Regulating Big Tech to Counter Online Disinformation: Avoiding Pitfalls while Moving Forward

Paolo Cesarini

Summary

Keywords

Disinformation - social media - Digital Services Act - ISP liability - content moderation

1. Introduction

The role of large social media platforms as privileged vectors of online disinformation became evident during the US Presidential elections and the Brexit referendum of 2016, and many other events thereafter have brought to the public attention the political, social and economic risks arising from an online environment that is still largely unregulated and dominated by a handful of powerful companies. Many governments around the world are searching for solutions to curb the phenomenon, but no silver bullet has been found yet. The European Commission has been one of the first political institutions to take policy action in this area, back in 2018. Following up on a number of successive initiatives, the European Democracy Action Plan (hereafter "EDAP")[1] and the Digital Services Act (hereafter "DSA")[2], both adopted in December 2020, mark a turning point in the policy that the Commission has been building up during the last years. The legislative debate that has now been set in motion before the European Parliament and the Council will have to ensure that these new regulatory tools will effectively tame the power of big tech companies and restore safety and trust

[1] Communication from the Commission to the European Parliament, the Council, the Economic and Social Committee, and the Committee of the Regions on the *European Democracy Action Plan*, 3 December 2020, COM(2020) 790 final.

[2] Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, 15 December 2020, COM(2020) 825 final.

in the online informational space.

The first step of this process was the adoption of a Commission Communication in April 2018[3], based on the advice provided by a High Level Expert Group on Fake News[4]. This initiative paved the way to the world-first self-regulatory framework for online platforms, the Code of Practice on Disinformation[5], to which adhered the main social media companies (Google, Facebook, Twitter, joined later by Microsoft and TikTok), as well as other important players (Mozilla and a number of trade associations representing the advertising industry). The Code adopted a definition characterising disinformation as «verifiably false or misleading information which is created, presented and disseminated for economic gain or to intentionally deceive the public, and which may cause public harm». The commitments taken by the industry in this context aimed at five main, broad objectives. First, limiting the risk of manipulations by actors using platforms' services to amplify false or misleading information. Second, introducing features to empower users, including adjustments to platforms' algorithms in order to prioritise information coming from different authoritative sources. Third, preventing the placements of ads liable to monetise purveyors of disinformation. Fourth, ensuring transparency of political advertising. And fifth, enabling access to platforms' data by independent researchers and fact-checkers, for a better understanding of the phenomenon and increased public awareness of the threats posed by it.

Of course, the problems arising from disinformation cannot be attributed only to the "gatekeeper" role of global online platforms. Many other factors influence the spread of disinformation in modern societies and its corrosive effects on democratic values. For this reason, the Action Plan[6] adopted by the Commission and the High Representative for Foreign Affairs and Security Policy in December 2018 prompted also several other initiatives, designed in particular to boost strategic communication capabilities, increase cooperation among Member States and with international partners, and enhance societal resilience through support to media literacy, technological innovation and fact-checking activities, in combination with actions in support of professional and independent journalism.

While all these aspects are crucial, one key issue emerges: how to set obligations for global online platforms that could be effective in countering disinformation, while balancing safety of users and freedom of speech and opinion. From this perspective, it is important to recall that, based on the lessons learned through targeted monitoring actions in the run up to the 2019 European Parliament elections and during the outbreak of the Covid-19 "infodemic", and in cooperation with the European Group

[3] Communication from the Commission to the European Parliament, the Council, the Economic and Social Committee, and the Committee of the Regions on *Tackling Online Disinformation: a European Approach*, of 26 April 2018, COM(2018) 236 final.

[4] Report of the independent High Level Expert Group on fake news and online disinformation (chairman Madeleine de Cock Buning), *A multi-dimensional approach to disinformation*, 12 March 2018.

[5] See *ec.europa.eu/digital-single-market/en/code-practice-disinformation*.

[6] Joint Communication from the Commission to the European Parliament, the Council, the Economic and Social Committee, and the Committee of the Regions, *Action Plan against Disinformation*, 5 December 2018, JOIN(2018) 36 final.

of Audiovisual Regulators[7], the Commission re-assessed the Code of Practice in September 2020[8] and came to the conclusion that, despite considerable improvements, the slow progress shown by online platforms in assuming their responsibilities called for a stronger regulatory initiative. While covering many other questions of general interest, The EDAP and the DSA followed up on these conclusions.

