Algorithmic content moderation and the LGBTQ+ community's freedom of expression on social media: insights from the EU Digital Services Act

Sergio Sulmicelli*

ABSTRACT: The essay explores the use of artificial intelligence as a tool for content moderation by social media platforms and how it poses risks to the freedom of expression of LGBTQ+ individuals, perpetuating online the marginalization experienced by queer individuals offline. The EU's Digital Services Act (DSA) regulation attempts to address this issue by balancing the need for content moderation with safeguards to freedom of expression. However, in order to mitigate the risks to the queer community's freedom of expression, a regulatory approach must include transparency obligations on platforms that use algorithmic content moderation, but also avoid incentivizing social media to adopt a more aggressive and generalized moderation approach to elude the risk of being sanctioned, generating the so-called 'better safe than sorry' effect.

KEYWORDS: Algorithmic content moderation; freedom of expression; LGBTQ+; Artificial Intelligence; Digital Services Act

1. The queer community in the digital space: two narratives and a problem

Digital technologies have revolutionized the way people communicate.[1] It is obvious the benefit that Internet, and in particular social media, has brought to people's freedom of expression[2] having helped people to connect with each other and share ideas, offering a *forum* for sharing information and opinions as an essential part of the democratic function.[3]

Social media platforms, such as Facebook, Instagram, Twitter, and YouTube, are the place where the democratic dialogue takes shape:[4] they have provided individuals with access to information that they

[*] *Dottorando di ricerca, Università di Trento. Mail: sergio.sulmicelli@unitn.it. A first draft of the research was presented at ICON-S "Il Futuro dello Stato", Università di Bologna, September 17, 2022. The article was subject to a double-bind peer review process.*

[1] L. FLORIDI, *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*, Oxford, 2014.

[2] S.J. BRISON, K. GELBER, *Free Speech in the Digital Age*, Oxford, 2019.

[3] H. MARGETTS, *The Internet and Democracy*, in W. Dutton (ed.), *The Oxford Handbook of Internet Technologies*, Oxford, 2013; M. MARGOLIS, G. MORENO-RIANO, *The Prospect of Internet Democracy*, Surrey, 2009.

[4] C.K. JHA, O. KODILA-TEKIDA, *Does social media promote democracy? Some empirical evidence*, in *Journal of Policy Modeling*, 42, 2020, 271-290. This study shows evidence of a "positive impact" deriving from Facebook

otherwise would not have had access to. Social media are also the 'place' where battles for liberty, as well as political and social revolutions, originate and are organized nowadays.[5] At the same time, it is undeniable how this view of the internet as a "freedom-enhancing" and "democracy-enabling" tool, advocated by "digital-evangelists", clashes with a "techno-realistic" point of view according to which internet acts conversely as a "freedom-infringing" and "democracy-undermining" creature.[6]

As a matter of fact, there is increasing evidence of how Internet has negatively impacted the spread of previously recognized types of harmful speech, such as hate speech,[7] and, at the same time, it has given rise to new types of harmful forms of expression such as cyber harassment, revenge porn, cyberstalking[8] which impact more on minority and victimized segments of the population (women, members of the LGBTQ+ community, black people and so on).[9]

penetration (a proxy for social media) on democracy. According to the authors, "the correlation between social media and democracy is even stronger for low-income countries than high-income countries". The empirical study indicates that "a one-standard deviation increase in Facebook penetration is associated with an improvement in the democracy index". More broadly, on the relation between social media and democracy, *see*: E. PRICE, *Social Media and Democracy*, in *Australian Journal of Political Science*, 48, 2013, 519-527.

Some authors have also reflected on the negative impact of social media on democratic discourse, *see*: C.R. SUNSTEIN, *Is Social Media Good or Bad for Democracy?*, in *Meta series on democracy and social media*, Jan. 22, 2018, available at: https://about.fb.com/news/2018/01/sunstein-democracy. ID., *#Republic. Divided Democracy in the Age of Social Media*, Princeton, 2018.

[5] J.A. TUCKER, Y. THEOCHARIS, M.E. ROBERTS, P. BARBERÁ, *From Liberation to Turmoil: Social Media And Democracy*, in *Journal of Democracy*, 28, 2017, 46-59. One of the clearest examples is the role Twitter played in the Arab Spring revolutions in 2011: G. WOLFSFELD, E. SEGEV, T. SHEAFER, *Social Media and the Arab Spring: Politics Comes First*, in *International Journals of Press and Politics*, 2013.

