 2018-20 Topic ICT-28-2018: Future Hyper-connected Sociality, one of whose funding lines was for projects that 'Provide measures against disinformation online and stimulate trust and a positive vision as to the role of Social Media & Networks'.

[195] European Parliament Resolution on Distributed Ledger Technologies and Blockchains: Building Trust with Disintermediation (P8_TA-PROV(2018)0373 B8-0397/2018), http://www.europarl.europa.eu/sides/getDoc.do?type=TA&reference=P8-TA-2018-0373&language=EN&ring=B8-2018-0397

[196] Code of Practice Agora (2014) *The Principles for Better Self- and Co-Regulation*, https://ec.europa.eu/digital-single-market/en/best-practice-principles-better-self-and-co-regulation#Article

[197] MIT-Harvard Ethics and Harvard Berkman-Klein Center for Internet and Society (2018) *The Ethics and Governance of Artificial Intelligence Initiative*, https://aiethicsinitiative.org/

3.2.2. Analysis of Facebook, YouTube and Twitter content policies

In the following table, this taxonomy of technology-based solutions is used to map the actions taken by three of the major intermediaries, Facebook, Google and Twitter.

Table 3.6 Facebook, YouTube and Twitter content policies

Content/account moderation	**Facebook**[198]	**YouTube**[199]	**Twitter**[200]
Flagging	Both machine and human-driven Team of 20.000 staff dealing with content review Use of external counsel	Both machine and human-driven Users, trusted flaggers Team of 10,000 staff dealing with content review for violations of user agreement Use of external counsel	Both machine and human-driven Users, trusted reporter Twitter Trust and Safety Council provides input on safety products, policies and programs[201]
Filtering and removal of content and advertising	Both filtering and removal Content is removed where it can cause physical harm, e.g. terrorist content, hate speech, self-harm/suicide	Both filtering and removal E.g. YouTube Content ID and hashing technologies to prevent re-uploading	Both filtering and removal E.g. no ads on Irish abortion referendum from outset (in line with its policy to ban advertising on

[198] Analysis primarily based on Facebook (2018) *Community Standards,* https://www.facebook.com/communitystandards/introduction; Facebook (2018)*Written Evidence for Lords Communications Committee - The Internet: To Regulate or Not To Regulate?* http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/communications-committee/the-internet-to-regulate-or-not-to-regulate/written/83287.html; Facebook (2018*) German NetzDG Transparency Report (Jan-Jun 2018),* https://fbnewsroomus.files.wordpress.com/2018/07/facebook_netzdg_july_2018_english-1.pdf ; EU Code of Practice on Disinformation. Annex II Current Best Practices from Signatories of the Code of Practice (2018) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[199] Analysis primarily based on expert interview with Jon Steinberg (Public Policy and Government Relations Manager for EMEA at Google, 30 August 2018); YouTube (2018) *Community Guidelines,* https://www.youtube.com/yt/about/policies/#community-guidelines; Google UK (2018) *Written Evidence for Lords Communications Committee - The Internet: To Regulate or Not To Regulate?* http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/communications-committee/the-internet-to-regulate-or-not-to-regulate/written/83086.pdf; Google (2018) *German NetzDG Transparency Report (Jan-Jun 2018),* https://transparencyreport.google.com/netzdg/youtube; EU Code of Practice on Disinformation. Annex II Current Best Practices from Signatories of the Code of Practice (2018) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[200] Analysis primarily based on expert interview with Stephen Turner (Head of Public Policy for Belgium at Twitter, 25 September 2018); Twitter (2018) *Rules and Policies*, https://help.twitter.com/en/rules-and-policies#research-and-experiments; Twitter (2018) *Oral Evidence for Lords Communications Committee - The Internet: To Regulate or Not To Regulate?* https://parliamentlive.tv/Event/Index/2cd62e7a-d3cf-4605-8d39-4fbaa0adaa76#player-tabs

Twitter (2018) *German NetzDG Transparency Report (Jan-June 2018)* https://cdn.cms-twdigitalassets.com/content/dam/transparency-twitter/data/download-netzdg-report/netzdg-jan-jun-2018.pdf ; EU Code of Practice on Disinformation. Annex II Current Best Practices from Signatories of the Code of Practice (2018) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[201] Twitter (2018) *The Twitter Trust and Safety Council,* https://about.twitter.com/en_us/safety/safety-partners.html

Stephen Turner (Head of Public Policy for Belgium at Twitter, 25 Sept 2018) discussed the role of Twitter's Trust and Safety Council in our expert interview. He describes: '[a]s part of our policy development process, Twitter works with a Trust and Safety Council which consists of a range of NGOs, academics and civil society groups from across the world, feeding into a range safety issues from digital rights to child welfare. The Council inputs feedback on new Twitter products and policies, for example by factoring local cultural contexts into the decision-making process. It is a very communal approach on how we deal with policies, how we can rebalance, how we take our enforcement measures forward, making sure that we are trying our best to strike the right balance.'

			certain health issues, including abortion) When action is taken on existing content, Twitter blocks, but requires user to remove Recent emphasis on proactively identifying and challenging problematic accounts
Blocking of content and advertising	E.g. cultural/national restrictions on content; decision to ban foreign spending on political ads during Irish abortion referendum Users can hide, snooze pages/people	E.g. cultural/national restrictions on content; no advertising on pornographic websites; decision to ban ads on Irish abortion referendum Users can enable YouTube Restricted Mode	At direct message-level: stop conversation, place message behind an interstitial At tweet-level: provide warning message, require deletion, meanwhile *place tweet behind an interstitial i.e. block* Users can block and mute accounts
(De)prioritisation of content and advertising	Prioritisation of content from sources the community rates as trustworthy Clickbait is tackled by reducing prominence of content with a headline that 'withholds information or if it exaggerates information separately'[202] Facebook adds context by placing fact-checked articles underneath disinformation Users can prioritise pages/people (see first in feed) Source labels for (political) advertising and pages In Spring 2019, roll out of political ads library across EU[203]	Prioritisation of authoritative sources Other than take down, content may be restricted through a.o. age restrictions In Spring 2019, roll out of new political advertising transparency tools[204]	At tweet-level: *provide warning message,* require deletion, meanwhile place tweet behind an interstitial i.e. block Twitter Global Ads Transparency Center[205] In Spring 2019, launch of Twitter EU Elections Center[206]
Disabling or suspension of accounts	Graduated approach: warning, limited features, suspension of accounts Ability to see and change which advertising categories you have been put in	Graduated approach (strikes): warning, limited features, suspension of accounts Ability to see and change which advertising categories you have been put in	Graduated approach: require changes to profile/media, place in read-only mode, verify account ownership, permanent suspension

[202] Babu, A., Lui, A., and Zhang, J. (17 May 2017) 'New Updates to Reduce Clickbait Headlines', *Facebook Newsroom,* https://newsroom.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/

[203] European Commission (16 October 2018), *Roadmaps to Implement the Code of Practice on Disinformation -- Facebook*, Press Release, https://ec.europa.eu/digital-single-market/en/news/roadmaps-implement-code-practice-disinformation

[204] European Commission (16 October 2018), *Roadmaps to Implement the Code of Practice on Disinformation – Google*, Press Release, https://ec.europa.eu/digital-single-market/en/news/roadmaps-implement-code-practice-disinformation

[205] Twitter (2018), *Ads Transparency Center,* https://ads.twitter.com/transparency

[206] European Commission (16 October 2018), *Roadmaps to Implement the Code of Practice on Disinformation – Twitter*, Press Release, https://ec.europa.eu/digital-single-market/en/news/roadmaps-implement-code-practice-disinformation

			Ability to switch off algorithm, show which advertising categories that you have been put in
Appeal	Notice of action, possibility to appeal, notification of decision Changes to appeals process: previously appeal only possible for profiles, pages, groups, roll out for individual posts as well (first for nudity/sexual activity, hate speech or graphic violence)[207]	Notice of action, possibility to appeal, notification of decision	Notice of action, possibility to appeal, notification of decision
Support for fact-checking, journalism, literacy	E.g. Facebook Journalism Project[208]	E.g. Google News Initiative, Digital News Innovation Fund[209], YouTube Fund to Support Journalism[210], Be Internet Legends and Be Internet Citizens[211] In Spring 2019, launch of Fact Check Explorer[212]	E.g. support for UNESCO[213], First Draft, Atlantic Council's Digital Forensic Research Lab, City University, Global Health Initiative[214] [215]

Facebook, Google and Twitter published roadmaps for implementing the EU Code of Practice on Disinformation in October 2018.[216] The roadmaps emphasise training of political groups in light of elections, updating policies on fake accounts and launching political advertising transparency tools. When scrutinising the actions of the three online intermediaries through the lens of freedom of expression, several encouraging practices are noted. In terms of increasing transparency, Facebook

[207] Bickert, M. (24 April 2018) 'Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process', *Facebook Newsroom*, https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/

[208] Facebook (2018) *Facebook Journalism Project*, https://www.facebook.com/facebookmedia/solutions/facebook-journalism-project

[209] Google News Initiative (2018) *Digital News Innovation Fund*, https://newsinitiative.withgoogle.com/dnifund/report/european-innovation-supporting-quality-journalism/

[210] Mohan, N., and Kyncl, R. (9 July 2018) 'Building A Better News Experience on YouTube, Together', *YouTube Official Blog*, https://youtube.googleblog.com/2018/07/building-better-news-experience-on.html

[211] Google (2018) *Be Internet Legends*, https://beinternetlegends.withgoogle.com/en-gb

YouTube (2018) *Be Internet Citizens*, https://internetcitizens.withyoutube.com

Public Policy and Government Relations Manager for EMEA at Google, Jon Steinberg emphasized the need for user education in our expert interview (30 August 2018), stating '[t]here was a saying in the US when I was growing up that 'you can't believe everything that you hear on television'. Well, this certainly applies to the internet today. Savvy users are less susceptible to being misled, so we should work in partnership to be educate users in their media practices."

[212] European Commission (16 October 2018), *Roadmaps to Implement the Code of Practice on Disinformation – Google*, Press Release, https://ec.europa.eu/digital-single-market/en/news/roadmaps-implement-code-practice-disinformation

[213] UNESCO (25 October 2018) *UNESCO Partners with Twitter on Global Media and Information Literacy Week 2018*, https://en.unesco.org/news/unesco-partners-twitter-global-media-and-information-literacy-week-2018

[214] Gadde, V.and Gasca, D. (2018) *Measuring Healthy Conversation*, Twitter Blog, https://blog.twitter.com/official/en_us/topics/company/2018/measuring_healthy_conversation.html

[215] For a full list of partners, see Twitter (2018) *Partnerships*, https://about.twitter.com/en_us/values/elections-integrity.html#Partnerships

[216] European Commission (16 October 2018), *Roadmaps to Implement the Code of Practice on Disinformation – Google*, Press Release, https://ec.europa.eu/digital-single-market/en/news/roadmaps-implement-code-practice-disinformation

published its internal enforcement guidelines in April 2018[217], and Twitter unveiled its Global Ads Transparency Center in June 2018[218]. In dealing with content moderation, all three companies seek to minimise harm by notifying those whose content has been taken down or account has been suspended. There is opportunity to appeal the decision, and a graduated approach in enforcing the rules (from warning, limited features to suspension of accounts) is in place.

Action is taken on the basis of existing law and/or community guidelines. Community guidelines are more restrictive, and thus limit speech, beyond what is legally required. During the Irish referendum, for instance, all three intermediaries banned in different ways referendum or abortion-related ads.[219] Private companies with global reach are determining, in a currently uncoordinated manner, what is acceptable expression, under their Terms of Service enforcement. Further, the tension between moderating content and media pluralism became evident in an exchange with Jon Steinberg (Public Policy and Government Relations Manager for EMEA at Google)[220] on the prioritisation of 'authoritative sources':

> *Steinberg: I think the point at which platforms feel uncomfortable (indeed, at which everyone should feel uncomfortable) is when we rely on platforms to determine what is truth. We prioritise authoritative sources, we will give more voice to NYT, Fox News, Washington Post, etc., but we cannot determine the veracity of the content.*
>
> *Interviewer: are there drawbacks to Google's approach?*
>
> *Steinberg: sure, I think the prioritisation of authoritative sources makes it particularly difficult for new new brands to emerge. So it can affect the plurality of voices in particular.*

More controversial even than algorithmic ranking is automated content recognition and removal. It is not clear how often and under which circumstances *ex ante* filtering or blocking take place on the platforms. Some is machine-driven[221], but it is unclear how the illegality of the content or its violation with the community guidelines is determined, nor what the safeguards are in place to prevent over-censoring of content. Transparency on the frequency and categories of content filtering is absent to external audiences, nor does it seem that appeal is possible. In line with the recommendation included in the EC Communication on Tackling Illegal Content Online, an opportunity for a counter-notice should be provided, also when the content removal is automated.[222] In terms of best practices on transparency and accountability, the authors point to the Santa Clara Principles. One useful recommendation they provide on appeals is to ensure 'human review by a person or panel of persons

[217] Babu, A., Lui, A., and Zhang, J. (17 May 2017) 'New Updates to Reduce Clickbait Headlines', *Facebook Newsroom,* https://newsroom.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/

[218] Falck, B. (28 June 2018) 'Providing More Transparency Around Advertising on Twitter', *Twitter Blog,* https://blog.twitter.com/official/en_us/topics/company/2018/Providing-More-Transparency-Around-Advertising-on-Twitter.html

[219] During the Irish abortion referendum, Facebook banned foreign spending on political ads; Google banned referendum-related ads; and Twitter's ad policy on certain health issues prohibits abortion-related advertising. See O'Brien, C., and Kelly, F. (9 May 2018) Google Bans Online Ads on Abortion Referendum, *The Irish Times,* https://www.irishtimes.com/business/media-and-marketing/google-bans-online-ads-on-abortion-referendum-1.3489046

In our expert interview, Joe McNamee (Executive Director at European Digital Rights – EDRI, 6 Sept 2018) expressed strong criticism of the lack of government action in the Irish referendum, leading to different self-regulatory actions by the online platforms.