2. Towards a EU co-regulatory framework

Tabled on 15 December 2020, the long-awaited proposal for the DSA represents an ambitious effort to create a safer and trusted digital environment where fundamental rights are effectively protected and the integrity of the internal market is safeguarded through fully harmonised rules and a EU-wide governance structure. These new rules cover a large variety of online services, from digital marketplaces and collaborative economy platforms to app stores, video-sharing platforms and social networks.

Building on the existing e-commerce Directive[9], the proposal does not change, but rather reinstates, the principle whereby online intermediaries are not liable for the transmission or storage of information provided by third parties, unless they obtain actual knowledge of its illegal nature and do not act expeditiously to remove it (Articles 3 to 5)[10]. Moreover, they are not subject to any general monitoring obligation in respect of the information transmitted or stored on their services (Article 7). The fact that an online intermediary carries out own-initiative investigations with a view to detecting and removing illegal information does not affect its immunity status (Article 6). The novelty of the DSA lies in the fact that this wide liability exemption is made conditional upon a range of due diligence obligations (detection mechanisms, remedial actions and referral procedures) which are tailored to the size of, and the nature of the services provided by different types of digital intermediaries, including online platforms. Due to their potentially higher societal impacts, very large online platforms (hereafter "VLOPs") - defined as services reaching 45 million of active monthly users in the EU, or 10% of the EU population - are subject to stricter conditions. Moreover, appropriate checks and balances, including an internal complaint-handling system and out-of-court redress procedures, are woven into this framework in order to ensure that fundamental rights, in particular freedom of expression, are duly upheld in case platforms would wrongly remove or disable access to legitimate content. Extensive investigatory and enforcement powers, mirroring those commonly used in antitrust investigations and including the power to impose hefty fines (up to 6% of global turn-

[7] *ERGA* *Report on disinformation: Assessment of the implementation of the Code of Practice*, 4 May 2020.

[8] Commission Staff Working Document, *Assessment of the Code of Practice on Disinformation - Achievements and areas for further improvement*,10 September 2020, SWD(2020) 180 final.

[9] 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, 17.7.2000, p. 1).

[10] On the concept of limited liability of online intermediaries, see Van Hoboken J. et al, *Hosting Intermediary Services and Illegal Content Online*, in *op.europa.eu*, 2018; S. Schwemer – T. Mahler – H. Styri, *Legal analysis of the intermediary service providers of non-hosting nature*, 2020.

over), are vested with national authorities and the Commission to prosecute cases of non-compliance.

In broad strokes, the primary objective of the DSA is to counter the relentless online proliferation of *illegal* content (e.g. hate speech, incitement to violence or defamatory information) or illegal activities (e.g. sale of dangerous or counterfeited goods). Nevertheless, the Commission's proposal also acknowledges that online platforms, especially VLOPs, may be «used in a way that strongly influence […] the shaping of public opinion and discourse» (Recital 56), and that the online dissemination of *harmful* - but not necessarily illegal - content may equally endanger users' safety. Without defining the notion of harmful content, the DSA expressly refers to disinformation as one of the serious harms emerging from the current online environment, and tackles disinformation-related harms by prudently combining three elements: (i) a limited set of relevant due diligence requirements, (ii) a legal basis for self-regulation by industry, and (iii) independent oversight and public scrutiny mechanisms.

3. Due diligence requirements

According to the proposal, VLOPs' accountability relies, firstly, on a risk management system consisting of *regular self-assessments of systemic risks and mitigation measures*. In particular, Article 26 imposes on VLOPs the obligation to self-assess, on a yearly basis, certain specific categories of «systemic risks», one of which covers specifically disinformation by referring to risks of «intentional manipulation of [the platform's] service, including by means of inauthentic use or automated exploitation of the service», which may harm public health, minors, civic discourse, electoral processes or public security. Recital 57 expressly mentions «the creation of fake accounts, the use of bots, and other automated or partially automated behaviours» as examples of service manipulations.

Moreover, in order to prevent any systemic risk, Article 27 requires VLOPs to «put in place reasonable, proportionate and effective mitigation measures». Such measures may include adjustments to content moderation or recommender systems, adaptations to the service's terms and conditions, restrictions to the display of ads on the services' online interfaces, the strengthening of internal security processes, and participation to codes of conducts.