[6] S.J. BRISON, K. GELBER, *op. cit.,* 5. The expressions 'digital evangelists' and 'techno-realists' are owed to: F. COMUNELLO, G. ANZERA, *Will the revolution be tweeted? A conceptual framework for understanding the social media and the Arab Spring*, in *Islam and Muslim Relations*, 4, 2012, 453-470. Some Authors have also reflected on the impact of artificial intelligence systems that, by profiling users, filter and suggest news and information: M. FASAN, *Intelligenza artificiale e pluralismo: uso delle tecniche di profilazione nello spazio pubblico democratico*, in *BioLaw Journal*, 1, 2019, 101-113; C.M. REALE, M. TOMASI, *Libertà d'espressione, nuovi media e intelligenza artificiale: la ricerca di un nuovo equilibrio nell'ecosistema costituzionale*, in *DPCE Online*, 1, 2022, 325 ss.

[7] S.J. BRISON, K. GELBER, *op. cit.,* 5. Hate speech can be understood as "as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin", see Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law, 28 Nov., 2008, available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:328:0055:0058:en:PDF.

[8] *Ibidem*. See also: D.K. CITRON, *Hate Crimes in Cyberspace*, Cambridge, 2014.

[9] E. JANE, *'Your a Ugly, Whorish, Slut': Understanding e-bile*, in *Feminist Media Studies*, 14, 2014, 531-546; S. FYFE, *Tracking hate speech acts as incitement to genocide in international criminal law*, in *Leiden Journal of International Law*, 30, 2017, 523–548; J. DANIEL, *Race and Racism in Internet Studies*, in *New Media & Society*, 15, 2012, 695-719; L. HUBBARD, *Online Hate Crime Report: Challenging online homophobia, biphobia and transphobia*, London, 2020; I. AWAN, I. ZEMPI, *The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts*, in *Aggression and violent behavior*, 27, 2016, 1-8.

Downloaded from www.biodiritto.org.
ISSN 2284-4503

To this point, while the digital world was first celebrated as a location where the "free marketplace of ideas"[10] could be encouraged and facilitated,[11] the new obstacles and risks posed by these platforms were quickly recognized. The necessity for action on the spread of online hate speech, in particular, has become increasingly necessary, and social media companies have taken steps to filter the content uploaded online.[12] These challenges soon proved particularly challenging for the online life of marginalized people, such as the queer community.[13]

Indeed, in order to provide a response to harmful content posted online, most social media platforms adopt algorithms that automatically filter or flag content that is deemed to be inappropriate.[14] However, these algorithms often fail to properly identify content that is aimed to offend LGBTQ+ people,[15] leading to a feeling of isolation and exclusion for members of the queer community who use social media.

Moreover, the tension between preventing online harmful speech and safeguarding freedom of expression[16] is proved by the conflict between content moderation operated by digital platforms and the risk that this form of control ends up affecting the LGBQT+ community's expression online,[17] the same

[10] The marketplace of ideas is a concept that was first proposed by John Stuart Mill in the 19th century. It is the idea that the truth will eventually prevail if all ideas are freely available and open to debate. This concept has been eventually adopted by Justice Oliver Wendell Holmes's dissenting in *Abrams v United States* 250 US 616, 624 (1919) to interpret the First amendment of the U.S. Constitution. See J.S. MILL, *On* Liberty, London, 1859. *Ex* multis, D.C. NUNZIATO, *The Marketplace of ideas online*, in *Notre Dame Law Review*, 94, 2018, 1520-1546.

[11] The U.S. Supreme Court itself, based on this doctrine, has seen Internet as a 'facilitator' of the free marketplace of ideas.

[12] See P. DUNN, *Moderazione automatizzata e discriminazione algoritmica: il caso dell'hate speech*, in *Rivista Italiana di Informatica e Diritto*, 1, 2022. The Author broadly analyzes the risks of balancing the need for automated moderation for fighting hate speech and the risks for minorities from a European antidiscrimination point of view.

