15

When the Algorithm Is Not Fully Reliable

The Collaboration between Technology and Humans in the Fight against Hate Speech

Federica Casarosa

15.1 INTRODUCTION

Our lives are increasingly inhabited by technological tools that help us with delivering our workload, connecting with our families and relatives, as well as enjoying leisure activities. Credit cards, smartphones, trains, and so on are all tools that we use every day without noticing that each of them may work only through their internal 'code'. Those objects embed software programmes, and each software is based on a set of algorithms. Thus we may affirm that most of (if not all) our experiences are filtered by algorithms each time we use such 'coded objects'.[1]

15.1.1 *A Preliminary Distinction: Algorithms and Soft Computing*

According to computer science, algorithms are automated decision-making processes to be followed in calculations or other problem-solving operations, especially by a computer.[2] Thus an algorithm is a detailed and numerically finite series of instructions which can be processed through a combination of software and hardware tools: Algorithms start from an initial input and reach a prescribed output, which is based on the subsequent set of commands that can involve several activities, such as calculation, data processing, and automated reasoning. The achievement of the solution depends upon the correct execution of the instructions.[3] However, it is

The contribution is based on the analysis developed within a DG Justice supported project e-NACT (GA no. 763875). The responsibility for errors and omissions remains with the author.

[1] See Ben Wagner 'Algorithmic Regulation and the Global Default: Shifting Norms in Internet Technology' (2016) *Etikk i praksis: Nord J Appl Ethics* 5; Rob Kitchin and Martin Dodge *Code/Space Software and Everyday Life* (MIT Press, 2011).

[2] See Jane Yakowitz Bambauer and Tal Zarsky 'The Algorithm Game' (2018) 94 *Notre Dame Law Review* 1.

[3] The set of instructions can include different type of mathematical operations, ranging from linear equations to polynomial calculations, to matrix calculations, and so forth. Moreover, each instruction

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

important to note that, contrary to the common perception, algorithms are neither always efficient nor always effective.

Under the efficiency perspective, algorithms must be able to execute the instructions without exploiting an excessive amount of time and space. Although technological progress allowed for the development of increasingly more powerful computers, provided with more processors and a better memory ability, when algorithms execute instructions that produce great numbers which exceed the space available in memory of a computer, the ability of the algorithm itself to sort the problems is questioned.

As a consequence, under the effectiveness perspective, algorithms may not always reach the exact solution or the best possible solution, as they may include a level of approximation which may range from a second-best solution,[4] to a very low level of accuracy. In this case, computer scientists use the definition of 'soft computing' (i.e., the use of algorithms that are tolerant of imprecision, uncertainty, partial truth, and approximation), due to the fact that the problems that they are addressing may not be solved or may be solved only through an excessive time-consuming process.[5]

Accordingly, the use of these types of algorithms involves the possibility to provide solutions to hard problems, though these solutions, depending on the type of problems, may not always be the optimal ones. Given the ubiquitous use of algorithms processing our data and consequently affecting our personal decisions, it is important to understand in which occasions we may (or should) not fully trust the algorithm and add a human in the loop.[6]

15.1.2 *The Power of Algorithms*

According to Neyland,[7] we may distinguish between two types of power: one exercised *by* algorithms, and one exercised *across* algorithms. The first one is the traditional one, based on the ability of algorithms to influence and steer particular effects. The second one is based on the fact that 'algorithms are caught up within

can be another algorithm, which increases the level of complexity of the overall procedure. See Erika Giorgini 'Algorithms and Law' (2019) 5 *Italian Law Journal* 144.

[4] A well-known example of this case is the Knapsack problem, where the goal is to select among a number of given items the ones that have the maximum total value. However, given that each item has a weight, the total weight that can be carried is no more than some fixed number X. So, the solution must consider weights of items as well as their value. Although in this case a recursive algorithm can find the best solution, when the number of items increases, the time spent to evaluate all the possible combinations increases exponentially, leading to suboptimal solutions.

[5] See the definition at https://en.wikipedia.org/wiki/Soft_computing accessed 13 March 2020.

[6] Council of Europe 'Algorithms and Human Rights – Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications' (2018) https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html accessed 13 March 2020.

[7] Daniel Neyland, *The Everyday Life of an Algorithm* (Palgrave Macmillan, 2019).

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

a set of relations through which power is exercised'.[8] In this sense, it is possible to affirm the groups of individuals that at different stages play a role in the definition of the algorithm share a portion of power.

In practice, one may distinguish between two levels of analysis. Under the first one, for instance when we digit a query over a search engine, the search algorithm activates and identifies the best results related to the keywords inserted, providing a ranked list of results. These results are based on a set of variables that are dependent on the context of the keywords, but also on the trust of the source,[9] on the previous history of searches of the individual, and so forth. The list of results available will then steer the decisions of the individual and affect his/her interpretation of the information searched for. Such power should not be underestimated, because the algorithm has the power to restrict the options available (i.e., avoiding some content because evaluated as untruthful or irrelevant) or to make it more likely to select a specific option. If this can be qualified as the added value of algorithms able to improve the flaws of human reasoning, which include myopia, framing, loss aversion, and overconfidence,[10] then it also shows the power of the algorithm over individual decision-making.[11]

Under the second level of analysis, one may widen the view taking into account the criteria that are used to identify the search results, the online information that is indexed, the computer scientist that set those variables, the company that distributes the algorithm, the public or private company that uses the algorithm, and the individuals that may steer the selection of content. All these elements have intertwining relationships that show a more distributed allocation of power – and, as a consequence, a subsequent quest for a shared type of accountability and liability systems.

