Platform values and democratic elections: how can the law regulate digital disinformation?

Chris Marsden, Trisha Meyer, Ian Brown

Publication date
01-04-2020

Document Version
Published version

Citation for this work (American Psychological Association 7th edition)
Marsden, C., Meyer, T., & Brown, I. (2020). *Platform values and democratic elections: how can the law regulate digital disinformation?* (Version 1). University of Sussex.
https://hdl.handle.net/10779/uos.23483480.v1

Published in
Computer Law and Security Review

Link to external publisher version
https://doi.org/10.1016/j.clsr.2019.105373

COMPUTER LAW & SECURITY REVIEW 36 (2020) 105373

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/CLSR

Computer Law & Security Review

Platform values and democratic elections: How can the law regulate digital disinformation?

Chris Marsden[a,*], *Trisha Meyer*[b], *Ian Brown*[c,1]

[a] *The University of Sussex, United Kingdom*
[b] *The Vrije Universiteit Brussel, Belgium*
[c] *Research ICT Africa, South Africa*

ARTICLE INFO

Keywords:
Disinformation
Artificial intelligence
Co-regulation
Self-regulation
Internet law
Social media regulation
Platform regulation
Elections
Fake news

ABSTRACT

This article examines how governments can regulate the values of social media companies that themselves regulate disinformation spread on their own platforms. We use 'disinformation' to refer to motivated faking of news. We examine the effects that disinformation initiatives (many based on automated decision-making systems using Artificial Intelligence [AI] to cope with the scale of content being shared) 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. We particularly focus on the responses of the member states and institutions of the European Union. In Section 1, we argue that the apparent significance of the threat has led many governments to legislate despite this lack of evidence, with over 40 national laws to combat disinformation chronicled by March 2019. Which types of regulation are proposed, which actors are targeted, and who is making these regulations? Regulating fake news should not fall solely on national governments or supranational bodies like the European Union. Neither should the companies be responsible for regulating themselves. Instead, we favour co-regulation. Co-regulation means that the companies develop – individually or collectively – mechanisms to regulate their own users, which in turn must be approved by democratically legitimate state regulators or legislatures, who also monitor their effectiveness. In Section 2, we explain the current EU use of Codes of Conduct. In Section 3, we then explain the relatively novel idea that social media content regulation, and specifically disinformation, can be dealt with by deploying AI at massive scale. It is necessary to deal with this technological issue in order to explain the wider content of co-regulatory policy options, which we explain and for which we argue in Section 4. In Section 5 we explain

* Corresponding author.
E-mail address: c.marsden@sussex.ac.uk (C. Marsden).

[1] This article is based on their expert report, Marsden, C. & Meyer, T. *Regulating Disinformation with Artificial Intelligence. The Effects of Disinformation Initiatives on Freedom of Expression and Media Pluralism*, (Brussels: European Parliament, 2019) doi: 10.2861/003689. See also Meyer, T., Marsden, C. & Brown, I. "Regulating disinformation with technology: analysis of policy initiatives relevant to illegal content and disinformation online in the European Union" in E. Kużelewska, G. Terzis, D. Trottier & D. Kloza (eds.) *Disinformation and digital media as a challenge for democracy*, European Integration and Democracy Series, Vol. 6. (Cambridge: Intersentia 2020, in print). A summary open access article was earlier published, see Marsden, C. and T. Meyer (2019) "How can the law regulate removal of fake news?" Computers and Law, https://www.scl.org/articles/10425-how-can-the-law-regulate-removal-of-fake-news.

https://doi.org/10.1016/j.clsr.2019.105373

what this means for technology regulation generally, and the socio-economic calculus in this policy field.

1. 'Fake news' and disinformation on social media platforms

The digitization of disinformation via social media platforms has been blamed for skewing the results of elections and referenda and amplifying hate speech in many other nations.[2] The evidence of harm, and legislative and judicial responses to deliberate disinformation, are growing but nascent.[3] The apparent significance of the threat has led many governments to legislate despite this lack of evidence, with over forty national laws to combat disinformation chronicled by March 2019.[4] 'Fake news' – more properly termed disinformation – has recently become endemic to social networking on the Internet.

We use 'disinformation' to refer to motivated faking of news, in line with the European Union (EU)'s institutions[5] and High Level Expert Group on disinformation,[6] and the regional and global United Nations (UN) rapporteurs on freedom of information's use of the term.[7] It is a problem at least as old as written media, but has become more controversial as evidence of state-sponsored domestic and foreign influence peddling online, and micro-targeted political influence marketing via social media, has become ubiquitous. Particularly, the problem of state-sponsored social media inaccuracy was first identified in the Ukraine in 2011, when Russia was accused of deliberately faking news of political corruption.[8]

This article examines how governments can regulate the values of companies (Facebook, YouTube, Twitter in particular) that themselves regulate disinformation spread on their own platforms. We examine the effects that disinformation initiatives (many based on automated decision-making systems using AI to cope with the sheer scale of content being shared) 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.[9] We particularly focus on the national and supranational responses of the member states and institutions of the European Union, which in the European Parliament elections of May 2019 held the largest global exercise in democracy behind Indian general elections, with a highly networked electorate, and serious concerns about foreign interference.

In international human rights law, such as Article 10 of the European Convention on Human Rights and Fundamental Freedoms 1950,[10] restrictions to freedom of expression must be provided by law, legitimate and proven necessary and the least restrictive means to pursue the aim. In this article, we argue that governments should not push this difficult judgement exercise in disinformation onto online intermediaries, who are inexpert in and not incentivized to judge fundamental rights, and not bound by States' international human rights commitments.[11] The UN Special Rapporteur on Freedom of Opinion and Expression recently called for assessments of the impact of technology-based solutions on human rights in general,[12] and freedom of expression and media pluralism in particular.[13]

What can be done, by whom, to whom, to address these problems? Which types of regulation are proposed, which actors are targeted, and who is making these regulations? Who should regulate fake news shared online? We argue that regulating fake news should not fall solely on national gov-

[2] Euronews *How Can Europe Tackle Fake News in the Digital Age?*, (9 Jan 2019) https://www.euronews.com/2019/01/09/how-can-europe-tackle-fake-news-in-the-digital-age.

[3] 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.

[4] Funke, Daniel (2019) *A guide to anti-misinformation actions around the world*, Poynter Institute, dynamically updated, at https://www.poynter.org/ifcn/anti-misinformation-actions/.

[5] Bentzen, Naja, *Understanding Propaganda and Disinformation*, (European Parliament Research Service 2015), http://www.europarl.europa.eu/RegData/etudes/ATAG/2015/571332/EPRS_ATA(2015)571332_EN.pdf.

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

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

[8] *See* Sanovich, Sergey *Computational Propaganda in Russia: The Origins of Digital Misinformation* (Oxford Computational Propaganda Research Project, Working Paper No. 2017.3, 2017), http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Russia.pdf.

[9] 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.

[10] Convention for the Protection of Human Rights and Fundamental Freedoms as amended by Protocols No. 11 and No. 14, European Treaty series No.5, Signed Rome, 4 November 1950, at https://www.coe.int/en/web/conventions/full-list/-/conventions/rms/0900001680063765.

[11] Brown, I. and Korff, K. *Digital Freedoms in International Law: Practical Steps to Protect Human Rights Online* (Global Network Initiative, 2012) https://globalnetworkinitiative.org/wp-content/uploads/2016/10/GNI-Digital-Freedoms-in-International-Law.pdf.

[12] UNHRC, *Report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression* (29 August 2018) UN Doc A/73/348.