[13] Although aware of the specificity of the term 'queer', in this essay this word and the expression 'LGBTQ+' will be used interchangeably. Specifically, I will use 'queer' as an umbrella term that can provide recognition for all members of the LGBTQ+ community. Against my methodological choice: A. RALLING, *A Provocation: Queer is Not a Substitute for Gay/Lesbian*, in *Harlot*, 1, 2008. On the definition of 'queer' see: I. BARNARD, *Queer Race: Cultural Interventions in the Racial Politics of Queer Theory*, New York, 2003 and F. VALDES, *Queers, Sissies, Dykes, and Tomboys: Deconstructing the Conflation of "Sex," "Gender," and "Sexual Orientation"*, in *California Law Review*, 1, 1995. This specificity is also reflected on the distinction between LGBT+ legal studies and queer legal studies. I consider this study as belonging to the former. See B. COSSMAN, *Queering Queer Legal Studies: An Unreconstructed Ode to Eve Sedgwick (and Others)*, in *Critical Analysis of Law,* 6, 2021.

[14] H. BLOCH-WEHBA, *Automation in Moderation*, in *Cornell International Law Journal*, 53, 2020, 41-96.

[15] *See* S.K. KATYAL, J.Y. JUNG, *The Gender Panopticon: AI, Gender, and Design Justice*, in *U.C.L.A. Law Review*, 68, 2021, 735.

[16] The topic has been broadly discussed by constitutional legal scholars. See, among others, J.M. BALKIN, *Digital speech and democratic culture: A theory of freedom of expression for the information society*, in A. SARAT, P. SCHIFF BERMAN (eds.), *Law and Society Approaches to Cyberspace*, London, 2007; G. DE GREGORIO, *Democratising Online Content Moderation: A Constitutional Framework*, in *Computer Law and Security Review*, 36, 2020; A.P. HELDT, *Content Moderation by Social Media Platform: The Importance of Judicial Review*, in E. CELESTE, A.P. HELDT, C. IGLESIAS KELLER (eds.), *Constitutionalising Social Media*, Oxford, 2022, 251-266.

[17] J. CASTELLO, *Why are Tumblr, Twitter and YouTube blocking LGBTQ+ content?*, in www.theestablishment.com, July 20, 2017. See *infra* sec. 3.

community that should be protected by those policies.[18] Indeed, there is a growing concern that social media platforms are using artificial intelligence tools that risk censoring LGBTQ+ content.[19] In particular, artificial intelligence can be used to flag and remove LGBTQ+-related content without human intervention. This could have both a direct and a chilling effect on freedom of expression for LGBTQ+ people.[20] Additionally, it could lead to self-censorship as people become aware of the risks of posting LGBTQ+-related content online.

As scholars criticizing "neoliberal Internet governance"[21] noted, two narratives about the impact of the Internet on the queer community must be acknowledged: a mainstream liberatory one and a critical counter narrative. According to this approach, there is a dominant narrative – "the celebrated dominant narrative"– that describes Internet as a safe place for queer community life, as well as a key space for LGBTQ+ people's participation in democratic discourse[22] and for their freedom of expression.[23]

The "emerging counter-narrative", on the other hand, sees the Internet governance, operated through the increased possibility of surveillance,[24] monitoring systems,[25] and content moderation,[26] as an attempt to replicate online the marginalization that the LGBTQ+ community experiences in the offline life.[27]

At a deeper look, these two narratives are not alternatives to each other. In fact, on the one hand, it is possible to argue how the queer community has found in digital space a place that potentially can help to reduce isolation, connect its members, provide information and representation, and promote their

[18] J. Mchangama, N. Alkiviadou, R. Mendiratta, *Thoughts on the DSA: Challenges, Ideas and the Way Forward through International Human Rights Law*, The Future of Free Speech, 2022, 13: "While regulating hate speech online, policymakers and social media platforms must be wary of broadly-worded bans of hate speech as they may be used to target dissenting views and the very groups such speech restrictions are supposed to protect".

[19] S. Allen, *Social media giants have a big LGBT problem. Can they solve it?*, in *Daily Beast*, October 12, 2019, available at: https://www.thedailybeast.com/social-media-giants-have-a-big-lgbt-problem-can-they-solve-it.

[20] O.L. Haimson, D. Delmonaco, P. Nie, *Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: marginalization and moderation gray areas*, in *Proceedings of the ACM on Human-Computer Interaction*, 5, 2021, 466.