[220] Expert interview with Jon Steinberg (Public Policy and Government Relations Manager for EMEA at Google, 30 August 2018)

[221] Based on databases with previously detected content, to our knowledge, terrorist content, child sex abuse images and copyright infringing content are regularly prevented from upload.

[222] The Communication goes on to specify that '[i]n certain circumstances, informing the content provider and/or allowing for a counter-notice would not be appropriate – in particular in cases where this would interfere in the investigative powers of Member States' authorities necessary for the prevention, detection and prosecution of criminal offences, such as in the case of child sexual abuse material.' EC Communication on Tackling Illegal Content Online (COM(2017) 555 final), p.17

that was not involved in the initial decision'.[223] The Principles start to flesh out the high-level human-rights based approach advocated by the UN Special Rapporteur on Freedom of Opinion and Expression[224].

3.3. Disinformation initiatives per regulatory type

At the end of Chapter 2, a range of regulatory options (from self-regulation to legislation) are identified. These will be elaborated upon further in the policy options. Here, this study maps them against four types of disinformation (public, private; electoral, foreign). Our focus on public, private, electoral, and foreign disinformation is deliberate. The study seeks to illustrate differences in the available approaches depending on the destination and origin of the disinformation. Public/private explains differences between exchanges that are posted in public or private fora. Electoral/foreign are both strategic forms of political influence. The study views the former as originating primarily from domestic political actors, while the latter is foreign political influence, whether government or private.

The comparison demonstrates that (a) there are existing basic measures that can be implemented/enforced in order to tackle the issue of disinformation. These may, however, be more difficult to use in the context of private exchanges. Moreover, there are additional sets of legislation that can be used in specific situations, such as elections to regulate the actions of both domestic and foreign actors. Further, the mapping illustrates that (b) technical standards are primarily about exchange of databases and best practices. So, for instance, there is not one standardised way to implement 'Notice and Action' across similar technical intermediaries in Europe. Approaches are at this point in time 'tailored' to the specificity of the intermediary in question. This creates flexibility, yet also incoherence. The diversified approach to content moderation has also been raised on multiple occasions in public consultations on the review of the E-Commerce Directive. It did not, however, result in legislative change.[225] The Santa Clara principles are an attempt to overcome some of the un-level playing fields that have been created in online content moderation through self-regulation.

[223] ACLU Foundation of Northern California, Center of Democracy and Technology, Electronic Frontier Foundation, New America's Open Technology Institute et.al. (2018) *Santa Clara Principles on Transparency and Accountability in Content Moderation,* https://santaclaraprinciples.org/

[224] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) *Report to the United Nations Human Rights Council on A Human Rights Approach to Platform Content Regulation,* A/HRC/38/35, https://freedex.org/wp-content/blogs.dir/2015/files/2018/05/G1809672.pdf; UN Special Rapporteur on Freedom of Opinion and Expression et. al. (2017) *Joint Declaration on Freedom of Expression and 'Fake News,' Disinformation and Propaganda,* UN Document FOM.GAL/3/17, https://www.osce.org/fom/302796?download=true

[225] See Meyer, T.(2017) *The Politics of Online Copyright Enforcement in the EU: Access and Control,* Cham: Palgrave Macmillan for detailed assessment of intermediary liability discussions in Europe.

Table 3.8 Mapping of disinformation types per regulatory archetype

Examples of *Destination / Origin of Disinformation* per Regulatory Archetype	**Media literacy / trust**	**Technical standards**	**Self-regulation**	**Co-regulation**	**Legislation**
Public	Literacy on clickbait Debunking and fact-checking, e.g. First Draft[226], Full Fact[227], Faktisk[228], AFP fact-checking[229] Developments of trustworthiness indicators, e.g. Santa Clara University Trust Project trust indicators[230], AFP, EBU & GEN's Journalism Trust Initiative[231]	Sharing of hash databases and practices, e.g. EU Internet Forum on Terrorist Content Online[232], Global Internet Forum to Counter Terrorism[233] Sharing of fact-checked sources and practices, e.g. International Fact Checking Network[234] Human/machine learning to detect disinformation (networks), e.g. EU	Filtering of content, e.g. spam, hashed terrorist content Removal and blocking of content, accounts based on terms of service Prioritisation of authentic, authoritative content Proposed codes of conduct for technical intermediaries, such as Santa Clara Principles on Transparency and	Removal and blocking of content, accounts (notice and action) based requests from users, third parties and law enforcement (Proposed) codes of conduct for technical intermediaries, such as EU Code of Conduct on Countering Illegal Hate Speech Online[238], EU Code	Legislation on intermediary liability, misuse of electronic networks Legislation against defamation, incitement to hatred, violence Legislation on consumer protection, e.g. against misleading advertising Legislation on data protection

[226] Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy (2018) *First Draft*, https://firstdraftnews.org/about/

[227] Full Fact (2018) *The UK's Independent Factchecking Charity*, https://fullfact.org

[228] Funke, D.(3 October 2017) *Three Months After Launching, Faktisk is Already Among the Most Popular Sites in Norway*, Poytner Institute, https://www.poynter.org/news/three-months-after-launching-faktisk-already-among-most-popular-sites-norway

[229] Agence France Press (2018) *Fact Check,* https://factcheck.afp.com/fact-checking-afp

[230] Santa Clara University Markkula Center for Applied Ethics (2017) *The Trust Project*, https://thetrustproject.org

[231] Reporters Without Borders (3 April 2018) *RSF and its Partners Unveil the Journalism Trust Initiative to Combat Disinformation*, https://rsf.org/en/news/rsf-and-its-partners-unveil-journalism-trust-initiative-combat-disinformation

[232] European Commission (3 December 2015) *EU Internet Forum: Bringing Together Governments, Europol and Technology Companies to Counter Terrorist Content and Hate Speech Online*, Press Release (IP/15/6243) http://europa.eu/rapid/press-release_IP-15-6243_en.htm

[233] Global Internet Forum to Counter Terrorism (2018) *Vision*, https://gifct.org

[234] Poytner Institute (2018) *Fact-Checking*, https://www.poynter.org/channels/fact-checking

[238] EU Code of Conduct on Countering Illegal Hate Speech Online (2016) http://europa.eu/rapid/press-release_IP-15-6243_en.htm

	Transparency on why ads are shown Enable users to prioritise, block certain content, users Enable users to view alternative feed	DisinfoLab[235], University of Michigan algorithm[236]	Accountability in Content Moderation[237]	of Practice on Disinformation[239]	
Private	Enable users to prioritise, block certain content, users Coverage on e.g. mob lynching or anti-immunisation messages	Sharing of hash databases (note: not possible for encrypted services, such as WhatsApp)	Filtering of content (note: not possible for encrypted services, such as WhatsApp) Removal and blocking of content, accounts based on terms of service Limits on redistribution of messages, e.g. on WhatsApp[240]	Removal and blocking of content, accounts (notice and action) based on requests from users and law enforcement (note: no insight for third parties)	*In addition to above (public):* Specific interpretation of existing legislation to target private messaging, such as Indian Penal Code (IPC) and IT Act[241]
Electoral	*In addition to above (public):* Coverage of electoral practices	Sharing of fact-checked sources and practices	Removal and blocking of content, accounts based on terms of service	Notice and action based on requests from electoral commissions	*In addition to above (public):* Electoral regulation e.g. ad spending caps, distinction between editorial content and advertising

[235] EU DisinfoLab (2018) *About Us,* http://disinfo.eu/aboutus/

[236] University of Michigan (21 August 2018) *Fake News Detector Works Better than A Human*, https://news.umich.edu/fake-news-detector-algorithm-works-better-than-a-human/

[237] ACLU Foundation of Northern California, Center of Democracy and Technology, Electronic Frontier Foundation, New America's Open Technology Institute et.al. (2018) *Santa Clara Principles on Transparency and Accountability in Content Moderation,* https://santaclaraprinciples.org/

[239] EU Code of Practice on Disinformation (2018) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[240] Funke, D. (20 July 2018) *WhatsApp Is Limiting Message Forwarding to Cut Down on Fake News*, Poytner Institute, https://www.poynter.org/news/whatsapp-limiting-message-forwarding-cut-down-fake-news

[241] India has used its Indian Penal Code (IPC) and IT Act to pursue administrators of WhatsApp groups, as well as the technical intermediary itself. See for instance, Phartiyal, S. (28 August 2018) *India's Top Court Seeks WhatsApp's Response on Petition Alleging It Breaches Law*, Reuters.com, https://www.reuters.com/article/us-whatsapp-india/indias-top-court-seeks-whatsapps-response-on-petition-alleging-it-breaches-law-idUSKCN1LD0SL and BBC.com (23 July 2018) *WhatsApp 'Admin' Spends Five Months in an Indian Jail,* https://www.bbc.com/news/technology-44925166

	Transparency on source and spending of political ads Sharing of election practices, e.g. Alliance of Democracies' Transatlantic Commission on Election Integrity[242]		Prioritisation of authentic, authoritative content Blackout periods prior to elections, e.g. as took place during Irish abortion referendum[243]		Criminal law if conspiracy
Foreign	*In addition to above (public):* Examples of debunking and fact-checking, specific to foreign actors: EEAS East Stratcom Task Force[244], Atlantic Council's DisinfoPortal.org[245] Coverage of foreign interference, e.g. Ukrainian Prism's Disinformation Resilience Index[246]	*In addition to above (public):* Human/machine learning to detect bots and cyberattacks, e.g. Internet Governance Project's development of a Transatlantic Attribution Institution[247]	*See above (public)*	*See above (public)*	*In addition to above:* Multilateral standards e.g. Cybercrime Treaty

[242] Alliance of Democracies (18 May 2018) *Transatlantic, Bi-Partisan Commission Launched to Prevent Election Meddling*, Press Release, http://www.allianceofdemocracies.org/initiatives/the-campaign/press_release_tcei/

[243] Facebook decided during the referendum to ban all related foreign ads. Google/YouTube decided during the referendum to ban all related ads. Twitter banned abortion ads from outset based on its advertising policy. See for instance, Satariano, A. (25 May 2018) *Ireland's Abortion Referendum Becomes a Test for Facebook and Google,* New York Times, https://www.nytimes.com/2018/05/25/technology/ireland-abortion-vote-facebook-google.html

[244] EEAS (2017) *Questions and Answers about the East StratCom Task Force* https://eeas.europa.eu/headquarters/headquarters-homepage/2116/-questions-and-answers-about-the-east-stratcom-task-force_en

[245] Atlantic Council Eurasia Center (2018) *Disinfoportal.org,* https://disinfoportal.org/about-disinfo-portal/

[246] Ukrainian Prism's Foreign Policy (2018) *Disinformation Resilience in Central and Eastern Europe*, http://prismua.org/en/dri-cee/

[247] Badii, F. (21 August 2018) *Is It Time to Institutionalize Cyber Attribution,* Internet Governance Project, https://www.internetgovernance.org/2018/08/21/new-igp-white-paper-is-it-time-to-institutionalize-cyber-attribution/ We discussed the Transatlantic Attribution Institution during an expert interview with Milton Mueller (Professor at Georgia Institute of Technology School of Public Policy and Director of the Internet Governance Project, 6 August 2018).

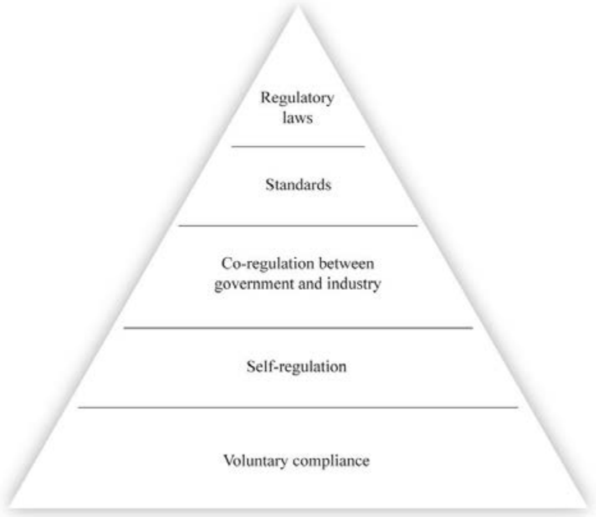
4. Policy options

Chapter 4 presents policy options, paying particular attention to interactions between technological solutions, freedom of expression and media pluralism. The opportunities and drawbacks of various self-regulatory to legislation legislative options are explored.

4.1. Options for regulating AI in disinformation introduced

There are several levels of AI disinformation policy options open to European policymakers, ranging from Option 0 (no new regulation but, further research and analysis into current self- and state regulation) to Option 5 (specific legislative instruments). Given the absence of European level regulation beyond those instruments, there is no deregulation option presented. Following the Millwood-Hargrave scale discussed in Chapter 2, six options are provided for technical means to moderate and remove disinformation. This study considers – without prejudice to the Parliament and its committees' work – the status quo/no-regulation option (0) and the pair that propose formal regulation (4-5) to be the least likely to be deployed, and focus on the Options 1-3 that propose self-regulatory outcomes. The authors believe legislation for freedom of expression may be premature and potentially hazardous with regard to fundamental rights: collaboration between different stakeholder groups with public scrutiny is preferable, where effectiveness can be independently demonstrated. Most importantly, options are interdependent – where regulation is proposed, it sits atop a pyramid of activities including co-regulation, self-regulation, technical standards and individual company/NGO/academic initiatives. There is no single option to solve the problem of disinformation.