[13] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression *Report to the United Nations Human Rights Council on A Human Rights Approach to Platform Content Regulation,* A/HRC/38/35, (6 April 2018) https://undocs.org/A/HRC/38/35.

ernments or supranational bodies like the European Union. Neither should the companies be responsible for regulating themselves and ourselves.[14] Instead, we favour co-regulation. Co-regulation means that the companies develop – individually or collectively – mechanisms to regulate their own users, which in turn must be approved by democratically legitimate state regulators or legislatures, who also monitor their effectiveness.[15] The article proceeds as follows. In the following Section 2, we explain the current use of Codes of Conduct. In Section 3, we then explain the relatively novel idea that social media content regulation, and specifically disinformation, can be dealt with by deploying AI at massive scale. It is necessary to deal with this technological issue in order to explain the wider content of policy options, for which we argue in Section 4 that co-regulation is the most likely and appropriate outcome. In Section 5 we explain what this means for technology regulation generally, and the socio-economic calculus in this policy field.

2. Current state of play for European disinformation policy

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, or the ban on certain misleading advertising techniques. 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. The best known example at national legislation is Germany's Network Enforcement Law 2017 ('NetzDG').[16] This article however focuses on European level responses. Different aspects of the disinformation problem merit different types of regulation. More broadly, institutional support can be provided to safeguard media pluralism, encourage fact-checking and enhance media literacy. We note that all proposed policy solutions explained in Section 4 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.

The most important European policy document dealing with disinformation was the result of policy formation in 2018. The 2018 EU-orchestrated 2018 self-regulatory *Code of Practice on Online Disinformation* followed from the EU High Level Group report, and examined technology-based solutions to disinformation, focused on the actions of online intermediaries (social media platforms, search engines and online advertisers) to curb disinformation online.[17] Though criticized by its own Sounding Board for not stipulating any measurable outcomes,[18] Nielsen argued the Code of Practice produced "three potentially major accomplishments" :[19]

- Signatories 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".
- Signatories must submit their efforts to counter disinformation to external scrutiny by an 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".
- 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".[20]

Other EU initiatives also call for pro-active measures by intermediaries through use of AI to aid removal of illegal content. The proposed EU Regulation on the Prevention of Dissemination of Terrorist Content Online[21] targets rapid removal terrorist content by online intermediaries. Article 17 of the recently passed Copyright in the Digital Single Market Directive 2019 suggests changing intermediary liability protections with a requirement to use filtering technologies.[22] The European Commission has used the overarching phrase "a fair deal for consumers" .[23] These policy developments fit in a context where social media platforms and search engines

[14] Belli, L., Francisco, Pedro Augusto P.; Zingales, N.eds. *Platform regulations: how platforms are regulated and how they regulate us* (FGV, Rio de Janeiro 2017).

[15] Marsden, C. 'Internet Co-Regulation and Constitutionalism: Towards European Judicial Review', *International Review of Law, Computers and Technology 26(2)* (2012) 212–228.

[16] Netzwerkdurchsetzungsgesetz (German Network Enforcement Act) 2017, see 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.

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

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

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

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

[21] 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.

[22] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market, OJ L 130, 17.5.2019, p. 92–125.

[23] Vestager, M. '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.

are increasingly scrutinized on competition grounds[24] and requested to take more responsibility for content removal.

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. 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), authorizing new licensees 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.[25]

Broadcasting is defined in Article 1 of the Audio Visual Media Services (AVMS) Directive as: "editorial responsibility of a media service provider [for the] principal purpose of providing programmes, in order to 'inform, entertain or educate' to general public, conveyed by electronic communications networks".[26] That is distinguished from Internet communication by its specific audience and that fact that the user chooses the content.[27] Broadcasting law has extensive statute and case law, including that interpreted by the European Convention on Human Rights,[28] appealed to the European Court of Human Rights at Strasbourg,[29] in regulating election advertising, there was until recently a paucity of case law for the Internet. Baroness Hale has argued that: "In the United Kingdom, and elsewhere in Europe, we do not want our government or its policies to be decided by the highest spenders. …We have to accept that some people have greater resources than others with which to put their views across. But we want to avoid the grosser distortions which unrestricted access to the broadcast media will bring."[30]

Application of broadcast rules to the Internet in the case of video on demand (VOD) services was applied by courts in Belgium nearly two decades ago.[31] Were the same reasoning applied more broadly to the Internet, specific electoral spending law would be rigorously applied by regulators and upheld by courts. The dissenting Strasbourg judges in *Animal Defenders International* argued strongly in 2013 that this should not be so: "Given the comparative potency of newer media such as the Internet, a distinction based on the particular influence of the broadcast media was not relevant. Information obtained through the use of the Internet and social networks is gradually having the same impact, if not more, as broadcasted information."[32] The majority might argue that the onus of regulation needs to be reversed based on the precautionary principle, and that broadcast electoral advertising rules could be extended to the Internet without infringing Article 10.

[24] For a scholarly overview and discussion of ongoing platform and search engine competition cases, see Mandrescu, D. 'Applying EU Competition Law to Online Platforms: The Road Ahead – Part I', Competition Law Review 38(8) (2017) 353–365; Mandrescu, D. 'Applying EU Competition Law to Online Platforms: The Road Ahead – Part II', Competition Law Review 38(9) (2017) 410-422. For an earlier call to co-regulation, see Marsden, C. (2012) n.15.

[25] See e.g. C-288/89 *Stichting Collectieve Antennevoorziening Gouda and others* judgment of 25 July 1991 [1991] ECR I-4007); 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, OJ L 298, 17.10.1989, pp.23–30 (particularly Recital 17).

[26] Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) OJ L 95, 15.4.2010, pp.1–24. Note this Directive is Consolidated, having been adopted in 2007 with transposition date 19 December 2009. Specifically excluded are services including: private correspondence such as email; games of chance, on-line games, search engines (Recital 22); stand-alone text based services (Recital 23); electronic newspapers (Recital 28); services 'where any audiovisual content is merely incidental to the service and not its principal purpose' (Recital 22). See Valcke, P, and Stevens, D. 'Graduated regulation of "regulatable" content and the European Audiovisual Media Services Directive: one small step for the industry and one giant leap for the regulator?' 24 Telematics & Informatics (2007) 285, 295.

[27] COM(96) 483, Green Paper on the protection of minors and human dignity in audiovisual and information services, 16.10.97; COM(97) 487, Communication on Illegal and Harmful content on the Internet, 16.10.97; Recommendation 98/560/EC on the development of the competitiveness of the European audiovisual and information services industry by promoting national frameworks aimed at achieving a comparable and effective level of protection of minors and human dignity OJ L 270, 7.10.1998. See further COM(2004)0341 European Parliament legislative resolution on the proposal for a recommendation of the European Parliament and of the Council on the protection of minors and human dignity and the right of reply in relation to the competitiveness of the European audiovisual and information services industry– C6-0029/2004 – 2004/0117(COD).

[28] Convention for the Protection of Human Rights and Fundamental Freedoms, Article 10.

[29] See for instance *Animal Defenders International* (2013) ECHR 362, (2013) 57 EHRR 21. This was a European Court of Human Rights Grand Chamber majority (4-3) judgment in the case, which upheld the Ofcom UK ban on television or radio advertising by the animal-protection organization, on grounds that objectives were "wholly or mainly of a political nature": held that there was no breach of Article 10 in applying Communications Act 2003 Act S.321(2).

[30] See Hale, Brenda (2012) "Argentoratum Locutum: Is Strasbourg or the Supreme Court Supreme?" Human Rights Law Review 12 (1): 65-78 doi: 10.1093/hrlr/ngs001 Hale argued: "This was clearly an interference with freedom of speech, indeed freedom of political speech, which is the most important of the kinds of speech protected by Article 10. It would certainly not be tolerated in the United States".

[31] *Mediakabel BV v. Commissariaat voor de Media*, case C-89/04, 6 November 2002 Belgian Constitutional Court judgment, appealed in Judgment of the European Court of Justice, decision of 2 June 2005, [2005] ECR I-04891.