[21] M. Zalnieriute, *The Anatomy of Neoliberal Internet Governance: A Queer Critical Political Economy Perspective*, in D. Otto (ed.), *Queering International Law: Possibilities, Alliances, Complicities and Risks*, London, 2017, chapter 3.

[22] Among others: C. Pullen, M. Coope, *LGBT Identity and Online New Media*, New York, 2017, 100; R. Walker, *Lesbian Deliberation: The Constitution of Community in Online Lesbian Forums*, PhD Thesis, Wollongong, 2010.

[23] On the idea of the Internet as a surveillance tool, see: S.K. Katyal, J.Y. Jung, *The Gender Panopticon*, *op. cit.*, 2021, 696. For a broad understanding of the concept see the monumental work by S. Zuboff, *The Age of Surveillance Capitalism*, London, 2019.

[24] R. Andreassen, *Social media surveillance, LGBTQ refugees and asylum: How migration authorities use social media profiles to determine refugees as 'genuine' or 'fraudulent'*, in *First Monday*, 26, 2021. The Author shows "the role of social media content in Danish asylum cases by examining verdicts from the years 2015–2019. In particular, it examines cases relating to LGBTQ+ refugees (i.e., asylum seekers who claim asylum on the basis of sexual orientation and/or gender identity) and how their credibility is determined, in part, by their social media profiles".

[25] T. Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*, Yale, 2018.

[26] M. Zalnieriute, *op. cit.*, 2017, 4.

[27] In this way is argued by A. Moena, *The Digital Closet: How the Internet Became Straight*, Cambridge, 2022.

freedom of expression.[28] On the other hand, however, it is undeniable how social media have reiterated internally restrictive biases and policies that already affect the LGBTQ+ community.

Indeed, while the queer community attempts to cope with hate speech, online harassment, cyberbullying, and violent content, the use by social media platforms of access filters to LGBTQ+-related material,[29] real name policies,[30] as well as restrictions on the publication of content and its moderation[31] risk to negatively impact the LGBTQ+ speech.[32]

This paper aims to analyze the role of artificial intelligence systems in the moderation of social media content and the impact that these tools may have on the LGBTQ+ community's freedom of expression. To this end, after an analysis of content moderation systems, and the intervention of artificial intelligence in content removal (section 2), the risks to the LGBTQ+ community's freedom of expression (section 3) will be accounted. At the end, normative remedies to the issue will be analyzed by specifically looking at the European Union's Digital Services Act[33] (section 4), underlining both its merits and pitfalls in protecting queer content (section 5). A brief conclusion will follow.

2. Technical approaches to social media content moderation

The amount of content that appears on social media every day has created a growing demand for moderation, which digital intermediary businesses have eventually taken over.[34] Generally speaking, moderation can be defined as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse".[35] It consists of "the screening, evaluation, categorization, approval or removal/hiding of online content according to relevant communications and publishing policies. It seeks to support and enforce positive communications behavior online, and to minimize aggression and anti-social behavior".[36] On social media, this can be done for a variety of reasons, but most often – and primarily – it is done for commercial reasons. Digital tech companies want to

[28] C. Pullen, *LGBT Identity and Online New Media, London*, 2010; more specifically: M. Cooper, K. Dzara, *The Facebook Revolution: LGBT Identity and Activism*, in C. Pullen (ed.), *op. cit.,* 2010.

[29] American Civil Liberties Union, *Don't Filter Me Initiative finds Schools in Four More States Unconstitutionally Censoring LGBT Website*, in www.aclu.org, April 11, 2021.

[30] S. Gunther, *Facebook's "Real Name" Policy: A Violation of the Corporate Responsibility to Respect Human Rights*, in www.humanrights.org, 2015. After a series of complaints from LGBTQ+ activists Facebook eventually removed the policy that required registration by ID real name.

[31] C. Southerton, D. Marshall, P. Aggleton, M. Rasmussen, R. Cover, *Restricted Modes, Social Media Classification and LGBTQ Sexual Citizenship*, in *New Media & Society*, 23, 2020, 920-921.

[32] M. Zalnieriute, *op. cit*., 2017, 5.

[33] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act).