Figure 4.1 Reeve model of regulatory pyramid[248]



With this context in mind, this study views the six options for regulating automated content recognition technology in disinformation as follows:

- **Option 0**: Status quo, noting that this would entail permitting both 'natural' technical experiments in moderation, research into creating evidence-based policy as outlined above, and the legislative responses that already exist.

[248] Reeve, B.(2011) 'The Regulatory Pyramid Meets the Food Pyramid: Can Regulatory Theory Improve Controls on Television Food Advertising to Australian Children?', *Journal of Law and Medicine 19(1),* 128-46.

- **Option 1**: Non-audited self-regulation, with increasing industry-government coordination, but no sanction on those companies choosing not to cooperate in standards.[249]
- **Option 2**: Audited self-regulation, under which for instance the code of practice on disinformation would be subjected to formal published audit by a commonly agreed self-regulator.[250]
- **Option 3:** A formal self-regulator, recognised by the European institutions and ideally with funding separate from the industry.
- **Option 4**: Formal co-regulation, in which the regulator is independent from government yet subject to prior approval of codes of conduct, systems for funding and arbitration.
- **Option 5**: Statutory regulation, in which a regulator is tasked to combat disinformation directly by licensing of content providers and their systems for content moderation. Current electoral and broadcast regulators already perform this function for offline media.

Given what AI use and abuse reveals about disinformation practices, potential actions are summarised in the table below.

Table 4.2 Typology of regulation and implications

Option and form of regulation	**Typology of regulation**	**Implications/Notes**
0 Status quo	Corporate social responsibility, single-company initiatives	Note that enforcement of the new General Data Protection Regulation and the proposed revised ePrivacy Regulation, plus agreed text for new AVMS Directive, would all continue and likely expand
1 Non-audited self-regulation	Industry code of practice, transparency reports, self-reporting	Corporate agreement on principles for common technical solutions and Santa Clara Principles
2 Audited self-regulation	European Code of Practice of September 2018; Global Network Initiative published audit reports	Open interoperable publicly available standard e.g. commonly engineered/designed standard for content removal to which platforms could certify compliance
3 Formal self-regulator	Powers to expel non-performing members, dispute resolution ruling/arbitration on cases	Commonly engineered standard for content filtering or algorithmic moderation. Requirement for members of self-regulatory body to conform to standard or prove equivalence. Particular focus on content 'put back' metrics and efficiency/effectiveness of appeal process
4 Co-regulation	Industry code approved by Parliament(s) or regulator(s) with statutory powers to supplant	Government-approved technical standard – for filtering or other forms of moderation. Examples from broadcast and advertising regulation
5 Statutory regulation	Formal regulation – tribunal with judicial review	National regulatory agencies – although note many overlapping powers between agencies on e.g. freedom of expression, electoral advertising and privacy

[249] Marsden (2011) supra xx, pp.107-113.

[250] Such as UK Safer Internet Centre (2018) for reporting and removing child sex abuse images online, https://www.saferinternet.org.uk/

4.2. Options for regulating AI in disinformation explained

In the following section, the options are laid out in more detail.

Option 0: Status quo

This study notes that this option would entail permitting both 'natural' technical experiments in moderation, and the legislative responses that already exist, such as that of Germany's Network Enforcement Law (NetzDG). However, it would also rely on individual corporate efforts to enforce, rather than an industry self-regulation scheme or democratically legitimate institutional oversight. Individual users would continue to rely on companies' terms of service enforcement for their own and others' freedom of expression (with widely varying content standards, definitions of abusive/harmful content etc.).

Individual companies would continue to pursue disparate aims according to their own judgement of brand interest (e.g. Google decided not to accept political advertising during the 2018 referendum on the Thirty-sixth Amendment of the Constitution Act in Ireland, whereas Facebook only banned foreign actors' adverts). Executive Director at European Digital Rights, Joe McNamee, is highly critical of the current reliance on terms of service enforcement. He points out that it can lead to avoiding the question whether an action is legal or illegal:

> *'Providers have two options: they can say 'it's a terms of service violation or it breaks the law'. If they say, 'it's the terms of service violation, they're not accusing anyone doing anything illegal'. In countries that would require a report to law enforcement, they would not have to report to law enforcement. The provider is happy because it's easy to delete, there's no liability, and they don't upset their user. The person who uploaded the latest incitement to violence is happy because 'phew, that could have been a problem, but I got away with it'. And law enforcement is saying 'Thank God, we're not getting those reports'. You are facilitating the circumvention or the non-application of the law, which is dangerous.'*[251]

The idea that a multinational public social media company acts as its own government with its own 'supreme court' was promulgated by Mark Zuckerberg in April 2018,[252] but is clearly a case of corporate social responsibility over-reach.[253]

However, much can be achieved using non-traditional regulatory tools to control AI use. We can classify the proposed solutions put forward by a highly influential Shorenstein Center for Media, Politics and Public Policy (at Harvard Kenney School) report for the Council of Europe, which is represented in the table below, with responses by platforms, news providers and governments identified separately.[254] This table provides insight into the many uses of (existing) self-regulatory approaches to the disinformation problem and shows how much might be achieved without formal regulatory intervention.

[251] Expert interview with Joe McNamee (Executive Director at European Digital Rights – EDRi, 6 Sept 2018)

[252] Kozlowska, H. (3 April 2018) 'Mark Zuckerberg Floated a 'Supreme Court' for Facebook. What Does That Mean?', *Quartz*, https://qz.com/1243203/mark-zuckerberg-floated-a-supreme-court-for-facebook-what-does-that-mean/

[253] On the role of multinationals in regulation generally, see Ruggie, J. (2018) 'Multinationals as Global Institution: Power, Authority and Relative Autonomy, *Regulation & Governance (2018)12*, 317–333, https://onlinelibrary.wiley.com/doi/pdf/10.1111/rego.12154

[254] Wardle, C.and Derakhstan, H. (2017) *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making* (DGI(2017)09), Shorenstein Center on Media, Politics and Public Policy at Harvard Kennedy School for the Council of Europe, https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking

Table 4.3 Shorenstein Center recommendations classified according to regulatory options (bold type indicates the authors' analysis of option)

What could technology companies do?	What could media organisations do?	What could national governments do?
Create an international advisory council. Members from a variety of disciplines that can (1) guide technology companies as they deal with information disorder and (2) act as an honest broker between technology companies. **Option 3 Formal self-regulation, similar to UN Rapporteur's call for a Social Media Council**	Collaborate. Newsrooms and fact-checking organisations should collaborate to prevent duplications of effort and free journalists to focus on other investigations. **Option 0/2 (Existing) audited self-regulatory collaboration (e.g. IFCN)**	Commission research to map information disorder. The methodology should be consistent across these research studies exercises, so that different countries can be accurately compared. **Option 0/4 Status quo, but regulatory attempt to standardise research**
Provide researchers with the data related to initiatives aimed at improving the quality of information. **Option 0/1 (Recent existing) non-audited self-regulation, included in EU Code of Practice**	Agree policies on strategic silence. News organisations should work on best practices. **Option 1 Non-audited self-regulation**	Regulate ad networks. **Option 5 Statutory regulation**
Provide transparent criteria for any algorithmic changes that down-rank content. Without this transparency, there will be claims of bias and censorship from different content producers. **At the very least, Option 2 Audited self-regulation**	Ensure strong ethical standards across all media. News organisations have been known to sensationalise headlines on Facebook in ways that would not be accepted on their own websites. News organisations should enforce the same content standards, irrespective of where their content is placed. **Option 0/1-2 (Existing) self-regulation**	Require transparency around Facebook ads. There is currently no oversight in terms of who purchases ads on Facebook, what ads they purchase and which users are targeted. National governments should demand transparency about these ads so that ad purchasers and Facebook can be held accountable. **Option 5 Statutory regulation**
Work collaboratively... encourage such collaboration, particularly when it involves sharing information about attempts to amplify content. **Option 0/2: (Existing) audited self-regulation (note: original report refers to terrorism and child abuse)**	Debunk sources as well as content. When content is being pushed out by bot networks, news organisations should identifying this as quickly as possible. This will require journalists to have computer programming expertise. **Option 0/1-2 (Existing) self-regulation**	Support public service media organisations and local news outlets. **Option 0/4 (Existing) co-regulation, but additional government support**
Highlight contextual details and build visual indicators. We recommend that social networks and search engines automatically surface contextual information and metadata that would help users ascertain the truth of a piece of content. **Option 1-2 Self-regulation**	The news media should produce more segments and features which teach audiences how to be critical of content they consume … they should explain to the audience how the process of verification was undertaken. **Option 1-2 Self-regulation**	Roll out advanced cyber-security training. **Option 0/2, Status quo, but reinforced collaboration. Included in platforms' roadmaps for implementing the EU code of practice**

Eliminate financial incentives. Technology companies as well as advertising networks more generally must devise ways to prevent purveyors of dis-information from gaining financially. **Option 0/1-2 (Existing) self-regulation among industries**	News and media organisations have a responsibility to educate audiences about the scale of information pollution worldwide, and the implications society faces because of it. **Option 0/1-2 (Existing) self-regulation**	Enforce minimum levels of public service news on platforms. Encourage platforms to work with independent public media organisations to integrate quality news into users' feeds. **Option 4-5: Co-regulation (as involves terms of licensing for PSBs) possible need for statutory change to platform regulation**
Crack down on computational amplification. Take stronger and quicker action against automated accounts used to boost content. **Option 0/1-2 (Existing) AI training**	Focus on improving the quality of headlines. Research using natural language processing techniques are starting to automatically assess whether headlines are overstating the evidence available in the text of the article. **Option 0/1-2 (Existing) self-regulation**	
Adequately moderate non-English content. Social networks need to invest in technology and staff to monitor mis-, dis- and mal-information in all languages. **Option 0/1-2 (Existing) AI training**	Do not disseminate fabricated content. Clickbait headlines, the misleading use of statistics, unattributed quotes adding to the polluted information ecosystem. **Option 0/1-2 (Existing) self-regulation**	
Pay attention to audio/visual forms of disinformation. **Option 1-2 AI training for 'deep fakes'**		
Provide metadata to trusted partners. **Option 0/1-2 (Existing) AI training**		
Build fact-checking verification tools. **Option 0/1-2 (Existing) AI training**		
Search engines to build out 'authenticity' engines and water-marking technologies to provide mechanisms for original material to be surfaced and trusted. **Option 1-2 AI training**		

The benefits of no regulation are the classic United States common law of the libertarian 'marketplace of ideas' to combat disinformation. However, the costs are that only research and evaluation could be carried out by government, with no carrot-and-stick threat to regulate. Sustainability would be jeopardised by any political calculation that disinformation has overwhelmed the media ecosystem's own established defences, and this study concludes that the 2016-17 electoral/referendum evidence shows substantial failures in the regulatory ecosystem for the media, notably with regard to bot accounts and unregulated online political advertising. The study also sees no future in Europe for an unregulated online free-for-all.[255] Much detailed internet

[255] Described by French former culture minister Jack Lang as 'the freedom of the fox in the barnyard': see Muravchik, J. (1998) *The Future Of The United Nations: Understanding The Past To Chart A Way Forward*, American Enterprise Institute

regulation is self-regulation despite such profound constitutional issues of fundamental rights. This is because US companies have implemented in terms of service the 'negative liberty' framework of the US First Amendment which stops 'Congress' intervening in the liberty of the press. By contrast, European law has 'positive obligations' including Paragraph 2 of Article 10 of the European Convention on Human Rights 1950 require states to intervene to protect rights. Despite US claims of the exceptionalism of free speech, Option zero is not an option for European legislators.

This study therefore argues that the initiatives identified by Shorenstein for the Council of Europe should be targeted and encouraged by European institutions, in the interests of a better approach to tackling disinformation. Option zero is only effective if the disinformation problem is held to be capable of self-healing by market actors and individuals without the need for more formal coordination, investment or even direct regulation.

Option 1: Non-audited self-regulation

This option would increase platform activity compared with Option zero in terms of preventing immediate regulatory intervention, with increasing industry-government coordination, but no sanction on those companies choosing not to cooperate. Many examples can be found in the Shorenstein table above. Government and private industry research funding could be increased to encourage machine learning-based and other forms of content moderation.[256] The EU code of practice on disinformation proposed by companies under the aegis of the European Commission would continue to be developed. However, the lack of formalised transparency processes (other than reporting) makes this option ineffective and potentially damaging to the European policy process, and thus it is an unsatisfactory hybrid option as compared to Option zero or Option 2.

The Santa Clara Principles for Content Moderation are a step towards Option 1. European Union funding for the World Wide Web Consortium is an example of technical sponsorship to help internet self-regulated standards.[257] In the AI space, standards for ethical algorithms are being developed by for instance the IEEE P7000 scheme,[258] but critics have pointed out that these ethical norms are the predecessor to legal standards.[259] Therefore, any ethical code that becomes an industry standard for certification, especially in an area affecting fundamental rights like algorithmically determined content recognition, is likely to lead to a call for legislative standards and enforcement.

for Public Policy Research, Washington, D.C., https://epdf.tips/the-future-of-the-united-nations-understanding-the-past-to-chart-a-way-forward.html, at p.85.