Secondly, Article 29 aims at increasing *transparency of recommender systems* thereby reducing the risks for users to be selectively exposed to content promoted by the platform's algorithms, and captured in filter bubbles. In particular, it obliges VLOPs to «set out in their terms and conditions, in a clear, accessible and easily comprehensible manner, the main parameters used in their recommender systems» and to provide users with options «to modify or influence those main parameters […] including at least one option which is not based on profiling».

Thirdly, the DSA provides for *transparency of online advertising,* encompassing political advertising. Pursuant to Article 24, all online platforms - large or small - that display ads on their interfaces are obliged to ensure that each ad is labelled as such, and to identify the sponsors and targeting criteria used for each individual recipient. In addition, under Article 30, VLOPs must keep dedicated repositories of all ads displayed on their online interfaces, make them publicly accessible during one year, and enable the identification of the sponsors, the parameters used to target specific groups of recipients and aggregated user engagement metrics.

4. A co-regulatory backstop

As a horizontal instrument, the DSA is not intended to set future-proof standards regulating in detail platforms' responsibilities in all possible areas. This is particularly relevant in an area like disinformation where fast evolving technologies, service-specific vulnerabilities and metamorphic information manipulation tactics can render too rigid rules quickly obsolete or ineffective. Cooperation from and across the tech industry is therefore necessary to design efficient counter-measures. In view of these challenges, Article 35 establishes a legal basis and minimum requirements (clarity as to the objectives and key performance indicators) for sector-specific codes of conducts, with a view to addressing «significant systemic risks within the meaning of Article 26». The Code of Practice on Disinformation is expressly mentioned in Recital 69, which also refers to the revision of this Code, as announced in the EDAP.

It should be noted that Article 35 merely gives the Commission the power to «invite» VLOPs, smaller online platforms, civil society organisations and other interested parties «to participate in the drawing up of codes of conducts». Participation to a code can only be made compulsory following a full-blown investigation leading to the finding of an infringement.

5. Independent audits, public scrutiny and enforcement

The third pillar of the co-regulatory system created by the DSA is a complex governance structure based on yearly transparency reporting by the companies (Article 33), yearly audits by independent entities (Article 28), mandatory platforms' data disclosure (Article 31), and public scrutiny and enforcement by specialised national authorities and the Commission, under the coordination of a new European Board for Digital Services (Chapter IV).

6. Gaps and issues for further consideration

The complexity of this legal architecture stems from its ambitious goal to cover, in a comprehensive and coherent manner, all types of online services, from digital marketplaces to social networks, and all types of illegal or harmful online activities. At the core of this complexity is the foundational principle (repeatedly confirmed by case-law)[11] whereby online platforms that are mere hosts of information provided by third parties are *a priori* exempted form liability. The risk for such a general framework, which operates by carving specific due diligence obligations out of this wide liability exemption, is to fail to capture all the (equally complex) dynamics that enable the diffusion of disinformation over the Internet, which may therefore entail certain regulatory gaps.

7. A too narrow definition of "systemic risks"

As pointed out above, disinformation-related systemic risks are defined in Article 26 as an « intentional manipulation of [the platform's] services», normally involving an « inauthentic use or automated exploitation of the service », terms which are not further clarified in the proposal. As this provision is the trigger for the application of all the other safeguards set out in the DSA (self-regulation, independent audits, public scrutiny and sanctions), it is important to reflect on the actual meaning of these terms. Tactics using fake accounts and bots to achieve virality are clearly captured by Article 26. The same can be said as regards the use of stolen identities, real accounts taken-over by inauthentic actors, or fake engagement. For example, as reported by NATO StratCom CoE[12], at the height of the US 2020 Presidential elections two US Senators agreed to participate to a test showing that their social media accounts could be easily boosted by using fake engagement bought from Russian social media manipulation outfits. Although the transaction did not involve directly the platform, its execution was aimed at gaming a specific feature of the service (i.e. the algorithm that prioritises content). Therefore, such a case would fall squarely under Article 26 definition of systemic risk.