[34] T. Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*, Yale, 2018. According to the Author: "[m]oderation is not an ancillary aspect of what platforms do. It is essential, constitutional, definitional. Not only can platforms not survive without moderation. They are not platforms without it".

[35] J. Grimmelmann, *The Virtues of Moderation*, in *Yale Journal of Law and Technology*, 17, 2015, 47.

[36] T. Flew, *Internet Regulation as Media Policy: Rethinking the Question of Digital Communication Platform Governance*, in *Journal of Digital Media & Policy*, 10, 2019, 40.

make sure that the content that is being posted on their behalf is appropriate and in line with their brand identity.[37]

This need was soon translated into contractual rules and technical architectures that stipulate what behaviors a user can engage in, within the so-called Term of Services (ToS) and community guidelines.[38] As it has been widely noted by scholars,[39] digital platforms concentrate on them a *quasi-public* powers,[40] *i.e.* the power to set the rules that govern online life, to enforce them, and ultimately to resolve any disputes,[41] both vertically, between user and platform, and horizontally between users. In this sense, they can be said having *quasi-normative*, *quasi-executive* and *quasi-judicial* functions.[42]

The Terms of Service (ToS) of social media platforms, in addition to substantively regulate users' online behavior, also provide procedures for the enforcement and justiciability of these rules. Specifically, enforcement can be 'human', entrusted to teams of moderators who apply the platform's policies, 'algorithmic', entrusted exclusively to artificial intelligence systems, or 'hybrid', when algorithmic tools are tasked with detecting illicit content and the human moderator is tasked with deciding whether or not to remove it.[43] Social media also provides dispute resolution and complaint mechanisms for the user whose content is deemed to be in violation of the ToS and so subjected to restriction and removal.[44]

[37] S. ZUBOFF, *Big Other: Surveillance Capitalism and the Prospects of an Information Civilization*, in *Journal of Information Technology*, 30, 2015, 75. As observed by De Gregorio: "The activity of content moderation is performed to attract revenues by ensuring a healthy online community, protect the corporate image and show commitments with ethic values". G. DE GREGORIO, *Democratising Online Content Moderation: A Constitutional Framework*, in *Computer Law and Security Review*, 36, 2020.

[38] L. BELLI, J. VENTURINI, *Private ordering and the rise of terms of service as cyber-regulation*, in *Internet Policy Review*, 5, 2016.

[39] L. BELLI, P.A. FRANCISCO, N. ZINGALES, *Law of the Land or Law of the Platform? Beware of the Privatisation of Regulation and Police*, in L. BELLI, N. ZINGALES (eds.), *Platform regulations: how platforms are regulated and how they regulate us*, Geneva, 2017, 42-44.

[40] G. DE GREGORIO, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society*, Cambridge, 2022, 80-120. See, on this topic, A. VEDASCHI, *Intelligenza artificiale e misure antiterrorismo alla prova del diritto costituzionale*, in *Liber amicorum per Pasquale Costanzo*, 2020, 501.

[41] L. BELLI, P.A. FRANCISCO, N. ZINGALES, *op. cit.*, 2017, 43. The Authors argues that digital platforms are increasingly undertaking "regulatory and police functions", which are traditionally considered "a matter of public law", explaining how "such functions have been growingly delegated to platforms by public authorities, while at the same time platforms are self-attributing such functions to avoid liability, *de facto* becoming private cyber-regulators and cyber-police". See, also: L. BELLI, P. De FILIPPI, N. ZINGALES, *Recommendations on Terms of Service & Human Rights*, 2015 arguing that "besides to reduce the imbalance between platform users and platforms owners when it comes to litigation, it is recommendable that the ToS be negotiated beforehand with consumer associations or other organizations representing Internet users" and J.M. BALKIN, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, in *University of California Davis*, 51, 2018, 1163.

[42] *Ibidem*.

[43] *See*, for a broad understanding of the topic: K. KLONICK, *The New Governors: The People, Rules, and Processes Governing Online Speech*, in *Harvard Law Review*, 131, 2017, 1598-1670.

[44] For instance, the Meta Oversight Board is the group of independent experts who review cases involving content that has been removed from the platform. The Board was created in 2020 in response to concerns about how Meta moderates content. The Board can overturn decisions made by Facebook by relying both on human rights legal basis and internal policy. *See* O. POLLICINO, G. DE GREGORIO, *Shedding Light on the Darkness of Content Moderation: The First Decisions of the Facebook Oversight Board*, in *VerfBlog*, May 2, 2021.

