Manuscript

State Sovereignty, Disinformation and Speech Governance

Julia Hörnle, Queen Mary University of London

1. Introduction

A powerful example of the ominous connection between disinformation and the dangers of unchecked social media companies is the recent decision by *Meta* to withdraw the system of fact-checkers from Facebook and Instagram in the US, and replace the fact-checking with community notes à la X, currying political favours[1] with the incoming Trump administration.[2] A political alliance between an incoming government and media organisations is not likely to improve free speech in respect of criticizing the new government, and it is undermining the role of media in holding those in power to account. The apparent reason given for this move is greater freedom of speech and removing subjectivity from fact-checking by news organisations.

At least from a European perspective, this argument is paradoxical: how can there be free speech if we do not ensure that political speech is based on factual truth? In other words, in a rational world, free speech without factual truth is useless. One important component of the right to free expression under a European conception is the right to receive information[3] and disinformation threaten this part of the equation. Of course, not all speech is grounded in facts, but where speech in news information sources[4] is grounded in facts, should we not ensure that the facts themselves are checked? Disinformation in Europe is regarded as a fundamental threat to freedom of expression as it endangers the information ecosystem. It is therefore seen as an existential threat to the liberal, democratic ordering of the state.[5]

The significance of fact-checking has to be put in the context of the ongoing ideology war in the US around free speech, disinformation and the increasing gulf between liberal Democrat and right-wing Republican world view and the corresponding perception of reality. This ideology war instrumentalizes free speech in the fight for power, starting with the so-called

---

[1]According to news reports, Meta has donated $1 million to Trump's inauguration fund and replaced the President of Global Affairs Nick Clegg (a former British Liberal Democrat politician) with the Republican Joel Kaplan (a former White House Deputy Chief of Staff for Policy under George W Bush). Zuckerberg might hope that in return Trump is able to curb social media regulation in Brazil and Europe, for example in the context of trade negotiations.

[2] BBC News (7. January 2025) https://www.bbc.co.uk/news/articles/cly74mpy8klo; New York Times (7. January 2025) https://www.nytimes.com/live/2025/01/07/business/meta-fact-checking

[3] Art 10 ECHR

[4]According to Ofcom's research in the UK, just over half (52%) of UK adults use social media for obtaining news. Younger generations have turned away from traditional news sources- the 16-24 year old cohort use social media as a main source for news (82%). However, the detailed picture is more complicated, 70% of adults (but only 49% of 16-24 year olds) continue to rely on TV news, and traditional media have scored high for trust, accuracy and impartiality- see Ofcom, News Consumption in the UK (10. September 2024) https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/tv-radio-and-on-demand-research/tv-research/news/news-consumption-2024/news-consumption-in-the-uk-2024-report.pdf?v=379621

[5] E Brogi, G De Gregorio "From the code of practice to the code of conduct? Navigating the future challenges of disinformation regulation" (2024) 16 (1) *Journal of Media Law* 38-46, 40-41; see also the EU Report of the independent High level Group on fake news and online disinformation (2018) https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271

"Twitter files" of 2022[6], allegations by the right-wing movement that Twitter, before the *Musk* take-over, had been biased in their content moderation towards left-wing views. International journalists such as the Philippine Nobel Peace Prize winner Maria Resser who has risked her life in uncovering government wrongdoing in her country and who is a staunch supporter of free speech has stated in response to *Zuckerberg's* announcement that "extremely dangerous times" lied ahead for journalism, democracy and social media users and that a "world without facts" was a "world that's right for a dictator".[7] These comments emphasize the need for (social) media organisation's independence from government and the need for news being based on facts.

This Article examines the connections between disinformation, media regulation, the power of social media, sovereignty and jurisdiction. It proceeds by first presenting two brief case-studies on how disinformation can destabilize democracy, followed by a brief, comparative analysis of the regulation of disinformation in the EU and the US and linking this to questions of jurisdiction and sovereignty. It argues that the concept of sovereignty should ground a principle under international law that states should be able to fight disinformation on big tech platforms as an assertion of their right to govern.

2. Disinformation case-studies

2.1 What is disinformation?

Academic literature usually makes a distinction between misinformation and disinformation. Disinformation is the intentional spreading of factually incorrect, false news with the intention to mislead the recipients. By contrast, misinformation is false news which was not intentionally spread as such, and the falsity may be due to a honest, but mistaken belief.[8] For the purposes of this article the distinction is not important and I use the term disinformation to mean information presented as news based on false facts and misleading in its message.

2.2 Two case-studies

The Article uses two illustrative disinformation case-studies in the context of the summer 2024 riots in the UK[9] and the US 2024 elections.

In July 2024 three little girls were brutally murdered and several others seriously injured by a loner in a mass stabbing incident during a dance class. Immediately after the attack a number of social media posts falsely claimed that the attacker was a Muslim asylum seeker and illegal migrant.[10] This disinformation sparked a series of violent, far-right wing, anti-immigration riots across the UK targeting mosques, shops and immigration hostels with

---

[6] TRT- The Newsmakers (14. December 2022) https://www.youtube.com/watch?v=sEpIPLzRkZE

[7] The Guardian (8. January 2025) https://www.theguardian.com/world/2025/jan/08/facebook-end-factchecking-nobel-peace-prize-winner-maria-ressa, Rappler (9. January 2025) https://www.rappler.com/technology/maria-ressa-profit-over-safety-meta-ends-fact-check-program-united-states/

[8] C Tan The Curious Case of Regulating False News on Google (2022) 46 *Computer Law and Security Review* 1-14, 2; European Commission, Code of Practice on Disinformation (European Commission, September 2022); N Bontridder, Y Poullet "The role of artificial intelligence in disinformation" (2021) *Data & Policy doi:10.1017/dap.2021.20* e32-2

[9] The Guardian (22. October 2024) https://www.theguardian.com/media/2024/oct/22/social-media-algorithms-must-be-adjusted-to-prevent-misinformation-ofcom

[10] BBC Verify (25. October 2024) https://www.bbc.co.uk/news/articles/c99v90813j5o

extensive violence, arson and injuries to police officers for several days. The riots led Elon Musk to further stoke the unrests by his claim that "civil war in the UK is now inevitable".[11] The media regulator in the UK, Ofcom later published a letter stating that there was a clear link between social media disinformation claiming that the stabbings were committed by an asylum seeker and the violent riots which followed.[12]

Disinformation played a role in the outcome of the 2024 US Election[13]. While disinformation may or may not be overwhelming in terms of quantity of information out there on social media it has influenced political opinions and voting behaviour.[14]

Generative AI and deepfakes have increased the risks in respect of disinformation during elections in every country.[15] Generative AI allows for the speedy and easy generation of inauthentic information, as it has the tendency to include false information and/or generate hallucinated facts.[16] Authentic looking deepfake videos are persuasive in the sense of the old adage that "an image speaks a thousand words" and the same applies to AI's ability to realistically imitate the human voice. A deepfake has been legally defined in the EU's AI Act as "AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful".[17]

The use of AI to imitate *President Biden's* voice in a misleading robocall made to many voters in the New Hampshire primaries exhorting them not to vote is an example of a mysterious deepfake campaign designed to influence voters through false information.[18]

Famous unsubstantiated stories circulated during the US 2024 election were for example: one (repeated by *Trump*) about Haitian immigrants stealing and eating pets[19], another the misspending by the Federal Emergency Management Agency of hurricane disaster relief funds on "taking in illegal immigrants"[20], or another a photo that *Kamala Harris* once embraced convicted sex offender Jeffrey Epstein in beach wear, which was posted on "X"

---

[11] The Telegraph (4. August 2024) https://www.telegraph.co.uk/news/2024/08/04/southport-latest-news-rioting-disorder-arrests-liverpool/

[12] https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/public-correspondence/2024/letter-from-dame-melanie-dawes-to-the-secretary-of-state-22-october-2024.pdf?v=383693

[13] I'm not claiming that disinformation was decisive for the outcome or that disinformation was only perpetrated by one group- all I'm saying here is that it happened.