However, false or misleading narratives can go viral on a social media platform as a result of information manipulation tactics which, technically, may be implemented also *outside* the platform's service. Several authors[13], including Bontcheva & Posetti (2020), François (2019) and Wardle & Derakhshan (2017), have argued that, to be effective, policy responses should address the entire lifecycle of disinformation, by focusing on the agents, their instigators, the message, the intermediaries and intended targets. From this broader perspective, the concept of systemic risk in Article 26, if narrowly construed, may fail to tackle a number of harmful conducts, notably when purveyors

[11] CJEU C-236/08 to C-238/08, *Google France and Google v. Vuitton* (2020); CJEU C-324/09, *L'Oréal v. eBay* (2011); CJEU C-70/10, *Scarlet v. SABAM* (2011); CJEU C-360/10, *SABAM v. Netlog* (2012); CJEU C-314/12, *UPC Telekabel Wien v. Constantin Film and Vega* (2014); CJEU C-484/14, *Tobias McFadden v. Sony Music* (2016); CJEU C-18/18, *Glawischnig* (2019).

[12] S. Bay - A. Dek - I. Dek - R. Fredheim, *How Social Media Companies are Failing to Combat Inauthentic Behaviour Online*, NATO StratCom Centre of Excellence, 21 December 2020.

[13] K. Bontcheva – J. Posetti ed., *Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression,* in Broadband Commission research report on 'Freedom of Expression and Addressing Disinformation on the Internet', 2020; C. François, *Actors, Behaviors, Content: A Disinformation ABC Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses*, Transatlantic Working Group 2019; Claire Wardle, Hossein Derakhshan, *Information Disorder: Toward an interdisciplinary framework for research and policy making*, Council of Europe report DGI, September 2017.

of disinformation make a *strategic use* of a platform, without necessarily engaging in an inauthentic or artificial exploitation of the service itself. The following examples may help illustrate this point.

Attention hacking techniques. Field research conducted by Marwick & Lewis (2017)[14] showed how far-right groups and various Internet sub-cultures in the US have taken advantage of the whole online media ecosystem to manipulate information. Typically, the process starts with the creation of manipulative content (sometimes called "drop"[15]) in darker places of the Internet (e.g. 4chan or kun8, notoriously linked to QAnon conspiracies), which is then declined in different formats (pseudo-news, manipulated images, memes) and planted on a variety of online resources (imposter websites, fringe media outlets, discussion forums, blogs, etc.) by activists operating within and across like-minded communities. The content may then be posted on one or more social media (e.g. by harnessing existing hashtags on Twitter), picked-up by followers and further shared on the platform, even without the help of fake accounts or social bots. The objective of this strategy is that, once "normalised" through authentic users interactions, the narrative may be imported into hyper-partisan media and sometimes unwittingly amplified by mainstream media. Strictly speaking, in this case, the platforms' services are not "artificially exploited", but only used as strategic vectors in view of a harmful purpose.

Information laundering. Building on the seminal work of Klein (2012)[16] who analysed the slip of racist narratives into the mainstream, this form of dissemination of disinformation has been defined in a recent NATO StratCom CoE report[17] as a process whereby «false or deceitful information is legitimised through a network of intermediaries that gradually apply a set of techniques to distort it and obscure the original source». This report analysed recent cases of information laundering in Germany and unveiled that Covid19-related disinformation (notably the story that the pandemic was the result of Bill Gates' depopulation policies and plans for world domination) was layered into the media ecosystem through a network of foreign and domestic proxy platforms (e.g. NewsPunch, News-Front, Connectiv.event, watergate.tv) and then integrated into YouTube and Telegram through their accounts, to be further shared by real users. Similarly, an investigation conducted by the Institute for Strategic Studies[18] ahead of the 2020 US Presidential elections detected cases of information laundering on Facebook and highlighted the difficulty to determine with certainty what constituted illegitimate, deceptive behaviour on this platform.

[14] A. Marwick – R. Lewis, *Media Manipulation and Disinformation*, Data & Society Research Institute, May 2017.

[15] An example can be found at *qalerts.app*.

[16] A. Klein, *Slipping Racism into the Mainstream: A Theory of Information Laundering*, in Communication Theory, November 2012.

[17] B. Carrasco Rodríguez, *Information Laundering in Germany*, NATO Stratcom Centre of Excellence, October 2020.

[18] C. Colliver - M. Hart – E. Maharasingam-Shah - D. Maki, *Spin Cycle: Information Laundering on Facebook*, ISD, December 2020.