[256] See for instance publications of the European Union funded ENCASE Social Computing project: https://encase.socialcomputing.eu/publications

Zinonos, S., Tsirtsis, A., and Tsapatsoulis, N. (2018) 'Twitter Influencers or Cheated Buyers?', *IEEE Cyber Science and Technology Congress*; Mariconti, E. et al. (2018) ''You Know What to Do'': Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks', *ArXiv*; Zannettou, S. et al. (2018) 'On the Origins of Memes by Means of Fringe Web Communities', *ACM Internet Measurement Conference (IMC)*; Zannettou, S. et al. (2018) 'The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans', *ArXiv*; Founta, A-M. et al. (2018) 'A Unified Deep Learning Architecture for Abuse Detection', *ArXiv*; Founta, A-M. et al. (2018) 'Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior', *International AAAI Conference on Web and Social Media (ICWSM)*; Zannettou, S. et al. (2018) 'The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube', *1st Deep Learning and Security Workshop, co-located with the 39th IEEE Symposium on Security and Privacy.*

[257] Marsden, C. (2011) *Internet Co-Regulation,* Cambridge: Cambridge University Press, pp. 107-113.

[258] IEEE (2018) *Global Initiative on Ethics of Autonomous and Intelligent Systems,* https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

[259] @rcalo: 'Now that I'm on my high horse, let me *specifically disavow* @IEEEorg's efforts to create an ethical certification program. IEEE is an important organisation we should look to for thought leadership. But offering an ethical certification is as dangerous as it is premature.' (23 October 2018) https://twitter.com/rcalo/status/1054834789570633729

Option 2: Audited self-regulation

Under audited self-regulation, the self-regulatory scheme is subject to regular (even annual) independent audit to ascertain the degree to which members are cohering to the criteria. For instance, the code of practice would be subjected to formal published audit by a commonly agreed self-regulator; an example is INHOPE, the pan-European hotline associated co-funded originally under the safer internet action plan.[260] Members of the EU High Level Expert Group on Disinformation, Clara Jiménez Cruz, Alexios Mantzarlis, Rasmus Kleis Nielsen and Claire Wardle, argue that: '[f]act-checking technology has an important role to play, provided it is independent and free from any political influence. Platforms can provide client-based interfaces for control and guidance on selecting, for example, priorities in news searches and news feeds, diversity of opinions on consumer time lines and the re-posting of fact-checked information. Platforms need to be transparent about their algorithms'.[261]

In the AI disinformation scheme, this audit could be undertaken by the industry body, as with the GNI, or by a self-regulator from an associated industry, for instance broadcasting or games classification (see Option 3). The HLEG members argue that: 'Google, Facebook and Twitter have now taken a public commitment to work with researchers who can independently assess the spread and impact of disinformation. The [EC disinformation] report specifically calls on major technology companies to provide data that would allow the independent assessment of efforts like Google's fact-check tags, Facebook's use of fact-checks as Related Articles or the downgrading of disinformation in the News Feed'. Jiménez Cruz et al. argue for: '[t]he creation of a network of Research Centers focused on studying disinformation across the EU, [as the] current knowledge base is almost entirely focused on the United States data'.[262] This is a vital area for further funded research by the European institutions.

The cost-benefit of audited self-regulation depends on the level of independence and rigour of the auditor function. It allows for flexible regulation, though efficiency depends on industry actors' commitment to the independence and rigour of the auditor in the absence of any penalty for lack of compliance, often a fatal failing.[263] Lower costs and more responsive regulation are possible, 'free riders' are very likely to exist, though the scale of the larger platforms and the existing code of practice commitments may ensure greater scrutiny. In essence, Jiménez Cruz et al. argue that Option 2 is best suited to the current evidence, for a 'structured process ahead that will document progress made and expose anyone not taking their responsibilities seriously'.[264]

Feasibility and effectiveness depend on the implementation of audit. Sustainability of audited self-regulation is very low, given the possibilities for non-compliance identified above. Human rights challenges will exist even with an independent multistakeholder board, so that self-audit is inevitably judged inadequate and may be supplanted by more formal regulatory bodies. Risks and future uncertainties are thus very high, and there is no satisfactory example of audited self-regulation on the internet without the backstop of formal regulation. Take for example the time-limited Google Advisory Council on the Right to be Forgotten,[265] a legal right which is subsequently

[260] UK Safer Internet Centre (2018) supra xx

[261] Jiménez Cruz, C., Mantzarlis, A., Nielsen, R.K., and Wardle, C. (12 March 2018), 'Six Points from the EU Commission's New Report on Disinformation, *Medium,* https://medium.com/@hlegresponse/six-key-points-from-the-eu-commissions-new-report-on-disinformation-1a4ccc98cb1c

[262] Ibidem. Note a network of Centres on Internet and Society already exists, and is currently studying this area, with circa 35 European centres, chaired over time by Politecnico de Torino (NEXA Centre) and Humblodt University: see https://networkofcenters.net/centers

[263] In the expert interview, Monique Goyens (Director-General at European Consumer Organisation – BEUC, 31 August 2018) expressed it in the following way: 'I have been in the job of consumer activism for more than thirty years. I have seen a lot of self-regulation. I have not seen much that has worked.'

[264] Jiménez Cruz, C., Mantzarlis, A., Nielsen, R.K., and Wardle, C. (12 March 2018), *supra 260.*

[265] Google (2015*) Google Advisory Council on the Right to be Forgotten*, https://archive.google.com/advisorycouncil/

subject to regulatory and court enforcement and was thus not an example of audited self-regulation. The GNI claims such an audit function, but annual reports do not give detail such that it would satisfy these criteria.[266]

Option 3: Formal self-regulator

This regulator would be recognised by the European institutions and ideally with funding separated from the industry. Recognition does not signal statutory power to intervene or to direct the regulator, but does indicate that the institutions wish to guide the choice of self-regulatory scheme employed, short of intervention via legislation.

An example is the Pan European Game Information (PEGI) scheme, under which 30 000 computer game products have been labelled and classified to indicate violence, sexual content, and other types of content that may give human dignity/child protection concerns, using the graphical warnings of the Netherlands *Kijkwijzer* scheme implemented by the Netherlands Institute for the Classification of Audio-visual Media, and the UK Video Standards Council.[267] App store games are regulated using the International Age Rating Coalition system.[268] PEGI is not formally regulated, but claims: 'PEGI is used and recognised throughout Europe and has the enthusiastic support of the European Commission. It is considered as a model of European harmonisation in the field of the protection of children'.[269]

Applied to AI and disinformation, this schematic would suggest a multistakeholder or at least EU institutions-industry dialogue establishing general principles applying to an AI regulator, while the self-regulator would set out details of the scheme design. Such principles may include, for instance, the principle that no account can be suspended without human intervention to correct for false positive identification of a bot account, and the potential for account holder appeal against such a deletion. As noted in Chapter 3, the UN Special Rapporteur on Freedom of Opinion and Expression has recommended such a body to deal with online content moderation.

However, note that human-regulated AI is more likely to be guaranteed with robust co-regulation than self-regulatory schemes (see following section).

The cost-benefit of self-regulation is held in general to allow for very flexible regulation, though efficiency depends on industry actors confirming to the rating scheme. Lower costs and more responsive regulation are possible, though 'free riders' who fail to conform fully may exist.

Feasibility and effectiveness depend on the initial design, as well as the implementation of that design by the self-regulator. A problem can be that the lack of sanctions for inappropriate labelling or failure to conform to standards may not be subject to a robust system of audit and correction.

Sustainability of self-regulation is always an issue. Internet regulation is often implemented directly by legislatures due to particularly profound constitutional and human rights challenges including freedom of expression and prevention of harm, so that self-regulation is judged inadequate and supplanted by state regulatory bodies. Risks and future uncertainties are thus closely tied to the regulatory commitment to making self-regulation an end state (subject to satisfactory independent audit of procedures) rather than an interim measure.

Coherence with EU objectives are easier to assess with co-regulation than with self-regulation because the national statutory criteria establishing the co-regulator must conform to European law principles, and ex-post comparative evaluation across Member States can more easily be

[266] Global Network Initiative (2018) *Annual Report 2017: Reinforcing a Global Standard,* https://globalnetworkinitiative.org/global-network-initiative-annual-report-2017-reinforcing-a-global-standard/

[267] *Kijkwijzer* (2018) *Netherlands Institute for the Classification of Audio-visual Media,* http://www.kijkwijzer.nl/nicam and Pan European Game Information (2018) *How We Rate Games*, https://pegi.info/page/how-we-rate-games

[268] International Age Rating Coalition (2018) *How IARD Works*, http://www.globalratings.com/how-iarc-works.aspx

[269] Marsden, C. (2011) supra xx and PEGI (2018) *PEGI Age Ratings*, https://pegi.info/page/pegi-age-ratings

undertaken given these common criteria. The divergence of regulatory means used for areas such as child protection and video on demand over the two decades of European consumer internet law show that a level of co-existence of different regulatory schemes is possible with national differences.

Potential ethical, social and regulatory impacts revolve around the media pluralism dilemma. The fundamental rights issues with co-regulation are similar to those for less direct regulatory interventions – freedom of expression as a fundamental right may be held inappropriate for anything but state regulation, a constant issue in internet regulation.

Option 4: Formal co-regulation

Formal co-regulation comprises a regulatory system in which the regulator is independent from government, making regulation subject to prior approval of codes of conduct, systems for funding and independent appeal. In Germany, this is known as 'regulated self-regulation'.[270] This is a hybrid system subject to statutory control. Examples from the internet regulatory ecosystem are:

- the largest European Domain Name System Registry operator, Nominet, which operates the .uk domain since 1996, under ultimate control by government via the Digital Economy Act 2010;[271]
- EURID which regulates and operates registries under the .eu domain since 2003.[272]

Note that this body would censor citizens directly, so the right to appeal to an independent adjudicator must be built in. The regulator could be associated with and certified/approved by state regulatory bodies, such as the EU Fundamental Rights Agency or European Data Protection Board.

Co-regulation offers the statutory underpinning and legitimacy of parliamentary approval for regulatory systems, together with general principles of good regulation, such as independence from regulatees, appeal processes, audit and governance principles. It also devolves the responsibility for these practices to an independent body, which theoretically gives agility and flexibility to the regulator within these general principles. As the Regulation establishing the .EU domain explains:

> *'Internet management has generally been based on the principles of non-interference, self-management and self-regulation…implementation of the.eu TLD may take into consideration best practices in this regard and could be supported by voluntary guidelines or codes of conduct where appropriate'*[273].

Co-regulation is therefore a good example of the pyramid of regulation, with a statutory tip of regulatory principles and authorisation for the regulator, a co-regulator layer that sets out regulatory design, and industry-shaped rules and codes to provide the detailed implementation.

Applied to AI and disinformation, this schematic would suggest a statute laying out the general principles applying to an AI regulator, while the regulator would set out details of the scheme design. Such principles may include, for instance, the principle that no account can be suspended

[270] See Hoffmann-Riem, W. (2001) *Modernisierung in Recht und Kultur,* Frankfurt: Suhrkamp; Huyse, L., and Parmentier, S. (1990) 'Decoding Codes: The Dialogue between Consumers and Suppliers through Codes of Conduct in the European Community', *Journal of Consumer Policy 13(3),* 253–272, at 260; Joerges, C., Meny, Y. and Weiler, J.H.H. (Eds., 2001) *Responses to the European Commission's White Paper on Governance,* European University Institute; Kleinstuber, H. (2004) 'The Internet between Regulation and Governance', in *Organisation for Security and Co-operation in Europe, The Media Freedom Internet Cookbook*, pp61-100; Latzer, M., Just, N., Saurwein, F., and Slominski, P. (2003) 'Regulation Remixed: Institutional Change through Self- and Co-Regulation in the Mediamatics Sector', *Communications and Strategies, 50(2),* 127-157.

[271] See Marsden, C. (2011) supra xx, at p. 61. By 2018, there were 12 million UK domains registered, see Nominet (2018), *UK Domains*, https://www.nominet.uk/uk-domains/

[272] Regulation (EC) No 874/2004 Laying Down Public Policy Rules concerning the Implementation and Functions of the .eu Top Level Domain and the Principles governing Registration

[273] Regulation (EC) No 733/2002 on the Implementation of the .eu Top Level Domain, at Recital 9.

without human intervention to correct for false positive identification of a bot account or egregious content, and the potential for account holder appeal against such a deletion. This would be a minimum requirement to maintain freedom of expression for social media users, to ensure accounts are not deleted without due process. A civil society stakeholder argues:

> *'Any measure to tackle the complex topic of online disinformation must not be blindly reliant on automated means, artificial intelligence or similar emerging technologies without ensuring that the design, development and deployment of such technologies are individual-centric and respect human rights'*[274].

This human-regulated AI is more likely to be guaranteed with robust co-regulation than self-regulatory schemes. The parallels with domain names are instructive, as accounts cannot be removed from owners without a formal process (even if the owner is deceased).

The cost-benefit of such co-regulation is held in general to allow for more efficient and flexible regulation. That theoretically can provide both lower costs and more responsive regulation, though in practical terms exceptions may exist. Feasibility and effectiveness depend on the initial statutory design as well as the implementation of that design by the co-regulator. There are many examples of successful internet co-regulation, though disinformation is a particularly rapidly moving target. Experience with another open internet issue, that of network neutrality, shows that such feasibility challenges can be overcome with appropriate multistakeholder engagement[275].

Sustainability of co-regulation is always an issue. While it is more robust than less interventionist regulatory designs, internet co-regulation is often chosen due to the particularly profound constitutional and human rights challenges, so that self-regulation is judged inadequate. Thus, a frequent failing of co-regulation is that it is eventually supplanted by state regulatory bodies, as for instance with video on demand under the Audiovisual Media Services Directive. Though the direction of travel from self-regulation to state regulation is not inevitable, it can be made due to pressure from both government and from regulates seeking regulatory certainty. In such situations, the costs of co-regulation can escalate as the scheme attempts to shadow state regulation. Risks and future uncertainties are thus closely tied to the regulatory commitment to making co-regulation an end state rather than an interim measure. As explained for Option 3, coherence with EU objectives are easier to assess with co-regulation than with self-regulation.