[14] Elaine C Kamarck, Darrell M West *Lies That Kill- A Citizen's Guide to Disinformation* (Brookings Institution Press 2024) 19-20

[15] N Bontridder, Y Poullet "The role of artificial intelligence in disinformation" (2021) Data & Policy doi:10.1017/dap.2021.20 e32-3; Brookings "How Disinformation Defined the 2024 Election Narrative" (7. November 2024) https://www.brookings.edu/articles/how-disinformation-defined-the-2024-election-narrative/

[16] Seb Butcher "Disinformation, Generative AI and Why Our Laws Need Urgent Reform" (2024) 30 (4) *Computer and Telecommunications Law Review* 85-85, 85

[17] Regulation (EU) 2024/1689 of 13. June 2024, Art 3 (60)

[18] BBC News (22. January 2023) https://www.bbc.co.uk/news/world-us-canada-68064247

[19] The Guardian (14. September 2024) https://www.theguardian.com/us-news/2024/sep/14/racist-history-trump-pet-eating-immigrant

[20] NBC News (4. October 2024) https://www.nbcnews.com/politics/donald-trump/false-claims-fema-disaster-funds-migrants-pushed-trump-rcna173955

shortly after she had announced that she would campaign for the Presidency[21]. She has also been falsely and repeatedly maligned as a "prostitute" or as incompetent for the role.[22]

Examples of Russian false information propaganda have also been uncovered[23], such as the entirely baseless allegations that the Minnesota Governor and vice-presidential candidate in the 2024 elections *Tim Walz* sexually assaulted one of his students, which have been tracked to a Russian disinformation campaign utilising deepfake whistleblower videos.[24] These stories went viral on social media and some of them were repeatedly mentioned or posted by politicians increasing their reach and impact.

It is incidents like the examples mentioned which risk that we are increasingly living in a post-truth society, where truth is replaced by the power of big tech, weakening elections and other democratic processes, and thereby the security of the democratic state, which is precisely what tech disrupters are trying to achieve.

2.3 Targeting of content, manipulation, disinformation

The starting point of this article are the power struggles between social media platforms manipulating citizens globally, by using complex algorithms to target content to citizens, and amplifying that content with a virality never seen before.[25] This mechanism-perfected to maintain engagement and maximise advertising revenue[26] confers enormous influence and thus power to social media companies. There are two fundamental problems with the virality of social media: first that disinformation frequently has greater reach and velocity[27] and secondly that content targeted for maximum engagement frequently appeals to the subliminal part of the human psyche and emotions, which encourages irrationality and thereby, ultimately popular politics. Algorithm based targeting of content based on behavioural profiling manipulates users.

The targeting of users based on psychological profiles and correlations as to their behaviour has come under much criticism in respect of its negative impacts on society, including privacy infringements, surveillance and lack of autonomy. It is for this reason that the EU Digital Services Act mandates that very large social media platforms and very large search engines have to give users the option to use their recommender systems without profiling,

---

[21] Snopes.com (fact-checking site) https://www.snopes.com/fact-check/kamala-harris-jeffrey-epstein-beach-photo/

[22] DW Factcheck (11. July 2024) https://www.dw.com/en/fact-check-what-role-did-disinformation-play-in-the-us-election/a-70729575

[23] Similar disinformation campaigns are now targeted at German politicians in the February 2025 elections, EDMO (27. January 2025) "Influence Operation Exposed: How Russia Meddles in Germany's election campaign" https://edmo.eu/publications/influence-operation-exposed-how-russia-meddles-in-germanys-election-campaign/

[24] Wired (21. October 2024)

[25] N Bontridder, Y Poullet "The role of artificial intelligence in disinformation" (2021) *Data & Policy doi:10.1017/dap.2021.20* e32-3

[26] Shoshana Zuboff Surveillance Capitalism (Profile Books 2019)

[27] There are a number of studies which show that disinformation travels faster and leads to more engagement, see for example S Vosoughi, D Roy, S Aral "The Spread of True and False News Online" (2020) 359 *Science* 1146-1151; American Psychology Association Consensus Statement "Using Psychological Science to Understand and Fight Health Misinformation" Report 23. November 2023, explaining why we spread disinformation and funded by the Centers for Disease Control and Prevention of the U.S. Department of Health and Human Services https://www.apa.org/pubs/reports/misinformation-consensus-statement.pdf

for example a paid subscription option.[28] Moreover it obliges all service providers to be transparent about their recommender systems and the parameters used.[29] In the US too there is evidence that the law on the divesture of TikTok was motivated by concerns how a foreign adversary (China) would manipulate the 170 million Tik Tok users in the US through recommender systems targeting content to them.[30]

Disinformation amplified by powerful algorithms has divided societies and challenges rational decision-making and free elections.[31] This state of affairs implies an urgent need to rethink media regulation in order to safeguard the sovereignty of democratic states.

3. Content Regulation and the Law on Disinformation

This paper compares the interface between platform regulation and national laws (comparing the EU and US) on disinformation.

3.1 US Law, Content Regulation, Disinformation

In the US tradition, the Constitution largely prohibits the regulation of content (with exceptions) by the state.

Content decisions have been left to private media, including online platforms. Private media are in charge of setting their own content policies and regulating speech through content moderation according to standards described in terms and conditions. Online services are not liable for restricting access to "obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable information" provided by another person, if the provider decides to moderate content on their own initiative (as opposed to being compelled by the law).[32]

Freedom of speech is guaranteed through competition of many different media outlets between which users can *choose according to their own political convictions*. The essence of free speech has been described in *Tik Tok v Garland* as "the principle that each person should decide for himself or herself the ideas and beliefs deserving of expression, consideration, and adherence."[33]

Harmful speech is reined in through counter-speech, and the notion of the free marketplace of ideas. The theory here is that many viewpoints are juxtaposed and debated, so that ultimately reason and common sense will prevail.[34] Therefore, speech restriction by government is only allowed, if it is absolutely necessary to prevent a serious harm which

---

[28] Art 38, Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services

[29] Arts 27, 40 (3): "explain the design, the logic, the functioning and the testing of their algorithmic systems, including their recommender systems" to the Commission/Digital Services Co-ordinators of the Member States (very large online services only).

[30] *TikTok Inc. v. Garland* Supreme Court of the United States 17. January 2025, p.9

[31] Tim Clement-Jones *Living With the Algorithm- Servant or Master?* (Unicorn Publishing Group 2024) 35-46

[32] 47 US Code §230 (c) (2) (A)

[33] Quoting *Turner Broadcasting System, Inc. v. FCC*, 512 U.S. 622, 641, 114 S.Ct. 2445, 129 L.Ed.2d 497 (1994) on p.5

[34] But it is precisely this mechanism which the behavioural profiling and the individualised micro-targeting of content undermines through echo chambers, and why the "marketplace of ideas" no longer functions see x-ref

cannot be prevented in any other way ("strict scrutiny")[35]. Under the strict scrutiny test, the government must show that the law is narrow tailored, that it advances a compelling governmental interest and that it is the least restrictive means of implementing this interest, *i.e.* it is necessary in a strict sense.[36] This necessity test is almost impossible to prove.[37] Thus, strict scrutiny means that most harmful speech (such as hate speech) cannot be criminalised by federal and state criminal law. But even criminal speech which is prohibited in conformity with the 1st Amendment (such as explicit and realistic child pornography[38]), does not trigger any liability for social media providers hosting this speech. The person who is responsible for the speech (the speaker, the editor) may be criminally liable, but interactive computer services have *absolute immunity from civil liability and state criminal laws* under Section 230 Communications Decency Act 1996[39] and need not take any measures (such as blocking or filtering of content) to prevent the illegal speech from spreading.[40]

This raises the question whether the algorithms actively driving content to users based on their behavioural profiling defeat section 230 CDA. An argument could be made here that social media platforms are not merely passively hosting user provided content, but actively target users with illegal content or particular items of disinformation based on the users' profile and the algorithms calculating that this specific user is *likely to engage with this particular item of content*. This raises the question whether the service goes beyond just being an interactive computer service.[41] Interactive computer services are "any information service, system, or access software provider that provides or enables computer access by multiple users to a computer server, including specifically a service or system that provides access to the Internet and such systems operated or services offered by libraries or educational institutions". [42] This includes all information and computer access systems, as well as hosting services. But arguably the algorithms targeting content specifically based on user profiling are new and may not have been envisaged in 1996- they are not mere information services and access systems.[43] This issue was relevant in two 2023 US Supreme Court cases, but ultimately has not yet been decided by the US Supreme Court. It was effectively side-stepped in *Twitter v Taamneh*[44] and *Gonzalez v Google*[45] where claims that Twitter or Youtube, respectively, were aiding and abetting terrorist acts through enabling terrorist radicalisation were unsuccessful. The decision was reached on the basis that social media platforms' activities did not amount to aiding and abetting so that the Court did not even reach the question of immunity from criminal liability under section 230 CDA. The minimalist ruling has been criticized on the basis that it is a postponement of this important

---

[35] ACLU v Reno caselaw

[36] *Reed v. Town of Gilbert*, 576 U.S. 155, 163, 135 S.Ct. 2218, 192 L.Ed.2d 236 (2015)

[37] Ref cases

[38] refs

[39] 47 US Code §230 (e) (3)

[40] This immunity, however, does not apply to Federal criminal law, 47 US Code §230 (e) (1). If a particular post implicates Federal criminal law, online service providers could theoretically be liable, if they do not take action on being notified.

