Opportunities and Limits of European Social Network Regulation

Marco Almada, Andrea Loreggia, Juliano Maranhão, Giovanni Sartor

1 Introduction

Social networks are a distinctive feature of modern society. As of 2022, people in almost every country of the world rely on one network or another for a multitude of tasks: to get information about local and global affairs, to interact with acquaintances old and new, to find—and even carry out—work, among other relevant aspects of social life. By creating spaces that lend themselves to such diverse uses, the companies running the largest social networks have managed to position themselves among the largest businesses in the world.[1] Yet, the sheer diversity of the interactions ongoing in social networks means some of such interactions are relevant to the law in one form or another, either for the prevention and repression of potentially harmful activities, or for the promotion of beneficial services and interactions. Therefore, the regulation of social networks is a problem that legislators and courts worldwide have to face, and the European Union (EU) is no exception.

Regulating social networks is a complex issue for a variety of factors. Some of the complexity stems from the global reach of platforms, which have users in various countries and are, accordingly, subject to various jurisdictions.[2] Moreover, regulation has to take into account the business model adopted by social networks: users normally can join and use networks for free,[3] but companies use the content they generate to attract new

1 See, e.g., 'Facebook Reports Third Quarter 2021 Results', Meta Investor Relations, 25 October 2021, https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Third-Quarter-2021-Results/default.aspx.

2 On the challenges of global governance, see, e.g., Robert Fay, 'A Model for Global Governance of Platforms', ed. Martin Moore and Damian Tambini (Oxford: Oxford University Press, 2021), 255–79, https://doi.org/10.1093/oso/9780197616093.003.0016.

3 Some networks, however, have experimented with tiered subscription models, in which users pay for having access to features not available to a general audience: Sara Beykpour and Smita Gupta, 'Introducing Twitter Blue - Twitter's First-Ever

consumers and render these users legible to various forms of marketing, notably targeted advertising.[4] As such, strategies used to regulate other kinds of business might not be as effective when directed towards social networks. A final challenge inheres in the technological complexity of social networks. These networks rely on sophisticated technical infrastructures that enable user communication and render users legible by storing the data they provide and drawing inferences from such data,[5] a practice that is compounded by the ongoing development of artificial intelligence (AI) technologies. On the one hand, legibility allows the use of AI systems directed at influencing user behaviour in ways that are not necessarily in their best interest, ranging from selling products[6] to shaping political behaviour through targeted propaganda[7] and forgeries that are indistinguishable from real content.[8] On the other hand, AI systems may be used to protect users' rights online, for example, by contributing to the detection and elimination of these kinds of influence.[9] Consequently, the debates on social networks are increasingly tangled with the present and future of AI.

Regulating social networks is a task that involves multiple levels. Competition law sets up rules meant to prevent social networks from abusing

---

Subscription Offering', Company Blog, *Twitter* (blog), 3 June 2021, https://blog.twitter.com/en_us/topics/company/2021/introducing-twitter-blue.

4 On this point, see Julie E. Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford, New York: Oxford University Press, 2019), chap. 2.

5 On the role of inferences as a source of data, see Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI', *Columbia Business Law Review* 2019, no. 2 (2019): 494–620.

6 See, e.g., Federico Galli, 'Online Behavioural Advertising and Unfair Manipulation Between the GDPR and the UCPD', in *Algorithmic Governance and Governance of Algorithms: Legal and Ethical Challenges*, ed. Martin Ebers and Marta Cantero Gamito, Data Science, Machine Intelligence, and Law (Cham: Springer International Publishing, 2021), 109–35, https://doi.org/10.1007/978-3-030-50559-2_6.

7 See, e.g., Ronan Ó Fathaigh et al., 'Microtargeted Propaganda by Foreign Actors: An Interdisciplinary Exploration', *Maastricht Journal of European and Comparative Law* 28, no. 6 (1 December 2021): 856–77, https://doi.org/10.1177/1023263X211042471.

8 Bobby Chesney and Danielle Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security', *California Law Review* 107, no. 6 (2019): 1753–1820.

9 Giovanni Sartor and Andrea Loreggia, 'The Impact of Algorithms for Online Content Filtering or Moderation. Upload Filters', Study for the committee on Citizens' Rights and Constitutional Affairs (Brussels: European Parliament, 2020).

dominant positions.[10] Other norms govern user-generated content as a source of data, notably data protection law.[11] Finally, some norms can be said to establish social network regulation in a narrow sense, as they establish what networks can or cannot do in their everyday operation. This latter set is the object of the present chapter.

This chapter argues that social networks are currently undergoing a turn towards adopting procedural safeguards and duties of care regarding the substantive rights of users. Section 2 presents the backdrop for this argument. The current EU regulatory framework, centred on the eCommerce Directive,[12] was thought for a different online environment. Therefore, it is strained by social networks in ways legislators and courts are currently trying to address. Some of these strains are produced by the institutional design of the regulatory framework, but these institutional factors only become a problem in light of the harms that social networks introduce or amplify, which are the subject of Section 3. Despite the fact that harmful user behaviour may sometimes be advantageous to social networks (e.g., by attracting certain groups of users), social networks may be induced to adopt content moderation approaches not only in the interest of the users that could be harmed or repelled by such behaviour, but also to avoid losing the liability exemption they enjoy as intermediary carriers of user-generated content. As Section 4 shows, content moderation may itself introduce risks to users' rights, and EU courts and legislators have sought to constrain the range of discretion available to moderators. In this context, we argue the regulation of social networks should be perceived as a socio-technical problem, in which neither technical approaches nor general law alone are conducive to socially desirable outcomes. Instead, regulation needs to be aware of the social impacts of platforms, and the role technology can play in amplifying or mitigating them.

---

10 In the European Union, see Nicolas Petit, 'The Proposed Digital Markets Act (DMA): A Legal and Policy Review', *Journal of European Competition Law & Practice* 12, no. 7 (1 September 2021): 529–41, https://doi.org/10.1093/jeclap/lpab062.

11 See, e.g., Paul Nemitz, 'Constitutional Democracy and Technology in the Age of Artificial Intelligence', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (28 November 2018), https://doi.org/10.1098/rsta.2018.0089.

12 European Union, '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')' (2000), https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32000L0031.

2 The European regulatory landscape

The eCommerce Directive,[13] adopted in 2000, provides the general framework for the regulation of the online environment in the European Union. This Directive harmonises the rules applicable to information society services, that is, to "service[s] normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of services".[14] As outlined in the introduction, a social network meets all elements of this definition: it provides services to users who voluntarily join the network through electronic means. Since these services are usually provided through a by-profit model, social networks fall into the scope of the existing regulatory framework for information society services.