[32] Dissenting justices Tulkens, Spielmann, Laffranque argued: "The more convincing the general justifications for the general measure, the less importance the Court will attach to its impact" citing *Murphy v. Ireland* (2003) [2003] ECHR 352, (2004) 38 EHRR 13, and *TV Vest AS v. Norway*, no. 21132/05 (2009) 48 EHRR 51 (Chamber, First Section). Further: "Government justified the contested measure by, in particular, need to protect electoral process as part of the democratic order, (*Bowman v. UK* (1998–I) ECHR 4, Court accepted that a statutory control of the public debate was necessary given the risk posed to the right to free elections…But prohibition in question is not limited to electoral periods, the Bowman judgment and reasoning based on electoral process are of little bearing in this case (TV Vest §66)".

The extension of broadcast rules to non-broadcast content, whether text-based or in any case at the user's individual choice, would be a significant step that would in all likelihood increase the concentration of online communication in the hands of the largest platforms that can employ economies of scale: deploying proprietary filters to remove harmful content. Examples from the Internet include the attempts to prevent child pornography and terrorist video distribution, as well as copyrighted files. In each case, the use of technologies such as checking hash values in theory permitted removal before publication by the platforms deploying the technology, specifically YouTube and Facebook. In practice, the proliferation of content was restricted but by no means prevented by such technological intervention.[33]

The European Commission has recognized that the platform liability position is becoming very uncertain for platforms, and has agreed to review Notice and Take Down procedures in the E-Commerce Directive (ECD).[34] It issued a 2013 working paper on its progress,[35] 2016 proposals on the Digital Single Market Strategy,[36] which may result in legislative proposals in the 2019–24 legislative period.

As platforms await a legislative option, they are also threatened by the possibility of filtering offered by the CJEU case of *Eva Glawischnig-Piesczek v Facebook Ireland*, decided on 3 October 2019.[37] It is useful to briefly explain the decade of prior CJEU case law in legal liability of intermediary platforms, whether as now for disinformation laws or earlier cases involving copyright violations. In *Scarlet Extended*, the CJEU had to balance rights holders against access providers and users' rights.[38] The CJEU recognized that the risk of preventing access to lawful content through over-blocking or over-filtering is a relevant factor to take into account. *Scarlet Extended* is an extension of the earlier CJEU reasoning in *Promusicae*[39]: the Belgian court ordered an access provider to filter all traffic for copyright infringement and 'pay for the privilege' of enforcing copyright on behalf of rights holders. Paragraphs 43–44 in *Scarlet Extended* are critical for general guidance:

> *The protection of the right to intellectual property is indeed enshrined in Article 17(2) of the Charter of Fundamental Rights of the European Union ('the Charter'). There is, however, nothing whatsoever in the wording of that provision or in the Court's case-law to suggest that right is inviolable and must for that reason be absolutely protected. As paragraphs 62 to 68 of the judgment in Promusicae make clear, the protection of the fundamental right to property, which includes the rights linked to intellectual property, must be balanced against the protection of other fundamental rights.*

The CJEU stated that the Belgian injunction in issue would be a serious infringement of the freedom of the access provider concerned to conduct its business, since it would require it to install a complicated, costly, permanent computer system at its own expense. The Belgian court's order would have emasculated Article 15 of the ECD. This reasoning was reconfirmed and extended from access to social networking platforms by the February 2012 decision in *SABAM v Netlog*.[40] *Scarlet Extended* is a short decision (as is *Netlog*), and the question asked was set at the most extreme end of the scale, an injunction that was: (a) preventative; (b) entirely at the access provider's expense; (c) for an unlimited period; (d) applied to all customers indiscriminately; (e) for all kinds of communications. Useful guidance for national courts in the judgment included that the complexity/cost of the proposed Belgian system weighed against it, that Internet Protocol addresses are personal data, that the Belgian injunction was overbroad and could interfere with lawful as well as unlawful use.[41] Confirmation that IP addresses may be considered personal data arrived in 2016.[42] The remedy of URL blocking in *Scarlet* is indiscriminate, whereas the United Kingdom 'Cleanfeed' system deployed in the *Newzbin2* judgment of the English High Court was already in place and the cost essentially negligible.[43] As a result UK rights holders requested ISPs to block access to the file sharing website Pirate Bay.[44] Other EU countries have also seen successful applications for injunctions against

[33] See for example the difficulties encountered by platforms attempting to restrict sharing of video footage of the Christchurch shootings: Herne, A. & Waterson, J. (2019) 'Social Media Firms Fight to Delete Christchurch Shooting Footage', The Guardian, 19 March, https://www.theguardian.com/world/2019/mar/15/video-of-christchurch-attack-runs-on-social-media-and-news-sites.

[34] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') OJ L 178/1 pp.1–16, 17 July 2000.

[35] European Commission "Report on the implementation of the e-commerce action plan" 23/04/2013 SWD(2013) 153 final.

[36] European Commission (2016) Proposal for a Regulation of the European Parliament and of the Council on addressing geo-blocking and other forms of discrimination based on customers' nationality, place of residence or place of establishment within the internal market and amending Regulation (EC) No 2006/2004 and Directive 2009/22/EC, 25 May.

[37] Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* ECLI:EU:C:2019:458 decided 3 October 2019.

[38] Case C-70/10 *Scarlet Extended SA v Société Belge des auteurs, compositeurs et éditeurs (SABAM)* OJ C 113, 1 May 2010: 20–20. Decided 24 November 2011, OJ C 25/6, 28 January 2012.

[39] Case C-275/06 *Productores de Música de España (Promusicae) v Telefónica de España SAU*, judgment of 29 January 2008 [2008] ECR I271 at para 70.

[40] Case C-360/10 *SABAM v Netlog*, 2 C.M.L.R. 18. 3 (2012) decided 16 February.

[41] See C-70/10 *Scarlet*, at paragraph 52: 'injunction could potentially undermine freedom of information since that system might not distinguish adequately between unlawful content and lawful content, with the result that its introduction could lead to the blocking of lawful communications'. See further *SABAM v Netlog*, at paragraphs 36–38.

[42] Case 582/14 *Patrick Breyer v Germany*, decided 19 October 2016, ECLI:EU:C:2016:779, 1 WLR 1569 (2017).

[43] *Twentieth Century Fox Film Corporation and Others v British Telecommunications Plc (No 2)* [2011] EWHC 2714 (Ch) 26 October 2011. This was the first order in the UK under Section 97A of the Copyright Designs and Patents Act 1988, which implements Article 8(3) of Directive 2001/29/EC, though it was already possible under UK law to seek injunctions against intermediaries.

[44] *Dramatic Entertainment Limited v British Sky Broadcasting Limited* [2012] EWHC 1152 (Ch) 2 May at: <http://www.bailii.org/ew/cases/EWHC/Ch/2012/1152.html>.

ISPs.[45] The law as it stands awaits the European institutions' decisions in 2020 on how to reform the ECD, which will affect the liability environment for platforms in their disinformation polices, together with other content issues such as copyright.

While many previous media law techniques are inappropriate for online social media platforms, and some of these measures were abused by governments against the spirit of media pluralism, legislators need to consider which regulatory measures may protect freedom of information and expression by providing a bulwark against disinformation.

In Section 3, we will argue that AI is in the short to medium term highly unlikely to replace human judgement, and there is no possibility of restricting disinformation at source such that no-one views it. A key lesson from the peer-to-peer file sharing discussion has been that alternate forms of offline digital distribution are powerful replacements for online sharing: copyrighted material and other forms of content can easily be shared via a cheap ubiquitous technology such as a USB key or other form of portable storage, with much less user education than using a 'dark net' encrypted service such as a Tor relay.