[41] Casenote Review "Communications Decency Act of 1996" (2023) 137 *Harvard Law Review* 400-409, 405

[42] 47 US Code §230 (5) (f)

[43] Ref lit and cases

[44] 598 U.S. 471

[45] 598 U.S. 617

question.[46] So while it is clear that online providers are not liable for illegal content, and that content moderation is largely left to the providers themselves, is state legislation which interferes with service providers' content moderation compliant with the 1st Amendment? There are two variants to this question, namely first whether (state) legislation can oblige online service to carry out certain types of content moderation and secondly whether (state) legislation can prohibit service providers from moderating certain types of content on their own private initiative.

As to the first variant, there are a number of current legal challenges concerning state laws *mandating content moderation* in respect of children's online safety[47], i.e. online safety legislation vaguely similar to that found in the EU and UK, if not that restrictive.[48] As to the second variant, legal challenges concerning laws *restricting private platform's content moderation*[49] have reached the US Supreme Court. Broadly generalising the law[50], currently the position seems to be that both laws mandating content moderation and laws restricting content moderation are unconstitutional, unless they meet the strict scrutiny test[51].

3.1.1 Online Safety Laws Mandating Content Moderation

Several states are in the process of enacting online child safety legislation, termed "Kids Code Bills" or "Age-Appropriate Design Codes".[52] California led the way in 2021 with its Age-Appropriate Design Code Act. These Bills have been successfully opposed by big tech companies though lobbying and First Amendment challenges in the courts. Net Choice[53] is bringing constitutional challenges in the courts, which are resulting in preliminary injunctions against the enforcement of these Acts in several states.[54] This litigation is ongoing at the time of writing. While the take-down of disinformation is not related to child safety, the First Amendment challenges, if successful, make it extremely unlikely that states will successfully enact more restrictive online safety legislation, dealing with child safety, let alone disinformation.

3.1.2 Laws Limiting Content Moderation

Moody v NetChoice arose out of out of a partisan Republican allegation that social media companies (Google, Meta, Twitter) had been moderating content in an imbalanced way, biased against views on the right and their deletion of accounts of politicians.[55] In response,

---

[46] Casenote Review "Communications Decency At of 1996" (2023) 137 Harvard Law Review 400-409, 409
[47] *NetChoice v Bonta* 113 F.4 1101 (9th Cir. California August 2024), *NetChoice v Reyes* 2024 WL 4135626 (US District Court D.Utah September 2024; appeal to 10. Cir pending)
[48] Discussed below, ref
[49] *Moody v NetChoice; Paxton v Net Choice* 144 S.Ct. 2383 (US Supreme Court,1. July 2024)
[50] This broad characterization is inaccurate, since the cases are currently making their way through the courts and the assessment depends on the precise provisions in state legislation challenged, but the generalisation helps to conceptualise the broader direction of travel. More detail is provided below.
[51] See FN ref 37
[52] California, Maryland, Vermont, Minnesota, Hawaii, Illinois, New Mexico, Nevada, South Carolina, Utah
[53] Net Choice is the trade association for the big tech companies, including Alphabet/Google, Meta, X, Snap and Pinterest, tasked with minimising regulation of their activities, see https://netchoice.org/about/
[54] *NetChoice v Bonta* 113 F.4 1101 (9th Cir. California August 2024), *NetChoice v Reyes* 2024 WL 4135626 (US District Court D.Utah September 2024; appeal to 10. Cir pending)
[55] BBC News (date) https://www.bbc.co.uk/news/world-us-canada-55597840

Florida and Texas passed state legislation prohibiting the deplatforming of politicians and viewpoint discrimination, thereby restricting platforms' mechanisms of content moderation. NetChoice brought a legal challenge against both laws in the respective US District Courts[56] both of which issued a preliminary injunction on the basis that these laws triggered strict scrutiny under the 1st Amendment which neither law satisfied. These conflicting decisions were eventually joined before the US Supreme Court in *Moody v NetChoice*[57], which unanimously vacated the appeal judgments and remanded the cases back to the lower courts[58], without deciding the issues fully. The US Supreme Court held that social media platforms produced their own compilations of expressions which were protected under the 1st Amendment, so that the prohibition of viewpoint discrimination might engage the *free speech rights of the social media companies* (as opposed to the users of the platforms).[59] Thus, even though in practice the platforms removed very little speech and were not exercising editorial functions as such, the fact that they compiled the expressions of users on their platform through algorithms gave the social media companies 1st Amendment protection.[60] The Supreme Court held that it was not the government's role to balance the free marketplace of ideas.[61] But ultimately it is for the lower courts to examine whether these Laws are in breach of the 1st Amendment[62] and commentators have pointed out that final resolution of these issues might be years away.[63]

These two aspects of the US Supreme Court view, namely that algorithmic content curation through micro-targeting of users and that governments does not interfere with the marketplace of ideas in the name of pluralism and diversification of views is precisely what distinguishes US law from EU law, where it is part of the state's role to ensure pluralism of ideas and where algorithmic targeting is seen as a threat to freedom of expression, and in particular the right to receiving information.[64]

3.1.3 Informal Nudging by the State to Encourage Online Services to Limit the Spread of Disinformation

Another tension between freedom of speech and content moderation was raised before the Supreme Court in 2024 in *Murthy v Missouri*[65]. The allegation was that the Biden administration had directed the social media companies to take down disinformation. In the wake of these allegations, the then Attorney Generals of Missouri and Louisiana and right-wing user groups applied for an injunction mandating the Biden administration not to urge the social media companies to act on disinformation. The injunction was initially granted[66],

---

[56] *Moody v NetChoice* US District Court for the Northern District of Florida 546 F.Supp.3d 1082 and *NetChoice v Paxton* US District Court for the Western District of Texas 573 F.Supp.3d 1092

[57] *Moody v NetChoice* 144 S.Ct. 2383; 219 L.Ed.2d 1075 (1. July 2024)

[58] Justice Kagan gave the Opinion of the Court, with four concurring opinions double-check

[59] At 2393, this was criticised by the concurring judgment of Justice Alito at 2431-2432

[60] At 2402

[61] At 2402-2403

[62] At 2409

[63] E Goldman, "Speech Nirvanas on the Internet: an Analysis of the U.S. Supreme Court's *Moody v NetChoice* Decision" 2024 *Cato Supreme Court Review* 125-155, 126

[64] Cross-ref and compare later discussion

[65] 144 S.Ct. 1972; 219 L.Ed.2d 604

[66] 680 F.Supp.3d 630; US District Court for the Western District of Louisiana

then narrowed on appeal to the Fifth Circuit[67], and finally stayed with an appeal to the US Supreme Court[68]. The US Supreme Court[69] held that the injunction was invalid for the reason that the claimants had no standing under Article III of the US Constitution[70] which only allowed judicial review where this was "necessary to *redress or prevent actual or imminently threatened injury* to persons caused by official violation of law" (my emphasis).[71] Since the claimants relied on *past* allegations of government interference and could not prove a chain of *causation* between the administrations' communications and the take-down or blocking of material there was no *actual or imminent threat* to their freedom of speech justifying the imposition of an injunction.[72] The Court emphasized that content moderation by social media companies was not as such unconstitutional and it found that they moderate content, so it could not be proven that content was removed *because of* state interference.[73]

The Supreme Court held in *Murthy* that a degree of political pressure can be applied by the state, as long as the ultimate responsibility for content moderation is free from actual government interference. The *Murthy* decision is significant in that it disallowed an injunction against Government influencing social media content moderation practices to a degree, including the *flagging of posts constituting disinformation information*.