Social networks are part of a well-defined regulatory environment, which contains not only a broad set of applicable norms but also enforcement structures at the national and EU levels.[15] But, as the short name of the Directive suggests, this regulatory framework was originally designed to deal with a different set of concerns than the ones raised by social network's current role in European society.[16] While eCommerce services profit from enabling the acquisition of goods through a virtual environment, and platforms such as newspapers act themselves as sources of content, social networks are doubly dependent on the information produced by the users in different ways: user-generated content makes the platform relevant to content-consuming users, while information about users allows for the monetisation strategies described above and for individualised strategies aimed at keeping users engaged with the platform. As a result, the frame-

13 European Union.

14 Article 1(1)(b) of European Union, 'Directive (EU) 2015/1535 of the European Parliament and of the Council of 9 September 2015 Laying down a Procedure for the Provision of Information in the Field of Technical Regulations and of Rules on Information Society Services (Text with EEA Relevance)' (2015), http://data.europa.eu/eli/dir/2015/1535/oj/eng. This directive repealed and replaced Directive 98/34/EC, to which Article 2(a) of the eCommerce Directive referred when defining "information society services".

15 Alexandre de Streel and Martin Husovec, 'The E-Commerce Directive as the Cornerstone of the Internal Market. Assessment and Options for Reform', Study for the committee on Internal Market and Consumer Protection (Luxembourg: European Parliament, 2020), sec. 2.3.4.

16 For a historical overview of the evolution of platform regulation in the European Union, see Giovanni De Gregorio, 'The Rise of Digital Constitutionalism in the European Union', *International Journal of Constitutional Law* 19, no. 1 (2021): 41–70, https://doi.org/10.1093/icon/moab001.

work established by the eCommerce Directive shows some signs of strain as it attempts to fit social networks into rules conceived for a different moment of the Internet.

The first issue demanding attention is that of regulatory fragmentation. By their very digital nature, social networks can reunite under the same virtual environment users physically located in different countries. From a legal perspective, geographical dispersion brings at least two challenges to the regulatory system. The first one is that two or more legal systems may have a claim to apply their laws to a given event, for example, in the case of a dispute between users based in two different countries. Such situations are in principle covered by existing rules on conflicts of law and court jurisdiction.[17] However, these rules are complicated subjects in their own right,[18] so their application to the context of online platforms may pose practical problems to lawyers and courts. Moreover, a single harmful act may have effects that are relevant to multiple jurisdictions.

Thus, users of the same network may be covered by different norms regarding the same conduct. Social networks are thus required to consider a user's location in the physical world to identify which laws apply to them, and possibly also other locations in which harmful effects were produced. Within the European Union, the eCommerce Directive reduces regulatory complexity, as it provides various requirements that the EU Member States must observe when designing their own laws for information society services. But the harmonisation provided by a Directive is only partial, as each Member State can choose the form and methods it will use to comply with the requirements imposed by EU legislation.[19] This partial harmonisation allows Member States to adopt regulation beyond the minimum guidelines set at the Union level. Indeed, Germany has done so in its own approach to network regulation.[20] As a result, EU nationals using the

17 In fact, the eCommerce Directive explicitly rejects the creation of new rules on these matters: see Article 1(4) and the accompanying Recital 23.

18 See, e.g., Pedro de Miguel Asensio, *Conflict of Laws and the Internet* (Edward Elgar Publishing, 2020), chap. 2; Ilaria Pretelli, 'Protecting Digital Platform Users by Means of Private International Law', *Cuadernos de Derecho Transnacional* 13, no. 1 (2021): 574–85.

19 On EU directives and their legal effects, see, e.g., Robert Schütze, 'Direct Effect', in *An Introduction to European Law*, 3rd ed. (Oxford: Oxford University Press, 2020), 109–32, https://doi.org/10.1093/he/9780198858942.003.0005.

20 See, in addition to the relevant chapters in this book, Robert Gorwa, 'Elections, Institutions, and the Regulatory Politics of Platform Governance: The Case of the German NetzDG', *Telecommunications Policy*, Norm entrepreneurship in Internet Governance, 45, no. 6 (1 July 2021): 102145, https://doi.org/10.1016/j.telpol.20

same network—and potentially interacting with the same content—might be subject to substantively different norms.

Fragmentation in European network regulation is not produced just by the Member States. Within the European Union legal order itself, various *lex specialis* instruments govern specific practices at the core of how social networks operate. This chapter engages directly with two such instruments— the Copyright Directive[21] and the Terrorist Content Regulation.[22] This fragmentation is not necessarily harmful to regulation, especially if it supplies an effective response to harms that would be ill-addressed by changes to general legislation. Yet, by definition, the adoption of specialised norms[23] may increase compliance costs for social networks and make users less certain about the rules that apply to their circumstances.

However, we should not overestimate the level of fragmentation seen in EU social network regulation. After all, the eCommerce Directive establishes various requirements for Member State legislation. Some of these are directed at ensuring harmonised conditions for the information society services themselves, such as the functioning of the internal market for such services,[24] their establishment,[25] or the possibility of relying on out-of-court dispute settlement.[26] Other provisions provide guarantees for the users of such services, such as the minimum standards for information to

---

21.102145; Patrick Zurth, 'The German NetzDG as Role Model or Cautionary Tale? – Implications for the Debate on Social Media Liability', *Fordham Intellectual Property, Media & Entertainment Law Journal* 31, no. 4 (2021): 1084–1153, https://doi.org/10.2139/ssrn.3668804.

21 'Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC (Text with EEA Relevance.)' (n.d.).

22 'Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance)' (2021).

23 On generality as a legal value, see, e.g., Gregor Kirchhof, 'The Generality of the Law: The Law as a Necessary Guarantor of Freedom, Equality and Democracy and the Differentiated Role of the Federal Constitutional Court as a Watchdog', in *Rational Lawmaking under Review: Legisprudence According to the German Federal Constitutional Court*, ed. Klaus Meßerschmidt and A. Daniel Oliver-Lalana, Legisprudence Library (Cham: Springer International Publishing, 2016), 89–127, https://doi.org/10.1007/978-3-319-33217-8_5.

24 Article 3 eCommerce Directive.

25 Article 4 eCommerce Directive excludes any need for prior authorisation before offering an information society service.

26 Article 17 eCommerce Directive.

be provided by the service[27] or the specific rules for commercial communications,[28] contracts concluded through electronic means,[29] and the liability of intermediary service providers for the content they provide.[30] Regardless of how the Member States exercise their legislative power with regard to platforms, they are still required to *at least* comply with the Directive and—more than that—cooperate actively in rendering it effective.[31] The eCommerce Directive thus provides users and platforms with a regulatory baseline, setting expectations for how social networks function.

Yet, this baseline is somewhat thin. While adequate transposition of the eCommerce Directive leads to various requirements being imposed upon social networks, these still have considerable leeway to determine the conditions for providing their service. Indeed, large social networks are notorious for adopting extensive terms of service,[32] which empower them with vast discretion regarding content removal, monetisation of user data, and various other aspects.[33] This discretion is somewhat reduced by the *sp*ecialised norms mentioned above, as their strict rules on content removal are accompanied by requirements that mandate procedural safeguards that users can invoke in case of removed content.[34] But the Directive itself has little to say about how platforms should set up their Terms of Service, leaving them considerable room for manoeuvre within the general constraints of the legal system to private autonomy. Given the centrality of social networks in modern social life, these decisions may have a considerable impact upon a person's social life or even their livelihood, thus prompting users to resort to judicial or administrative authorities to assert their rights liberties, and interests.