State-sponsored propaganda, including foreign interference. A research conducted by Bradshaw & Howard (2017)[19] looked at State-driven social media manipulations across 28 countries, comprising democratic and authoritarian regimes. It found that so-called cyber-troops, including volunteers and paid citizens, were often used by many governments to influence the civic discourse on social media through authentic interactions with users (e.g. comments on social media posts, smearing campaigns against opinion leaders or prominent journalists or bloggers). In particular, it noted that «as bots become increasingly political, social media platforms have become stricter in their take-down policies. As a result, many people have gone back to operating the accounts themselves, rather than automating them». A more recent study by Bradshaw, Bailey and Howard (2020)[20] highlights that «cyber-troops activity continues to increase around the world», finding evidence in 81 countries.

Organic virality. Information manipulations may be prompted in some cases by the intervention of *influencers*. For instance, as reported by The Guardian[21], popular channels with millions of followers on YouTube have been able to push content questioning the results of the 2020 Presidential elections through vast, organic audiences. The same has been observed regarding Covid-19 conspiracy theorists exploiting YouTube culture[22]. In other cases, *statements by politicians* themselves are the direct cause of the viral sharing of false or misleading information on social media. For example, as reported by The New York Times[23], Mr. Trump's false declarations about electoral fraud helped unite «hyper-partisan conservative activists and the standard-bearers of the right-wing media, such as Breitbart, with internet trolls and QAnon supporters behind a singular viral message: #StopTheSteal». Further evidence is provided by a report released by the The Atlantic Council's Digital Forensic Research Lab[24], which compared certain disinformation tactics deployed in the context of the 2018 elections in Latin America. It found that, differently from the Mexican elections, which were affected primarily by automation and artificial amplification on social media, «disinformation in Colombia's elections largely comprised organic disinformation at times amplified by media outlets and political leaders». It should be stressed that, as technology creates new capabilities to forge deceptive content (e.g. deep-fakes), the risks of organic disin-

[19] S. Bradshaw – P. N. Howard, *Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation*, University of Oxford, OII,July 2017.

[20] S. Bradshaw – H. Bailey – P. N. Howard, *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation*, University of Oxford,OII, January 2021.

[21] L. Beckett – J. C. Wong, *The misinformation media machine amplifying Trump's election lies*, in *The Guardian*, 10 November 2020.

[22] A. Ohlheiser, *How covid-19 conspiracy theorists are exploiting YouTube culture*, MIT Technology Review, 7 May 2020.

[23] M. Rosenberg – J. Rutenberg – N. Corasaniti, *The Disinformation Is Coming From Inside the White House*, in *The New York Times*, 5 November 2020.

[24] L. Bandeira – D. Barojan – R. Braga – J. L. Peñarredonda – M. F. Pérez Argüello, *Disinformation in Democracies: Strengthening Digital Resilience in Latin America*, Atlantic Council, Digital Forensic Research Lab, March 2019.

formation may be expected to increase in the future.

Cross-platform migrations. A recent report by The Washington Post [25]warned about the increasing tendency amongst conspiracy theorists to migrate towards smaller platforms in reaction to the take-down measures decided by Twitter, Facebook and YouTube in the context of the Covid-19 infodemic. Unexpected sites such as Internet Archives may become conduits for keeping disinformation online and eventually re-injecting it back into the main social networks. Most recently, The Guardian[26] and Time[27] reported how, after the main social media platforms took the unprecedented decision to block the President's accounts, Mr. Trump's followers massively moved to Parler, and migrated further to Telegram when Google, Amazon and Apple forced Parler offline for hosting threats of violence and racist slurs. These examples show how difficult may be to operationalise the concept of systemic risk if referred to a single service, and if applicable only to VLOPs. The interlinkages enabling a piece of false information banned on one site to reappear in another suggest that disinformation-related risks are endemic to the whole ecosystem and should therefore be addressed in a holistic manner.

The above list does not pretend to be exhaustive. It points, however, to the need to reconsider the scope of Article 26 in light of a broader notion of *information manipulations aimed at the users of the service*. In other words, VLOPs should assess systemic risks not only by reference to what may directly affect the security and technical integrity of their services, but also by taking into account exogenous factors, such as content and source-related manipulations occurring outside their services but liable to spin disinformation across their user base in various ways. Requiring VLOPs to exercise a degree of due diligence over content, sources and overall propagation patterns would have no bearing on the protection of free speech, which would rather depend on the type of mitigation measures eventually adopted. It would instead allow a more comprehensive and deeper understanding of the actual or potential risks incurred by users.