15.1.3 *The Use of Algorithms in Content Moderation*

In this chapter, the analysis will focus on those algorithms that are used for content detection and control over user-generated platforms, the so-called content moderation. Big Internet companies have always used filtering algorithms to detect and

[8] Ibid. at 6.

[9] As, for instance, the well-known algorithm used at the beginning by Google, namely Pagerank. See Larry Page et al. 'The PageRank Citation Ranking: Bringing Order to the Web' (1999) http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf accessed 13 March 2020.

[10] David Stevens 'In defence of "Toma": Algorithmic Enhancement of a Sense of Justice' in Mireille Hildebrandt and Keiran O'Hara (eds.) *Life and the Law in the Era of Data-Driven Agency* (Edward Elgar, 2010), analysing Mireille Hildebrandt, *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology* (Edward Elgar Publishing, 2015).

[11] Kevin Slavin 'How Algorithms Shape Our World' (2011) www.ted.com/talks/kevin_slavin_how_algorithms_shape_our_world.html accessed 13 March 2020; Frank Pasquale 'The Algorithmic Self' (2015) *The Hedgehog Review*, Institute for Advanced Studies in Culture, University of Virginia. Note that this aspect is the premise of so-called surveillance capitalism as defined by Shoshana Zuboff in 'Big Other: Surveillance Capitalism and the Prospects of an Information Civilization' (2015) 30 *Journal of Information Technology* 75.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

classify the enormous quantity of uploaded data daily. Automated content filtering is not a new concept on the Internet. Since the first years of Internet development, many tools have been deployed to analyse and filter content, and among them the most common and known are those adopted for spam detection or hash matching. For instance, spam detection tools identify content received in one's email address, distinguishing between clean emails and unwanted content on the basis of certain sharply defined criteria derived from previously observed keywords, patterns, or metadata.[12]

Nowadays, algorithms that are used for content moderation are widely diffuse, having the advantage of scalability. Such systems promise to make the process much easier, quicker, and cheaper than would be the case when using human labour.[13]

For instance, the LinkedIn network published the update of the algorithms used to select the best matches between employers and potential employees.[14] The first steps of the content moderation are worth describing: at the first step, the algorithms check and verify the compliance of the content published with the platform rules (leading to a potential downgrade of the visibility or complete ban in case of incompliance). Then, the algorithms evaluate the interactions that were triggered by the content posted (such as sharing, commenting, or reporting by other users). Finally, the algorithms weigh such interactions, deciding whether the post will be demoted for low quality (low interaction level) or disseminated further for its high quality.[15]

As the example of the LinkedIn algorithm clearly shows, the effectiveness of the algorithm depends on its ability to accurately analyse and classify content in its context and potential interactions. The capability to parse the meaning of a text is highly relevant for making important distinctions in ambiguous cases (e.g., when differentiating between contemptuous speech and irony).

For this task, the industry has now increasingly turned to machine learning to train their programmes to become more context sensitive. Although there are high expectations regarding the ability of content moderation tools, one should not underestimate the risks of overbroad censorship,[16] violation of the freedom of speech

[12] Thamarai Subramaniam, Hamid A. Jalab, and Alaa Y. Taqa 'Overview of Textual Anti-spam Filtering Techniques' (2010) 5 *International Journal of Physical Science* 1869.

[13] Christoph Krönke 'Artificial Intelligence and Social Media' in Thomas Wischmeyer and Timo Rademacher (eds.) *Regulating Artificial Intelligence* (Springer, 2019).

[14] For a description of the LinkedIn platform, see Jian Raymond Rui 'Objective Evaluation or Collective Self-Presentation: What People Expect of LinkedIn Recommendations' (2018) 89 *Computers in Human Behavior* 121.

[15] See the wider procedure described at https://engineering.linkedin.com/blog/2017/03/strategies-for-keeping-the-linkedin-feed-relevant accessed 13 March 2020.

[16] See, for instance, the wide debate regarding the effectiveness of filtering systems adopted at national level against child pornography. See Yaman Akdeniz *Internet Child Pornography and the Law – National and International Responses* (Routledge, 2016), and T. J. McIntyre and Colin Scott 'Internet Filtering – Rhetoric, Legitimacy, Accountability and Responsibility' in Roger Brownsword and Karen Yeung (eds.) *Regulating Technologies: Legal Futures, Regulatory Frames and Technological Fixes* (Bloomsbury Publishing, 2008).

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

principle, as well as biased decision-making against minorities and non-English speakers.[17] The risks are even more problematic in the case of hate speech, an area where the recent interventions of European institutions are pushing for more human and technological investments of IT companies, as detailed in the next section.

15.2 THE FIGHT AGAINST HATE SPEECH ONLINE

Hate speech is not a new phenomenon. Digital communication may be qualified only as a new arena for its dissemination. The features of social media pave the way to a wider reach of harmful content. 'Sharing' and 'liking' lead to a snowball effect, which allows the content to have a 'quick and global spread at no extra cost for the source'.[18] Moreover, users see in the pseudonymity allowed by social media an opportunity to share harmful content without bearing any consequence.[19] In recent years, there has been a significant increase in the availability of hate speech in the form of xenophobic, nationalist, Islamophobic, racist, and anti-Semitic content in online communication.[20] Thus the dissemination of hate speech online is perceived as a social emergency that may lead to individual, political, and social consequences.[21]

15.2.1 A *Definition of* Hate Speech

Hate speech is generally defined as speech 'designed to promote hatred on the basis of race, religion, ethnicity, national origin' or other specific group characteristics.[22]

[17] Natasha Duarte, Emma Llansó, and Anna Loup 'Mixed Messages? The Limits of Automated Social Media Content Analysis, Proceedings of the 1st Conference on Fairness, Accountability and Transparency' (2018) 81 *PMLR* 106.