A final source of tension between social networks and regulation based on older models of information society services is data governance. Tradi-

27 Article 5 eCommerce Directive.
28 Articles 6–8 eCommerce Directive.
29 Articles 9–11 eCommerce Directive.
30 Articles 12–15 eCommerce Directive, which Section 4 below examines in further detail.
31 Article 19 eCommerce Directive.
32 These terms are often opaque, in the sense they are difficult reading even for a trained lawyer: Marco Lippi et al., 'CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service', *Artificial Intelligence and Law* 27, no. 2 (1 June 2019): 117–18, https://doi.org/10.1007/s10506-019-09243-2.
33 See, *inter alia*, Dan Wielsch, 'Private Law Regulation of Digital Intermediaries', *European Review of Private Law* 27, no. 2 (1 April 2019), http://kluwerlawonline.com/journalarticle/European+Review+of+Private+Law/27.2/ERPL2019013.
34 See Section 4 below.

tional information society services produced—and made use of—substantial volumes of data about users and their transactions. As a result, data protection law was already a key factor in their governance.[35] For social networks, however, users' data is not just an instrument for controlling their operation but also a central element in their business models. Acknowledging this new reality, the EU has substantially revamped its data governance framework, most notably by adopting a General Data Protection Regulation.[36] Those norms are directly applicable to the operations of social networks and provide safeguards to the rights of platform users and third parties that might be affected by content shared on the networks or by inferences made from it.[37] Yet, data protection law, by construction, focuses on individuals rights, thus failing to account for the systemic effects that data may have within social networks.[38]

35 Accordingly, the CJEU has produced a considerable volume of case law on information society services. For an overview, see Giovanni De Gregorio, 'From Constitutional Freedoms to the Power of the Platforms: Protecting Fundamental Rights Online in the Algorithmic Society', *European Journal of Legal Studies* 11 (2019): sec. III.1.

36 At the same time the GDPR supplies a stricter framework for the governance of personal data, other pieces of EU legislation—such as the proposed Data Governance Act—seek to create favourable conditions for the circulation of non-personal data. For an overview of data governance in the European Union, see Thomas Streinz, 'The Evolution of European Data Law', in *The Evolution of EU Law*, ed. Paul Craig and Gráinne de Búrca, 3rd ed. (Oxford: Oxford University Press, 2021), 902–36, https://doi.org/10.1093/oso/9780192846556.003.0029. It is important to keep in mind, however, that the distinction between personal data and non-personal data is not always clearcut: Marco Almada, Juliano Maranhão, and Giovanni Sartor, 'Article 4 Para. 5. Pseudonymisation', in *General Data Protection Regulation. Article-by-Article Commentary*, ed. Indra Spiecker gen. Döhmann et al. (Munich; Baden-Baden; Oxford: Beck; Nomos; Hart Publishing, 2022).

37 Pedro A. de Miguel Asensio, 'Data Protection in the Internet: A European Union Perspective', in *Data Protection in the Internet*, ed. Dário Moura Vicente and Sofia de Vasconcelos Casimiro, Ius Comparatum - Global Studies in Comparative Law (Cham: Springer International Publishing, 2020), 457–77, https://doi.org/10.1007/978-3-030-28049-9_18.

38 For general analyses of the limits of this individualistic framework, see Przemysław Pałka, 'Data Management Law for the 2020s: The Lost Origins and the New Needs', *Buffalo Law Review* 68, no. 2 (1 April 2020): 559–640; Cohen, *Between Truth and Power*, chap. 2. For an example, consider how data protection law offer little remedy against the production of filter bubbles through algorithmic recommender systems: Marco Almada, Juliano Maranhão, and Giovanni Sartor, 'Article 6 Para. 1. Content Personalisation', in *General Data Protection Regulation. Article-by-Article Commentary*, ed. Indra Spiecker gen. Döhmann et al. (Munich; Baden-Baden; Oxford: Beck; Nomos; Hart Publishing, 2022).

Considering these challenges posed by social networks to the governance of the information society, the EU legislator is currently seeking to update this overall framework. The key idea beyond the changes to social network regulation is *digital constitutionalism*,[39] that is, the extension to the digital environment of the constitutionalist ideals of separation of powers and protection of fundamental rights.[40] In the context of social networks, these ideals are translated into a double movement: introducing substantive requirements for the protection of rights online[41] and adopting due process considerations regarding network decisions on whether to remove online content.[42]

This movement towards digital constitutionalism has been reflected in the specialised instruments mentioned above, but it is particularly salient in the Digital Services Act package proposed by the European Commission.[43] At the core of this package lie two pieces of legislation. The first one is the eponymous legal instrument, which amends the framework of the eCommerce Directive to extend its principles to a context marked by different technologies and the substantial power of very large online platforms.[44] This proposal is complemented by Digital Markets Act, which includes a broad range of measures to restrict the power of so-called gatekeeper services, such as advertising services and the social networks themselves.[45] While these legal instruments focus on different legal challenges posed by platforms such as social networks, they nevertheless share the two elements of digital constitutionalism presented above, as they impose limits to what platforms can do and forces them to adopt formal

39 De Gregorio, 'The Rise of Digital Constitutionalism in the European Union'.

40 Edoardo Celeste, 'Digital Constitutionalism: A New Systematic Theorisation', *International Review of Law, Computers & Technology* 33, no. 1 (2 January 2019): 76–99, https://doi.org/10.1080/13600869.2019.1562604.

41 See, e.g., De Gregorio, 'From Constitutional Freedoms to the Power of the Platforms', V.II.

42 See, e.g., De Gregorio, V.I.

43 https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

44 See European Commission, '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' (Brussels: European Commission, 15 December 2020).

45 For an introduction to the DMA as it stands as of December 2021, see Filomena Chirico, 'Digital Markets Act: A Regulatory Perspective', *Journal of European Competition Law & Practice*, no. lpab058 (2021), https://doi.org/10.1093/jeclap/lpab058; Petit, 'The Proposed Digital Markets Act (DMA)'; Natalia Moreno Belloso, 'The Proposal for a Digital Markets Act (DMA): A Summary', 3 January 2022, https://papers.ssrn.com/abstract=3999966.

procedures for handling complaints. Still, they retain the core element of the governance regime described above: treating social network liability as the exception and not the rule.

3 User-generated content and online harms

Social networks, as seen in the Introduction, are not in the business of producing content. Instead, they provide their users with a digital environment to interact with other users.[46] This interaction, in turn, produces *user-generated content* of various forms, such as private messages to other users, texts aimed at a general audience, memes, or live streams of audiovisual content. User-generated content may benefit users: they may learn new things from online sources, find joy in meeting new people and reconnecting with old acquaintances, and so on. However, online interactions may also negatively affect users, leading to psychological or even material harm. This section provides a brief overview of the various mechanisms through which users may be harmed within social networks and how these networks respond to harmful content within the current EU regulatory framework.