As mentioned earlier, control and removal can be entrusted to content moderation algorithms, to a human moderator, or to hybrid forms that integrate both of them.[45]

The combination of these factors makes it possible to distinguish three approaches in the moderation of content posted on social media: an "industrial approach", an "artisanal approach", and, finally, an approach referred to as "community-reliant".[46]

In the industrial approach, moderation relies mainly on automatic and algorithmic content flagging, filtering, and removal systems. It is characterized by modest involvement of users, as the moderation is based solely on internal policies and guidelines.[47] These rules are incorporated and structured into the algorithm (following the principle of "regulation by architecture"[48]). This approach is employed in content moderation by major social media platforms: Facebook, Instagram, TikTok, Twitter, and YouTube. The reliance on automated moderation systems, on the one hand, offers the advantage of being able to quickly process the enormous amount of data and content published in these digital platforms. On the other hand, it discloses the long-standing issue of the opacity of algorithm operation,[49] as well as the risk of technical impossibility of detecting the context and tone of the content.[50]

In the artisanal approach to moderation, content removal relies mainly on *ex post* and manual human control through specific moderation teams. Clearly, this approach carries the disadvantage of lower efficiency and speed, as well as higher costs for the platform.[51] This approach is used by medium-large platforms, and it is increasingly being abandoned in favor of automated and hybrid moderation systems (Change, Patreon; as well as by major dating apps: Grindr, Romeo, Tinder).

Lastly, some platforms, such as Wikipedia, offer a community-reliant approach to moderation: any user can moderate as well as edit and remove content.[52]

As already explained, large social media platforms, such as Facebook, Twitter, YouTube, employ algorithmic tools relying on artificial intelligence in moderation activities, due to the huge amount of content posted by users.[53]

[45] P. DUNN, *op. cit.*, 2022, 135.

[46] R. KAPLAN, *Content or Context Moderation. Artisanal, Community-Reliant and Industrial Approaches*, in *Data Society*, 2018, available at: https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf.

See also C. PERSHAN, *Moderation our (dis)content: renewing the regulatory approach*, in *Renaissance Numérique*, 2020, 16-22.

[47] C. PERSHAN, *op. cit*., 2020, 16.

[48] L. BELLI, P.A. FRANCISCO, N. ZINGALES, *op. cit.*, 2017, 44. According to the principle of 'regulation by architecture' the contractual rules and technical architecture "establish what behaviors are allowed in the online world. In this perspective, digital platforms may be considered as cyberspaces in the sense of true virtual territories whose frontiers are defined by their technical architecture". *See* also: L. BELLI, *De la gouvernance à la régulation de l'Internet*, Berger-Levrault, 2016.

[49] For a general overview on algorithmic opacity, *see*: J. BURRELL, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, in *Big Data & Society*, 2016, 1-12. *Infra*, section 3.

[50] *Infra*, section 3.

[51] C. PERSHAN, *op. cit*., 2020, 21.

[52] *Id.*, 22. See also, C. DINAR, *The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act*, Washington D.C., 2021.

[53] R. GORWA, R. BINNS, C. KATZENBACH, *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*, in *Big Data & Society*, 7, 2020, 2.

Some data deriving from social media reports and press release serve as an excellent example of the considerable impact that automation has already begun to play in enforcing content regulation.

For instance, algorithms currently control more than 95% of content removal and bans on Facebook (up from 23% in 2017);[54] YouTube now reports that "98% of the videos removed for violent extremism are flagged by machine-learning algorithms",[55] and Twitter revealed that 93% of 'terrorist content' is reported by proprietary internal tools (*i.e.*, algorithms for detecting terrorist content) and removed.[56]

From a technical point of view, there are many different types of algorithms that can be used for content moderation purposes. Automation can be used at the stage of proactive identification of potentially illicit content and automated evaluation and execution of a decision to remove, tag, demonetize, demote, or prioritize content.[57]

In the analysis of textual content, different types of algorithms can be deployed. Less technologically advanced algorithmic moderation tools are based on the technique of keywords analysis. These algorithms filter or delete content based on certain words or phrases, totally detached from the context.[58]

However, moderation of content on large social media platforms relies more often on machine learning algorithmic techniques. This type of technology allows for a more sophisticated textual analysis that enables a more comprehensive evaluation of the content.