Potential ethical, social and regulatory impacts revolve around the media pluralism dilemma, that increasing pluralism and diversity with regulation risks regulatory capture and the danger that the regulated diversity does not satisfy the users' needs in a free society. The fundamental rights issues with co-regulation are similar to those for less direct regulatory interventions – freedom of expression as a fundamental right may be held inappropriate for anything but state regulation, a constant issue in internet regulation.

Option 5: Statutory regulation

In Option 5, a regulator would be tasked to combat disinformation directly by licensing of content providers and their systems for content moderation. Current electoral and broadcast regulators already perform this function for offline media. The UK Parliament states that '[i]n this rapidly changing digital world, our existing legal framework is no longer fit for purpose'[276] and has

[274] EDRi (19 October 2018) *Civil Society Calls for Evidence-Based Solutions to Disinformation*, https://edri.org/civil-society-calls-for-evidence-based-solutions-to-disinformation/, quoting Statement of Hidvégi, Fanny, European Policy Manager with Access Now.

[275] See Marsden C. (2017) Network Neutrality, supra xx

[276] UK House of Commons (2018) *Interim Report on Disinformation and 'Fake News'*, Select Committee on Media, Culture and Sport, https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/36302.htm

suggested this option. Hearings are ongoing on the role of the UK Information Commissioner and communications regulator Ofcom in such a scheme.[277]

Chris Marsden, co-author of this study, has previously suggested a new merged regulator should be formed.[278] Each national context will differ, but in general they will encompass:

- reformed/strengthened powers for the
    - electoral commission,
    - data protection authority,
    - advertising regulator, and
    - communications regulator (broadcast, newspaper);
- with police enforcement of criminal law regarding fraud (bot accounts) and other malicious (illegal) communications.

However, it is unclear what such a regulator could achieve without invoking direct censorship of non-conforming organisations. AI systems may be forced to conform to a mandatory national or regional standard, which could lead to dominant standards being enforced anti-competitively. While this was overcome in, for instance, the 3G standard for mobile telephony, there is no convincing example of content moderation subject to technical standards being successfully mandated. The UK government's example of mandatory age rating that it is introducing in 2018 is not a promising approach.[279]

It is also notable that such a merger of many regulators is not necessary to combine the functions via coordinated federated networks of those regulators. The UK Information Commissioner 'Democracy Disrupted' report makes this clear as the most effective and sustainable method in the short- to medium-term: 'The Government should conduct a review of the regulatory gaps in relation to the content, provenance and jurisdictional scope of political advertising online'. Best practice from the various Member States should be collated, analysed and disseminated, ideally by the European Parliament with assistance from the EU Fundamental Rights Agency.[280] The Digital Rights Clearinghouse set up by the EU Data Protection Supervisor with data protection, consumer protection and competition authorities is another example.

Given the speed and flexibility of response demanded by the political priority to combat disinformation, it may be that the reform of existing legislation is a more effective and sustainable form of regulation. For instance, electoral advertising rules can be brought within the ambit of the existing regulator without necessarily reforming primary legislation. The removal of bot accounts is ongoing, and appeal processes could be built into the removal of disinformation, ideally within Option 3. A raft of incremental improvements will be more compatible with the mission to control disinformation and the uses of AI therein, than a more disruptive change at this stage.

[277] UK Information Commissioner's Office (2018) *Democracy Disrupted? Personal Influence and Political Influence*, https://ico.org.uk/media/action-weve-taken/2259369/democracy-disrupted-110718.pdf, Recommendation 10 at p. 46

[278] Marsden (2018) 'Towards OffData', Georgetown Law Review, supra xx

[279] UK Department for Digital, Culture, Media and Sport (2018) *Explanatory Memorandum To The Online Pornography (Commercial Basis) Regulations 2018*, http://www.legislation.gov.uk/ukdsi/2018/9780111173183/pdfs/ukdsiem_9780111173183_en.pdf

For criticism, see Hill, R. (17 October 2018) 'UK.gov To Press Ahead with Online Smut Checks (but expects £10m in Legals in Year 1)', *The Register,* https://www.theregister.co.uk/2018/10/17/age_verification_legislation_bbfc/

[280] For FRA activities in this area, see European Union Agency for Fundamental Rights (2018), *Enabling Human Rights and Democratic Space in Europe*, http://fra.europa.eu/en/event/2018/enabling-human-rights-and-democratic-space-europe

4.3. Focus on freedom of expression and media pluralism

The impacts of policies in this area are universally high, and Option 1 remains the least favourable option throughout. The costs of uncertainty are much higher for the less regulatory options, and regulatory sustainability and protection of fundamental rights (including freedom of expression/media pluralism) is more strongly supported for the more regulatory Options 4/5.

Noting that the objective of free and fair European parliamentary elections in May 2019 is a high political priority, regulatory Option 5 is scored highly, specifically to ensure electoral online advertising is regulated online, as it currently is offline. However, that is **not** a proposal for any kind of super-regulator or 'OffData'. Overall, the authors believe legislation may be premature and potentially hazardous for freedom of expression: collaboration between different stakeholder groups with public scrutiny is preferable, where effectiveness can be independently demonstrated via audit.

Furthermore, noting that Option zero means a lack of protection of fundamental rights, including appeal against account suspension, as well as exposure to unregulated disinformation:

1. This study argues that options to ensure **independent appeal and audit** of platforms' regulation of their users be introduced as soon as feasible. When technical intermediaries need to moderate content and accounts, detailed and transparent policies, notice and appeal procedures, and regular reports are crucial. It is believed this is also valid for automated removals.
2. This study advises against regulatory action that would encourage increased use of AI for content moderation purposes, without **strong human review and appeal processes**.
3. There is scope for standardising (the basics of) notice and appeal procedures and reporting, and creating **a self-regulatory multistakeholder body**, such as the UN Special Rapporteur's suggested 'social media council'.[281] As recommended by the Special Rapporteur, this multistakeholder body could, on the one hand, have competence to deal with industry-wide appeals and, on the other hand, work towards a better understanding and minimisation of the effects of AI on freedom of expression and media pluralism. It is believed this would best fit Option 3 classification.
4. This study emphasises that disinformation is best tackled through **media pluralism and literacy initiatives**, as these allow diversity of expression and choice. **Source transparency indicators** are preferable over (de)prioritisation of disinformation, and users need to be given the opportunity to understand how their search results or social media feeds are built, and edit their search results/feeds where desirable.
5. Finally, noting the lack of independent evidence or even **detailed research** in this policy area, the risk of harm remains far too high for any degree of regulatory certainty. The authors reiterate that **far greater transparency** must be introduced into the variety of AI and disinformation reduction techniques used by online platforms and content providers.

[281] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) Report on A Human Rights Approach to Platform Content Regulation, supra xx, pars 58, 59, 63, 72

5. References

A.T. Kearney (2010) *Internet Value Chain Economics. Gaining a Deeper Understanding of the Internet Economy,* https://www.atkearney.com/documents/20152/434237/internet-value-chain-economics.pdf/285d1a4d-a49c-43d1-5966-9fcca69aa55a

AALEP (19 April 2011) *Biblical Accounts of Lobbying,* http://www.aalep.eu/biblical-accounts-lobbying

Access Now, Civil Liberties Union for Europe, and EDRI (2018) *Informing the 'Disinformation' Debate*, https://edri.org/files/online_disinformation.pdf

ACLU Foundation of Northern California, Center of Democracy and Technology, Electronic Frontier Foundation, New America's Open Technology Institute et.al. (2018) *Santa Clara Principles on Transparency and Accountability in Content Moderation,* https://santaclaraprinciples.org/

Agence France Press (2018) *Fact Check,* https://factcheck.afp.com/fact-checking-afp

Akdeniz, Y. (2011) '*Freedom of Expression on the Internet: Study of Legal Provisions and Practices related to Freedom of Expression, the Free Flow of Information and Media Pluralism on the Internet in OSCE Participating States,* Vienna: Office of the Representative on Freedom of the Media, Organisation for Security and Co-operation in Europe, http://www.osce.org/fom/80723

Alexander J., and Smith, J. (2011) 'Disinformation: A Taxonomy', *IEEE Security & Privacy 9(1),* 58-63, doi: 10.1109/MSP.2010.141

Alliance of Democracies (18 May 2018) *Transatlantic, Bi-Partisan Commission Launched to Prevent Election Meddling*, Press Release, http://www.allianceofdemocracies.org/initiatives/the-campaign/press_release_tcei/

American Civil Liberties Union v Reno (1997) 21 US 844 of 27 June

Angelopoulos, C., and Quintais, J.P. (30 August 2018), 'Fixing Copyright Reforms: How to Address Online Infringement and Bridge the Value Gap', *Kluwer Copyright Blog,* http://copyrightblog.kluweriplaw.com/2018/08/30/fixing-copyright-reform-address-online-infringement-bridge-value-gap/

Article 19 (14 June 2018) *Google: New Guiding Principles on AI Show Progress But Still Fall Short on Human Rights Protections*, https://www.article19.org/resources/google-new-guiding-principles-on-ai-show-progress-but-still-fall-short-on-human-rights-protections/

Atlantic Council Eurasia Center (2018) *Disinfoportal.org,* https://disinfoportal.org/about-disinfo-portal/

Ayala, D. (31 July 2018) *Introducing the Dweb*, Mozilla Blog, https://hacks.mozilla.org/2018/07/introducing-the-d-web/

Azevedo, L. (2018) 'Truth or Lie: Automatically Fact Checking News', in *Companion Proceedings of The Web Conference 2018 (WWW '18),* International World Wide Web Conferences Steering Committee, Geneva, Switzerland, pp. 807-811, DOI: https://doi.org/10.1145/3184558.3186567

Babu, A., Lui, A., and Zhang, J. (17 May 2017) 'New Updates to Reduce Clickbait Headlines', *Facebook Newsroom,* https://newsroom.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/

Badii, F. (21 August 2018) *Is It Time to Institutionalize Cyber Attribution,* Internet Governance Project, https://www.internetgovernance.org/2018/08/21/new-igp-white-paper-is-it-time-to-institutionalize-cyber-attribution/

Barnes, R., Cooper, A., Kolkman, O. Thaler, D., and Nordmark, E. (2016) *RFC 7754 – Technical Considerations for Internet Service Blocking and Filtering, March 2016*, https://tools.ietf.org/html/rfc7754#page-27

BBC.com (23 July 2018) *WhatsApp 'Admin' Spends Five Months in an Indian Jail*, https://www.bbc.com/news/technology-44925166

Benet, J. (2014) *IPFS - Content Addressed, Versioned, P2P File System* (*DRAFT* 3), https://ipfs.io/ipfs/QmR7GSQM93Cx5eAg6a6yRzNde1FQv7uL6X1o4k7zrJa3LX/ipfs.draft3.pdf

Benkler, Y. et al (2017) *Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election,* Harvard Berkman-Klein Center, https://cyber.harvard.edu/publications/2017/08/mediacloud

Benkler, Yochai, Robert Faris, and Hal Roberts (2018) *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*, Oxford: Oxford University Press

Bentzen, N. (2015) *Understanding Propaganda and Disinformation,* European Parliament Research Service At a Glance, http://www.europarl.europa.eu/RegData/etudes/ATAG/2015/571332/EPRS_ATA(2015)571332_EN.pdf

Bhaskaran, H., Harsh, M., and Pradeep, N. (2017) 'Contextualizing Fake News in Post-Truth Era: Journalism Education in India', *Asia Pacific Media Educator 27(1)* 41–50

Bickert, M. (24 April 2018) 'Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process', *Facebook Newsroom*, https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/

Bishop, C. (1995) *Neural Networks for Pattern Recognition*, Gloucestershire: Clarendon Press

Brown, I. (2013) *Online Freedom of Expression, Association, Assembly and the Media in Europe*, Council of Europe MCM(2013)007, Strasbourg: Council of Europe

Brown, I. (2013) *Transparency to Protect Internet Freedom: a Shared Commitment*, Strasbourg: Council of Europe

Brown, I. and Korff, D. (2012) *Digital Freedoms in International Law*, Global Network Initiative

Burshtein, S. (2017) 'The True Story on Fake News', *Intellectual Property Journal 29(3)*

C-288/89 (judgment of 25 July 1991, Stichting Collectieve Antennevoorziening Gouda and others [1991] ECR I-4007)

C(2018) 1177 final EC Recommendation on Measures to Effectively Tackle Illegal Online Content https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online

C(2018) 5949 final EC Recommendation on Election Cooperation Networks, Online Transparency, Protection against Cybersecurity Incidents and Fighting Disinformation Campaigns in the Context of Elections to the European Parliament

Chander, A. and Vivek, K. (2018) 'The Myth Of Platform Neutrality', *Georgetown Law Technology Review 2(2)* 400-416

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. (2009) 'Reading Tea Leaves: How Humans Interpret Topic Models', in Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.) *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, pp. 288–96

Chatzakou, D. et al. (2017), 'Mean Birds: Detecting Aggression and Bullying on Twitter', *WebSci*

Christie, E.H. (2018) 'Political Subversion in the Age of Social Media', *CES Policy Brief, Octobe*r; Access Now, Civil Liberties Union For Europe

Clayton, R. (2005) *Anonymity and Traceability in Cyberspace*, Cambridge Computer Lab Technical Report 653, http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-653.pdf

Code of Practice Agora (2014) *The Principles for Better Self- and Co-Regulation*, https://ec.europa.eu/digital-single-market/en/best-practice-principles-better-self-and-co-regulation#Article