Online harm may take various forms. In some cases, harm comes from practices much older than social networking. Scammers can use social networks to identify and contact potential victims, bullies can expose their victims to ridicule or worse, and racists and other hate groups can direct their vitriol against vulnerable individuals and groups. While these practices are long-standing social issues, social networks transform how they take place. Through social networks, users with harmful intent can contact a larger number of victims simultaneously, even if these targets are geographically distant from one another. Social networking may also amplify the effect of harms committed in public, such as bullying: given the difficulties of removing content from the Internet,[47] targeted users may be forced to revisit the pain and humiliation of what they have been through. .

46 These users might be natural persons or collective profiles standing for a legal person or other groupings of people.

47 Not just from the technical issues of removal, but also because the very attempt of removing something might call attention to the original content, in the so-called Streisand Effect: Daphne Keller, 'Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling', *GRUR International* 69, no. 6 (1 June 2020): 622, https://doi.org/10.1093/grurint/ikaa047.

User-generated content may also be directed towards forms of harm with no clear offline analogue. One such phenomenon is *doxing*, that is, the disclosure of personal information about a user within a network.[48] This practice is often, though not always, directed towards users that express controversial opinions online[49] as an attempt to highlight to these users that their opinion will have offline consequences. In fact, the information disclosure is often accompanied by pressure towards real-world acquaintances of the targeted user, such as calls for their employer to fire them for their online expression.[50]

The recent developments in artificial intelligence technologies, combined with the vast amounts of data available in social networks,[51] introduce new avenues for harm. Artificial Intelligence (AI) is a field of Computer Science whose aim is studying and developing methodologies to build artefacts that can engage in intelligent behaviour. A formal definition of AI that may satisfy everyone does not exist due to the absence of a definition of intelligence. One of the founding fathers of the discipline, Marvin Minsky, defines "artificial intelligence" as "the science of making machines do things that would require intelligence if done by men".[52] As you can notice, this does not provide a clear definition of the discipline but rather defines what artificial means, that is, something done by a machine.

Recently, the High-Level Expert Group on AI ventured a definition: "AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is

48 Caroline Cauffman and Catalina Goanta, 'A New Order: The Digital Services Act and Consumer Protection', *European Journal of Risk Regulation* 12, no. 4 (2021): 767, https://doi.org/10.1017/err.2021.8.

49 Here, it is important to keep in mind that what counts as "controversial" for a xenophobe might be simply called "respect to human rights" for most of us.

50 A particularly gruesome example was the case of Samuel Paty, a French teacher murdered in 2020 after being the target of a social media campaign that, among other issues, publicised his home address: Bahar Makooi, '"The Violence Shook Me Profoundly": Teachers, Students Remember Samuel Paty's Murder', *France 24*, 15 October 2021, sec. france, https://www.france24.com/en/france/20211015-the-violence-shook-me-profoundly-teachers-students-remember-samuel-paty-s-murder.

51 Francesca Lagioia and Giovanni Sartor, 'Artificial Intelligence in the Big Data Era: Risks and Opportunities', in *Legal Challenges of Big Data*, ed. Joe Cannatacci, Valeria Falce, and Oreste Pollicino (Northampton: Edward Elgar, 2020), 280–307.

52 Marvin Minsky, ed., *Semantic Information Processing* (Cambridge, Mass.: MIT Press, 1968), v.

affected by their previous actions".[53] Thus, AI is a wide area that comprehends a heterogeneous set of methodologies that can be divided into two macro-categories: symbolic AI and sub-symbolic AI. The former focuses on top-down approaches that leverage high-level symbolic representation of problems. Symbolic AI is based on logical representation coupled with reasoning processes. This approach makes the functioning of such systems comprehensible to humans, but it has difficulties in scaling up, given the difficulty of capturing complex real-life scenarios through human-generated formalisations. Instead, sub-symbolic AI is based on bottom-up approaches that learn from data how to reach particular objectives. This reliance on machine learning tasks allows sub-symbolic AI to generalise to extraordinarily complex situations, but it requires a huge amount of data to train the systems.

During the last few years, we witnessed the rise of machine learning techniques. Due to the impressive performance that these technologies can get in many different domains, they were also adopted in moderation to filter unwanted content. A machine learning model learns from data a probabilistic model that generalises to unseen scenarios. Let us consider a standard classification model, for instance, one based on a neural network (many models in machine learning are based on neural networks and their variants). A classification model has as many inputs as the number of features representing the input sample, and it has as many outputs as the number of classes or categories. For each sample, the model computes the probability that the input belongs to each class, returning as the model prediction the class with the highest probability. To do that, the model must be trained. During the training phase, the model is fed with samples and the corresponding real label, thus allowing the system to compare its prediction with the correct one and compute the error. This comparison is used to adjust the internal state to minimise the error. If the data is representative of the domain, this process teaches the model how to generalise its predictions also to input that is not seen during the training phase.

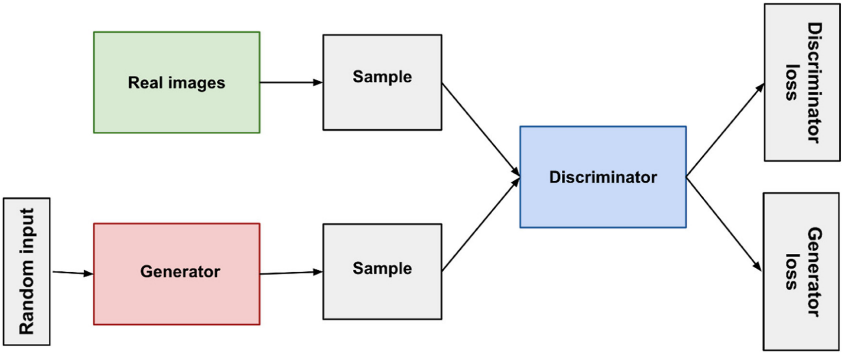
Recently, Generative Adversarial Networks (GAN)[54] have come to the attention of many researchers, practitioners, and to the public audience as an immensely promising tool and very risky threat at the same time. A GAN is a model made by two machine learning modules (usually two

53 AI HLEG, 'Ethics Guidelines for Trustworthy AI', Independent High-Level Expert Group on Artificial Intelligence (Brussels: European Commission, 2019).

54 Ian Goodfellow et al., 'Generative Adversarial Nets', in *Advances in Neural Information Processing Systems*, vol. 27 (NIPS, 2014).

neural networks), one is called the generator, and the other is called the discriminator. The aim of the generator is to produce synthetic data that can be used as an input to the discriminator. The latter aims at identifying whether a given input is fake (i.e., generated by the generator module) or genuine. The generator gets a positive reward when the discriminator is fooled. Similarly, the discriminator gets a positive reward when it correctly classifies an input. During the first part of the training phase, the generator produces low-quality data. Still, if the model is configured correctly and there is enough training data, at the end of the training phase the generator becomes really good at generating data such that it is almost impossible to distinguish fake contents from the real ones. Figure 1 shows a schema of the architecture of a standard GAN.[55]

Figure 1 Overview of a standard GAN schema.