One important consequence of such a revision would be to strengthen the legal basis of the co-regulatory backstop foreseen in Article 35. As this provision exhorts platforms to participate in the drawing up of codes of conduct to address any « significant systemic risk within the meaning of Article 26 », a broader definition of systemic risk would reduce current loopholes and pave the way for an improved Code of Practice on Disinformation, as foreseen in the Commission's EDAP.

[25] E. Dwoskin, *Misinformation about coronavirus finds new avenues on unexpected sites*, in *The Washington Post*, 20 May 2020.

[26] M. Townsend, *How Trump supporters are radicalised by the far right*, in *The Guardian*, 17 January 2021.

[27] B. Perrigo, *Big Tech's Crackdown on Donald Trump and Parler Won't Fix the Real Problem With Social Media*, in *Time*, 12 January 2021.

8. Need for stronger detection mechanisms

To ensure effective detection of systemic risks, the current DSA proposal relies only on VLOPs' *yearly self-assessments*. Even if subject to regular independent audits, it is questionable whether such a mechanism is sufficient. Audit reports can represent a strong incentive for VLOPs to better tackle, in the future, grievances for harms already occurred, but they can do little to prevent the deployment of evolving forms of information manipulation.

Many analyses point to the need to further develop «identification responses» based on independent fact-checking and investigative research. In particular, a recent study by Teyssou, Posetti & Gregory (2020)[28] insists on the importance of integrating content verification and source-checking approaches with «insights into the dynamics of disinformation campaigns, including such elements as the networks conducting them, the targets, the mediums used, the methods used, budgets available, along with attribution and intent». The need for effective identification responses was acutely felt during the Covid-19 infodemic[29]. Therefore, the DSA should cater for this need and include additional due diligence obligations by requiring VLOPs (i) to continuously re-assess systemic risks following *alerts issued by independent fact-checking and research organisations,* and (ii) to ensure *privacy-compliant access to their data* through adequate collaborative frameworks with vetted researchers and fact-checkers. As regards the latter point, it should be noted that, under Article 31, access to platforms's data for vetted researches is made conditional upon a request by the Digital Service Coordinator of the Member State of establishment (in most cases Ireland). Regrettably, this excludes the possibility for vetted researchers to obtain *direct* and *independent* access to platforms' data, which may severely hamper a timely identification of disinformation threats across the EU.

The need for a more structured collaborative framework with fact-checkers and independent researchers, including a role for the new European Digital Media Observatory (EDMO), is expressly acknowledged in the EDAP, where the Commission has pledged to provide guidance to steer the revision of the Code of Practice on Disinformation also as regards access to platforms' data. However, while the details of such collaborative frameworks can be set through a self-regulatory process involving all stakeholders concerned, it is clear that a stronger legal basis in the DSA itself would help ensure legal certainty and consistency. As an alternative, some scholars including Vermeulen (2020)[30], have suggested the use of Article 40 of the General Data Protection Regulation as a basis for issuing guidelines on this matter.

Moreover, good practices should also include other detection mechanisms. First, as for illegal content, *users and civil society organisations should have the possibility to provide notices* to

[28] D. Teyssou - J. Posetti – S. Gregory, *Identification Responses*, in *Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression*, Broadband Commission for Sustainable Development, September 2020.

[29] P. Ball – A. Maxmen, *The epic battle against coronavirus misinformation and conspiracy theories*, in *Nature*, 27 May 2020.

[30] M. Vermeulen, *The keys to the kingdom: Overcoming GDPR concerns to unlock access to platform data for independent researchers*, paper submitted to the Knight First Amendment Institute at Columbia University, October 2020.

VLOPs in order to flag content that they regard as false or misleading, without prejudice to the possibility for the authors of the disputed content to contest any decision eventually taken by the platform against them. Second, online platforms should be encouraged to promote *exchanges of information between their security teams,* notably to facilitate an early detection of covert coordinated networks or "cross-platform migration" cases.

9. More clarity regarding risk mitigation measures

Disinformation is a catch-all term describing a multi-faceted phenomenon involving "fifty shades of harm". At one end of the spectrum, it may simply consist in the inadvertent sharing of false or misleading information amongst social media users (generally referred to as misinformation). At the other end, disinformation can be an intrinsic - and therefore illegal - component of a hate campaign or messaging designed to instigate violence or terrorist attacks. In between these extremes, it can take different forms, from the intentional spread of deceptive information by individuals pursuing political or economic aims, to sophisticated disinformation campaigns carried out by state or non-state actors to influence a targeted domestic audience, or foreign interference operations using hybrid strategies in adversarial scenarios to disrupt the free formation and expression of citizens' political will. Given the variety of possible situations, tracing a clear line between what is lawful or unlawful would entail a too high risk of restricting free and legitimate speech. Attempts to pass «fake news» laws in several countries since the outbreak of Covid-19[31] have often been motivated by the intent to suppress political opposition and social critique.