President Trump in his first days in office has passed an Executive Order "Restoring Freedom of Speech and Ending Federal Censorship"[74]. Already in 2022, he vowed to "shatter the left-wing censorship regime".[75] This Executive Order, in its first part, criticised the Biden administration for trampling on

> "free speech rights by censoring Americans' speech on online platforms, often by exerting substantial coercive pressure on third parties, such as social media companies, to moderate, deplatform, or otherwise suppress speech that the Federal Government did not approve".[76]

While this first part gives context, it does not have retroactive effect, it simply criticises what has happened in the past. In the second part the Executive Order directs the federal (Trump) administration to "ensure that no Federal Government officer, employee, or agent engages in or facilitates any conduct that would unconstitutionally abridge the free speech of any American citizen"[77]. While the political intention of this exhortation is clear, its legal significance is at least doubtful. Either a federal act is an infringement of free speech, then this would be illegal under the 1st Amendment anyway, or it does not engage free speech

---

[67] 83 F.4th 350
[68] 144 S.Ct. 7
[69] Opinion written by Justice Amy Coney Barrett
[70] U.S. Const. art. 3, § 2, cl. 1; Fn 65 at 1985-97
[71] Citing *Clapper v. Amnesty Int'l USA*, 568 U.S. 398, 409, 133 S.Ct. 1138, 185 L.Ed.2d 264
[72] At 1987
[73] Fn 65 at 1992-96
[74] January 20., 2025 see https://www.whitehouse.gov/presidential-actions/2025/01/restoring-freedom-of-speech-and-ending-federal-censorship/
[75] Sara Dorn "Trump Vows To Dismantle 'Censorship Cartel' If He's Re-Elected—An Apparent Nod To Musk's 'Twitter Files' Release" (Forbes 15. December 2022) https://www.forbes.com/sites/saradorn/2022/12/15/trump-vows-to-dismantle-censorship-cartel-if-hes-re-elected-an-apparent-nod-to-musks-twitter-files-release/
[76] Clause 1 Purpose
[77] Clause 2 (b) and 3 (a)

rights, in which case the clause is not applicable. It is difficult to see how this Executive Order adds to 1st Amendment Rights or their interpretation by the courts. More significant are Clauses 2 (c) and 3 (a) which state that no federal resources must be deployed for fighting disinformation online which means that no federal agency is able to use resources and manpower to liaise with social media companies and search engines to carry out that task.[78] Even more concerningly, these Clauses may mean that any third party, independent of the federal government, but supported by federal funding, such as non-profits, universities and colleges may have their funding withdrawn if they engage in activities such as content flagging or fact-checking, or even research which may impact the free flow of disinformation.[79]

Furthermore Clauses 2 (d) and 3 (a) provide that it is the policy of the Trump administration to "identify and take appropriate action to correct past misconduct by the Federal Government related to censorship of protected speech". This raises the question whether officers who were fighting disinformation (for example during the Pandemic) will be dismissed or subjected to some disciplinary action in their employment. Such disciplinary action would be retroactive and raises questions about its legality. The wording is "to correct past misconduct"- this is an impossible task as disinformation repressed in the past cannot be effectively corrected, possibly years later. The correction may be a right of reply of the person whose speech had been suppressed or some duty to correct the record- but how effective this would be is the question- a real Don Quixote remedy. Therefore, ultimately it is not clear what the correction entails. However, for the purposes of remedial action, the Executive Order mandates that the Attorney Generals should investigate and prepare a Report submitted to the President with recommendations for remedial action.[80] The Executive Order makes clear that it does not confer any rights of action to persons who claim that their speech has been curtailed illegally.[81]

3.2 European Laws: Media Regulation in the Service of Freedom of Expression

On the other side of the Atlantic, in democratic European countries, speech regulation is an accepted form of media regulation and involves the careful balancing of freedom of expression with harm stemming from speech. This is reflected in the European Convention of Human Rights, which in Article 10 protects the right to freedom of expression which is conceived as both a right to speak as well as a right to receive information[82], and arguably it is the right to receive truthful information which is engaged by the spreading of disinformation, or the state not preventing the spread of disinformation. But the right to freedom of expression is constrained where this is absolutely necessary to protect the rights

[78] "ensure that no taxpayer resources are used to engage in or facilitate any conduct that would unconstitutionally abridge the free speech of any American citizen."
[79] Sara Dorn "Trump Vows To Dismantle 'Censorship Cartel' If He's Re-Elected—An Apparent Nod To Musk's 'Twitter Files' Release" (Forbes 15. December 2022) https://www.forbes.com/sites/saradorn/2022/12/15/trump-vows-to-dismantle-censorship-cartel-if-hes-re-elected-an-apparent-nod-to-musks-twitter-files-release/
[80] Clause 3 (b)
[81] Clause 4 (c)
[82] Art 10 (1) "Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This Article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises."

of others as prescribed by the law, including laws covering false and harmful speech.[83] European media law, by way of generalisation is characterised by arms-length regulation through independent[84] regulators[85] and ultimately independent courts who carry out this balance. Media regulation in Europe is broadly characterised by the goal to ensure plurality of media voices.[86] Additionally it has been argued that users of media services should be exposed to a number of different voices and viewpoints to ensure media plurality is effective and that this should be one of the goals of regulation.[87] *Koltay* has pointed out that mass media should be a forum enabling public debate, but in practice commercial interests have negatively impacted this function, such that the media "treat the public less as a participant in public debate and more as a revenue-generating audience".[88]

In view of this conception of media law, it is only logical that European laws have instituted regulation of social media services in the shape of the Digital Services Act (EU)[89] and the Online Services Act (UK[90]), which both regulate social media platforms and search engines.

In the EU media regulation is not focused on restricting illegal or harmful content, but rather on the dynamics of disinformation, including the algorithms which amplify and target content and the responsibility of gatekeepers.[91]

3.2.1 The EU Digital Services Act

The Digital Services Act[92] imposes obligations on social media companies to moderate content based on the notification of illegal content (including by trusted flaggers such as the police/state authorities or victim organizations whose notifications should be prioritised[93]). Furthermore, online service providers have to restrict illegal content and notify illegal content, which they are aware of, to the (police) authorities of the relevant Member State. Online service providers have to operate systems which allow users to notify illegal content, and instituting action to prevent continued access ("notice and take-down", "notice and

---

[83] Art 10 (2) "The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary."

[84] Independent of the government, albeit that practice may not always follow the principle (which is another discussion for another article)

[85] This is also reflected in the Digital Services Act, Art 52 (1) and (2): "When carrying out their tasks and exercising their powers in accordance with this Regulation, the Digital Services Coordinators shall act with complete independence. They shall remain free from any external influence, whether direct or indirect, and shall neither seek nor take instructions from any other public authority or any private party."