This technology has rapidly spread on the Internet as it generates data for research purposes, data augmentation,[56] or the generation of computational art.[57] Unfortunately, this technology has many nefarious uses. For instance, it is possible to employ the tool to change the tone of a recorded voice to make it resemble somebody else's voice.[58] With some adjustments,

55 https://developers.google.com/machine-learning/gan/gan_structure.

56 Data augmentation refers to the expansion of existing data sets through synthetic data. GANs contribute to this task as they produce "realistic" data, in the sense that the data generated by the network resembles the properties of the original data set.

57 For an example, see the "Dream" application: https://www.wombo.art/.

58 In 2019, this kind of new attack has been used to impersonate the CEO of a company voice and demand a fraudulent transfer: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

this approach can be applied to different media to generate fake videos, images, text, and their combinations.

Generating fake content can be harmful in multiple ways. For example, one can create a false image, or edit an existing one, by generating faces of individuals that do not exist but are nevertheless remarkably similar to real faces.[59] This verisimilitude raises the question of how these contents can be spotted to prevent the spread of fake content,[60] especially considering the potentially harmful uses that can be made of such content. Despite the novelty of these technologies, some examples of harmful uses have already been spotted, such as using real photos of people as source material to generate fake pornographic videos involving those people, which can be used for blackmail or revenge.[61] In these cases, social networks can be both the source of the material used for generating the fakes and the means for potentially spreading the fake content.

Social networks are not—at least in most cases—the producers of this harmful content. They nevertheless play a pivotal role in shaping the various forms through which harm may come to pass in digital environments, both through their decisions regarding which types of content to carry. Accordingly, these networks often rely on content moderation approaches to remove or constrain the reach of potentially harmful content, either to comply with legal requirements or to ensure users are not driven away from their platforms. As they do so, social networks are subject to various legal constraints, which we examine in the following section.

4 Content moderation and the challenges of automation

The term "content moderation" covers a broad range of interventions platforms may adopt towards user-generated content. Some types of intervention are directed at specific content items. For example, a network may take down a post that does not comply with its Terms of Service or add geographical restrictions to content that is lawful in some jurisdictions but not in others. Other interventions target the users that produce unaccept-

59 https://thispersondoesnotexist.com/.

60 Article 52(3) of the AI Act proposal seeks to set up a disclosure requirement: any uses of deep fake must disclose the artificial generation or manipulation of the content.

61 https://www.technologyreview.com/2021/09/13/1035449/ai-deepfake-app-face-swaps-women-into-porn/. For a legal analysis of deep fakes, see Chesney and Citron, 'Deep Fakes'.

able content: banning or suspending them from the network altogether, restricting the visibility of their posts, flagging them with some warning regarding the content of their profile, or adding relevant content to the user's feed to correct or highlight possible disinformation, among other approaches.[62] While the details of each intervention may differ, they all require platforms to adopt a proactive approach to identifying potentially harmful content and responding to it.

Why might social networks want to do so? After all, the eCommerce Directive treats social networks as intermediaries rather than content producers,[63] a decision that restricts their liability for user-generated content. In fact, the general rule is that social networks can only be held liable for this kind of content if they fail to act expeditiously after learning that a user is using the network to store illegal information or conduct illegal activities.[64] Since, as a rule, they are not required to actively pursue this knowledge,[65] social networks are exempted from most forms of liability regarding harms produced through them.

Yet, content moderation is a sensible practice even in the absence of an obligation to that effect. From a business perspective, users might be less inclined to remain in a social network in which they are exposed to scams, hate speech, toxic debates, and other forms of harmful content. By fostering a healthy online environment,[66] content moderation allows networks to offer users a more interesting value proposition, thus retaining their engagement and content production. But the implementation of moderation policies requires a more proactive position regarding user content, thus raising questions on whether the social network is a mere host of user-generated content—and thus excluded from liability—or a co-creator that can be held liable by harms ensuing from that content.

---

62 Social networks may also exercise controls toward the content that is provided to each specific user, for example by ensuring a diversity of viewpoints to avoid filter bubbles. Full coverage of this topic would exceed the scope of this chapter, but we point the interested reader towards Almada, Maranhão, and Sartor, 'Content Personalisation'; Lucien Heitz et al., 'Benefits of Diverse News Recommendations for Democracy: A User Study', *Digital Journalism* 0, no. 0 (8 February 2022): 1–21, https://doi.org/10.1080/21670811.2021.2021804.

63 Sartor and Loreggia, 'The Impact of Algorithms for Online Content Filtering', 30–31.

64 Article 14(1) eCommerce Directive. This provision is retained in Article 5(1) DSA.

65 Article 15 of the eCommerce Directive, preserved in Article 7 DSA.

66 Sartor and Loreggia, 'The Impact of Algorithms for Online Content Filtering', sec. 2.1.

Under current CJEU case law, hosting services—such as social networks—only become liable for content if they turn out to play "an active role of such a kind as to give it knowledge of, or control over, the data" they host.[67] It is not *prima facie* implausible to say that moderation gives platforms control over specific items of user-generated content, as they may decide whether any such item remains available or not.[68] But even if one is willing to grant this point, such control would only exist with regard to the small fraction of user-generated content that is effectively moderated, not to their operations as a whole.[69] Furthermore, holding networks liable due to moderation would substantially reduce a network's incentives to address online harms, as a strong legal pull towards inaction would counter the business rationales described above. Instead, the European Commission has adopted a "good Samaritan" approach, which acknowledges that addressing some categories of harm requires proactive measures and considers this activity is not enough, in itself, to remove the liability exemption.[70] To consolidate this possibility, Article 6 of the DSA explicitly states that voluntary own-initiative investigations for complying with legal requirements do not render a network ineligible for the liability exemption. We welcome this provision, as it increases legal certainty regarding proactive content moderation, thus contributing to a safer online environment.

This is not to say there are not several fault lines between content moderation and the framework of the eCommerce Directive. The first challenge for regulation is determining the proper scope of content moderation. Current instruments oblige platforms to remove illegal content, as liability exemptions only apply when platforms expeditiously remove illegal content or activities they are made aware of.[71] However, online harm is not solely produced by unlawful activity: for example, users may

67 *L'Óreal* (Case C-324/09), para. 116.

68 After all, the liability exclusion in Article 14(1) eCommerce Directive does not apply if service providers fail to act against unlawful content they know about.

69 Increasing the share of content that undergoes moderation, in turn, might be problematic, given the prohibition of general monitoring duties under Article 15 eCommerce Directive.

70 Sartor and Loreggia, 'The Impact of Algorithms for Online Content Filtering', 30–31.

71 Article 14(1)(b) eCommerce Directive conditions the protection from liability to the expeditious removal (or disabling) of unlawful content. Article 17(4)(c) of the Copyright Directive and Article 3 of the Terrorist Content Regulation establish similar duties, but with additional obligations a platform must follow after removal to preserve their protection from liability.

engage in toxic debate, even within the reasonable limits of their freedom of expression, as a result of political polarisation[72] or other forms of echo chambers.[73] Users may also be harmed not by a single post, but by the cumulative product of various lawful practices.[74] To the extent platforms currently address such lawful harms, they do so based on their Terms of Service rather than any general empowerment stemming from the law. As a result, there are several questions about the legitimacy of platforms grounding their content moderation decisions—which impact fundamental rights, notably freedom of expression—on private law instruments,[75] especially considering such instruments are notoriously opaque to the end-user.[76] The moderation of lawful content may thus be a source of tension between users, platforms, and the legal system.