Against this backdrop, Article 27 does not seem to shed much light as to what is actually required from VLOPs in order to mitigate risks emerging from disinformation. At the same time, the debate sparked by the decision by major social media to block Mr. Trump's accounts amid violent protests at Capitol Hill[32] has brought to the forefront the problem of leaving global platforms alone with the power to decide what can or cannot stay online. Inaction by policy-makers and legal vacuum would make large online platforms the ultimate arbiters of democracy. Aware of this problem, and as announced in the EDAP, the Commission will issue guidance to steer the upcoming revision of the Code of Practice. In this context, the following issues should deserve special attention.

Content moderation. Obviously, when disinformation constitutes an intrinsic component of a piece of illegal content, a take-down measure is warranted. As recently stressed by Commissioner Breton «what is illegal offline should be illegal online»[33]. But across the many shades of harm arising from disinformation, things get blurred and balancing user safety with the imperative of protecting freedom of speech and

[31] J. Wiseman, *Rush to pass 'fake news' laws during Covid-19 intensifying global media freedom challenges*, in *International Press Institute*, 22 October 2020.

[32] M. Scott, *Trump's social media ban reignites fight over how to police online content*, in *Politico*, 12 January 2021,

[33] T. Breton, *Capitol Hill — the 9/11 moment of social media*, in *Politico*, 10 January 2021,

opinion, as strongly advocated by Commissioner Jourova[34], becomes critical. To this end, the principle of proportionality should shape the upcoming Commission's guidance for a range of mitigation measures, tailored to the different types of information manipulation liable to cause public harms. For example, in case of information manipulations affecting the security or technical integrity of the services (e.g. bots, fake accounts, fake engagement, etc.), the deceptive intent could be presumed and, as for any other form of fraudulent behaviour, content removals or account take-downs could be justified. In case of suspected influence operations (attention hacking, information laundering or use of cyber-troops) the primary objective should be to ensure that platforms apply effective detection processes and take effective action as soon as the suspicion is confirmed. In particular, when vetted fact-checkers and researchers provide them with *prima facie* evidence of a possible influence operation, platforms should immediately activate their internal security teams to ascertain the possible existence of covert coordinated networks (within or outside their services) in collaboration, where appropriate, with the security teams of other platforms. If the investigation confirms the suspicion, they should instantly inform the competent public authorities of the Member States targeted by the operation and disable access to the disputed content. For other types of manipulation, such as in case of organic disinformation, mitigation measures should be rather geared towards raising awareness and providing tools for user empowerment. Thus, building on existing practices, such measures could include features giving automatic prominence to official sources of information and fact-checks, demotion or demonetisation of content assessed by independent fact-checkers as false or misleading, systematic warnings or labels identifying fact-checked content and discouraging further sharing. But they should also include pro-active collaborations with independent media outlets and media literacy experts to prevent unwitting amplification and enable early exposure of disinformation by journalists.

Responsible algorithmic design. As shown by many studies, including Acerbi (2019)[35], algorithms embedded in recommender and content ranking systems may give prominence to information that, because of its outrageous, shocking or misleading nature, maximises users' attention by exploiting individual cognitive preferences. Their aim, of course, is to optimise the advertising-driven business model that often underpins platforms' service. Article 29 acknowledges in part such risks by setting out algorithmic transparency and accountability rules, including the possibility for users to opt out from algorithms using profiling data. However, to fully empower users, additional adjustments should be considered. As pointed out in the EDAP, one of the objectives of the revised Code of Practice should be to «support adequate visibility of reliable information of public interest and maintain a plurality of views». In this perspective, an effective integration of trustworthiness indicators[36] for information sources into platforms' algorithms should be one of the core elements to be considered in view of the revision of the Code.

[34] Florian Eder, *Jourová: Big Tech's Trump ban 'dangerous for free speech'*, in *Politico*, 13 January 2021.