Cohen, M. (2017) 'Fake News and Manipulated Data, the New GDPR, and the Future of Information'*, Business Information Review 34(2)* 81-85

COM (2015) 192 final EC Communication on A Digital Single Market Strategy for Europe

COM (2016) 288 EC Communication on Online Platforms and the Digital Single Market: Opportunities and Challenges for Europe

COM(2016) 593 final – 2016/0280(COD) Proposed EU Directive on Copyright in the Digital Single Market https://ec.europa.eu/digital-single-market/en/news/proposal-directive-european-parliament-and-council-copyright-digital-single-market

COM(2017) 555 final EC Communication on Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms

COM(2017)0481 final - 2017/0219 (COD) Proposed EU Revised Regulation on the Statute and Funding of European Political Parties and European Political Foundations

COM(2017)10 final – 2017/0003(COD) Proposed EU Regulation concerning the Respect for Private Life and the Protection of Personal Data in Electronic Communications and Repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-privacy-and-electronic-communications

COM(2018) 638 final EC Guidance on the Application of Union Data Protection Law in the Electoral Context

COM(2018) 640 final - 2018/0331 (COD) Proposed EU Regulation on Prevention of Dissemination of Terrorist Content Online https://ec.europa.eu/commission/sites/beta-political/files/soteu2018-preventing-terrorist-content-online-regulation-640_en.pdf

Conroy, N, Rubin, V., and Chen, Y. (2015) 'Automatic Deception Detection: Methods for Finding Fake News', in *Proceedings of the Association for Information Science and Technology 52(1),* pp. 1–4

Council Directive 89/552/EEC on the Coordination of Certain Provisions Laid Down by Law, Regulation or Administrative Action in Member States concerning the Pursuit of Television Broadcasting Activities

Cubby v CompuServe (1991) 766 F Supp 135

Davis, E. (2017) *Why We Have Reached Peak Bullshit and What We Can Do About It,* Little: Brown

de Cock Buning, M. (10 Sept 2018) 'We Must Empower Citizens In The Battle Of Disinformation', *International Institute for Communications,* http://www.iicom.org/themes/governance/item/we-must-empower-citizens-in-the-battle-of-disinformation

DFRLab (2018) 'Fake News: Defining and Defeating Real Techniques for Identifying Fake News and Disinformation', *Medium,* https://medium.com/dfrlab/fake-news-defining-and-defeating-43830a2ab0af?_branch_match_id=553166123622124243

Directive 2002/58/EC concerning the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector (Directive on Privacy and Electronic Communications) https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:HTML

Directive 97/7/EC on the Protection of Consumers in Respect of Distance Contracts, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A31997L0007

Doering, D. and Neff, G. (2018) 'Fake News as a Combative Frame: Results from a Qualitative Content Analysis of the Term's Definitions and Uses on Twitter', *4th Annual International Journal of Press/Politics Conference*, Oxford, 12 October

Dunbar, R. I. M. (2016) 'Do Online Social Media Cut Through the Constraints that Limit the Size of Offline Social Networks?', *Royal Society Open Science 2016(3)*, DOI: 10.1098/rsos.150292

EDRi (19 October 2018) *Civil Society Calls for Evidence-Based Solutions to Disinformation*, https://edri.org/civil-society-calls-for-evidence-based-solutions-to-disinformation/

Edwards, L. and Veale, M. (2017) *Slave to the Algorithm? Why a 'Right to Explanation' is Probably Not the Remedy You are Looking for*, https://ssrn.com/abstract=2972855

EEAS (2017) *Questions and Answers about the East StratCom Task Force* https://eeas.europa.eu/headquarters/headquarters-homepage/2116/-questions-and-answers-about-the-east-stratcom-task-force_en

EEAS (2017) *Questions and Answers about the East StratCom Task Force* https://eeas.europa.eu/headquarters/headquarters-homepage/2116/-questions-and-answers-about-the-east-stratcom-task-force_en

Enriques, L. (9 Oct 2017) 'Financial Supervisors and RegTech: Four Roles and Four Challenges', *Oxford University, Business Law Blog*, http://disq.us/t/2ucbsud

EP Resolution on Distributed Ledger Technologies and Blockchains: Building Trust with Disintermediation (P8_TA-PROV(2018)0373 B8-0397/2018), http://www.europarl.europa.eu/sides/getDoc.do?type=TA&reference=P8-TA-2018-0373&language=EN&ring=B8-2018-0397

Epstein, R. & Robertson, R.E. (2015) 'The Search Engine Manipulation Effect (SEME) and its Possible Impact on the Outcomes of Elections', *112 Proc Nat'l Acad. Sci.* E4512

Erdos, D. (2016) 'European Data Protection Regulation and Online New Media: Mind the Enforcement Gap', *Journal of Law and Society 43(4)* 534-564, http://dx.doi.org/10.1111/jols.12002

EU Code of Conduct on Countering Illegal Hate Speech Online (2016), http://europa.eu/rapid/press-release_IP-15-6243_en.htm

EU Code of Practice on Disinformation (2018) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

EU Code of Practice on Disinformation. Annex II Current Best Practices from Signatories of the Code of Practice (2018) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

EU Directive 2000/31/EC on Certain Legal Aspects of Information Society Services, in particular Electronic Commerce, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031

EU Directive 2006/114/EC concerning Misleading and Comparative Advertising

EU Directive 2010/13/EU on Audiovisual Media Services (the Coordination of Certain Provisions Laid Down by Law, Regulation or Administrative Action in Member States concerning the Provision of Audiovisual Media Services)

EU DisinfoLab (2018) *About Us*, http://disinfo.eu/aboutus/

EU Human Rights Guidelines on Freedom of Expression Online and Offline (2014) https://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/EN/foraff/142549.pdf

EU Regulation 2016/679 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data, and repealing Directive 95/46/EC (General Data Protection Regulation) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

European Commission (12 Jan 2018) *Experts Appointed to the High-Level Group on Fake News and Online Disinformation*, Press Release, https://ec.europa.eu/digital-single-market/en/news/experts-appointed-high-level-group-fake-news-and-online-disinformation

European Commission (12 September 2018) *State of the Union 2018: Commission Proposes New Rules to Get Terrorist Content Off the Web,* Press Release (IP/18/5561), http://europa.eu/rapid/press-release_IP-18-5561_en.htm

European Commission (16 October 2018), *Roadmaps to Implement the Code of Practice on Disinformation*, Press Release, https://ec.europa.eu/digital-single-market/en/news/roadmaps-implement-code-practice-disinformation

European Commission (18 Sept 2018) *The European Union Strengthens its Support to Media Freedom and Young Journalists in the Western Balkans*, Press Release (IP/18/5789), http://europa.eu/rapid/press-release_IP-18-5789_en.htm

European Commission (2015) *Public Consultation on the Regulatory Environment for Platforms, Online Intermediaries, Data and Cloud Computing and the Collaborative Economy* https://ec.europa.eu/digital-single-market/news/public-consultation-regulatory-environment-platforms-online-intermediaries-data-and-cloud

European Commission (2016) *Full Report on the Results of the Public Consultation on the Regulatory Environment for Platforms, Online Intermediaries and the Collaborative Economy,* https://ec.europa.eu/digital-single-market/en/news/full-report-results-public-consultation-regulatory-environment-platforms-online-intermediaries

European Commission (2018) *Countering Illegal Hate Speech Online #NoPlace4Hate,* Press Release, http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300

European Commission (2018) *Expert Group to the EU Observatory on the Online Platform Economy, Team responsible E-Commerce and Platforms (Unit F.2),* https://ec.europa.eu/digital-single-market/en/expert-group-eu-observatory-online-platform-economy

European Commission (2018) *Media Pluralism Monitor,* https://ec.europa.eu/digital-single-market/en/media-pluralism-monitor-mpm

European Commission (2018) *State of the Union 2018: European Commission Proposes Measures for Securing Free and Fair European Elections,* Press Release (IP/18/5681), http://europa.eu/rapid/press-release_IP-18-5681_en.htm

European Commission (26 September 2018) *Code of Practice on Disinformation,* Press Release, https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

European Commission (27 June 2017) Antitrust: Commission Fines Google €2.42 Billion for Abusing Dominance as Search Engine by Giving Illegal Advantage to Own Comparison Shopping Service, Factsheet, http://europa.eu/rapid/press-release_MEMO-17-1785_en.htm

European Commission (3 December 2015) *EU Internet Forum: Bringing Together Governments, Europol and Technology Companies to Counter Terrorist Content and Hate Speech Online,* Press Release (IP/15/6243) http://europa.eu/rapid/press-release_IP-15-6243_en.htm

European Commission (6 Dec 2016) *Fighting Illegal Online Hate Speech: First Assessment of the New Code of Conduct*, Press Release, http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50840

European Commission (6 Feb 2018) *Launch of the #SaferInternet4EU Initiatives on Safer Internet Day,* Press Release, https://ec.europa.eu/digital-single-market/en/news/launch-saferinternet4eu-initiatives-safer-internet-day

European Parliament (2001) *Final Report on the Existence of a Global System for the Interception of Private and Commercial Communications (ECHELON interception system)*, Temporary Committee on the ECHELON

European Parliament Resolution on Media Pluralism and Media Freedom in the European Union (P8_TA(2018)0204) http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2018-0204+0+DOC+PDF+V0//EN

European Union Agency for Fundamental Rights (2018), *Enabling Human Rights and Democratic Space in Europe*, http://fra.europa.eu/en/event/2018/enabling-human-rights-and-democratic-space-europe

Europol (22 July 2016) *Europol Internet Referral Unit One Year On,* Press Release, https://www.europol.europa.eu/newsroom/news/europol-internet-referral-unit-one-year

Facebook (2018) *Community Standards,* https://www.facebook.com/communitystandards/introduction

Facebook (2018) *Facebook Journalism Project,* https://www.facebook.com/facebookmedia/solutions/facebook-journalism-project

Facebook (2018*) German NetzDG Transparency Report (Jan-Jun 2018),* https://fbnewsroomus.files.wordpress.com/2018/07/facebook_netzdg_july_2018_english-1.pdf

Facebook (2018) *Written Evidence for Lords Communications Committee - The Internet: To Regulate or Not To Regulate?* http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/communications-committee/the-internet-to-regulate-or-not-to-regulate/written/83287.html

Falck, B. (28 June 2018) 'Providing More Transparency Around Advertising on Twitter', *Twitter Blog,* https://blog.twitter.com/official/en_us/topics/company/2018/Providing-More-Transparency-Around-Advertising-on-Twitter.html

Fletcher, R., and Nielsen, R.K. (2018) 'Are People Incidentally Exposed to News on Social Media? A Comparative Analysis', *New Media & Society 20(7)* 2450–2468, https://doi.org/10.1177/1461444817724170

Founta, A-M. et al. (2018) 'A Unified Deep Learning Architecture for Abuse Detection', *ArXiv*; Founta, A-M. et al. (2018) 'Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior', *International AAAI Conference on Web and Social Media (ICWSM)*

Frosio, G. (2017) 'From Horizontal to Vertical: An Intermediary Liability Earthquake in Europe, *Journal of Intellectual Property Law and Practice 12(7)* 565-575

Full Fact (2018) *The UK's Independent Factchecking Charity*, https://fullfact.org

Funke, D. (20 July 2018) *WhatsApp Is Limiting Message Forwarding to Cut Down on Fake News*, Poytner Institute, https://www.poynter.org/news/whatsapp-limiting-message-forwarding-cut-down-fake-news

Funke, D.(3 October 2017) *Three Months After Launching, Faktisk is Already Among the Most Popular Sites in Norway*, Poytner Institute, https://www.poynter.org/news/three-months-after-launching-faktisk-already-among-most-popular-sites-norway

Gadde, V.and Gasca, D. (2018) *Measuring Healthy Conversation*, Twitter Blog, https://blog.twitter.com/official/en_us/topics/company/2018/measuring_healthy_conversation.html

Geiger, C. & Izyumenko, E. (2016) 'The Role of Human Rights in Copyright Enforcement Online: Elaborating a Legal Framework for Website Blocking'*, American University International Law Review 32(1),* 43

Gibbons, T. (2000) 'Pluralism, Guidance and the New Media', in C. Marsden (Ed.) *Regulating the Global Information Society*, Abingdon: Routledge, pp. 304-315

Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., and Crowcroft. J. (2017) 'Of Bots and Humans (on Twitter)', in *ASONAM '17 Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* pp. 349-354

Glaser, A. (23 Oct 2018) 'Facebook Will Never Run Out of Moles to Whack: The Latest Disclosure of Russian Election Meddling Reveals the Limits of Social Media's New Dedication to Fighting False News', *Slate,* https://slate.com/technology/2018/10/project-lakhta-facebook-russia-election-meddling-midterms.html

Global Internet Forum to Counter Terrorism (2018) *Vision,* https://gifct.org

Global Network Initiative (2018) *Annual Report 2017: Reinforcing a Global Standard,* https://globalnetworkinitiative.org/global-network-initiative-annual-report-2017-reinforcing-a-global-standard/

Godfrey v Demon Internet Service [2001] QB 201

Google (2015*) Google Advisory Council on the Right to be Forgotten*, https://archive.google.com/advisorycouncil/

Google (2018) *Be Internet Legends*, https://beinternetlegends.withgoogle.com/en-gb

Google (2018) *German NetzDG Transparency Report (Jan-Jun 2018),* https://transparencyreport.google.com/netzdg/youtube

Google (2018) *How Content ID Works* https://support.google.com/youtube/answer/2797370?hl=en

Google (2018) *YouTube Transparency Report*, https://transparencyreport.google.com/youtube-policy/overview