Issues also appear when content moderation follows legal requirements. The eCommerce Directive and the Copyright Directive both require social networks to act "expeditiously" when it comes to unlawful content. Still, the definition of what counts as expeditious action is left to each Member State. For example, Germany's *NetzDG* requires the removal of manifestly illegal content within 24 hours of receiving notice.[77] This tendency to

---

72 See, e.g., Mathias Osmundsen et al., 'Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter', *American Political Science Review* 115, no. 3 (2021): 999–1015, https://doi.org/10.1017/S0003055421000290; Richard Fletcher, Alessio Cornia, and Rasmus Kleis Nielsen, 'How Polarized Are Online and Offline News Audiences? A Comparative Analysis of Twelve Countries', *The International Journal of Press/Politics* 25, no. 2 (1 April 2020): 169–95, https://doi.org/10.1177/1940161219892768.

73 See, e.g., C. Thi Nguyen, 'Echo Chambers and Epistemic Bubbles', *Episteme* 17, no. 2 (June 2020): 141–61, https://doi.org/10.1017/epi.2018.32.

74 For a case study on this kind of harm, see Burkhard Schafer, 'Death by a Thousand Cuts: Cumulative Data Effects and the Corbyn Affair', *Datenschutz und Datensicherheit - DuD* 45, no. 6 (1 June 2021): 385–90, https://doi.org/10.1007/s11623-021-1456-8.

75 For an introduction to such critiques, see Naomi Appelman, João Pedro Quintais, and Ronan Fahy, 'Using Terms and Conditions to apply Fundamental Rights to Content Moderation: Is Article 12 DSA a Paper Tiger?', *Verfassungsblog* (blog), 1 September 2021, https://verfassungsblog.de/power-dsa-dma-06/; Cauffman and Goanta, 'A New Order', 768. On the legitimacy issues stemming from regulation by code, see Laurence Diver, *Digisprudence: Code as Law Rebooted* (Edinburgh: Edinburgh University Press, 2021).

76 Lippi et al., 'CLAUDETTE'.

77 NetzDG, § 3 para. 2, n. 2. Note, however, that this timeframe is not applicable to all content, but only to items in which unlawfulness can be assessed without an in-depth examination: see Zurth, 'The German NetzDG as Role Model or Cautionary Tale?', 1113.

narrow timeframes is also seen in the deadlines set at the EU level, notably in the one-hour deadline for giving effect to a removal order relating to terrorist content.[78] Social networks are thus required to make decisions within a very narrow timeframe, a duty they largely comply with. This compliance, however, introduces risks not only for the workers involved in the moderation process, who may be subject to excessive pressure,[79] but also to the proper assessment of the fundamental rights of the users in particular cases.

Content moderation arrangements must also cope with a broad range of requirements to remove specific types of content. One of the key ideas behind the current regulatory platform is that information society services cannot be subject to any general obligation to moderate the content they carry or actively pursue facts or circumstances relating to illegal activity. In one form or another, this prohibition appears in all EU instruments on social networks.[80] Still, the notion of a "general obligation" is not seen as incompatible with various monitoring duties, some of them constructed very broadly. Within the regulatory sub-system defined by the Copyright Directive, social networks are required to not only remove specific content items deemed to violate copyright protection but also to ensure the unavailability of some works even before there is any complaint[81] and to prevent future uploads of any content deemed to be equivalent to a content item already subject to a removal order.[82] Member State courts have ordered similar measures under the general regime of the eCommerce Directive, mandating the remotion of any content equivalent to specific posts which were deemed unlawful, and the CJEU has found such decisions do not amount to a general obligation to remove content.[83] Furthermore, even the duty to remove "equivalent" content would not amount to a general duty of removal, as platforms are required to remove only content items that can be deemed equivalent to the target of the original order without an in-depth assessment.[84] Social networks can thus be obliged, by

78 Article 3(3) Terrorist Content Regulation.

79 See, e.g., Queenie Wong, 'Facebook Content Moderation Is an Ugly Business. Here's Who Does It', CNET, 19 June 2019, https://www.cnet.com/tech/mobile/facebook-content-moderation-is-an-ugly-business-heres-who-does-it/.

80 See, e.g., Article 15(1) eCommerce Directive, Article 17(8) Copyright Directive, Article 5(8) Terrorism Content Regulation.

81 Article 17(4)(b) Copyright Directive.

82 Article 17(4)(c) Copyright Directive.

83 *Glawischnig-Piesczek* (Case C-18/18), paras. 31–37.

84 *Glawischnig-Piesczek* (Case C-18/18), paras. 38–47. For an in-depth analysis of the decision, see Keller, 'Facebook Filters, Fundamental Rights, and the CJEU's

legislation and courts, to actively pursue specific kinds of content in *all* posts made in a platform, so long as this duty is defined in narrow enough terms to avoid the label of a "general obligation".

Social networks have adopted multiple approaches to the sources of strain described above, which share two major features. When it comes to choosing the means for moderation, platforms are increasingly relying on automated tools, such as systems based on machine learning.[85] This turn is partially driven by other factors, such as the Covid-19 pandemic[86] or the growing capabilities of natural language processing systems. However, it is also a response to legal demands,[87] as using AI technologies may be *de facto* unavoidable to evaluate a large amount of content potentially covered by broad-but-technically-not-general monitoring obligations.[88] Faced with such demands, platforms have embraced the promise of efficiency represented by automated moderation techniques.

Despite its immense potential, automation of content moderation practices may fail to deliver satisfactory results in practice. Sometimes, these failures stem from technical limitations of the existing technologies available for moderation. One of the first applications of automation to moderation relies on the fixed representation of contents of interest—e.g., copyrighted, unlawful, or specific harmful content items—, using these representations to compare new information from digital platforms to find unwanted data. This goal can be achieved through various techniques, such as blacklists, fingerprinting, hash-functions, which aim at creating a fixed and unique representation of input. When two inputs have the same representation, they are deemed to refer to the same content. Unfortunate-

---

Glawischnig-Piesczek Ruling'. Drawing from this rationale, Advocate General Øe has argued that Article 17(4) of the Copyright Directive provides sufficient safeguard to freedom of expression online, thus recommending the dismissal of the action for annulment Poland has proposed with regard to this provision (Case C-401/19). As of February 2022, the CJEU has not ruled on the matter.

85 Robert Gorwa, Reuben Binns, and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', *Big Data & Society* 7, no. 1 (2020), https://doi.org/10.1177/2053951719897945.

86 Tarleton Gillespie, 'Content Moderation, AI, and the Question of Scale', *Big Data & Society* 7, no. 2 (2020): 2, https://doi.org/10.1177/2053951720943234.

87 This is the case even though none of the applicable directives and regulations mandate the use of automated moderation techniques. In fact, Article 5(8) Terrorist Content Regulation explicitly states compliance with the specific measures required under the remainder of this article does not require the adoption of automated tools.