3. Machine learning and AI as 'Solutions' for disinformation

Written evidence of disingenuous or 'fake news' is as old as the cuneiform tablets of Hammurabi.[46] Technological solutions risk threatening freedom of speech and media pluralism. The problem has become far more visible and arguably acute online; social networks such as Facebook, YouTube, Twitter and WhatsApp allow information, authentic or otherwise, to spread globally and instantly. 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 A/B testing, targeted advertising, and automated, targeted recycling of fake content (manipulation).*[47]

She 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".[48]

Within machine learning techniques that are advancing towards AI, automated content recognition technologies are textual and audio-visual analysis programmes that are trained to identify potential 'bot' accounts and unusual potential disinformation material.[49] In this article, AI refers to the use of automated techniques in the recognition and moderation of content and accounts, to assist human judgement.[50] Moderating content at larger scale requires AI as a supplement to human moderation (editing).[51] The shorthand 'AI' is used in the remainder of the article to refer to both these technologies. Where necessary, we specify which of the two (recognition or moderation) is implied.

Can AI solve the fake news problem? One argument being put forward by the owners of online platforms is that new technologies can solve the very problems they create. Chief among those technologies is machine learning or AI, alongside user reporting of abuse. However the notion that AI is a 'miracle cure', the panacea for fake news, is optimistic at best. Platforms argue that the use of automated content filtering systems, that use algorithmic processes to identify harmful content, provide a means for effective self-regulation by platforms. AI algorithms cannot be the only way to regulate content in future. Automated technologies such as AI are not a silver bullet for identifying illegal or "harmful" content. They 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. The European Commission consultations on online platforms, assessment of the formally self-regulatory Code of Conduct fighting hate speech, and its over-

[45] Court of Appeal of Antwerp Case 2010/AR/2541 *VZW Belgian Anti-Piracy Federation v NV Telenet* 26 September 2011.

[46] Discussed in Enriques, L. *Financial Supervisors and RegTech: Four Roles and Four Challenges* (Oxford University, Business Law Blog, 9 Oct 2017) http://disq.us/t/2ucbsud.

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

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

[49] 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, see Klinger, J., Mateos-Garcia, J.C., and Stathoulopoulos, K. *Deep Learning, Deep Change? Mapping the Development of the Artificial Intelligence General Purpose Technology*, (2018) doi: http://dx.doi.org/10.2139/ssrn.3233463. See also Zuckerberg, M. "A Blueprint for Content Governance and Enforcement", *Facebook Notes*, (15 Nov 2018) 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". See also Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. "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 2009), pp. 288–96; Monroe, B., Colaresi, M., and Quinn, K. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict", *Political Analysis 16(4)* (2008) 372-403; Azevedo, L. "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, (2018) pp. 807-811, doi: https://doi.org/10.1145/3184558.3186567.

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

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

all Digital Single Market strategy all arguing for greater "responsibility" by online platforms.[52] This means faster private enforcement which can also be seen as censorship given the lack of 'put back' appeal guarantees.[53]

One of the problems is that they are responding to a perceived need from politicians to remove more content, rather than addressing fair process and due process. The informal incentive structure may require platforms to demonstrate to politicians how much content they have removed, when a very important factor in accountability for legal content posted may be "examples of successful appeals to put content back online". While the involvement of private parties in EU administrative governance has the clear advantage of delivering policies which are based on the expertise of the regulatees, private-party rule-making raises significant concerns in terms of its legitimacy. The EU response has been to effectively ignore the inconvenience that Internet self- or co-regulation is private censorship of free speech. The UK Parliament Artificial Intelligence (AI) Committee reported on some of these issues in 2017.[54] There are an enormous number of false positives in taking material down, which need human intervention to analyse. Google and Facebook announced in 2018 their intention to employ 50,000 more people as content moderators (subcontracted to so-called Mechanical Turks). 'Mechanical Turks' are people employed—subcontracted, typically—to carry out these activities,[55] in parts of the world where their own cultural understanding of the content they are dealing with may not be ideal.[56] 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.

Technical research into disinformation has followed several tracks:

- identifying and removing billions of bot as distinct from human accounts[57];
- identifying the real world effects of Internet communication on social networks[58];
- assessing the impact of disinformation via media consumption and electoral outcomes[59];
- researching security threats from disinformation;
- researching discrimination and bias in the algorithms used to both propagate and increasingly to identify and/or disable disinformation.[60]

Online disinformation consumption includes that of video news and newspapers, whose readerships have largely migrated online,[61] 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. Partial evidence of AI effectiveness is supplied by corporates. Facebook stated in 2018 that its automated systems detect 99% of the terrorism-related content it removes, as well as 96% of nude images and 52% of hate speech.[62] It has also been reported to automatically detect "nearly 100% of spam… 98.5%

[52] COM (2015) 192 A Digital Single Market Strategy for Europe, final, 6 May 2015, para. 3.3. EC (2015) Public Consultation on the Regulatory Environment for Platforms, Online Intermediaries, Data and Cloud Computing and the Collaborative Economy available at https://ec.europa.eu/digital-single-market/news/public-consultation-regulatory-environment-platforms-online-intermediaries-data-and-cloud; EC (2016) Fighting Illegal Online Hate Speech: First Assessment of the New Code of Conduct, available at http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50840 COM (2016) 288 Online Platforms and the Digital Single Market: Opportunities and Challenges for Europe, 25 May 2016, p. 9.

[53] See Frosio, Giancarlo F. (2017) 'From horizontal to vertical: an intermediary liability earthquake in Europe' 12 Journal of Intellectual Property Law and Practice 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' 25 May.

[54] 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).

[55] Hara, Kotaro; Adams, Abi; Milland, Kristy; Savage, Saiph; Callison-Burch, Chris; Bigham, Jeffrey (2017) "A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk" eprint arXiv:1712.05796 Conditionally accepted for inclusion in the 2018 ACM Conference on Human Factors in Computing Systems (CHI'18) Papers program.

[56] See Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In CHI'10 extended abstracts on Human factors in computing systems (pp. 2863-2872). Association of Computing Machinery. See effects in Youtube Transparency Report (2018) https://transparencyreport.google.com/youtube-policy/overview.

[57] 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*.

[58] 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.

[59] Zannettou, S. et al. (2018) *Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web*, arXiv:1801.09288v1.

[60] 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.

[61] 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.

[62] Zuckerberg, M. (2018) 'A Blueprint for Content Governance and Enforcement', 15 November, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/?hc_location=ufi.

of fake accounts... and 86% of graphic violence-related removals".[63]

This evidence of AI removals is only unaudited company claims. Note Facebook's AI claims to detect "just 38% 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".[64] 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 of both fact verification and data-driven machine learning judgments can be integrated".[65] This is a sensible approach where resources allow for such a wide spectrum of solutions.

AI therefore cannot be the only way to regulate content in future.[66] 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 current incentive structure is 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.[67] Content moderation at scale still needs human intervention to interpret AI-flagged content.

A satisfactory solution to algorithmic transparency might be the ability to replicate the result that has been achieved by the company producing the algorithm. Transparency and explanation is necessary, but it is a small first step towards better regulation.[68] Veale, Binns and 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.[69] 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.

Fig. 1 – Reeve model of regulatory pyramid.
***Note:* Reeve, B'. "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) (2011) 128–46.**

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 judgement, mistaking the imitation for what is imitated".[70] Some of the claims that AI can 'solve' the problem of disinformation do just that. Over time, AI solutions to detect and remove illegal/undesirable content are becoming more effective, but they raise questions about who is the 'judge' in determining what is legal/illegal, and undesirable in society. Underlying AI use is a difficult choice between different elements of law and technology, public and private solutions, with trade-offs between judicial decision-making, scalability, and impact on users' freedom of expression.

In Section 4 we explore the options for dealing with this tool in analysing disinformation, where an imitation game is insufficient to identify truth and falsehood, and human intervention on a large scale is required.