In particular, the so-called Natural Language Processing (NLP) tools are machine learning software that can be trained to predict whether a text is expressing a positive or negative emotion (sentiment analysis) and to classify it as belonging or not belonging to some category (such as hate speech, online harassment, cyberbullying, violent content and so on).[59]

However, the harmfulness of content posted on social media depends on the context.[60] Algorithmic tools, even those based on the most advanced forms of machine learning, fail to take all the contextual data into account.[61] It is in fact particularly difficult from a technological point of view to train

[54] P. DUNN, *op. cit.*, 2022, 136.

[55] I. LAPOWSKY, *After sending content moderators home, YouTube doubled its video removals*, in www.protocol.com, August 25, 2020, available at: https://www.protocol.com/youtube-content-moderation-covid-19.

[56] *See* Twitter Transparency Report, 2018, available at: https://blog.twitter.com/official/en_us/topics/company/2018/twitter-transparency-report-12.html

[57] *See* E. LLANSÓ, J. VAN HOBOKEN, P. LEERSSEN, J. HARAMBAM, *Artificial Intelligence, Content Moderation, and Freedom of Expression*, Working paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression, 2020, available at https://www.ivir.nl/twg/.

[58] E. ENGSTROM, N. FEAMSTER, *The Limits of Filtering: A Look at the Functionality and Shortcomings of Content Detection Tools*, in *Engine*, March 2017, available at: http://www.engine.is/the-limits-of-filtering/.

[59] A. MARSOOF, A. LUCO, H. TAN, S. JOTY, *Content-filtering AI systems–limitations, challenges and regulatory approaches*, in *Information & Communications Technology Law*, 2022 (forthcoming).

[60] J. COBBE, *Algorithmic Censorship by Social Platforms: Power and Resistance*, in *Philosophy and Technology*, 34, 739–766, 2021.

[61] N. DUARTE, E. LLANSO, A. Loup, *Mixed messages? The limits of automated social media content analysis*, Center for Democracy and Technology, Nov. 2017, 5: "Among studies using NLP to judge the meaning of text (including hate speech detection and sentiment analysis), the highest accuracy rates reported hover around 80%, with most of the high-performing tools achieving 70 to 75% accuracy. These accuracy rates may represent impressive advancement in NLP research, but they should also serve as a strong caution to anyone considering the use of such tools in a decision-making process. An accuracy rate of 80% means that one out of every five people is treated 'wrong' in such decision-making; depending on the process, this would have obvious consequences for civil liberties and human rights.".

algorithms with data that informs historical, cultural, or political context, or that can distinguish the satirical, educational, informational use of a term or expression.[62] For this reason, the use of algorithms will always produce the risk of erroneous outcomes, i.e., false negative or false positive results.[63] A false positive outcome represents an unjustified burden on the user's freedom of expression, as content that is actually permitted would be removed or restricted.[64] At the same time, false negatives grant 'impunity' to those who disseminate hate speech, online harassment, violence, and other disreputable content. This could result in a chilling effect on targeted individuals and groups.

Second, even when discussing the algorithms used in online content moderation, the age-old issue of algorithmic biases comes up. Machine learning tools develop their ability to identify and distinguish content types based on the datasets used in the training phase. However, if these datasets, trained by humans, do not include data from different languages and from certain groups, communities, and minorities, the tools will not be able to parse the communications of these users.[65]

In particular, if the datasets are influenced by pre-existing biases and inequalities in the offline world, then the model will reflect or amplify these inequalities.[66]

When platforms adopt automated content analysis tools, the algorithms behind the tools can become the *de facto* rules for enforcing a web site's terms of service. The disparate enforcement of terms of service by biased algorithms that disproportionately censor marginalized groups, such as the queer community, raises obvious concerns.