88 Gillespie, 'Content Moderation, AI, and the Question of Scale', 2.

ly, it is quite easy to fool these approaches, as simple and minor changes in the input lead to different representations.[89]

As the moderation problems become more complex, AI technologies face additional challenges. For example, posts on social networks often involve parodies, jokes, memes, and other humoristic content, but humour is a very contextual form of human communication that current linguistic models do not capture well.[90] As such, automated filters may produce erroneous results in dealing with uses of humour within posts, and those errors may, in turn, impinge upon the rights of platform users.[91] There is also the risk that automatic filters produce discriminatory decisions[92] or produce other forms of harm.[93] To address such risks, EU legislation

89 For assessments of technologies used for content filtering, see Felipe Romero Moreno, '"Upload Filters" and Human Rights: Implementing Article 17 of the Directive on Copyright in the Digital Single Market', *International Review of Law, Computers & Technology* 34, no. 2 (3 May 2020): 153–82, https://doi.org/10.1080/13600869.2020.1733760; Sartor and Loreggia, 'The Impact of Algorithms for Online Content Filtering'. For a case study, see Hal Abelson et al., 'Bugs in Our Pockets: The Risks of Client-Side Scanning', *ArXiv:2110.07450 [Cs]*, 14 October 2021, http://arxiv.org/abs/2110.07450.

90 For a primer on the difficulties in automating humour, see also Julia Taylor Rayz and Victor Raskin, 'Fuzziness and Humor: Aspects of Interaction and Computation', in *Fuzzy Techniques: Theory and Applications*, ed. Ralph Baker Kearfott et al., Advances in Intelligent Systems and Computing (Cham: Springer International Publishing, 2019), 655–66, https://doi.org/10.1007/978-3-030-21920-8_58; Tony Veale, *Your Wit Is My Command Building AIs with a Sense of Humor* (The MIT Press, 2021).

91 On online humour as a legal problem, see Joao Paulo Capelotti, 'The Dangers of Controlling Memes through Copyright Law', *The European Journal of Humour Research* 8, no. 3 (12 October 2020): 115–36, https://doi.org/10.7592/EJHR2020.8.3.Capelotti; Renata Vaz Shimbo and Marco Almada, 'A Robot and a Moderator Walk into a Bar: The Use of AI in Online Moderation of Humoristic Content' (Artificial Intelligence: The New Frontier of Business and Human Rights, The Hague: T.M.C. Asser, 2021).

92 On algorithmic discrimination, see, *inter alia*, Alexander Tischbirek, 'Artificial Intelligence and Discrimination: Discriminating Against Discriminatory Systems', in *Regulating Artificial Intelligence*, ed. Thomas Wischmeyer and Timo Rademacher (Cham: Springer International Publishing, 2020), 103–21, https://doi.org/10.1007/978-3-030-32361-5_5; Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI', *Computer Law & Security Review* 41 (July 2021), https://doi.org/10.1016/j.clsr.2021.105567.

93 For an assessment of the shortcomings of large language models, see Emily M. Bender et al., 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability,*

has increasingly added safeguards regarding the use of automation in social network contexts, such as requiring platforms to disclose the use of content moderation algorithms[94] and removing certain kinds of decisions from the reach of automation.[95] Consequently, even advanced AI techniques are not a sure-fire response to content moderation challenges.

Regardless of the extent to which they automate content moderation procedures, social networks face a strategic challenge: how much content should they remove? As examined above, failure to remove unlawful content in a timely fashion may expose platforms to liability for user-generated content. But, in some contexts, determining the lawfulness of a content item might not be a straightforward task. For example, moderators might find themselves needing to evaluate whether a post by a user is an anti-immigration discourse or, in fact, a satire against this kind of discourse.[96] Since the decision on whether a content item should or not stay up must be taken in a short window of time, moderators often tend to engage in *over-removal*, that is, in the removal of any content items that have anything beyond a minimal probability of being unlawful.[97] In doing so, they

---

and Transparency, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.

94 At the EU level, Article 7(1) and 7(3) of the Terrorist Content Regulation require that platforms be transparent about their use of automated tools for moderation, a duty Article 23(1)(c) of the DSA would extend to platforms in general. In addition, Article 15(2)(c) DSA establishes a duty to explain the role automated means played in a specific decision. The Parliament position at first reading broadens this requirement by replacing "decision" with "action", thus encompassing all uses of AI as a guide for moderation practices.

95 Article 17(5) DSA proposal precludes the automation of decisions about complaints submitted by users to the platform.

96 In New Year's Day, 2018, the German comedian Sophie Passmann made a post mocking the national tradition of airing "Dinner for One" on TV, which was taken down after it was construed as a joke targeted at immigrants: Kristen Chick and Sara Miller Llana, 'Is Germany's Bold New Law a Way to Clean up the Internet or Is It Stifling Free Expression?', *Christian Science Monitor*, 8 April 2018, https://www.csmonitor.com/World/Europe/2018/0408/Is-Germany-s-bold-new-law-a-way-to-clean-up-the-internet-or-is-it-stifling-free-expression.

97 As an example, YouTube's first transparency report found that more than 60% of the disputed claims on copyright it adjudicated in the first half of 2021 were resolved in favour of the claimant, meaning that the original decision to remove the content item was unwarranted: 'YouTube Copyright Transparency Report H1 2021' (YouTube, December 2021), https://blog.youtube/news-and-events/access-all-balanced-ecosystem-and-powerful-tools/. However, general evidence on over-removal is hard to come by, given the various challenges in collecting and assessing metrics on content moderation: Daphne Keller and Paddy Leerssen, 'Facts and

reduce the risk of non-compliance with legal requirements while acting within the margin of the discretion afforded by the network's Terms of Service.

Over-removal is a risk-mitigating strategy for social networks, but it may affect users by impinging on their freedom of expression. If moderators are likely to remove content at the slightest whiff of a problem, users might be prompted to self-censorship, as users try to avoid posts that might cause problems with moderators.[98] This is particularly true in cases where platforms do not offer clear mechanisms for questioning or obtaining information about removals; in these cases, a user can either accept the removal decision or seek to strike it down through judicial means, in a procedure that takes much more time than the original decision-making by the network.[99] Without clear guidance on acceptable content or channels to contest removal decisions,[100] users thus find themselves at the mercy of opaque decision-making by platforms.

As it reforms the social network regulatory framework, the EU addresses the concerns mentioned above through the digital constitutionalist turn mentioned in Section 2. Separation of powers is translated to the context of content moderation by the creation of procedural requirements for moderation decisions, such as the need to provide internal channels for receiving complaints about takedown decisions[101] and the information

---

Where to Find Them: Empirical Research on Internet Platforms and Content Moderation', in *Social Media and Democracy: The State of the Field, Prospects for Reform*, ed. Joshua A. Tucker and Nathaniel Persily, SSRC Anxieties of Democracy (Cambridge: Cambridge University Press, 2020), 220–51.