Google News Initiative (2018) *Digital News Innovation Fund*, https://newsinitiative.withgoogle.com/dnifund/report/european-innovation-supporting-quality-journalism/

Google UK (2018) *Written Evidence for Lords Communications Committee - The Internet: To Regulate or Not To Regulate?* http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/communications-committee/the-internet-to-regulate-or-not-to-regulate/written/83086.pdf

Harambam, J. & Helberger, N., and van Hoboken, J. (2018) 'Democratizing Algorithmic News Recommenders: How to Materialize Voice in a Technologically Saturated Media Ecosystem,' *Philosophical Transactions of The Royal Society A: Mathematical Physical and Engineering Sciences 376(2133)*, DOI 10.1098/rsta.2018.0088

Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy (2018) *First Draft*, https://firstdraftnews.org/about/

Hautala, L. (16 Oct 2018) 'Hackers, Trolls and the Fight over Your Vote in the 2018 Midterm Elections', *CNET*, https://www.cnet.com/news/hackers-trolls-and-the-fight-over-your-vote-in-the-2018-midterm-elections/

High Level Expert Group (HLEG) on Fake News and Online Disinformation (2018) *A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation,* https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

Hildebrandt, M. (2018) 'Primitives of Legal Protection in the Era of Data-Driven Platforms', *Georgetown Law Technology Review 2(2)*

Hill, R. (17 October 2018) 'UK.gov To Press Ahead with Online Smut Checks (but expects £10m in Legals in Year 1)', *The Register,* https://www.theregister.co.uk/2018/10/17/age_verification_legislation_bbfc/

Hill, R. (25 April 2018) 'Academics: Shutting down Facebook API Damages Research, Oversight, Competition. Open Letter Throws Heavy Shade on Social Network's Research Initiative', *The Register*, https://www.theregister.co.uk/2018/04/25/shutting_down_facebook_api_damages_research_oversight_competition_warn_academics/

Hillard, D., Purpura, S., and Wilkerson, J. (2008) 'Computer-Assisted Topic Classification for Mixed-Methods Social Science Research', *Journal of Information Technology & Politics 4(4)* 31–46

Hine, G. et al. (2017) 'Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and its Effects on the Web', *ICWSM*

Hoffmann-Riem, W. (2001) *Modernisierung in Recht und Kultur,* Frankfurt: Suhrkamp

Hurst, A. (2017) 'Fake News: Striking a Balance between Regulation and Responsibility,' *Society of Computers & Law, August/September* https://www.scl.org/articles/8961-fake-news-striking-a-balance-between-regulation-and-responsibility

Husovec, M. (2017) *Injunctions Against Intermediaries in the European Union. Accountable But Not Liable?,* Cambridge, UK: Cambridge University Press

Huyse, L., and Parmentier, S. (1990) 'Decoding Codes: The Dialogue between Consumers and Suppliers through Codes of Conduct in the European Community', *Journal of Consumer Policy 13(3),* 253–272

Ibosiola, D. et al. (2018) 'Movie Pirates of the Caribbean: Exploring Illegal Streaming Cyberlockers', *ICWSM*

IEEE (2018) *Global Initiative on Ethics of Autonomous and Intelligent Systems,* https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

International Age Rating Coalition (2018) *How IARD Works,* http://www.globalratings.com/how-iarc-works.aspx

Jiménez Cruz, C., Mantzarlis, A., Nielsen, R.K., and Wardle, C. (12 March 2018), 'Six Points from the EU Commission's New Report on Disinformation, *Medium,* https://medium.com/@hlegresponse/six-key-points-from-the-eu-commissions-new-report-on-disinformation-1a4ccc98cb1c

Joerges, C., Meny, Y. and Weiler, J.H.H. (Eds., 2001) *Responses to the European Commission's White Paper on Governance,* European University Institute

Kijkwijzer (2018) *Netherlands Institute for the Classification of Audio-visual Media,* http://www.kijkwijzer.nl/nicam and Pan European Game Information (2018) *How We Rate Games*, https://pegi.info/page/how-we-rate-games

Kleinstuber, H. (2004) 'The Internet between Regulation and Governance', in *Organisation for Security and Co-operation in Europe, The Media Freedom Internet Cookbook*, pp61-100

Klinger, J., Mateos-Garcia, J.C., and Stathoulopoulos, K. (2018) *Deep Learning, Deep Change? Mapping the Development of the Artificial Intelligence General Purpose Technology,* DOI: http://dx.doi.org/10.2139/ssrn.3233463

Klonick, K. (2018) 'Why The History Of Content Moderation Matters', *Content Moderation at Scale 2018 Essays: Techdirt,* https://www.techdirt.com/articles/20180129/21074939116/why-history-content-moderation-matters.shtml

Knight Foundation (2018) *Misinformation in Graphics*, https://www.knightfoundation.org/features/misinfo/

Koebler, J., and Cox, J. (23 Aug 2018) 'The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People', *Motherboard*, https://motherboard.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works

Korff, D. with Brown, I. (2013) *The Use of the Internet & Related Services, Private Life & Data Protection: Trends & Technologies, Threats & Implications,* Council of Europe T-PD(2013)07, Strasbourg: Council of Europe.

Kotaro, H. et.al (2017) 'A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk', in *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI'18)*, Paper no. 449

Kozlowska, H. (3 April 2018) 'Mark Zuckerberg Floated a 'Supreme Court' for Facebook. What Does That Mean?', *Quartz*, https://qz.com/1243203/mark-zuckerberg-floated-a-supreme-court-for-facebook-what-does-that-mean/

Lamb, K. (23 July 2018) 'I Felt Disgusted: Inside Indonesia's Fake Twitter Account Factories', *The Guardian,* https://www.theguardian.com/world/2018/jul/23/indonesias-fake-twitter-account-factories-jakarta-politic

Lamo, M. and Calo, R. (2018) 'Regulating Bot Speech', *UCLA Law Review 2019*, http://dx.doi.org/10.2139/ssrn.3214572

Latzer, M., Just, N., Saurwein, F., and Slominski, P. (2003) 'Regulation Remixed: Institutional Change through Self- and Co-Regulation in the Mediamatics Sector', *Communications and Strategies, 50(2),* 127-157

Lemley, M. (2006) 'Terms of Use', *Minnesota Law Review 91(2)* 459-483

Mac Síthigh, D. (2008) 'The Mass Age of Internet Law', *Information & Communications Technology Law 17(2),* 79-94

Mandrescu, D. (2017) 'Applying EU Competition Law to Online Platforms: The Road Ahead – Part I', *Competition Law Review 38(8)* 353-365

Mandrescu, D. (2017) 'Applying EU Competition Law to Online Platforms: The Road Ahead – Part II', *competition Law Review 38(9)* 410-422

Mariconti, E. et al. (2018) ''You Know What to Do': Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks', *ArXiv*

Marsden (2018) 'The Log View Towards Creating OffData'*, Georgetown Law Review 2(2),* 376-398

Marsden, C. (1999) *Pluralism in Multi-Channel Digital Communications: Suggestions for Regulatory Scrutiny,* MM-S-PL 1999-12, Study prepared on behalf of the Committee of Specialists on Media Pluralism, Directorate of Human Rights Strasbourg: Council of Europe

Marsden, C. (2000) 'Not So Special? Merging Media Pluralism with Competition and Industrial Policy', *Info 2(1)* 9-15

Marsden, C. (2011) *Internet Co-Regulation*, Cambridge: Cambridge University Press

Marsden, C. (2012) 'Internet Co-Regulation and Constitutionalism: Towards European Judicial Review' *International Review of Law, Computers & Technology 26(2-3)* 215-216

Marsden, C. (2017) *Network Neutrality: From Policy to Law to Regulation,* Manchester: Manchester University Press

Marsden, C. (2018) 'Regulating Intermediary Liability and Network Neutrality', in I. Walden (Ed.) *Telecommunications Law and Regulation*, 5th Edition, Oxford: Oxford University Press, pp. 733-788

Marwick, A. & Lewis, R. (2017) 'Media Manipulation and Disinformation Online', *Data & Society,* https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformation

Marwick, A.E. (2018) 'Why Do People Share Fake News? A Sociotechnical Model of Media Effects', *Georgetown Technology Law Review 2(2)*, https://www.georgetownlawtechreview.org/issues/volume-2-issue-2/

Matsa, K.E. (8 June 2018) 'Across Western Europe, Public News Media are Widely Used and Trusted Sources of News', Pew Research Centre, http://www.pewresearch.org/fact-tank/2018/06/08/western-europe-public-news-media-widely-used-and-trusted/

McStay, Andrew (2011) *The Mood of Information: A Critique of Online Behavioural Advertising*, London: A&C Black.

Merton, R.K. (1948) 'The Self-Fulfilling Prophecy', *The Antioch Review 8(2),* 193-210

Meyer, T. (2012) Graduated Response In France: The Clash of Copyright and the Internet, *Journal of Information Policy 2,* 107-127, DOI: 10.5325/jinfopoli.2.2012.0107

Meyer, T. (2017) *The Politics of Online Copyright Enforcement in the EU: Access and Control,* Cham: Palgrave Macmillan

Michael, K. (2017) 'Bots Trending Now: Disinformation and Calculated Manipulation of the Masses [Editorial]', *IEEE Technology and Society Magazine 36(2),* 6-11, doi: 10.1109/MTS.2017.2697067

Millwood-Hargrave, M. (2007) *Report for Working Group 3 of the Conference of Experts for European Media Policy, More Trust in Content – The Potential of Co- and Self-Regulation in Digital Media*, Leipzig: 9-11 May

MIT-Harvard Ethics and Harvard Berkman-Klein Center for Internet and Society (2018) *The Ethics and Governance of Artificial Intelligence Initiative,* https://aiethicsinitiative.org/

Mitchell, T. (1997) *Machine Learning,* New York: McGraw-Hill Education, pp. 7–9

Mohan, N., and Kyncl, R. (9 July 2018) 'Building A Better News Experience on YouTube, Together', *YouTube Official Blog,* https://youtube.googleblog.com/2018/07/building-better-news-experience-on.html

Monroe, B., Colaresi, M., and Quinn, K. (2008) 'Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict', *Political Analysis 16(4)* 372-403

Muravchik, J. (1998) *The Future Of The United Nations: Understanding The Past To Chart A Way Forward*, American Enterprise Institute for Public Policy Research, Washington, D.C., https://epdf.tips/the-future-of-the-united-nations-understanding-the-past-to-chart-a-way-forward.html

Narayanan, V. et al. (2018) *'Polarization, Partisanship and Junk News Consumption over Social Media in the US', COMPROP Data Memo 2018(1),* Computational Propaganda Project, http://comprop.oii.ox.ac.uk/wpcontent/uploads/sites/93/2018/02/Polarization-Partisanship-JunkNews.pdf

Nielsen, R.K. (24 Oct 2018) 'Misinformation: Public Perceptions and Practical Responses', *Misinfocon London*, hosted by the Mozilla Foundation and Hacks/Hackers, https://www.slideshare.net/RasmusKleisNielsen/misinformation-public-perceptions-and-practical-responses/1

Nielsen, R.K. (26 Sept 2018) *Disinformation Twitter Thread*, https://twitter.com/rasmus_kleis/status/1045027450567217153

Nielsen, R.K. and Ganter, S. (2017) 'Dealing with Digital Intermediaries: A Case Study of the Relations Between Publishers and Platforms', *New Media & Society 20(4),* 1600-1617, doi: 10.1177/1461444817701318

Nilizadeh, S. et al. (2017) 'POISED: Spotting Twitter Spam Off the Beaten Paths', *CCS*

Noam, E. (2001) *Will the Internet Be Bad for Democracy?,* Columbia Institute for Tele Information, New York, http://www.citi.columbia.edu/elinoam/articles/int_bad_dem.htm

Nominet (2018), *UK Domains*, https://www.nominet.uk/uk-domains/

O'Brien, C., and Kelly, F. (9 May 2018) Google Bans Online Ads on Abortion Referendum, *The Irish Times,* https://www.irishtimes.com/business/media-and-marketing/google-bans-online-ads-on-abortion-referendum-1.3489046

O'Leary, S. (2018) 'Balancing Rights in a Digital Age,' *Irish Jurist 59,* 59

Osnos, E., Remnick, D., and Yaffa, J. (6 March 2017) 'Trump, Putin, and the New Cold War: What Lay Behind Russia's Interference in the 2016 Election—And What Lies Ahead?', *The New Yorker,* https://www.newyorker.com/magazine/2017/03/06/trump-putin-and-the-new-cold-war

Pariser E, (2011) *The Filter Bubble: What the Internet is Hiding From You*, London: Penguin Press

PEGI (2018) *PEGI Age Ratings*, https://pegi.info/page/pegi-age-ratings

Perez-Rosas, V., Kleinberg, B. Lefevre, A. and Mihalcea, R. (2018) *Automatic Detection of Fake News,* http://web.eecs.umich.edu/~mihalcea/papers/perezrosas.coling18.pdf

Perez, B., Musolesi, M., and Stringhini, G. (2018) 'You are Your Metadata: Identification and Obfuscation of Social Media Users using Metadata Information', *ICWSM*

Periñán, B., 'The Origin of Privacy as a Legal Value: a Reflection on Roman and English Law', *American Journal of Legal History 52(1)*

Pew Research Centre (27 Sept 2018) *Europe News Platforms, Topline Questionnaire,* Press Release, http://www.pewresearch.org/wp-content/uploads/2018/09/FT_18.09.27_EuropeNewsPlatforms_Topline.pdf

Phartiyal, S. (28 August 2018) *India's Top Court Seeks WhatsApp's Response on Petition Alleging It Breaches Law*, Reuters.com, https://www.reuters.com/article/us-whatsapp-india/indias-top-court-seeks-whatsapps-response-on-petition-alleging-it-breaches-law-idUSKCN1LD0SL

Playboy Enterprises, Inc. v. Frena, 839 F. Supp. 1552 (M.D. Fla. 1993)

Poytner Institute (2018) *Fact-Checking*, https://www.poynter.org/channels/fact-checking

Quercia, D., Lambiotte, R., Stillwell, D. Kosinski, M., and Crowcroft, J. (2012) 'The Personality of Popular Facebook Users', in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*, pp. 955-964, https://doi.org/10.1145/2145204.2145346

Reddit poster (2017) Artificial Intelligence Can't Tell Fried Chicken from Labradoodles, https://www.reddit.com/r/funny/comments/6h47qr/artificial_intelligence_cant_tell_fried_chicken/

Reeve, B.(2011) 'The Regulatory Pyramid Meets the Food Pyramid: Can Regulatory Theory Improve Controls on Television Food Advertising to Australian Children?', *Journal of Law and Medicine 19(1),* 128-46.