4. Options for regulating AI in disinformation introduced

In this section, we explore the policy options that are available in some depth, and which would form the basis of EU legal

[63] 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.

[64] Koebler and Cox (2018).

[65] 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

[66] Schaake, M. (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.

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

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

[69] 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.

[70] The imitation game is often known as the Turing test, after Turing, A.M. (1950) '*Computing Machinery and Intelligence*', *Mind* 49, 433-460.

policy towards disinformation.[71] Policy options have moved beyond the 'Goldilocks' theory of a three card trick (one too hot, one too cold, one just right) to encompass a range of self- and co-regulatory options.[72] (Fig. 1).

4.1. *Co-regulation as a regulatory technique explained*

Internet self- and co-regulatory arrangements have a legal foundation, and specific legal constraints or conditions to be respected. The Internet developed self-regulation based on the Codes of Conduct (CoC) and Terms of Use (ToU) that early Internet users employed, in the scientific institutions that first developed the protocols and social standards.[73] The use of such ethical standards is more corporate social responsibility than law.[74] Given the rapid growth, complex interrelationships and dynamic changes that have taken place in the current century, governments have broadly accepted that a more flexible and innovation-friendly model of regulation is required.[75] Cafaggi stated in regard to sanctioned regulation: "An intermediate hypothesis between delegated private regulation and ex post recognized private regulation is that in which private regulation, produced by the private or self-regulator, has to be approved by a public authority to become effective."[76] It is a pragmatic acceptance that the models used for regulation should be as flexible as possible, to permit significantly greater user innovation and freedom than with other types of communications (notably telecoms and broadcasting). This includes using both hard and 'soft law' forms of regulation.[77] The Internet self-regulatory paradigm has been increasingly challenged by the growth and evolution of the Internet and associated technologies including cloud computing, blockchains, smart contracts and Artificial Intelligence.[78]

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.[79] In Germany, this is known as regulated self-regulation.[80] This is a hybrid system subject to statutory control. Examples from the internet regulatory ecosystem are:

- Nominet, largest European Domain Name System Registry operator, which operates the .uk domain since 1996, under ultimate control by government via Digital Economy Act 2010 [81];

[71] European Union, Inter-institutional agreement on better law-making, OJ L 123, 12.5.2016, p. 1, at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016Q0512(01)&from=EN. See also Council of the European Union, Impact Assessment - Indicative guidance for Working Party Chairs, Brussels, 9 June 2016, 9790/16, at http://data.consilium.europa.eu/doc/document/ST-9790-2016-INIT/en/pdf.

[72] Dunlop, C., & Radaelli, C. "Impact Assessment in the European Union: Lessons from a Research Project" European Journal of Risk Regulation, 6(1) (2015). 27-34. doi:10.1017/S1867299X00004256; Damonte, A., Dunlop, C., & Radaelli, C. "Regulatory Reform: Research Agendas, Policy Instruments and Causation" European Journal of Risk Regulation, 8(1), (2017) 72-76. doi:10.1017/err.2016.12.

[73] Werbach, Kevin *Digital Tornado: The Internet and Telecommunications Policy*, Office of Plans and Policies Working Paper 29. (Washington: Federal Communications Commission 1997).

[74] Abbott K, Snidal D. "The Governance Triangle: Regulatory Standards Institutions and the Shadow of the State" Chapter 2 in Mattli W, Woods N (eds) *The Politics of Global Regulation*, (Princeton University Press 2009), pp.44-88. Abbott K, Snidal D. (2004) "Hard and soft law in international governance", International Organization 54, pp.421-422; Helin, S., & Sandström, J. "An inquiry into the study of corporate codes of ethics", Journal of Business Ethics 75 (2007), pp.253–271. Higgs-Kleyn, N., & Kapelianis, D. "The role of professional codes in regulating ethical conduct", Journal of Business Ethics, 19 (1999), 363–374. Vrielink, Mirjan Oude, Cor van Montfort, Meike Bokhorst Codes as hybrid regulation, ECPR Standing Group on Regulatory Governance, (June 17-19 2010), Dublin.

[75] Generally on the role of smart regulation, see Gunningham N., Rees J. "Industry Self-regulation: An Institutional Perspective", Law & Policy 19(4) (1997); Gunningham, N. and Grabosky, P. *Smart Regulation: Designing Environmental Policy*, (Oxford University Press 1998); Gaines, Sanford E. and Cliona Kimber "Redirecting Self-Regulation", Env. Law 13 (2001), p.157. More generally, see Black, J. "Constitutionalising Self-Regulation", Modern Law Review, Vol. 59, No. 1, (1996) pp. 24-55 at p.59; Black, J. *Managing the Financial Crisis – The Constitutional Dimension*, LSE Legal Studies Working Paper No. 12/2010 (2010).

[76] See Cafaggi, F. (2006) Rethinking private regulation in the European regulatory space, *EUI Working Paper LAW No. 2006/13*, at http://cadmus.eui.eu/bitstream/handle/1814/4369/LAW2006.13.PDF?sequence=1 at p.24.

[77] On the role of 'soft law' more generally, see Senden, L. (2005) Soft Law, Self-Regulation and Co-Regulation In European Law: Where Do They Meet? Electronic Journal of Comparative Law, vol. 9.1 at http://www.ejcl.org/91/abs91-3.html. Cosma, H. & Whish, R. (2003) Soft Law in the Field of EU Competition Policy, European Business Law Review, Vol. 14,. Pt. 1 pp.25-56. Hodson, Dermot and Imelda Maher (2004) Soft law and sanctions: economic policy co-ordination and reform of the Stability and Growth Pact, Journal of European Public Policy, Volume 11 Issue 5 pp.798–813.

[78] Werbach, Kevin (2018) The Blockchain and the New Architecture of Trust, MIT Press; Cohen, Julie E. (2016) The Regulatory State in the Information Age, 17 Theoretical Inq. L. 369-414; Finck, Michèle "Digital Co-Regulation: Designing a Supranational Legal Framework for the Platform Economy", European Law Review (2018): https://ssrn.com/abstract=2990043; Hildebrandt, Mireille "Law as Information in the Era of Data-Driven Agency", Modern Law Review, 79: (2016), 1–30, at doi:10.1111/1468-2230.12165; Werbach, Kevin "Contracts Ex Machina", 67 Duke LJ (2017) 101.

[79] Marsden, C. T. (2011) Internet Co-Regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace, Cambridge University Press, explores 25 such self- and co-regulatory schemes.

[80] 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.

[81] See Marsden, C. (2011), n.79 at p61. By 2018, there were 12 million UK domains registered, see Nominet (2018), *UK Domains*, https://www.nominet.uk/uk-domains/.

- EURID which regulates and operates registries under the .eu domain since 2003.[82]

European co-regulation in wider consumer protection legislation was detailed in 2002, and became official policy in December 2003, with the Inter-Institutional Agreement on Better Law-Making, which defined co-regulation.[83] Self-regulation is viewed as making standards and practices across industry that the Commission, or a Member State, views agnostically in pre-legislative or legislative terms. Government then analyse the extent to which self-regulation approaches the standards of 'representativeness' which co-regulation is meant to demonstrate as a best practice. The Inter-Institutional Agreement confirmed in 2003 that forms of regulation short of state regulation: "will not be applicable where fundamental rights or important political options are at stake or in situations where the rules must be applied in a uniform fashion in all Member States." The European Commission in 2005 went on to analyse co-regulation in terms of 'better regulation' (COM/2005/97). This was immediately made part of internal EC practice in the Impact Assessment Guidelines (SEC/2005/791) which the Commission must follow before bringing forward a new legislative or policy proposal, updated in 2015.[84]