[86] A Koltay *Media Freedom and the Law* (Routledge 2025) 141, 158, 161

[87] P M Napoli "Exposure Diversity Reconsidered" (2011) 1 *Journal of Information Policy* 246-259

[88] FN 86, 162

[89] In force since 17. February 2024

[90] The United Kingdom left the European Union after a referendum in 2016 on 31. January 2020. The referendum campaign to leave was influenced by serious misinformation, such as the famous campaign bus, which claimed that the state would save £350 million which could go to the National Health Service, BBC News (16. January 2018) https://www.bbc.co.uk/news/uk-42698981

[91] E Brogi, G De Gregorio "From the code of practice to the code of conduct? Navigating the future challenges of disinformation regulation" (2024) 16 (1) *Journal of Media Law* 38-46, 39

[92] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services; came into force on 17. February 2024, Art 93

[93] DSA Art 22

action").[94] Service providers[95] have to provide complaints systems, allowing both, for complaints that content has been unfairly restricted or, the reverse, where a notification of illegal content has been ignored.[96] Complainants must be provided with independent Alternative Dispute Resolution, if their complaint has not been solved to their satisfaction.[97] Complainants also have a right to complain to their national regulatory authority.[98] There are also obligations to prepare Transparency Reports about content moderation.[99] Very large online services[100] (both social media platforms and search engines) have to comply with additional obligations[101]. These companies have to carry out systemic risk assessments (including risks stemming from algorithms) as to harms arising from illegal content or illegal activities on their services[102] and mitigate these risks proactively.[103] The risk assessments must include the risks for "inauthentic use", ie the spreading of disinformation.[104] The risk mitigation requirements are extensive and include a whole range of measures in Art. 35 (1), including adapting the design and functioning of the service, adapting and increasing content moderation, adapting and testing algorithms, including content recommender systems, adapting advertising systems, demonetizing content, awareness raising measures such as explaining why content is being targeted or fact-checks in respect of false information, measures protecting children such as age-verification, parental controls or support measures (for example for mental health problems or eating disorders).[105] Deepfakes must be prominently and clearly marked as false or inauthentic (for example in respect of humour or parody) to ensure that they are not misleading. There must be an interface[106] for users enabling them to mark up inauthentic content.[107] The Digital Services Act additionally requires specific measure to respond to crises in public security or public health[108], to deal with the negative consequences of situations such as rioting, a major

---

[94] Take-down is only one option- other action can include restricting accounts, deprioritising or demonetizing content etc, Arts 10, 16-17 DSA

[95] Other than micro- and small enterprises, Art 19

[96] Art 20 DSA

[97] Art 21 DSA; I have suggested the introduction of ADR for such disputes already 23 years ago: J. Hornle "Internet Service Provider Liability – Let's Not Play Piggy in the Middle" (2002) 7(3) *Communications Law* 85–89

[98] Art 53

[99] Art 24 and Art 42 (for very large online services this includes details about human content moderators, broken down by language and accuracy) DSA

[100] Defined as having more than 45 million users, Art 33 (1) DSA

[101] The European Commission has designated the very large online platform services and very large online search providers, see https://digital-strategy.ec.europa.eu/en/library/designation-decisions-first-set-very-large-online-platforms-vlops-and-very-large-online-search (20. December 2023)

[102] The risks are not limited to illegality, and include risks to fundamental rights such as users' freedom of expression, viewpoint plurality, civic discourse and electoral processes, and public security and negative effects on gender based violence; negative effects on children and public health, Art 34 (1)

[103] Arts 34-35 DSA

[104] Art 34 (1) DSA

[105] Art 35 (1) DSA

[106] This can take the shape of Community Notes, see further the discussion x-ref

[107] Art 35 (1) (k) DSA: "ensuring that an item of information, whether it constitutes a generated or manipulated image, audio or video that appreciably resembles existing persons, objects, places or other entities or events and falsely appears to a person to be authentic or truthful is distinguishable through prominent markings when presented on their online interfaces, and, in addition, providing an easy to use functionality which enables recipients of the service to indicate such information. "

[108] A "crisis shall be deemed to have occurred where extraordinary circumstances lead to a serious threat to public security or public health in the Union or in significant parts of it", Art 36 (2)

terrorist incident or a pandemic further fanned, for example, by social media disinformation.[109] The Commission makes the decision whether such a crisis has arisen to kick the crisis management in action.[110] The crisis provisions require the service providers to conduct risk assessments specific to the crisis situation and take measures in dialogue with the European Commission, taking into account civil liberties, and then report on the measures taken to the Commission.[111] In terms of the crisis fighting measures required, the DSA refers back to Article 35 (1) and states that the restrictive measures have to be proportionate to the public security or public health risks[112]. Furthermore, the Commission is drawing up crisis management protocols in advance and in co-operation with very large (and other) online service providers.[113] The measures could involve restricting the accounts of certain groups or slowing down the amplification of content, speeding up notification systems for illegal content and beefing up the concomitant content moderation, or imposing fact-checking measures. The online service providers themselves determine the actual, specific measures to be adopted[114] in a "dialogue" with the European Commission[115]. The DSA envisages that the Commission effectively takes the role of protector of civil liberties (safeguarding privacy and freedom of expression), by checking that the measures are proportionate, subject to safeguards and time-limited.[116] The Commission in turn has to report to and consult the new European Board for Digital Services[117] in order to ensure that it does not fail in this role.[118]

As in the US the DSA leaves the technical decision on the measures and their implementation to online service providers. However, the initiation of the crisis management procedure is left the European Commission who also has to ensure it complies with civil liberties. This conception of the state (and here a supranational organisation) as the guarantor of civil liberties is fundamentally different to the US, where this is an alien notion. The European Commission is the main enforcer of the DSA as regards the obligations of very large online service providers, with co-operation of the Member States' regulatory authorities.[119]

Moreover, the DSA has stringent[120] and extensive enforcement provisions, including the requirement to have independent, annual compliance audits[121], a compliance function[122], investigatory powers, including against third parties, such as the independent auditors[123],

---

[109] Art 36 DSA
[110] Art 36 (1)
[111] Art 36 (1) (a), (b) and (c)
[112] "specific, effective and proportionate measures" Art 36 (1) (b)
[113] Art 48 (2)
[114] Art 36 (5)
[115] Art 36 (4), (6)
[116] Art 36 (1) (b), (3), (8) (b)
[117] Art 61 composed of the so-called Digital Services Co-ordinators, ie the national regulatory authorities
[118] Art 36 (4) (c), (7), (8), (10)
[119] Arts 56-57; this avoids the "Irish problem", *ie* the fact that most very large online service providers are headquartered in Ireland, and the Irish regulators having insufficient resources and willpower to effectively enforce the law ref; Arts 65-66 on the Commission's enforcement procedure.
[120] Some may say "draconian" enforcement powers, which extend to third parties holding relevant information, Art 51 (2)
[121] Art 37 (very large online services)
[122] Art 41 (very large online services)
[123] Art 51 (1) (national regulators); Arts 67-69 (European Commission)

and significant fines for non-compliance, including daily penalties[124]. Users can claim compensations from service providers for failure to comply with the DSA.[125] As a measure of last resort, the national regulators of the Member States may order the management of online services to set out an "action plan" to avert identified risks, and if this is not forthcoming, the relevant service may be temporarily blocked by a court order.[126] The court proceedings for these wholesale access blocking orders must allow *amicus curae* representations.[127]All enforcement measures, including the blocking order must weigh up the seriousness of the harms to be prevented with the restrictions on freedom of expression and the feasibility and impact on the business of online services[128]. In other words, enforcement measures have to comply with the principle of proportionality.[129]

The DSA additionally mandates access by vetted researchers to online service providers' data and systems to enable them to independently research the systemic risks posed by very large social media services and search engines.[130] It also encourages co-regulatory Codes of Conduct.[131]

3.2.2 Voluntary Code of Practice on Disinformation

One example of a Code, which precedes the DSA is the EU co-regulatory Code of Practice on Disinformation.[132] This Code contains *voluntary* commitments negotiated with online service platforms to fight disinformation through a number of actions. In particular, there are seven main areas on which the Code focuses, namely: identifying and demonetizing disinformation including rules on ad placements, disclosures related to political advertising, agreeing a definition of impermissible manipulative behaviours, enhanced co-operation with fact-checkers, user empowerment and digital literacy, guaranteeing that researchers are having access to data and, the creation of a new transparency centre providing data about disinformation and information sharing. The signatories[133] have entered into commitments in respect of measures with Key Performance Indicators attached in each of these seven areas. The Code led to the establishment of a permanent task force, supported by the European Digital Media Observatory (EDMO). EDMO uses the resources of 14 research hubs covering the entire 27 Member States of the EU. The "hubs" are consortia composed of research institutes and non-profit media and free speech organisations (including fact-checking organisations). Their role is to identify and analyse disinformation campaigns, playing a supporting role in investigative journalism to expose such campaigns, develop tools to analyse the provenance of information, including deepfakes, organising media literacy