[18] Katharina Kaesling 'Privatising Law Enforcement in Social Networks: A Comparative Model Analysis' (2018) *Erasmus Law Review* 151.

[19] Natalie Alkiviadou 'Hate Speech on Social Media Networks: Towards a Regulatory Framework?' (2019) 28 *Information & Communications Technology Law* 19.

[20] See Eurobarometer 'Special Eurobarometer 452 – Media Pluralism and Democracy Report' (2016) http://ec.europa.eu/information_society/newsroom/image/document/2016-47/sp452-summary_en_19666.pdf accessed 13 March 2020. See also Article 19 'Responding to "Hate Speech": Comparative Overview of Six EU Countries' (2018) www.article19.org/wp-content/uploads/2018/03/ECA-hate-speech-compilation-report_March-2018.pdf accessed 13 March 2020.

[21] See European Commission – Press Release 'A Europe That Protects: Commission Reinforces EU Response to Illegal Content Online' 1 March 2018 http://europa.eu/rapid/press-release_IP-18-1169_en.htm accessed 13 March 2020.

[22] Michel Rosenfeld 'Hate Speech in Constitutional Jurisprudence: A Comparative Analysis' (2002–2003) 24 *Cardozo L Rev* 1523; Alisdair A. Gillespie 'Hate and Harm: The Law on Hate Speech' in Andrej Savin and Jan Trzaskowski (eds.), *Research Handbook on EU Internet Law* (Edward Elgar, 2014); Natalie Alkiviadou 'Regulating Internet Hate: A Flying Pig?' (2016) 7 *Journal of Intellectual Property, Information Technology and E-Commerce Law* 3; Oreste Pollicino and Giovanni De Gregorio 'Hate Speech: una prospettiva di diritto comparato (2019) 4 *Giornale di Diritto Amministrativo* 421.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

Although several international treaties and agreements do include hate speech regulation,[23] at the European level, such an agreed-upon framework is still lacking. The point of reference available until now is the Council Framework Decision 2008/913/JHA on Combatting Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law.[24] As emerges from the title, the focus of the decision is the approximation of Member States' laws regarding certain offences involving xenophobia and racism, whereas it does not include any references to other types of motivation, such as gender or sexual orientation.

The Framework Decision 2008/913/JHA should have been implemented by Member States by November 2010. However, the implementation was less effective than expected: not all the Member States have adapted their legal framework to the European provisions.[25] Moreover, in the countries where the implementation occurred, the legislative intervention followed different approaches than the national approaches to hate speech, either through the inclusion of the offence within the criminal code or through the adoption of special legislation on the issue. The choice is not without effects, as the procedural provisions applicable to special legislation may be different to those applicable to offences included in the criminal code.

Given the limited effect of the hard law approach, the EU institutions moved to a soft law approach regarding hate speech (and, more generally, also illegal content).[26] Namely, EU institutions moved toward the use of forms of co-regulation where the Commission negotiates a set of rules with the private companies, under the assumption that the latter will have more incentives to comply with agreed-upon rules.[27]

As a matter of fact, on 31 May 2016, the Commission adopted a Code of Conduct on countering illegal hate speech online, signed by the biggest players in the online market: Facebook, Google, Microsoft, and Twitter.[28] The Code of Conduct requires

[23] Note that the definitions of *hate speech* provided at international level focus on different facets of this concept, looking at content and at the manner of speech, but also at the effect and at the consequences of the speech. See the Rabat Plan of Action adopted by the United Nations in 2013, Annual report of the United Nations High Commissioner for Human Rights, A/HRC/22/17/Add.4.

[24] Council Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law, [2008] O.J. (L 328) 55 (Framework Decision 2008/913/JHA).

[25] European Parliament 'Study on the Legal Framework on Hate Speech, Blasphemy and Its Interaction with Freedom of Expression' (2015) www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU%282015%29536460 accessed 13 March 2020.

[26] See also the recent interventions on fake news and illegal content online, respectively the EU Code of Practice on Disinformation http://europa.eu/rapid/press-release_STATEMENT-19-2174_en.htm accessed 13 March 2020, and Commission Recommendation of 1.3.2018 on measures to effectively tackle illegal content online (C(2018) 1177 final https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online accessed 13 March 2020.

[27] Chris Marsden *Internet Co-regulation – European Law, Regulatory Governance, and Legitimacy in Cyberspace* (Cambridge University Press, 2011).

[28] European Commission Press Release IP/16/1937 'European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech' (May 30, 2016); see also European Commission 'Countering Illegal Hate Speech Online #NoPlace4Hate' (2019) https://ec.europa.eu/

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

that the IT company signatories to the code adapt their internal procedures to guarantee that 'they review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary'.[29] Moreover, according to the Code of Conduct, the IT companies should provide for a removal notification system which allows them to review the removal requests 'against their rules and community guidelines and, where necessary, national laws transposing the Framework Decision 2008/913/JHA'.