98 Yenn Lee and Alison Scott-Baumann, 'Digital Ecology of Free Speech: Authenticity, Identity, and Self-Censorship', ed. Simeon Yates and Ronald E. Rice (Oxford University Press, 2020).

99 As of 2021, German courts took about 680 days to reach a decision in cases relating to takedown decisions: Jacob Mchangama, Natalie Alkiviadou, and Raghav Mendiratta, 'Rushing to Judgment: Are Short Mandatory Takedown Limits for Online Hate Speech Compatible with The Freedom of Expression?' (Copenhagen: Justitia, 2021).

100 There is, however, some non-binding guidance in the form of private standards and the EU Code of Conduct on countering illegal hate speech online, which counts with the participation of several of the largest platforms currently in operation: Didier Reynders, 'Countering Illegal Hate Speech Online. 6th Evaluation of the Code of Conduct', Factsheet (Brussels: European Commission, 7 October 2021).

101 Article 10 Terrorist Content Regulation and Article 17(9) of the Copyright Directive. Article 17 of the DSA extends this obligation beyond the scope of these legal instruments.

that a user needs to appeal against a decision,[102] such as the role automated systems played in it.[103] These procedural guarantees are accompanied by substantive limits to what platforms can do, as case law[104] and *lex specialis*[105] require platforms to protect the fundamental rights of their users. This duty of care is consolidated in the DSA, which renders it applicable to all forms of content moderation and commands platforms to interpret their Terms of Service in light of the protection of fundamental rights.[106] In the case of very large platforms, the DSA introduces a risk-based approach, under which platforms mitigate risks to these rights that may stem during their operation.[107] Finally, the Parliament position at first reading includes a new paragraph into Article 6 DSA, which requires voluntary own-initiative moderation to be "effective and specific", including a broad set of safeguards to "demonstrate that those investigations and measures are accurate, non-discriminatory, proportionate, transparent and do not lead to over-removal of content". Therefore, barring a radical change of course by legislators and courts, the future of content moderation in the EU moves towards the protection of fundamental rights in the online environment through substantive and procedural mechanisms.

5 Concluding remarks

The regulation of social networks is not a novel challenge for the European Union's legal order. While the framework established around the eCommerce Directive underwent various changes through case law and specialised legislation, its main tenets remain stable. Platforms are largely

102 Article 17 DSA.

103 Article 15(2)(c) DSA. Article 7(1) and 7(3) of the Terrorist Content Regulation provide a more abstract duty of transparency regarding the use of AI, which is not present in the Copyright Directive but also finds an analogue in.

104 For an overview of the applicable CJEU decisions, see De Gregorio, 'The Rise of Digital Constitutionalism in the European Union', sec. 3.

105 In the Copyright Directive, the opening to fundamental rights appears mostly in recitals (especially Recital 84, which states "this Directive should be interpreted and applied in accordance with those rights and principles"). The Terrorist Content Regulation explicitly refers to fundamental rights in Article 5(3), which sets conditions for the design of specific measures to address terrorist content.

106 Article 12 DSA explicitly states the need to observe fundamental rights in the application of the terms of service.

107 Articles 26 and 27 DSA require very large online platforms to assess and mitigate risks to fundamental rights.

protected from liability stemming from user-generated content, encouraged—but compelled only in a narrow set of cases—to adopt a proactive approach for maintaining a healthy online environment, and enjoy considerable discretion in addressing lawful but undesirable content. In this sense, the DSA promotes continuity within the regulatory regime rather than a radical rupture with the eCommerce Directive.

Nevertheless, the DSA—at least as of February 2022—brings substantial changes to this regulatory regime. Platforms only retain their protections against liability and their normative discretion to the extent they protect users' fundamental rights and ensure transparent and fair procedures for exercises of power such as banning users or removing content items. As a result of these changes, the safeguards introduced by case law and specialised legislation are extended to all aspects of a social network's operation, effectively establishing a duty of care towards users that modulates the exercise of private autonomy by platforms.

This procedural turn in social network regulation is a global phenomenon,[108] partly driven by the increased complexity of the online environment in which these networks operate. Since AI technologies are an important part of this environment, for good and for bad, the Copyright Directive, the Terrorist Content Regulation, and the DSA dedicate some attention to them. On the one hand, these systems are seen as sources of risk, which require tailored techno-social safeguards which cannot be directly provided by binding law, but rather require the active engagement of social networks themselves. On the other hand, the need to process large volumes of data, often in a narrow time frame, turns automation into a *de facto* requirement for legal compliance. There is a risk that legislation and platforms become overly confident on the efficacy of AI tools to monitor harmful content in social networks, as the tools available still face several technical challenges and may also incorporate biases, and consequently may affect human rights. It remains to be seen how this tension will be managed in practice.

Since many of the provisions examined above are directed at the internal procedures of social networks, their effectiveness in protecting users

108 Within the EU itself, see the aforementioned example of NetzDG. Outside the EU, debates around the reform of Section 230 of the Communications Decency Act in the US and the Internet Transparency and Responsibility Bill proposed in Brazil also reflect this trend: Juliano Maranhão et al., 'Nota Técnica sobre Procedimentos de Moderação de Conteúdo' (São Paulo: Instituto Legal Grounds, 10 September 2020), https://institutolgpd.com/blog/nota-tecnica-sobre-procedimentos-de-moderacao-de-conteudo/.

will depend on their implementation through such internal procedures. The examination of specific enforcement structures, many of them defined at the national and sub-national levels, exceeds the scope of this chapter. Nevertheless, we believe these structures would do well only if they are based on a socio-technical approach that understands technologies in terms of the social change they enable. If such an approach is effectively implemented in the internal processes of social networks, we believe that the EU approach which focuses in promoting and directing moderation may succeed in limiting harm to user and encouraging beneficial online interactions.

Acknowledgements

The authors would like to thank Renata Shimbo for her comments on the overarching legal framework and by pointing out the challenges involved in humour moderation.

Marco Almada is a doctoral researcher at the European University Institute, funded by a Fundación Carolina grant.

Andrea Loreggia is an assistant professor at the Department of Information Engineering of the University of Brescia, Italy.

Juliano Maranhão is associate Professor at the University of São Paulo Law School and associate researcher at the University of São Paulo's Center For Artificial Intelligence (C4AI).

Giovanni Sartor is professor in Legal theory and legal informatics at the University of Bologna and the European University Institute of Florence.

CompuLaw, that stands for 'Computable Law', is an ERC Advanced Grant proposal coordinated by Prof. Giovanni Sartor. The project was selected for funding by the European Research Council in the 2019 (G.A. 833647). Motivation for CompuLaw lies in the need for the law to govern intelligent computational entities, and the human-artificial social ecology in which they participate. These entities are so many, so fast, and so ubiquitous that it is impossible for humans to monitor them and anticipate illegal behaviour. The solution envisaged by CompuLaw is to make law computation-oriented. That is, to integrate, map and partially translate legal and ethical requirements into computable representations of legal knowledge and reasoning. The European University Institute (EUI) and the University of Bologna provide the main expertise for the five-year multi-disciplinary project. All the information of the project as well as news and publications are available at https://site.unibo.it/compulaw