[124] Art 52 (3)- (4) and Arts 74, 76: a maximum of 6% of annual worldwide turnover in the preceding financial year, or periodic penalties of 5 % of the average daily worldwide turnover
[125] Art 54
[126] Art 51 (3); Arts 75-76, 82 (European Commission in respect of very large online services)
[127] Art 51 (3) Second sentence
[128] Art 51 (5): "the economic, technical and operational capacity of the provider of the intermediary services"
[129] Art 51 (3) Second sentence, and (5); Art 36 (1) (b), (3), (8) (b)
[130] Art 40
[131] Arts 45, 46, 47
[132] "Strengthened Code of Practice on Disinformation" of 16. June 2022, which built on its predecessor, the 2018 self-regulatory Code of Conduct, see https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation
[133] The original signatories of the voluntary Code include Google, Meta, Microsoft and Tik-Tok; unsurprisingly "X" is not part of this system (16. June 2022), see https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation

activities to citizens in their region, monitoring the digital information system and feeding this intelligence to national media regulators and other public authorities.[134] The significance of the voluntary Code on Disinformation is that it has created an infrastructure for transparency reporting and framework for research to better understand the structural indicators of disinformation and the impact of measures taken. It has been argued that the metrics should be improved across platforms and that the audit requirements under the DSA, the access for researchers to the data of very large service providers and the Code of Practice on Disinformation should be better connected.[135] The following structural indicators have been suggested to measure the effectiveness of the voluntary Code: prevalence of a disinformation campaign, its sources, audience engagement, disinformation revenue, measuring cooperation and investments in fact-checking, and support for implementing the Code by platforms.[136]

3.2.3 EU Regulation of Artificial Intelligence and Disinformation

In addition to the DSA and the Code on Disinformation, the EU Artificial Intelligence Act[137] is relevant to the containment of disinformation. Artificial Intelligence plays a role in dissemination of disinformation and its containment in four respects. First of all, as already pointed out, Generative AI is enabling the creation of disinformation. Secondly, AI is used in the behavioural profiling of social media and search users and the targeting and matching of content. Thirdly the use deepfakes in disinformation also involves AI. Finally, tools detecting and labelling disinformation are increasingly AI based and their malfunctioning impacts freedom of expression. This raises the question of how the new AI regulation applies to these four uses of AI.

As to Generative AI, this falls into the category of General Purpose AI (GPAI) Models[138]. Where Generative AI models are used to generate information or advertising there is a risk that the content generated contains mistakes and so-called hallucinations[139], which leads to the spreading of disinformation.[140] Alternatively, AI may be created to purposively spread disinformation, and the Model is trained to generate the most fantastic and outlandish content with a view to increasing engagement and virality of news stories, whether the motive is income generation or ideology. The EU AI Act envisages that the issue of synthetically generated of manipulated information can be addressed by technological means of detection and labelling according to standards laid down in Codes of Conduct.[141]

---

[134] EDMO "About Us" https://edmo.eu/about-us/edmo-hubs/

[135] S. Lai, K Yadav "Operational Reporting in Practice: the EU's Code of Practice on Disinformation" Carnegie Research Article (21. November 2023) https://carnegieendowment.org/research/2023/11/operational-reporting-in-practice-the-eus-code-of-practice-on-disinformation?lang=en

[136] E Brogi, G De Gregorio "From the code of practice to the code of conduct? Navigating the future challenges of disinformation regulation" (2024) 16 (1) Journal of Media Law 38-46, 45

[137] Regulation (EU) 2024/1689 of 13. June 2024

[138] A GPAI Model "displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications", Art 3 (63) AI Act

[139] This occurs where Large Language Models identify non-existent patterns, see IBM, "What are AI hallucinations?" (1. September 2023) https://www.ibm.com/think/topics/ai-hallucinations

[140] The use of Generative AI for the spreading of disinformation is foreseen by the Regulation, see Recital 120

[141] Art 56 and Recitals 135-136

It regulates Generative AI in three ways. First it lays down disclosure requirements on *users*[142] and *providers*[143] of Generative AI.[144] *Users of AI* generating or manipulating text which is published with the purpose of informing the public on matters of public interest (news) must disclose that the text has been artificially generated or manipulated.[145] This disclosure obligation does not apply if there is human review and editorial responsibility.[146] The definition of users does not encompass private, personal users, so that the disclosure obligation also does not apply to private and personal use of Generative AI.[147] *Providers of* Generative AI must ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. These technical solutions must be effective, interoperable, robust and reliable as far as this is technically feasible.[148]

Secondly it provides specific standards and obligations on Generative AI systems with systemic risks.[149] A systemic risk is defined as "a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the [EU] market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale".[150] Generative AI systems made to disseminate disinformation may fit into this category but the classification of an AI system depends on classification by the EU Commission[151] according to high impact capabilities benchmarks[152]. If an AI model is classified as having systemic risks, such as the wide spreading of disinformation, the provider has to perform "model evaluation in accordance with standardised protocols and tools reflecting the state of the art", conduct and document adversarial testing, and mitigate the risks identified, such as a risk that the model leads to the spreading of disinformation.[153]

Thirdly it imposes documentation and transparency obligations on all Generative AI systems, including training and testing processes.[154] Providers of General Purpose AI models placed on the market in the EU must appoint a legal representative within the EU[155] to ensure compliance and co-operation with the AI Office.

---

[142] Called deployers of AI, defined in Art 3 (4) AI Act

[143] The legislative term is "providers of an AI system", Art 3 (3) AI Act: "natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark (...)"

[144] Art 50 (2) AI Act: "Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated."

[145] Art 50 (4) AI Act

[146] Art 50 (4) AI Act

[147] Art 3 (4) AI Act

[148] Art 50 (2) AI Act

[149] Art 55 AI Act

[150] Art 3 (65)

[151] Art 51 (1) (b) AI Act

[152] See Art 51 (1) and Annex XIII AI Act; in vague terms: the reach, size and autonomy of the AI

[153] Art 55 (1) (a)

[154] Art 53 (1) (a) and Annex XI Evaluation

[155] Art 54 (1)

As to deepfakes[156], the AI Act does not prohibit them, but imposes marking and disclosure requirements, essentially the same as for General Purpose AI.

First, *users*[157] *generating the deepfake content* must disclose that the content/output is artificially generated or manipulated.[158] However, if the content "forms part of an evidently artistic, creative, satirical, fictional or analogous work or programme", the disclosure obligations are limited in order to ensure that they do not interfere with the display or enjoyment of the work.[159]

Secondly, the *providers of the AI*[160] for generating the inauthentic or false content have to mark the outputs in a machine-readable format, so that its nature becomes detectable by AI.[161] This requirement applies regardless of the nature of the output or its use. Thus, the AI Act mainly relies on technological means to detect deepfakes and these technical methods must be "effective, interoperable, robust and reliable as far as this is technically feasible, taking into account (…) the costs of implementation and the generally acknowledged state of the art, as may be reflected in relevant technical standards."[162]

3.2.4 Immunity from Legal Liability for Online Service Providers in the EU

As in the US, the EU has legislative provisions giving online services, including social media, immunity for the content they host.[163] But unlike the US, the immunity is based on knowledge, in other words they are immune from civil and criminal liability, unless they have actual or constructive knowledge of the illegal content or activities.[164] EU Member States must not impose obligations on online services to filter *all* content for illegal items.[165] There is clearly a tension between the requirement for very large online service providers to carry out risk assessments and risk mitigation, on the one hand, and the exclusion of a general monitoring obligation, on the other hand.[166]

There are fundamental differences in media regulation between Europe and the USA. In Europe media plurality is achieved by state interference regulating media service providers and balancing commercial freedom with obligations as to plurality and content moderation, for example the prevention of disinformation. In the US media plurality is achieved by encouraging competition and leaving regulation to the marketplace of ideas, including technological innovation. These two approaches reflect how social media services are

---

[156] See FN 17X
[157] See FN 142
[158] Art 50 (4) AI Act
[159] Art 50 (4) AI Act
[160] See FN 143
[161] Art 50 (2) "detectable as artificially generated or manipulated"
[162] Art 50 (2) AI Act
[163] This had been first introduced by E-commerce Directive 2000/31/EC, and is now in the DSA
[164] Art 6 DSA: "the service provider shall not be liable for the information stored (…) on condition that the provider: (a) does not have actual knowledge of illegal activity or illegal content and, as regards claims for damages, is not aware of facts or circumstances from which the illegal activity or illegal content is apparent; or (b) upon obtaining such knowledge or awareness, acts expeditiously to remove or to disable access to the illegal content."
[165] Art 8 DSA
[166] Ref to cases

regulated and are fundamentally incompatible[167], hence creating a conflict of law on the internet.