As is evident, the approach taken by the European Commission is more focused on the timely removal of the allegedly hate speech than on the procedural guarantees that such private enforcement mechanism should adopt in order not to unreasonably limit the freedom of speech of users. The most recent evaluation of the effects of the Code of conduct on hate speech shows an increased number of notifications that have been evaluated and eventually led to the removal of hate speech content within an ever-reduced time frame.[30]

In order to achieve such results, the signatory companies adopted a set of technological tools assessing and evaluating the content uploaded on their platforms. In particular, they finetuned their algorithms in order to detect potentially harmful content.[31] According to the figures provided by the IT companies regarding the flagged content, human labour alone may not achieve such task.[32] However, such algorithms may only flag content based on certain keywords, which are continuously updated, but they always lag behind the evolution of the language. And, most importantly, they may still misinterpret context-dependent wording.[33] Hate speech is a type of language that is highly context sensitive, as the same word may radically change its meaning if used at different places over time. Moreover, algorithms may be improved and trained in one language, but not in other languages which are less prominent in online communication. As a result, an algorithm that works only through the classifications of certain keywords cannot attain the level of complexity of human language and

newsroom/just/item-detail.cfm?item_id=54300 accessed 13 March 2020. Note that since 2018, five new companies joined the Code of Conduct: Instagram, Google+, Snapchat, Dailymotion and jeuxvideo.com. This brings the total number of companies that are part of the Code of Conduct to nine.

[29] Ibid. at p. 2.

[30] See the Commission Factsheet '5th evaluation of the Code of Conduct', June (2020) https://ec.europa.eu/info/sites/default/files/codeofconduct_2020_factsheet_12.pdf accessed 28 June 2021. In particular, the document highlights that 'on average 90% of the notifications are reviewed within 24 hours and 71% of the content is removed'.

[31] See Sissi Cao 'Google's Artificial Intelligence Hate Speech Detector Has a "Black Tweet" Problem' (*Observer*, 13 August 2019) https://observer.com/2019/08/google-ai-hate-speech-detector-black-racial-bias-twitter-study/ accessed 13 March 2020.

[32] See EU Commission 'Results of the Fourth Monitoring Exercise' https://ec.europa.eu/info/sites/info/files/code_of_conduct_factsheet_7_web.pdf accessed 13 March 2020. The Commission affirms that the testing evaluation provided for little more than 4,000 notifications in a period of 6 weeks, with a focus on only 39 organisations from 26 Member States.

[33] Sean MacAvaney et al. 'Hate Speech Detection: Challenges and Solutions' (2019) 14(8) *PLOS One* 1.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

runs the risk of producing unexpected false positives and negatives in the absence of context.[34]

15.2.2 *The Human Intervention in Hate Speech Detection and Removal*

One of the strategies able to reduce the risk of structural over-blocking is the inclusion of some human involvement in the identification and analysis of potential hate speech content.[35] Such human involvement can take different forms, either internal content checking or external content checking.[36]

In the first case, IT companies allocate to teams of employees the task of verifying the sensitive cases, where the algorithm was not able to single out if the content is contrary to community standards or not.[37] Given the high number of doubtful cases, the employees are subject to a stressful situation.[38] They are asked to evaluate in a very short time frame the potentially harmful content, in order to provide a decision regarding the opportunity to take the content down. This will then provide additional feedback to the algorithm, which will learn the lesson. In this framework, the algorithms automatically identify pieces of potentially harmful content, and the people tasked with confirming this barely have time to make a meaningful decision.[39]

The external content checking instead involves the 'trusted flaggers' – that is, an individual or entity which is considered to have particular expertise and

[34] This is even more problematic in the case of image detection, as the recent case of the publication of the Led Zeppelin cover on Facebook was deemed contrary to community standards due to nudity and sexual images. See Rob Picheta 'Facebook Reverses Ban on Led Zeppelin Album Cover' (CNN, 21 June 2019) www.cnn.com/2019/06/21/tech/facebook-led-zeppelin-album-cover-scli-intl/index.html accessed 13 March 2020. For a wider analysis of the reasons to avoid the ubiquitous use of algorithms for decision-making, see Guido Noto la Diega 'Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information' (2018) 9 JIPITEC 3.

[35] Cambridge Consultants, 'The Use of AI in Content Moderation' (2019) www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf accessed 13 March 2020.

[36] James Grimmelmann 'The Virtues of Moderation' (2015) 17 *Yale J.L. & Tech.* 42.

[37] See the approach adopted by Facebook and Google in this regard: Issie Lapowsky 'Facebook Moves to Limit Toxic Content as Scandal Swirls' (Wired, 15 November 2018) www.wired.com/story/facebook-limits-hate-speech-toxic-content/ accessed 13 March 2020.; Sam Levin 'Google to Hire Thousands of Moderators after Outcry over YouTube Abuse Videos' (*The Guardian*, 5 December 2017), www.theguardian.com/technology/2017/dec/04/google-YouTube-hire-moderators-child-abuse-videos accessed 13 March 2020.

[38] Nicolas P. Suzor *Lawless: The Secret Rules That Govern Our Digital Lives (and Why We Need New Digital Constitutions That Protect Our Rights*) (Cambridge University Press, 2019).