This European regulatory activity in defining co- and self-regulation was matched by its continued research into the impact of the Internet and its own legislative and policy initiatives since 1998. The European Commission thus commissioned substantial independent research from 2001 onwards in assessing Internet regulation and the enforcement thereof by private actors.[85] Price and Verhulst examined private Internet enforcement via internal self-organisation: they identified increasing realism in recognising competition problems, emerging monopolies and dominance beginning to emerge in the early 2000s.[86] A 2004 report for the European Commission based on a three year study of private law enforcement concluded: "There is a danger that some aspects of internet self-regulation fail to conform to accepted standards. We recommend co-regulatory audit as the best balance of fundamental rights and responsive regulation"[87]. Latzer et al. provided excellent analysis of the types of co-regulation beginning to develop, and their institutional path dependency.[88] Self Regulatory Organisations (SROs) generally form as single issue bodies, often crisis-driven, but then develop according to their institutional environment, for instance broadcast content self-regulation bodies can develop into video games, video film, or Internet content self-regulation. They note that there are different economic as well as political incentives for self-regulation, and analysis is needed with attention to the loss of constitutional guarantees.[89] Transparency and explanation by the SRO is necessary, but small first steps towards greater co-regulation.[90] Digital information policy is critically concerned with relationships between existing government-industry actors and 'prosumer' groups, whose role in production, distribution and consumption is growing rapidly, and whose motivations and activism are often non-monetary, requiring a more sophisticated interdisciplinary method for assessing contributions, motivations and sustainability of the 'prosumption economy', the growth of the virtual polity and social communities online, and a new prosumer law and policy to govern the regulation of the digital information ecology. This calls for a new form of consumer and citizen protection, which Brown and Marsden termed 'prosumer law'.[91]

[82] 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.

[83] Inter-Institutional Agreement on Better Law-Making, Official Journal of the European Union December 2003, 2003/C 321/01 at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2003:321:0001:0005:EN:PDF See variously COM 2002/704 Towards a reinforced culture of consultation and dialogue - General Principles and minimum standards for consultation of interested parties by the Commission, 11 December.COM(2002) 275 European Governance: Better Lawmaking, 5 June; COM/2002/0278 Action plan: Simplifying and improving the regulatory environment. Later plans contained more detail: see COM(2009) 504 Report From The Commission On Subsidiarity And Proportionality (16th report on Better Lawmaking) at http://ec.europa.eu/governance/better_regulation/documents/com_2009_0504_en.pdf and COM (2005) 97 Better Regulation for Growth and Jobs in the EU.

[84] European Commission (2015) Better regulation for better results – An EU agenda, 19 May, SWD(2015) 110 final.

[85] Tambini, D., Leonardi, D., and Marsden, Christopher T. (2008) Codifying Cyberspace: Communications Self-Regulation in the Age of Internet Convergence, Routledge: London. European Community of Practice Agora on Better Self and Coregulation (2013–17) : https://ec.europa.eu/digital-single-market/en/newsroom-agenda/event/cop-better-self-and-coregulation.

[86] Price, M. and Verhulst, S. (2004) Self Regulation And The Internet, Kluwer Law International.

[87] Directorate-General for Communications Networks, Content and Technology (European Commission , Programme in Comparative Law and Policy (2004) *Self-Regulation of Digital Media Converging on the Internet: Industry Codes of Conduct in Sectoral Analysis, Final Report of IAPCODE Project for European Commission DG Information Society Safer Internet Action Plan*, 30 April, Section 12.7 at https://publications.europa.eu/en/publication-detail/-/publication/b7c998d9-75d6-464d-9d91-d59aa90a543c/language-en.

[88] Latzer, Michael, Price, Monroe E., Saurwein, Florian, Verhulst, Stefaan G. *Comparative Analysis of International Co- and Self-Regulation in Communications Markets*, Research report commissioned by Ofcom, September, Vienna: ITA (2007) at www.mediachange.ch/media//pdf/publications/latzer_et_al_2007_comparative_analysis.pdf.

[89] See regulators' analyses: Ofcom, *Criteria for promoting effective co and self-regulation: Statement on the criteria to be applied by Ofcom for promoting effective co and self-regulation and establishing coregulatory bodies* (2004) www.ofcom.org.uk/consult/condocs/coreg/promoting_effective_coregulation/co_self_reg.pdf. Office of Regulation Review *A Guide to Regulation, Second Edition,* December 1998 at www.pc.gov.au/orr/reguide2/reguide2.pdf.

[90] Edwards, Lilian and Veale, Michael, Slave to the Algorithm? Why a 'Right to Explanation' is Probably Not the Remedy You are Looking for: Edwards, Lilian and Veale, Michael, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" 16 Duke Law & Technology Review 18 (2017) http://dx.doi.org/10.2139/ssrn.2972855. Erdos, David (2016) European Data Protection Regulation and Online New Media: Mind the Enforcement Gap Journal of Law and Society, Vol. 43, Issue 4, pp. 534-564, http://dx.doi.org/10.1111/jols.12002.

[91] Brown, I. and Marsden, C. *Regulating code: good governance and better regulation in the information age* (Cambridge, MA: MIT Press, 2013) at Chapter 8. See also Jasmontaite, L. "The European Data Protection Supervisor (EDPS) Opinion 4/2015 towards new digital ethics". European Data Protection Law Review (EDPL) 2(1), (2016) 93–96 at 95.

Helberger et al. have called this 'networked consumer' law.[92] The European Commission has used the overarching objective of "a fair deal for consumers" online.[93] Commissioner Vestager has explained that Internet social networks are essentially addiction platforms.[94]

Can social media platforms, addictive or not, be left to privately enforce the online regime, or will determined co-regulatory intervention make this happen? Two judgments of the European Court of Human Rights shed light on this. The first was an Estonian reference to the Grand Chamber, *Delfi*,[95] in which a news website was made liable for the comments that were underneath the news article. It was fined for the comments, which led news websites across Europe to think that they would have to either pre-moderate, which would require a great deal of investment, or alternatively remove comments altogether. That case has since been followed by *MTE v. Hungary*, which restored some kind of balance, and came to a different conclusion on the facts, deciding that pre-moderation of comments was not required. As a lower chamber decision, it could not overturn *Delfi*.[96] If the law requires prior approval of comments, whether it be on Twitter, a news website or wherever else, that requires a great deal more investment, and websites may well choose to exclude all comments.[97] Revisiting the protections from liability for hosting third party content, as suggested by the new European Commission in 2019 inspired in part by disinformation threats,[98] may cause the entire co-regulatory structure to partially unravel.

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

4.2. *Options for regulating AI in disinformation explained*

Six options are provided for technical means to moderate and remove disinformation, ranging from Option 0 (no new regulation but, further research and analysis into current self- and state regulation) to Option 5 (specific legislative instruments):

- **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.
- **Option 1**: Non-audited self-regulation, with increasing industry-government coordination, but no sanction on those companies choosing not to cooperate in standards.[99]
- **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.[100]
- **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.

Note that the options are interdependent – where regulation is proposed, it sits atop a pyramid of activities including co-regulation, self-regulation, technical standards and individual company initiatives. There is no single option to solve the problem of disinformation. Given what AI use and abuse reveals about disinformation practices, potential actions are summarised in Table 1 below.

Option 0: status quo

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) .[101] 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

[92] Helberger, Natali, Borgesius, Frederik J. and Reyna, Agustin "The Perfect Match? A Closer Look at the Relationship between EU Consumer Law and Data Protection Law" Common Market Law Review, Vol. 54, No. 5 (2017).

[93] Vestager, M. (2018) "Competition and a fair deal for consumers online", Netherlands Authority for Consumers and Markets Fifth Anniversary Conference, The Hague, 26 April https://ec.europa.eu/commission/commissioners/2014-2019/vestager/announcements/competition-and-fair-deal-consumers-online_en.

[94] See https://www.b.dk/globalt/eu-commissioner-margrethe-vestager-facebook-is-designed-to-create-addiction-like.