4. How does this relate to jurisdiction and sovereignty?

The differences in the regulation of social media is increasingly leading to conflicts of laws between EU States, the UK and the US. It is here that the concepts of jurisdiction and sovereignty are relevant. The essence of jurisdiction is the authority to state the law, which is literally what this Latin term means. The concept concerns the competence of a state's authorities to pass laws, adjudicate disputes and enforce the laws and it is territorially bounded, or attaches to nationality or domicile of persons.[168] The source of the authority in democratic countries is the people conferring power on their government.[169] Jurisdiction is not limited to a state's territory and is overlapping in the sense that more than one state may assert jurisdiction over the same activity.[170] Furthermore, even if a state asserts jurisdiction, in practice the state may not be able to enforce the law it wishes to apply, so as well as jurisdiction being overlapping, it is frequently incomplete or ineffective.[171] This jurisdictional enforcement challenge impacts a state's sovereignty, as it frequently limits the state's ability to govern activities online.[172]

Sovereignty, then, as a concept of international law describes a state's *power* over its territory, and as a corollary, over its population and the activities located in its territory.[173] It is agnostic as to the source of the power and its legitimacy. It is disputed whether sovereignty is merely a concept to explain statehood, or confers a bundle of rights to states, including the right to non-interference.[174] The concept of state sovereignty has recently come under attack.[175] But while legal systems are still composed of states, statehood is inexorably associated with sovereignty, and to that extent, it is a necessary concept.[176] Arising from the concept of sovereignty is the notion of equality of states under international law regardless of their economic or political power, and that states have a right of non-interference with their domestic affairs, both of which are fundamental tenets of the modern international legal system, even if not always obeyed in practice. [177]

As a matter of *international law* what do these two concepts of jurisdiction as *authority* and sovereignty as *power* tell us about internet regulation, including how states and their societies deal with disinformation? Essentially there are two aspects to this: first, the jurisdictional challenge means that states find it increasingly difficult to effectively apply and

---

[167] See fn 86, 163
[168] See further J. Hörnle *Internet Jurisdiction Law and Practice* (OUP 2021) 4
[169] Eg Universal Declaration of Human Rights 1948, Art 21 (3)
[170] J. Hörnle *Internet Jurisdiction Law and Practice* (OUP 2021) 81-113
[171] ibid
[172] J. Hörnle *Internet Jurisdiction Law and Practice* (OUP 2021) 436
[173] See further J. Hörnle *Internet Jurisdiction Law and Practice* (OUP 2021) 7
[174] S Wheatley "Election hacking, the rule of sovereignty, and deductive reasoning in customary international law" (2023) 36 (3) *Leiden Journal of International Law* 675-689, 677
[175] D Herzog *Sovereignty RIP* (Yale University Press 2020)
[176] H Krüger "Of Zombies, Witches and Wizards- Tales of Sovereignty" (2022) 33 (1) *European Journal of International Law* 275-296, 294
[177] Arts 1, 3 of the Montevideo Convention on Rights and Duties of States, 26. December 1933, 165 LNTS 19 and Art 2 (1) of the UN Charter "principle of the sovereign equality of all its Members", see also H Krüger "Of Zombies, Witches and Wizards- Tales of Sovereignty" (2022) 33 (1) *European Journal of International Law* 275-296, 287

enforce their national laws against disinformation online. This entails that states lose sovereignty over the governance of disinformation. Secondly, where disinformation campaigns are actively launched by foreign states to interfere with elections and political processes, this raises the question whether this is a type of foreign interference, illegal under *international* law. The concept of sovereignty might help us to understand what measures states can take, and to understand global social media companies' power struggle with states. The UK Parliamentary Foreign Affairs Committee is currently holding an Inquiry mapping foreign interference by state and non-state actors, precisely in order to understand how this affects diplomatic relations and the actions a state can take.[178]

4.1 The Jurisdictional Challenge

The ubiquitous and cross-border nature of the internet has led to a massive jurisdictional challenge. In principle, speech and social interactions online are without borders and take place remotely without necessarily being tied to any particular territory. This means that jurisdictional claims overlap, several states may try to apply their national law, in practice it may be difficult to enforce the civil and criminal law across a border, and there is a real challenge in investigating illegal acts online.[179] This means that the power of states (sovereignty) to govern is reduced and conflicts of law abound (jurisdictional challenge to effectively apply and enforce the law). Multiple states may assert jurisdiction, laws differ in substance, there is confusion as to the applicable law(s) with significant overlap and states find it difficult to apply the law reflecting their constitutional values to the activities of their citizens.[180] As a consequence, states exert pressure on online service providers as the gatekeepers of the internet to recreate borders online by identifying the physical location of the users of their online services. Generally speaking, states force online service providers to apply EU law to users located in the EU, UK law to users located in the UK, and US law to users located in the US. This leads to the re-establishment of borders on the internet[181]. I argue, though that this is inevitable in our current legal system where the law is largely that of nation states and where the ubiquity of the internet clashes with a system of regulation and values based on nation states.

States' controlling of speech, such as disinformation online, has largely been presented negatively as extending state repression against free speech into cyberspace.[182] However the recreation of national delineations for determining jurisdiction for the purposes of state regulation over citizens' internet activities and data is neither repressive nor liberal *per se*. It can equally be represented as protecting national values, including protecting Constitutional rights, media pluralism, civil liberties and the rule of law, but it can equally constitute the extension of undemocratic political power and repression of opposition.[183] The extension of state regulation to disinformation can be explained by the concept of sovereignty as a

---

[178] UK Parliament (15. January 2025) https://committees.parliament.uk/committee/78/foreign-affairs-committee/news/204722/new-inquiry-disinformation-diplomacy-how-malign-actors-are-seeking-to-undermine-democracy/

[179] See further J. Hörnle *Internet Jurisdiction Law and Practice* (OUP 2021) Chapters 3-6

[180] ibid

[181] BA Simmons et al "Cyberborders Exercising State Sovereignty Online" (2023) 95 *Temple Law Review* 617-640, 624-626

[182] BA Simmons et al "Cyberborders Exercising State Sovereignty Online" (2023) 95 *Temple Law Review* 617-640, 624

[183] Ibid 625

national[184], societal preference for domestic regulation over *other* states' conflicting conception of free speech overreaching into the domestic sphere:

"(…) border orientation captures how the state/society governs forces emanating from the rest of the world ("horizontal" relations). As such, it demonstrates a preference for defining and protecting the home environment from the global one (…) Despite the technological challenges of regulating the internet, governments are "fencing" the internet for very traditional reasons: they want to maintain sovereign control over the information environment in their national territory."[185]

While both democratic and undemocratic states exercise sovereignty, the concept itself is speech neutral. Sovereignty is concerned with international relations, *ie* the power relationships between states and non-interference with the domestic matters of a foreign state.