[39] Sarah T. Roberts 'Commercial Content Moderation: Digital Laborers' Dirty Work' in S. U. Noble and B. Tynes (eds.) *The Intersectional Internet: Race, Sex, Class and Culture Online* (Peter Lang Publishing, 2016); Ben Wagner 'Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems' (2018) 11 *Policy & Internet* 104; Andrew Arsht and Daniel Etcovitch 'The Human Cost of Online Content Moderation' (2018) *Harvard Law Review* Online https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation accessed 13 March 2020.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

responsibilities for the purposes of tackling hate speech. Examples for such notifiers can range from individual or organised networks of private organisations, civil society organisations, and semi-public bodies, to public authorities.[40]

For instance, YouTube defines *trusted flaggers* as individual users, government agencies, and NGOs that have identified expertise, (already) flag content frequently with a high rate of accuracy, and are able to establish a direct connection with the platform. It is interesting to note that YouTube does not fully delegate the content detection to trusted notifiers but rather affirms that 'content flagged by Trusted Flaggers is not automatically removed or subject to any differential policy treatment – the same standards apply for flags received from other users. However, because of their high degree of accuracy, flags from Trusted Flaggers are prioritized for review by our teams'.[41]

15.3 THE OPEN QUESTIONS IN THE COLLABORATION BETWEEN ALGORITHMS AND HUMANS

The added value of the human intervention in the detection and removal of hate speech is evident; nonetheless, concerns may still emerge as regards such an involvement.

15.3.1 *Legal Rules versus Community Standards*

As hinted previously, both algorithms and humans involved in content detection and removal of hate speech evaluate content vis-à-vis the community standards adopted by each platform. Such distinction is clearly affirmed also in the YouTube trusted flaggers programme, where it is affirmed that 'the Trusted Flagger program exists exclusively for the reporting of possible Community Guideline violations. It is not a flow for reporting content that may violate local law. Requests based on local law can be filed through our content removal form'.

These standards, however, do not fully overlap with the legal definition provided by EU law, pursuant to the Framework Decision 2008/913/JHA.

Table 15.1 shows that the definitions provided by the IT companies widen the scope of the prohibition on hate speech to sex, gender, sexual orientation, disability or disease, age, veteran status, and so forth. This may be interpreted as the achievement of a higher level of protection. However, the width of the definition is not

[40] Flagging is the mechanism provided by platforms to allow users to express concerns about potentially offensive content. This mechanism allows to reduce the volumes of content to be reviewed automatically. See Kate Klonick 'The New Governors: The People, Rules and Processes Governing Online Speech', 131 *Harvard Law Review* 1598, at 1626 (2018).

[41] See 'YouTube Trusted Flagger Program' https://support.google.com/YouTube/answer/7554338?hl=en accessed 13 March 2020.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

TABLE 15.1 *Hate speech as defined by several major IT companies*

Facebook definition[42]	YouTube definition[43]	Twitter definition[44]	Framework Decision 2008/913/JHA
What does Facebook consider to be hate speech? Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability or disease is not allowed. We do, however, allow clear attempts at humour or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (example: jokes, stand-up comedy, popular song lyrics, etc.).	Hate speech refers to content that promotes violence against or has the primary purpose of inciting hatred against individuals or groups based on certain attributes, such as: - race or ethnic origin - religion - disability - gender - age - veteran status - sexual orientation/ gender identity.	*Hateful conduct*: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.	All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin.

always coupled with a subsequent detailed definition of the selected grounds. For instance, the YouTube community standards list the previously mentioned set of attributes, providing some examples of hateful content. But the standard only sets two clusters of cases: encouragement towards violence against individuals or groups based on the attributes, such as threats, and the dehumanisation of individuals or groups (for instance, calling them subhuman, comparing them to animals, insects,

[42] Facebook 'How Do I Report Inappropriate or Abusive Things on Facebook (Example: Nudity, Hate Speech, Threats)' www.facebook.com/help/212722115425932?helpref=uf_permalink accessed 13 March 2020.

[43] Google 'Hate Speech Policy' https://support.google.com/YouTube/answer/2801939?hl=en accessed 13 March 2020.

[44] Twitter 'Hateful Conduct Policy' https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy accessed 13 March 2020.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

pests, disease, or any other non-human entity).[45] The Facebook Community policy provides for a better example, as it includes a more detailed description of the increasing levels of severity attached to three tiers of hate speech content.[46] In each tier, keywords are provided to show the type of content that will be identified (by the algorithms) as potentially harmful.

As a result, the inclusion of such wide hate speech definitions within the Community Guidelines or Standards become de facto rules of behaviour for users of such services.[47] The IT companies are allowed to evaluate a wide range of potentially harmful content published on their platforms, though this content may not be illegal according to the Framework Decision 2008/914/JHA.

This has two consequences. First, there is an extended privatisation of enforcement as regards those conducts that are not covered by legal provisions with the risk of an excessive interference with the right to freedom of expression of users.[48] Algorithms deployed by IT companies will then have the power to draw the often-thin line between legitimate exercise of the right to free speech and hate speech.[49]

Second, the extended notion of harmful content provided by community rules imposes a wide obligation on platforms regarding the flow of communication. This may conflict with the liability regime adopted pursuant relevant EU law, namely the e-Commerce Directive, which imposes a three-tier distinction across intermediary liability and, most importantly, prohibits any general monitoring obligation over ISP pursuant art. 15.[50] As it will be addressed later, in the section on liability, striking the balance between sufficient incentives to block harmful content and over-blocking effects is crucial to safeguard the freedom of expression of users.

[45] Article 19, 'YouTube Community Guidelines: Analysis against International Standards on Freedom of Expression' (2018) www.article19.org/resources/YouTube-community-guidelines-analysis-against-international-standards-on-freedom-of-expression/ accessed 13 March 2020.

[46] Article 19, 'Facebook Community Standards: Analysis against International Standards on Freedom of Expression' (2018) www.article19.org/resources/facebook-community-standards-analysis-against-international-standards-on-freedom-of-expression/ accessed 13 March 2020.