[35] Alberto Acerbi, *Cognitive attraction and online misinformation*, in *Nature*, 12 February 2019.

[36] Christophe Leclercq, Marc Sundermann, Paolo Cesarini, *Time to act against fake news*, in *Euractiv*, 2

Demonetisation. The story[37] of the group of teenagers from the Macedonian city of Veles who created about a hundred of fake news websites to pump out sensationalist pieces during the 2016 US Presidential campaign, and earn cash from advertising, is an emblematic example of how disinformation can also be driven by an economic incentive. Article 27 refers, among others, to measures intended to limit the display of ads *on the platform's online interface* (e.g. YouTube removing ads from channels pushing debunked conspiracy theories). However, this provision does not limit the placement of ads on *third-party websites* (as in the Macedonian kids example). The difficulty here stems from the fact that the programmatic advertising industry is complex ecosystem encompassing a variety of online intermediaries that may (as is the case for Google) or may not be qualified as VLOPs. The EDAP has well identified this issue, which is now in the list of the outstanding gaps to be plugged in the revised Code of Practice.

10. Complementary regulation for political advertising

Online political advertising has become a political problem after the Facebook/Cambridge Analytica scandal revealed how personal data could be abused by grouping unsuspecting voters into clusters defined by specific psychological traits, and then by using online advertising as means to target them with well-tailored political messages. Since then, online political advertising has been part of a wider debate concerning the integrity of elections, often intersecting with other issues such as transparency of financing sources of political parties, cybersecurity, parity of treatment and balanced media coverage of different political programmes.

Drawing from the existing Code of Practice on Disinformation, the DSA sets out, in Articles 24 and 30, general transparency requirements covering also political advertising. The legally binding nature of these new rules will ensure in the future stricter compliance by platforms. However, the issues raised by online political advertising go far beyond the scope of the DSA. The responsibilities of advertising intermediaries and advertisers in a political context are subject also to other rules laid down in national and EU laws that protect free and fair elections. For this reason, the Commission has pledged in the EDAP to propose in the coming months, as a complement to the DSA, new legislation to ensure greater transparency for sponsored content in a political context, and to adopt support measures and guidance for political parties and Member States. Therefore, it will be important to revisit this topic once the new regulatory "package" will be on the table.

11. Conclusions

The above analysis suggests that the Commission's proposal for the DSA and the

December 2020.

[37] E. J. Kirby, The city getting rich from fake news, in *BBC News*, 5 December 2016.

complementary actions set out in the EDAP do establish a sound and robust legal architecture to counter online disinformation. Taking a legislative initiative to address this continuously evolving phenomenon and manage effectively the threat it represents for democratic societies, requires the design of rules that are sufficiently flexible to address not only well-known risks, but also those that could emerge in the future from the increasing digitalisation of the informational space: a degree of flexibility that can only be achieved by combining self-regulation with a powerful regulatory backstop. Moreover, an inclusive multi-stakeholder process, framed by clear guidance from the Commission, should ensure that the revised Code of Practice on Disinformation will incorporate specific criteria firmly anchored on the proportionality principle for counter-measures not to encroach on fundamental rights, in particular freedom of expression.

Nevertheless, as usual, the devil is in the details. The DSA starts from the foundational principle whereby online platforms are *a priori* exempted from liability in respect of the information provided by third parties and hosted on their services. By carving specific due diligence obligations out of this wide liability exemption, it may allow possible harmful conducts to slip through the net. Our analysis suggests that this unintended consequence may result, in particular, from a too narrow definition of the concept of "systemic risks", which is key to delineate, in connection with other provisions of the DSA, the actual scope of the mechanisms (mitigation measures, transparency reporting, independent audits, public scrutiny and enforcement) that should make online platforms effectively accountable. In addition, in order to provide for a solid and future-proof co-regulatory backstop, it would be important to consider other additional due diligence obligations to promote a more efficient detection and in-depth investigation of the evolving threats stemming from online disinformation. Finally, regarding the nature of the mitigation measures that platforms should be expected to adopt, clearer guidance from the Commission seems necessary to avoid that too much discretion is left in their hands, such as to vest them with the role of ultimate arbiters of democracy.

The DSA and EDAP have set out the high ambition for the EU to be the first regulator in the world to tame the extraordinary power of global digital players. The DSA is now for discussion before the European Parliament and the Council, but the legislative process will certainly be long and laborious.