However, it is clear that this reintroduction of borders in cyberspace is only partially effective. If online service providers ignore or partially ignore compliance with national law in respect of the services they provide to users in a jurisdiction, states may find it difficult to effectively enforce regulations against disinformation. While big tech companies initially were motivated by the legitimate business purpose of increasing their advertising income, in the current era we see direct political influence in social media companies and exploiting their technology for political power, with Elon Musk's "X" as the paradigmatic example. Likewise, in his announcement to replace fact-checkers *Zuckerberg* also stated that he would work with President-elect Trump to fight "censorship" in other jurisdictions including Europe and Brazil.[186] Given the ideology and political aims of these platforms, disinformation and the disruption caused is intentional, as they are seeking political power themselves and this course puts them in confrontation with national governments. So while states seek to introduce borders in internet communication this is not always effective, as states jurisdiction is challenged by limited investigative and enforcement jurisdiction. A number of US based platforms have failed to comply with the DSA and the European Commission has started enforcement proceedings[187]- this is a good example of the power struggle between big tech companies and the state. The outcome of this particular struggle is not yet clear.[188] Another example is of course the bipartisan US legislation[189] ordering that Tik-Tok be sold to a US company or banned in the US from 19. January 2025. While this law is based on concerns of data harvesting of US users and national security concerns[190], not disinformation, this is also an example of recreating borders and exerting data sovereignty,

---

[184] Or in the case of the EU a regional preference

[185] Ibid 625

[186] G Scofield "Alliance between Meta and Trump is likely to create informational, economic and geopolitical conflicts around the world" (8. January 2025) The Conversation https://theconversation.com/alliance-between-meta-and-trump-is-likely-to-create-informational-economic-and-geopolitical-conflicts-around-the-world-246872

[187] EU Commission Press Release (12. July 2024) in relation to X https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761 ; EU Commission Press Release (16. May 2024) in relation to Meta https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2664

[188] The European Commission has extensive investigatory and sanctioning powers, for example fines of up to 6% of the worldwide annual turnover of the company, as well as periodic penalties of up to 5% of the average daily worldwide turnover for each day of delay of compliance, see FN 124

[189] Protecting Americans from Foreign Adversary Controlled Applications Act, April 2024

[190] The US Supreme Court applied intermediate scrutiny on the basis that this was not a law *about content*, see p.5

in order to deal with foreign interference. The legislation has been upheld as compatible with the 1st Amendment by the US Supreme Court[191] and at the time of writing Tik Tok was blocked in the US, to the frustration of its 170 million users.

4.2 Foreign Interference as an Act against State Sovereignty?

Disinformation can stem from national bad actors, but equally can be an act of interference by foreign states with the political system and election integrity. Examples are *Elon Musk* interfering with misleading comments on British politics[192] or the election campaign in Germany[193] or Russian interference with elections in Western countries[194].

This raises the question whether this is illegal under international law as an act against the sovereignty of the states concerned. This is controversial.[195] The Talinn Manual 2.0 provides in Rule 4: "States must not conduct cyber operations that violate the sovereignty of another State."[196] However, Rule 4, and the scope of sovereignty over the internet is unclear and disputed.[197] Some states deny that cyber operations interfere with a state's sovereignty, whereas other states claim that cyber operations can indeed be an interference with a state's sovereignty.[198] Cyber operations in this context mean hacking, computer misuse and foreign interference.

Wheatley argues that usurping governmental functions[199] violates the rule of sovereignty[200], but merely interfering with the exercise of government processes (such as elections) does not.[201] This means that foreign state cyber-operations interfering with the outcome of elections through the spreading of disinformation in the target state are unlikely to be an interference of a state's sovereignty illegal *under international law*. Likewise, a state colluding with a social media platform to exert influence over a population in a foreign state would not count as an act of illegal interference under *international law*.[202] However the *national law* of a state may criminalise such conduct. The UK has created a criminal offence

---

[191] *TikTok v Merrick B Garland* US Supreme Court Westlaw 17. January 2025

[192] The Guardian (20. November 2024) https://www.theguardian.com/media/2024/nov/20/mps-summon-elon-musk-x-role-uk-summer-riots

[193] M Fitzpatrick "The far-right is rising at a crucial time in Germany, boosted by Elon Musk" (1. February 2025) The Conversation https://theconversation.com/the-far-right-is-rising-at-a-crucial-time-in-germany-boosted-by-elon-musk-247895

[194] BBC News article about RT paying a covert US operation to influence the outcome of the 2024 Presidential elections and criminal prosecutions (4. September 2024) https://www.bbc.co.uk/news/articles/c8rx28v1vpro

[195] S Wheatley "Election hacking, the rule of sovereignty, and deductive reasoning in customary international law" (2023) 36 (3) *Leiden Journal of International Law* 675-689, 676

[196] The Tallinn Manual 2.0 (2017) is a work of international law scholarship digesting the principles of international law on cyber-operations which do not use force or armed conflicts, published by Cambridge University Press, see https://www.cambridge.org/gb/universitypress/subjects/law/humanitarian-law/tallinn-manual-20-international-law-applicable-cyber-operations-2nd-edition?format=PB

[197] FN 195, 680

[198] FN 195, 682-683

[199] Using deductive reasoning, in the sense of "exclusive right to exercise sovereign authority with respect to a territory" p. 688, referring to *Island of Palmas* (Netherlands v. USA), Award of 4 April 1928, 2 RIAA 829 (1928), at 838

[200] For example, the *Corfu Channel* case, (United Kingdom of Great Britain and Northern Ireland v. Albania), Merits, Judgment of 9. April 1949, [1949] ICJ Rep. 4

[201] See FN

[202] TIK TOK

of foreign interference, where a foreign power spreads disinformation interfering with political processes such as elections in the UK.[203]

Arguably international law *should* be applied against illegal cyber-operations, including foreign interference and state practice should be changed in this respect. Of course it is uncertain whether state practice will ever emerge in this respect. But the concept of state sovereignty might help to explain the auhtority of the state to contain the power of social media companies and it is to this issue that I will turn next.

4.3 State Sovereignty and Social Media Companies

This article argues that a new powerful type of actor has emerged on the international political scene in addition to states, namely the social media companies themselves. International law may not yet recognise this new category of international actor. But state sovereignty should additionally be concerned with the power relationships between the state and multi-national big tech companies, given their power and influence over the world's citizens. The key question, though, in this power struggle is who is the better guarantor for our civil liberties, including but not limited to free speech: the (democratic) state with a government constrained by a Constitution, the rule of law and mechanisms protecting fundamental rights, or private social media companies having no fundamental rights guarantees and giving unedited, but highly manipulated voices[204] to the "man (and woman[205]) on the street" and increasingly turning into an oligarchy of unelected, but extremely rich and powerful big tech companies?

Civil liberties originally conceived against the all powerful state, are now threatened by entities more powerful than the state and controlled by CEOs with political power aspirations. This threat raises the question of whether the state should be the guarantor of fundamental rights and protect our civil liberties against unaccountable big tech and powerful social media platforms. This is how the rule of law and fundamental rights connect with questions of sovereignty. Democratic states possess institutions and mechanisms to safeguard civil liberties. If these institutions and mechanisms are undermined by the manipulation of big tech companies there is a danger that the democratic state itself will disappear. Social media platforms arguably are more powerful in terms of political influence than any state sovereign and the traditional conception of state sovereignty under international law does not take this into account. Therefore, platform regulation can be seen as a power struggle between nation states and platforms. This power-struggle is a threat to the constitutional order of states, the rule of law, and the guarantee of civil liberties.

The power struggle between big tech companies and the state is challenging democratic nation states' sovereignty over their citizens' activities and data. This power struggle may in the past have been conceived as greater liberty of cyberspace vis-à-vis the state, but citizens' civil liberties are now threatened by the power of big tech controlling their data and

[203] Sections 13-15 National Security Act 2023, for or on behalf of a foreign power, s.31 (1) (a)

[204] Manipulated, because they target content to us based on behavioural profiling, and maximising engagement through predicting our subliminal desires

[205] Given the current waves of misogyny and abuse directed at women, arguably women have much less of a voice than men in the social media space, creating new inequalities and serious questions about free speech for women at the expense of that of men. But this is a different topic to be explored in another article.

their online behaviour through opaque and unaccountable algorithms which manipulate the information citizens see.[206]

5. Conclusion

While the ubiquitous nature of the internet may have increased citizens' access to information and freedom of expression generally, this new freedom is increasingly being severely curtailed by a handful of powerful big tech companies. The purpose of content regulation is to ensure the equivalent protections against social media companies. It is in this light that we need to perceive the fight against disinformation. Therefore, state sovereignty should mean a right for the state to regulate social media companies to protect its citizens' right to receive information and to protect media pluralism. Rather than prohibiting "censorship", international law and human rights frameworks need to limit *how* states regulate social media companies and ensure that this regulation enables plurality, review of decisions, complaints mechanisms, transparency, access by researchers and other safeguards.

[206] See also the UK Parliament Science and Technology Select Committee's Social Media Inquiry