[47] Wolfang Benedek and Matthias C. Kettemann *Freedom of Expression and the Internet* (Council of Europe Publishing, 2013), 101. See the decision of Italian courts on this matter, as presented in F. Casarosa, 'Does Facebook get it always wrong? The decisions of Italian courts between hate speech and political pluralism', presented at Cyberspace conference, November 2020.

[48] Council of Europe, Draft Recommendation CM/Rec (2017x)xx of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, MSI-NET (19 September 2017).

[49] National and European courts are still struggling in identifying such boundary; see, for instance, the rich jurisprudence of the ECtHR, European Court of Human Rights Press Unit, Factsheet – Hate Speech (January 2020), www.echr.coe.int/Documents/FS_Hate_speech_ENG.pdf accessed 13 March 2020.

[50] Note that this principle is also confirmed by the Council of Europe (n 48).

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

15.3.2 *Due Process Guarantees*

As a consequence of the previous analysis, the issue of procedural guarantees of users emerges.[51] A first question is related to the availability of internal mechanisms that allow users to be notified about potentially harmful content, to be heard, and to review or appeal against the decisions of IT companies. Although the strongest position safeguarding freedom of expression and fair trial principle would suggest that any restriction (i.e., any removal of potentially harmful content) should be subject to judicial intervention,[52] the number of decisions adopted on a daily basis by IT companies does not allow either the intervention of potential victims and offenders, or the judicial system. It should be noted that the Code of Conduct does not provide for any specific requirement in terms of judicial procedures, nor through alternative dispute resolution mechanisms, thus it is left to the IT companies to introduce an appeal mechanism.

Safeguards to limit the risk of removal of legal content are provided instead in the Commission Recommendation on Tackling Illegal Content Online,[53] which includes within the wider definition of illegal content also hate speech.[54] The Recommendation points to automated content detection and removal and underlines the need for counter-notice in case of removal of legal content. The procedures involve the exchange between the user and the platform, which should provide a reply: in case of evidence provided by the user that the content may not be qualified as illegal, the platform should restore the content that was removed without undue delay or allow for a re-upload by the user; whereas, in case of a negative decision, the platform should include reasons for said decision.

Among the solutions, the signatories to the Code of Conduct proposed Google provides for a review mechanism, allowing users to present an appeal against the decision to take down any uploaded content.[55] Then, the evaluation of the justifications provided by the user is processed internally and the final decision is sent afterward to the user, with limited or no explanation.

A different approach is adopted by Facebook. In September 2019, the social network announced the creation of an 'Oversight Board'.[56] The Board has the task of providing the appeals for selected cases that address potentially harmful content.

[51] Giancarlo Frosio 'Why Keep a Dog and Bark Yourself? From Intermediary Liability to Responsibility' (2018) 26 *Oxford Int'l J. of Law and Information Technology* 1.

[52] See, for instance, the suggestion made by UN Rapporteur Frank La Rue, in Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2011) www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf, p. 13 accessed 13 March 2020.

[53] Commission Recommendation 2018/334 on measures to effectively tackle illegal content online, C/2018/1177, OJ L 63, 6.3.2018, pp. 50–61

[54] Ibid., at 3.

[55] See Google 'Appeal Community Guidelines Actions' https://support.google.com/YouTube/answer/185111 accessed 13 March 2020.

[56] For a detailed description of the structure and role of the Oversight Board, see Facebook 'Establishing Structure and Governance for an Independent Oversight Board' (Facebook Newsroom, 17 September 2019) https://newsroom.fb.com/news/2019/09/oversight-board-structure/ accessed

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

Although the detailed regulation concerning the activities of the board is still to be drafted, it is clear that it will not be able to review all the content under appeal.[57] Although this approach has been praised by scholars, several questions remain open: the transparency in the selection of the people entrusted with the role of adjudication, the type of explanation for the decision taken, the risk of capture (in particular for the oversight board), and so on. And, at the moment, these questions are still unanswered.

15.3.3 *Selection of Trusted Flaggers*

As mentioned previously in Section 15.2.2., the intervention of trusted flaggers in content detection and removal became a crucial element in order to improve the results of said process. The selection process to identify and recruit trusted flaggers, however, is not always clear.

According to the Commission Recommendation, the platforms should 'publish clear and objective conditions' for determining which individuals or entities they consider as trusted flaggers. These conditions include expertise and trustworthiness, and also 'respect for the values on which the Union is founded as set out in Article 2 of the Treaty on European Union'.[58]

Such a level of transparency does not match with the practice: although the Commission Monitoring exercise provides for data regarding at least four IT companies, with a percentage of notifications received by users vis-à-vis trusted flaggers as regards hate speech,[59] apart from the previously noted YouTube programme, none of the other companies provide a procedure for becoming a trusted flagger. Nor is any guidance provided on whether the selection of trusted notifiers is a one-time accreditation process or rather an iterative process whether the privilege is monitored and can be withdrawn.[60]

13 March 2020, and Facebook 'Oversight Board Charter' (Facebook Newsroom, 19 September 2019) https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf accessed 13 March 2020.

[57] The figures can clarify the challenge: the number of board members is currently set at 40 people, while the number of cases under appeal yearly by Facebook is 3.5 million (only related to hate speech), according to the 2019 Community Standards Enforcement Report https://transparency.facebook.com/community-standards-enforcement#hate-speech accessed 13 March 2020.

[58] Commission (2018) Recommendation 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, C/2018/1177, OJ L 63, 6.3.2018, pp. 50–61.