[95] Delfi AS v Estonia [GC], 64569/09 ECHR [2015].

[96] MTE v Hungary ECHR 22947/13. For case comment, see Bjarnadóttir, María Rún (2017) Case Law, Strasbourg: Einarsson v Iceland, Defamation on social media and Article 8, Inforrm Blog, 14 November, at https://inforrm.org/2017/11/14/case-law-strasbourg-einarsson-v-iceland-defamation-on-social-media-and-article-8-maria-run-bjarnadottir/.

[97] Kerr, A. & Musiani, F. & Pohle, J. "Communication and internet policy: a critical rights-based history and future" Internet Policy Review, 8(1) (2019) doi: 10.14763/2019.1.1395.

[98] von der Leyen, Ursula *A Union that strives for more: My agenda for Europe*, p.13, at https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf.

[99] Marsden, C. Internet Co-regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace Cambridge University Press (2011), pp.107–113.

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

[101] Netzwerkdurchsetzungsgesetz (German Network Enforcement Act) 2017, see 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.

Table 1 – 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 General Data Protection Regulation, AVMS Directive, and the proposed revised ePrivacy Regulation, 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
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

2018 referendum on the Thirty-sixth Amendment of the Constitution Act in Ireland, whereas Facebook only banned foreign actors' adverts). 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,[102] but is clearly a case of corporate social responsibility over-reach.[103] However, much can be achieved using non-traditional regulatory tools to control AI use. A highly influential Shorenstein centre report for the Council of Europe, outlines responses by platforms, news providers and governments identified separately.[104]

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 article 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. An unregulated online free-for-all is unappealing to European policy makers.[105]

Much detailed internet 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 Article 10 Paragraph 2 of the European Convention on Human Rights, permitting states to intervene to protect rights.[106] Despite US claims of the exceptionalism of free speech, Option Zero is not a realistic option for European legislators.

This article therefore argues that the initiatives identified by Wardle et al. should be targeted and encouraged by European institutions, in the interests of a better approach to tackling disinformation.[107] It also proposes additional regulatory measures explained in Options 1–5 below. 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.

[102] 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/.

[103] 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.

[104] 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.

[105] 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 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.

[106] Most recently, see Keller, Daphne, *Dolphins in the Net: Internet Content Filters and the Advocate General's Glawischnig-Piesczek v. Facebook Ireland Opinion*, Stanford Center for Internet and Society, September 4, 2019.

[107] Wardle et al (2017) n. 104.

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 co-operate. 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.[108] 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.[109] In the AI space, standards for ethical algorithms are being developed by for instance the IEEE P7000 scheme,[110] but critics have pointed out that these ethical norms are the predecessor to legal standards.[111] 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.

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.[112] Members of the EU High Level Expert Group on Disinformation 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".[113]

In the AI disinformation scheme, this audit could be undertaken by the industry body, 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".[114] 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.[115] 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

[108] 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.*

[109] Marsden, C. (2011) n.79, pp. 107–113.

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

[111] @****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.

[112] UK Safer Internet Centre (2018).

[113] 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.

[114] 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 Humboldt University: see https://networkofcenters.net/centers.

[115] 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.'

will document progress made and expose anyone not taking their responsibilities seriously".[116]

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 multi-stakeholder 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,[117] a legal right which is subsequently subject to regulatory and court enforcement and was thus not an example of audited self-regulation. The Global Network Initiative claims such an audit function, but annual reports do not give detail such that it would satisfy these criteria.[118]

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.[119] App store games are regulated using the International Age Rating Coalition system.[120] 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".[121]

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, the UN Special Rapporteur on Freedom of Opinion and Expression has recommended such a body to deal with online content moderation. 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 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

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

[116] Jiménez Cruz, C., Mantzarlis, A., Nielsen, R.K., and Wardle, C. (12 March 2018), n.113.

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

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

[119] *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.

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

[121] Marsden, C. (2011) n.79, p187 and PEGI (2018) *PEGI Age Ratings*, https://pegi.info/page/pegi-age-ratings.

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.*[122]

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 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 argued that: "Any measure to tackle the complex topic of online disinformation must not be blindly reliant on automated means, AI or similar emerging technologies without ensuring that the design, development and deployment of such technologies are individual-centric and respect human rights"'.[123]

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, network neutrality, shows that such feasibility challenges can be overcome with appropriate multistakeholder engagement.[124]

Sustainability of co-regulation is 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.[125] 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'[126] and has suggested this option. Hearings are ongoing on the role of the UK Information Commissioner and communications regulator Ofcom in such a scheme.[127] Each national context will differ, but in general a regulator would encompass: reformed, strengthened powers for the: electoral commission, data protection authority, advertising regulator, and communications regulator (broadcast, newspaper); police enforcement of criminal law regarding fraud (bot accounts) and other malicious (illegal) communications. It is unclear what such a regulator could achieve without invoking direct censorship of non-conforming organisations.[128]

AI systems may be forced to conform to a mandatory national or regional standard, which could lead to dominant

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

[123] 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.

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

[125] See European Commission, Proposal for a Council Directive amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services in view of changing market realities, COM(2016) 287 final.

[126] 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.

[127] 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.

[128] Marsden, C. (2018) "Prosumer Law and Network Platform Regulation: The Long View Towards Creating Offdata", 2 Georgetown Tech. L.R. 2, pp.376–398 at 387.

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.[129]

A merger of many regulators is not necessary to combine the functions via coordinated federated networks of those regulators. The UK Information Commissioner 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.[130] The Digital Rights Clearinghouse set up by the EU Data Protection Supervisor with data protection, consumer protection and competition authorities is another example.[131]

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.

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 parliamentary elections are the highest political priority, regulatory Option 5 would specifically ensure electoral online advertising is regulated online, as it currently is offline. However, that is not a proposal for any kind of super-regulator. Overall, we argue legislation may be premature and potentially hazardous for freedom of expression: co-regulation 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, we argue 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.

We advise against regulatory action that would encourage increased use of AI for content moderation purposes, without strong human review and appeal processes. There is scope for standardising (the basics of) notice and appeal procedures and reporting, and creating a self-regulatory multi-stakeholder body, such as the UN Special Rapporteur's suggested social media council.[132] As recommended by the Special Rapporteur, this multi-stakeholder 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.

This article 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. 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. We reiterate that far greater transparency must be introduced into the variety of AI and disinformation reduction techniques used by online platforms and content providers.[133]

4.4. *Summarizing the cost-benefit of disinformation regulation*

Fighting disinformation does have a cost. Unless European citizens are engaged to work independently on behalf of platform companies – this will be unpopular because this is expensive – policy cannot solve this problem in Europe. What European institutions, whether as a bloc or among its constituent national governments, need to do is to make sure that

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

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

[131] See EDPS (2019) *Big Data & Digital Clearinghouse*, at https://edps.europa.eu/data-protection/our-work/subjects/big-data-digital-clearinghouse_en On why regulatory intervention is needed to ensure legitimacy in this area, see European Data Protection Supervisor Opinion on online manipulation and personal data 3/2018 p.20; Article 29 Working Party Opinion: Guidelines on Consent under Regulation 2016/679 p.19.

[132] UN Special Rapporteur (2018) n.13, paragraphs 58, 59, 63, 72.

[133] See further Marsden, C. and R. Nicholls "Interoperability: A solution to regulating AI and social media platforms" Computers and Law (2019), at https://www.scl.org/articles/10662-interoperability-a-solution-to-regulating-ai-and-social-media-platforms.

what companies do is engage European fact-checkers to work with their AI programmes to properly resource their own attempts to stop 'fake news'. They also need European lawyers to work on appeals. Executives in California, ex-politicians such as Nick Clegg, or thousands of badly-paid contractors hired off the internet, from the Philippines or India, cannot regulate European fake news: it has to be Europeans. They must have training in journalism and European human rights law to make judgements on journalistic opinion and freedom of expression. That such a proposal appears highly optimistic is a sign of little regard platforms have thus far been required to show for European human rights standards.