Regulation (EC) No 733/2002 on the Implementation of the .eu Top Level Domain

Regulation (EC) No 874/2004 Laying Down Public Policy Rules concerning the Implementation and Functions of the .eu Top Level Domain and the Principles governing Registration

Reporters Without Borders (3 April 2018) *RSF and its Partners Unveil the Journalism Trust Initiative to Combat Disinformation*, https://rsf.org/en/news/rsf-and-its-partners-unveil-journalism-trust-initiative-combat-disinformation

Richardson, M. and Hadfield, G. (1999) *The Second Wave of Law and Economics*, Sydney: Federation Press; Lessig, L. (1998) 'The New Chicago School', *the Journal of Legal Studies* 27(2) 661-691

Rieder, B., Matamoros-Fernández, A., and Coromina, Ò. (2018) 'From Ranking Algorithms to 'Ranking Cultures': Investigating the Modulation of Visibility in YouTube Search Results', *Convergence 24(1)* 50–68, https://doi.org/10.1177/1354856517736982

Rosati, E. (2019) *Copyright and the Court of Justice of the European Union*, Oxford: Oxford University Press

Rubin, Vi., Chen, Y., and Conroy, N.(2015) 'Deception Detection for News: Three Types of Fake News', in *Proceedings of the Association for Information Science and Technology*, St. Louis, MO: ASIST, pp.1–4

Ruggie, J. (2018) 'Multinationals as Global Institution: Power, Authority and Relative Autonomy, *Regulation & Governance (2018)12*, 317–333, https://onlinelibrary.wiley.com/doi/pdf/10.1111/rego.12154

Russell, N.W. (2016) *The Digital Difference: Media Technology and the Theory of Communication Effects,* Cambridge, MA: Harvard University Press

Samuelson, P. (1999) 'A New Kind of Privacy? Regulating Uses of Personal Data in the Global Information Economy', *California Law Review 87,* 751-778

Sanchez, L. (14 Aug 2018) *Hadopi: Beaucoup d'Avertissements Mais Peu de Condamnations*, LeMonde.fr, https://www.lemonde.fr/les-decodeurs/article/2018/08/14/hadopi-beaucoup-d-avertissements-mais-peu-de-condamnations_5342325_4355770.html

Sanovich, S. (2017) 'Computational Propaganda in Russia: The Origins of Digital Misinformation', *Oxford Computational Propaganda Research Project, Working Paper No. 2017(3),* http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Russia.pdf

Sanovich, Sergey (2017) 'Computational Propaganda in Russia: the Origins of Digital Misinformation'

Santa Clara University Markkula Center for Applied Ethics (2017) *The Trust Project*, https://thetrustproject.org

Sartor, G. (2017) *Providers Liability: From the eCommerce Directive to the Future,* In-Depth Analysis for EP IMCO Committee, IP/A/IMCO/2017-07, Brussels: European Parliament

Satariano, A. (25 May 2018) *Ireland's Abortion Referendum Becomes a Test for Facebook and Google,* New York Times, https://www.nytimes.com/2018/05/25/technology/ireland-abortion-vote-facebook-google.html

Schaake, M. (4 April 2018) 'Algorithms Have Become So Powerful We Need a Robust, Europe-Wide Response', *The Guardian* https://www.theguardian.com/commentisfree/2018/apr/04/algorithms-powerful-europe-response-social-media

Shetland Times Ltd v Jonathan Wills and Another, 1997 FSR (Ct Sess. OH), 24 October 1996

Stephens, H. (26 March 2018) *Internet Platforms: It's Time to Step Up and Accepty Your Responsibility (Or Be Held Accountable),* https://hughstephensblog.net/2018/03/26/internet-platforms-its-time-to-step-up-and-accept-your-responsibility-or-be-held-accountable/

Stratton Oakmont Inc v. Prodigy 1995 NY Misc. 23 Media L. Rep. 1794

Syed, N. (2017) 'Real Talk About Fake News: Towards a Better Theory for Platform Governance', *Yale Law Journal 127(Forum)* 337-357

The Verge (2016) *Automated Systems Fight ISIS Propaganda, But At What Cost?,* https://www.theverge.com/2016/9/6/12811680/isis-propaganda-algorithm-facebook-twitter-google

Tobitt, C. (13 September 2018) 'National Newspaper ABCs: Free Evening Standard and Metro Only UK Papers to See Circulation Growth in August', *Press Gazette,* https://www.pressgazette.co.uk/national-abcs-free-evening-standard-only-uk-paper-to-see-circulation-growth-in-august/

Turing, A.M. (1950) 'Computing Machinery and Intelligence', *Mind 49*, 433-460

Tushnet, M. (1998) 'Everything Old is New Again: Early Reflections on the New Chicago School', *Wisconsin Law Review 579*

Twitter (2018) *German NetzDG Transparency Report (Jan-June 2018)* https://cdn.cms-twdigitalassets.com/content/dam/transparency-twitter/data/download-netzdg-report/netzdg-jan-jun-2018.pdf

Twitter (2018) *Oral Evidence for Lords Communications Committee - The Internet: To Regulate or Not To Regulate?* https://parliamentlive.tv/Event/Index/2cd62e7a-d3cf-4605-8d39-4fbaa0adaa76#player-tabs

Twitter (2018) *Partnerships,* https://about.twitter.com/en_us/values/elections-integrity.html#Partnerships

Twitter (2018) *Rules and Policies*, https://help.twitter.com/en/rules-and-policies#research-and-experiments

Twitter (2018) *The Twitter Trust and Safety Council,* https://about.twitter.com/en_us/safety/safety-partners.html

Twitter (2018), *Ads Transparency Center,* https://ads.twitter.com/transparency

UK Department for Digital, Culture, Media and Sport (2018) *Explanatory Memorandum To The Online Pornography (Commercial Basis) Regulations 2018,* http://www.legislation.gov.uk/ukdsi/2018/9780111173183/pdfs/ukdsiem_9780111173183_en.pdf

UK House of Commons Select Committee on Media, Culture and Sport (2018) *Interim Report on Disinformation and 'Fake News'*, https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/36302.htm

UK House of Lords (2017) *AI Select Committee: AI Report Published* https://www.parliament.uk/business/committees/committees-a-z/lords-select/ai-committee/news-parliament-2017/ai-report-published/ (note the report is published in non-standard URL accessed from this link)

UK Information Commissioner's Office (2018) *Democracy Disrupted? Personal Influence and Political Influence*, https://ico.org.uk/media/action-weve-taken/2259369/democracy-disrupted-110718.pdf

UK Information Commissioner's Office (25 Oct 2018) *ICO Issues Maximum £500,000 Fine to Facebook for Failing to Protect Users' Personal Information,* News, https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2018/10/facebook-issued-with-maximum-500-000-fine/

Ukrainian Prism's Foreign Policy (2018) *Disinformation Resilience in Central and Eastern Europe*, http://prismua.org/en/dri-cee/

UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) *Report to the United Nations Human Rights Council on A Human Rights Approach to Platform Content Regulation,* A/HRC/38/35, https://freedex.org/wp-content/blogs.dir/2015/files/2018/05/G1809672.pdf

UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) *Malaysia Reply of 18 June,* https://www.ohchr.org/Documents/Issues/Opinion/Legislation/ReplyMalaysiaOL.pdf

UNESCO (25 October 2018) *UNESCO Partners with Twitter on Global Media and Information Literacy Week 2018,* https://en.unesco.org/news/unesco-partners-twitter-global-media-and-information-literacy-week-2018

University of Michigan (21 August 2018) *Fake News Detector Works Better than A Human*, https://news.umich.edu/fake-news-detector-algorithm-works-better-than-a-human/

Veale, M., Binns, R., and Van Kleek, M. (2018) 'The General Data Protection Regulation: An Opportunity for the CHI Community? (CHI-GDPR 2018)', *Workshop at ACM CHI'18,* 22 April 2018, Montreal, arXiv:1803.06174

Vestager, M. (2018) 'Competition and A Fair Deal for Consumers Online', *Netherlands Authority for Consumers and Markets Fifth Anniversary Conference,* 26 April 2018, The Hague, https://ec.europa.eu/commission/commissioners/2014-2019/vestager/announcements/competition-and-fair-deal-consumers-online_en

Wagner, K. (14 April 2018) *Here's How to See, Edit and Delete the Topics that Facebook Advertisers Use to Target You,* Recode, https://www.recode.net/2018/4/14/17236072/facebook-mark-zuckerberg-ad-advertising-pixel-data

Wardle, C. and Derakhstan, H. (2017) *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making* (DGI(2017)09), Shorenstein Center on Media, Politics and Public Policy at Harvard Kennedy School for the Council of Europe, https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking

Winner, L. (1989) *The Whale And The Reactor: A Search For Limits In An Age of High Technology*, Chicago: University of Chicago Press

Woolley, S., and Howard, P.N. (Eds) Working Paper No.2017.3, *University of Oxford, UK: Project on Computational Propaganda*, http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Russia.pdf

Wylie, K. (9 October 2018) 'Hillary Clinton Sttacks Putin over Brexit as She Claims Democracy is 'Under Siege'', *The Independent,* https://www.independent.co.uk/news/world/americas/hillary-clinton-vladimir-putin-brexit-democracy-under-siege-a8575001.html

Yeung, K. (2017) 'Hypernudge': Big Data as a Mode of Regulation by Design' *Information, Communication & Society* 20(1) pp.118-136

YouTube (2018) *Be Internet Citizens*, https://internetcitizens.withyoutube.com

YouTube (2018) *Community Guidelines,* https://www.youtube.com/yt/about/policies/#community-guidelines

Zannettou, S. et al. (2017) 'The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources', *IMC*

Zannettou, S. et al. (2018) 'On the Origins of Memes by Means of Fringe Web Communities', *ACM internet Measurement Conference (IMC)*

Zannettou, S. et al. (2018) 'The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube', *1st Deep Learning and Security Workshop, co-located with the 39th IEEE Symposium on Security and Privacy*.

Zannettou, S. et al. (2018) 'The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans', *ArXiv*

Zannettou, S. et al. (2018) 'Understanding Web Archiving Services and Their (Mis)Use on Social Media', *ICWSM*

Zannettou, S. et al. (2018) *Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web*, arXiv:1801.09288v1

Zinonos, S., Tsirtsis, A., and Tsapatsoulis, N. (2018) 'Twitter Influencers or Cheated Buyers?', *IEEE Cyber Science and Technology Congress*

Zuckerberg, M. (15 Nov 2018) 'A Blueprint for Content Governance and Enforcement', *Facebook Notes*, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

6. Annex: expert interviews conducted during the study

6.1. Interview respondents

1 David Kaye: UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression

2 Naja Bentzen: Policy Analyst in External Policies Unit at European Parliament Research Service

3 Stephen Turner: Head of Public Policy for Belgium at Twitter

4 Jon Steinberg: Public Policy and Government Relations Manager for EMEA at Google

5 Renate Schroeder: Director at European Federation of Journalists

6 Jennifer Baker: Independent Reporter (Tech Policy and Digital Rights)

7 Monique Goyens: Director-General at European Consumer Organisation (BEUC)

8 Joe McNamee: Executive Director at European Digital Rights

9 Milton Mueller: Professor at Georgia Institute of Technology School of Public Policy; Director Internet Governance Project

10 Madeleine de Cock Buning: Professor of Law, Economics and Governance, Utrecht University School of Law/Molengraaff Institute for Private Law

6.2. Interview protocol

1 Context: Is there anything new about disinformation today?

2 Definition: How do you define 'fake news' or disinformation?

3 Problem/cause: What is the problem or cause of disinformation?

4 Best practices: What is the solution? Can you identify best practices?

5 Freedom of expression and media pluralism: Which solutions/discussions should be emphasized from the perspective of freedom of expression and media pluralism?

6 Technical solutions: Where is there room for improvement in using technical solutions to tackle disinformation online? Can the relationship between technological control and human rights be improved?

QA-01-19-194-EN-N

In this study, we examine the consequences of the increasingly prevalent use of artificial intelligence (AI) disinformation initiatives upon freedom of expression, pluralism and the functioning of a democratic polity.

The study examines the trade-offs in using automated technology to limit the spread of disinformation online. It presents options (from self-regulatory to legislative) to regulate automated content recognition (ACR) technologies in this context. Special attention is paid to the opportunities for the European Union as a whole to take the lead in setting the framework for designing these technologies in a way that enhances accountability and transparency and respects free speech. The present project reviews some of the key academic and policy ideas on technology and disinformation and highlights their relevance to European policy.

This is a publication of the European Science-Media Hub
EPRS | European Parliamentary Research Service

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

ISBN 978-92-846-3947-2 | doi: 10.2861/003689 | QA-01-19-194-EN-N