[59] See also the figures provided in Commission Factsheet, 'How the Code of Conduct Helped Countering Illegal Hate Speech Online', February (2019) https://ec.europa.eu/info/sites/info/files/hatespeech_infographic3_web.pdf accessed 13 March 2020. The Commission report affirms that 'The IT companies reported a considerable extension of their network of 'trusted flaggers' in Europe and are engaging on a regular basis with them to increase understanding of national specificities of hate speech. In the first year after the signature of the Code of conduct, Facebook reported to have taken 66 EU NGOs on board as trusted flaggers; and Twitter 40 NGOs in 21 EU countries.'

[60] Sebastian Schwemer 'Trusted Notifiers and the Privatization of Online Enforcement' (2018) 35 *Computer Law & Security Review*.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

This issue should not be underestimated, as the risk of rubberstamping the decisions of trusted flaggers may lead to over-compliance and excessive content takedown.[61]

15.3.4 *Liability Regime*

When IT companies deploy algorithms and recruit trusted flaggers in order to proactively detect and remove potentially harmful content, they may run the risk of losing their exemption of liability according to the e-Commerce Directive.[62] According to art. 14 of the Directive, hosting providers are exempted from liability when they meet the following conditions:

– Service providers provide only for the storage of information at the request of third parties;
– Service providers do not play an active role of such a kind as to give it knowledge of, or control over, that information.

According to the decision of the CJEU in *L'Oréal* v. *eBay*,[63] the Court of Justice clarified that whenever an online platform provides for the storage of content (in the specific case offers for sale), sets the terms of the service, and receives revenues from such service, this does not change the position of the hosting provider denying the exemptions from liability. In contrast, this may happen when the hosting provider 'has provided assistance which entail, in particular optimising the presentation of the offers for sale in question or promoting those offers'.

This indicates that the active role of the hosting provider is only to be found when it intervenes directly in user-generated content.[64] If the hosting provider adopts technical measures to detect and remove hate speech, does it fail its neutral position vis-à-vis the content?

The liability exemption may still apply only if two other conditions set by art. 14 e-Commerce Directive apply. Namely,

[61] Note that evidence from the SCAN project highlights that removal rates differs between the reporting channels used to send the notification, with an average of 15 per cent higher, with the exceptional case of Google+, where all the notified cases were accepted by the company. See SCAN 'Diverging Responsiveness on Reports by Trusted Flaggers and General Users – 4th Evaluation of the EU Code of Conduct: SCAN Project Results' (2018) http://scan-project.eu/wp-content/uploads/2018/08/sCAN_monitoring1_fact_sheet_final.pdf accessed 13 March 2020.

[62] 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, [2000] O.J. (L 178) 1, 16 (e-Commerce Directive). Note that the proposed Digital Services Act, COM(2020) 825 final, confirms that providers of intermediary services are not subject to general monitoring obligations.

[63] Case 324/09 *L'Oréal SA and Others* v. *eBay International* AG *and Others* [2011] ECR I-06011.

[64] Christina Angelopoulos et al. 'Study of Fundamental Rights Limitations for Online Enforcement through Self-Regulation' (2016) https://openaccess.leidenuniv.nl/handle/1887/45869 accessed 13 March 2020.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

- hosting providers do not have actual knowledge of the illegal activity or information and, as regards claims for damages, are not aware of facts or circumstances from which the illegal activity or information is apparent; or
- upon obtaining such knowledge or awareness, they act expeditiously to remove or to disable access to the information.

It follows that proactive measures taken by the hosting provider may result in that platform obtaining knowledge or awareness of illegal activities or illegal information, which could thus lead to the loss of the liability exemption. However, if the hosting provider acts expeditiously to remove or to disable access to content upon obtaining such knowledge or awareness, it will continue to benefit from the liability exemption.

From a different perspective, it is possible that the development of technological tools may lead to a reverse effect as regards monitoring obligations applied over IT companies. According to art. 15 of the e-Commerce Directive, no general monitoring obligation may be imposed on hosting providers as regards illegal content. But in practice, algorithms may already deploy such tasks. Would this indirectly legitimise monitoring obligations applied by national authorities?

This is the question posed by an Austrian court to the CJEU as regards hate speech content published on the social platform Facebook.[65] The preliminary reference addressed the following case: in 2016, the former leader of the Austrian Green Party, Eva Glawischnig-Piesczek was the subject of a set of posts published on Facebook by a fake account. The posts included rude comments, in German, about the politician, along with her image.[66]

Although Facebook complied with the injunction of the First Instance court across the Austrian country, blocking access to the original image and comments, the social platform appealed against the decision. After the appeal decision, the case achieved the Oberste Gerichtshof (Austrian Supreme Court). Upon analysing the case, the Austrian Supreme Court affirmed that Facebook can be considered as an abettor to the unlawful comments; thus it may be required to take steps so as to repeat the publication of identical or similar wording. However, in this case, the injunction regarding such a pro-active role for Facebook could indirectly impose a monitoring role, which is in conflict not only with art. 15 of the e-Commerce Directive but also with the previous jurisprudence of the CJEU. Therefore, the Supreme Court decided to stay the proceedings and present a preliminary reference to the CJEU. The Court asked, in particular, whether art. 15(1) of the e-Commerce Directive precludes the national court to make an order requiring a hosting provider,

[65] Case C-18/18, *Eva Glawischnig-Piesczek* v. *Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821.