While it would appear to be in the platform owners' best interests to reduce the dissemination of disinformation, the means of doing so could prove to be a sticking point. As ever, it comes down to a question of money. The platforms claim results from AI, not least because it is much cheaper than employing enough humans to solve the problem. The accurate way to deal with fake news is to have a hybrid model of trained humans working on problems that AI has identified. Humans have to make the value judgements. That is expensive for Facebook and YouTube, but absolutely essential to accuracy. They will only make those investments in qualified European values, fact-checkers and 'fake news' spotters if co-regulation is introduced, if they are forced to do so by governments.

Does the evidence support any further legal intervention to control disinformation? First, note that the evidence base is growing rapidly in 2019, and there is strong recent evidence that electoral outcomes have been affected by online disinformation. Second, the UK's Information Commissioner is engaged in auditing the activities of the Brexit campaigners in the 2016 UK referendum, having issued £120,000 fines on 1 February 2019 for three separate illegal uses of personal data.[134] Third, there remain significant questions about Online Behavioural Advertising (OBA), in electoral periods, as a campaign tool more widely, and as an effective and appropriate use of personal information more broadly under the General Data Protection Regulation (GDPR) .[135] We have just begun the investigation of regulating online disinformation and its uses in our democracies.

5. Conclusion: whom to regulate, why and how?

We conceptualised the value dimensions as the protection of representative and electoral democracy. We acknowledged that the size of the economic actors involved means that economic value creation is affected by their regulation, though issues concerning democratic and social values are paramount. Public choice theory theorizes that politicians will mundanely pursue their self-interested course, and the conduct of elections is their primary concern. Given that distinction between electoral regulation and all other forms of public policy, it is unsurprising that electoral reform is central to political concerns. We caution that elections are conducted in a multimedia environment that varies by nation, and that is converging on digital media, but that existing forms of media still predominate. Thus Internet browsing on smartphones even on social media platforms involves largely consumption of content created by existing media organizations that predate the Internet, whether that be television or radio news clips or online versions of newspaper articles.

When the political settlement of these older media was made in the period prior to the 1990s, the political concern with forms of representative democracy resulted in a regulatory settlement that placed political concerns alongside economic concerns. For instance, legislation introduced bans on political advertising on the dominant European forms of social media, broadcasting. The exception to these bans was the United States, where political ownership of local television and radio stations led to a very different policy outcome. The emergence of US-dominated social media platforms has led to a largely unquestioned adherence to the United States model of permitting political advertising as a form of free expression. This has only been effectively challenged in Canada (2019) and Ireland (2018), where the requirements for transparency and a ban on overseas donations to political campaigns led social media platforms to ban all political advertising on these media.

A second set of important questions concerns what kind of institutions and regulatory tools can identify, protect and uphold the policy values in electoral regulation Processes and mechanisms to restore democratic values and social justice and infuse them into digital platforms included transparency, media literacy, and the introduction of forms of human-centred co-regulation. The rents captured in the Internet advertising economic value chain were acknowledged, and regulation for the elimination of those rents has been proposed, by for instance requiring social media companies to redistribute revenue to media organisations, and to donate substantially to fact-checking and other forms of disinformation awareness campaigning. To "follow the value" in this case was the clearly preferred option, in this case to focus on the social media platforms themselves. To regulate access to that value in a way that aligns the incentives of economic operators with those of society was the explicit goal. To refrain from interfering with the economic value process was far less considered given the primary importance to politicians of preventing interference with democratic processes. The alternative for welfare improvement across the value chain would be the wholesale importation of United States 'richest takes all' political campaigning, a policy most vociferously opposed by one of the 'fathers of Internet regulation', Lawrence Lessig.[136] To think

[134] ICO *ICO to Audit Data Protection Practices at Leave.EU and Eldon Insurance after Fining Both Companies for Unlawful Marketing Messages*, (1 Feb 2019) https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2019/02/ico-to-audit-data-protection-practices-at-leaveeu-and-eldon-insurance-after-fining-both-companies-for-unlawful-marketing-messages.

[135] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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) OJ L 119, 4.5.2016, p.1–88 ELI: http://data.europa.eu/eli/reg/2016/679/2016-05-04.

[136] Lessig, L. Republic, *Lost: How Money Corrupts Congress—and a Plan to Stop It* (Twelve Publishing, 2011) ISBN 978-0-446-57643-7.

more clearly about what constitutes value creation and value extraction in this policy arena, in order not to recommend regulation that creates equal or greater economic rents, is to argue for the reform, abolition or illegality of 'recommender systems' (targeted advertising using personal data online) ,[137] or even for the abandonment of the capitalist model of digital information creation.[138] While that ambition lies outside the scope of this article, it is notable that several prominent experts now suggest that is the direction in which future policy should be oriented.

Finally, to what entities do we apply rules based on specific values? Who are the recipients of the regulation aimed at fostering the value(s) we chose and protecting the value we create? The answer is once again individuals, but also the social media platforms, and the electoral system itself. It is the uses and abuses of existing rules for elections and social media which have combined to produce a toxic disinformation environment online. In this respect, we note that social media regulation is an ongoing process that has built on earlier instances of Internet regulation, and call for more study of the history of Internet law. Phenomena such as Distributed Ledger Technology (using so-called 'blockchain'[139]), AI and disinformation can be regulated using many of the co-regulatory lessons learnt from Internet regulatory history, and such history should be researched, broadcast and applied.[140]

We also urge historical context: disinformation is as old as the written word, as explained in Section 2. It cannot be "solved", but its worst effects can be somewhat ameliorated using those policy options outlined in Section 4 and summarized in Section 5. As with so many technological regulatory problems, from railways to nuclear power to the Internet to AI, the lessons of regulatory history are important to adapting existing, and deploying new, regulation for new technology.[141] The complex socio-economic deployment of innovations is what creates regulatory issues, not the technology itself.[142] Elections have a long history, and fake news has played a role in outcomes. The regulation we apply to social media disinformation is a further layer to place over the existing layers of media and election regulation, a further bandage over a gaping wound in imperfect democratic processes. Recognizing the added complexity of local content on digital media, and more internationally sourced disinformation, adds a new and disparate element to the regulation of representative democracy.

Declaration of Competing Interests

There are no conflicts of interest associated with this submission.

This is the first of several book-length arguments about financial corruption of political processes on which Lessig has focussed since his research focus shifted from Internet policy.

[137] Briant, Emma L 'LeaveEU: Dark Money, Dark Ads and Data Crimes' (2019) in Paul Baines; Nancy Snow & Nicholas O'Shaughnessy (Eds) *Sage Handbook of Propaganda*, Sage: London; Cobbe, Jennifer and Singh, Jatinder, *Regulating Recommending: Motivations, Considerations, and Principles* (April 15, 2019). Available at http://dx.doi.org/10.2139/ssrn.3371830.

[138] Moglen, Eben, *The dotCommunist Manifesto* (January 2003) at http://moglen.law.columbia.edu/publications/dcm.html.

[139] Guadamuz, A. and Marsden, C. 'Blockchains and Bitcoin: regulatory responses to cryptocurrencies' First Monday, 20(12) (2015) ISSN 1396-0466.

[140] See K. Werbach ed. *After the Digital Tornado: Networks, Algorithms, Humanity*, (Cambridge University Press 2020, in press).

[141] Marsden, C. (2018) "Prosumer Law and Network Platform Regulation: The Long View Towards Creating Offdata" 2 Georgetown Tech. L.R. 2, pp.376–398 at p380.

[142] Guadamuz and Marsden (2015) supra n.139; Marsden C. (2017) *Net neutrality*, n.124 at Chapter 8.