[66] Ms Glawischnig-Piesczek requested Facebook to delete the image and the comments, but it failed to do so. Ms Glawischnig-Piesczek filed a lawsuit before the Wien first instance court, which eventually resulted in an injunction against Facebook, which obliged the social network not only to delete the image and the specific comments, but also to delete any future uploads of the image if it was accompanied by comments that were identical or similar in meaning to the original comments.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

who has failed to expeditiously remove illegal information, not only to remove the specific information but also other information that is identical in wording.[67]

The CJEU decided the case in October 2019. The decision argued that as Facebook was aware of the existence of illegal content on its platform, it could not benefit from the exemption of liability applicable pursuant to art. 14 of the e-Commerce Directive. In this sense, the Court affirmed that, according to recital 45 of the e-Commerce Directive, national courts cannot be prevented from requiring a host provider to stop or prevent an infringement. The Court then followed the interpretation of the AG in the case,[68] affirming that no violation of the prohibition of monitoring obligation provided in art. 15(1) of the e-Commerce Directive occurs if a national court orders a platform to stop and prevent illegal activity if there is a genuine risk that the information deemed to be illegal can be easily reproduced. In these circumstances, it was legitimate for a Court to prevent the publication of 'information with an equivalent meaning'; otherwise the injunction would be simply circumvented.[69]

Regarding the scope of the monitoring activity allocated to the hosting provider, the CJEU acknowledged that the injunction cannot impose excessive obligations on an intermediary and cannot require an intermediary to carry out an independent assessment of equivalent content deemed illegal, so automated technologies could be exploited in order to automatically detect, select, and take down equivalent content.

The CJEU decision tries as much as possible to provide a balance between freedom of expression and freedom to conduct a business, but the wide interpretation of art. 15 of the e-Commerce Directive can have indirect negative effects, in particular when looking at the opportunity for social networks to monitor through technological tools the upload of identical or equivalent information.[70] This

[67] Questions translated by the preliminary reference decision of the Oberste Gerichtshof, OGH, case number 6Ob116/17b.

[68] In his opinion, A. G. Szpunar affirmed that an intermediary does not benefit from immunity and can 'be ordered to seek and identify the information equivalent to that characterised as illegal only among the information disseminated by the user who disseminated that illegal information. A court adjudicating on the removal of such equivalent information must ensure that the effects of its injunction are clear, precise and foreseeable. In doing so, it must weigh up the fundamental rights involved and take account of the principle of proportionality'.

[69] The CJEU then defined information with an equivalent meaning as 'information conveying a message the content of which remains essentially unchanged and therefore diverges very little from the content which gave rise to the finding of illegality' (par. 39).

[70] See Agnieszka Jabłonowska 'Monitoring Duties of Online Platform Operators Before the Court – Case C-18/18 Glawischnig-Piesczek' (6 October 2019) http://recent-ecl.blogspot.com/2019/10/monitoring-duties-of-platform-operators.html; Eleftherios Chelioudakis 'The *Glawischnig-Piesczek* v. *Facebook* Case: Knock, Knock. Who's There? Automated Filters Online' (12 November 2019) www.law.kuleuven.be/citip/blog/the-glawischnig-piesczek-v-facebook-case-knock-knock-whos-there-automated-filters-online/ accessed 13 March 2020; Marta Maroni and Elda Brogi '*Eva Glawischnig-Piesczek* v. *Facebook Ireland Limited*: A New Layer of Neutrality' (2019) https://cmpf.eui.eu/eva-glawischnig-piesczek-v-facebook-ireland-limited-a-new-layer-of-neutrality/ accessed 13 March 2020.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press

approach safeguards the incentives for hosting providers to verify the availability of harmful content without incurring additional levels of liability. However, the use of technical tools may pave the way to additional cases of false positives, as they may remove or block content that is lawfully used, such as journalistic reporting on a defamatory post – thus opening up again the problem of over-blocking.

15.4 CONCLUDING REMARKS

Presently, we are witnessing an intense debate about technological advancements in algorithms and their deployment in various domains and contexts. In this context, content moderation and communication governance on digital platforms have emerged as a prominent but increasingly contested field of application for automated decision-making systems. Major IT companies are shaping the communication ecosystem in large parts of the world, allowing people to connect in various ways across the globe, but also offering opportunities to upload harmful content. The rapid growth of hate speech content has triggered the intervention of national and supranational institutions in order to restrict such unlawful speech online. In order to overcome the differences emerging at the national level and enhance the opportunity to engage international IT companies, the EU Commission adopted a co-regulatory approach inviting the same table regulators and regulates, so as to defined shared rules.

This approach has the advantage of providing incentives for IT companies to comply with shared rules, as long as non-compliance with voluntary commitments does not lead to any liability or sanction. Thus the risk of over-blocking may be avoided or at least reduced. Nonetheless, considerable incentives to delete not only illegal but also legal content exist. The community guidelines and standards presented herein show that the definition of hate speech and harmful content is not uniform, and each platform may set the boundaries of such concepts differently. When algorithms apply criteria defined on the basis of such different concepts, they may unduly limit the freedom of speech of users, as they will lead to the removal of legal statements.

The Commission approach explicitly demands proactive monitoring: ‘Online platforms should, in light of their central role and capabilities and their associated responsibilities, adopt effective proactive measures to detect and remove illegal content online and not only limit themselves to reacting to notices which they receive’. But this imposes de facto monitoring obligations which may be carried out through technical tools, which are far from being without flaws and bias.

From the technical point of view, the introduction of the human in the loop, such as in the cases of trusted flaggers or the Facebook Oversight board, does not reduce the questions of effectiveness, accessibility, and transparency of the mechanisms adopted. Both strategies, however, show that some space for stronger accountability mechanisms can be found, though the path to be pursued is still long.

https://doi.org/10.1017/9781108914857.016 Published online by Cambridge University Press