3. The queer community and online content moderation by algorithms: a risk for the LGBTQ+ freedom of expression

Given this technical theoretical background, it could be now analyzed how algorithmic moderation can infringe on LGBTQ+ people's freedom of expression. A first practical demonstration comes from a study conducted by Olivia, Antonialli and Gomes on the impact of artificial intelligence content moderation tools on LGBTQ+ speech.[67] The study starts from the assumption that, in queer linguistic,[68] the

[62] A. Reyes, P. Rosso, D. Buscaldi, *From humor recognition to irony detection: the figurative language of social media*, in *Data & Knowledge Engineering*, 74, 2012, 1-12.

[63] E. Llansó, J. van Hoboken, P. Leerssen, J. Harambam, *op. cit*., 2020, 9. A false positive outcome is a content that has been wrongly classified as illicit. In case of false negative outcome, the automated tool has missed a content that should have been classified as illicit.

[64] *Ibidem*.

[65] R. Binns, M. Veale, M. van Kleek, et al., *Like trainer, like bot? Inheritance of bias in algorithmic content moderation*, in G.L. Ciampaglia, A. Mashhadi, T. Yasseri (eds) *Social Informatics*, Berlin, 2017, 405-415.

[66] O.L. Haimson, D. Delmonaco, P. Nie, *Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: marginalization and moderation gray areas*, in *Proceedings of the ACM on Human-Computer Interaction*, 5, 2021, 466.

[67] T. D. Oliva, D.M. Antonialli, A. Gomes, *Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online*, in *Sexuality & Culture*, 25, 2021, 700-732.

[68] The assumption is referred to the following queer linguistic studies, cited by Oliva, Antonialli and Gomes: S.O. Murray, *The art of gay insulting*, in *Anthropological Linguistics*, 21, 1979, 211; J. Perez, *Word play, ritual insult, and volleyball in Peru*, in *Journal of Homosexuality*, 58, 2011, 834; R.G. Jones, *Drag queens, drama queens and friends: Drama and performance as a solidarity-building function in a gay male friendship circle*, in *Kaleidoscope*, 6, 2007, 61; B.L. Heisterkamp, *Control and desire: identity formation through teasing among gay men and lesbians*,

usage of "impolite" terms develops a hostility-averse copying mechanism, which queer people themselves utilize as a code of communal communication.[69]

The researchers used Perspective API,[70] a machine learning software (from Jigsaw, a Google subsidiary) which is employed to identify and classify textual content posted online based on their level of toxicity.[71] In order to understand the potential negative impact on LGBTQ+ speech, the study compares the software-calculated level of toxicity of tweets posted by some American drag queens[72] and that of prominent figures in the U.S. supremacist movement.[73] By analyzing 114.000 tweets, the research shows that a significant number of drag queen users are considered more toxic than the accounts of far-rightists: the level of toxicity of drag queen accounts ranges from about 17% to 38%, while that of white supremacists between 21% and 29%.[74]

Moreover, by having the software analyze certain terms detached from context, the research proves how words such as 'gay', 'lesbian', and 'queer' – that are neutral and descriptive terms – are considered by Perspective to be highly toxic, at about 76%, 61%, and 51%, respectively. This demonstrates an evident bias of the algorithmic tool toward neutral words from the LGBTQ+ community culture.[75]

Finally, the study considers words that are deemed to be 'impolite', and yet are often used by LGBTQ+ people themselves because they positively influence the internal code of communication of the queer community[76]. According to the researchers, "it seems the use of Perspective, as well as other similar technologies, to police content on internet platforms without human oversight could hinder the use of free speech by members of the LGBTQ+ community".[77]

As a matter of fact, there have been multiple claims from LGBTQ+ activists claiming that their content was removed as a result of what appears to be biased algorithmic enforcement of platforms' Term of Services.

in *Communication Studies*, 51, 2010, 388; S. McKinnon, '*Building a thick skin for each other': The use of 'reading' as an interactional practice of mock impoliteness in drag queen backstage talk*, in *Journal of Language and Sexuality*, 6, 2017, 90.

[69] T.D. Oliva, D.M. Antonialli, A. Gomes, *op. cit.*, 2021, 703.

[70] Perspective is a software that uses machine learning to identify 'toxic' comments. See more on https://perspectiveapi.com.

[71] Drag queens, in the LGBTQ+ culture, can be defined as "gay individuals who don female clothing with the explicit goal of performing in front of audiences". *See* M. Moncrieff, P. Lienard, *A Natural History of the Drag Queen Phenomenon*, in *Evolutionary Psychology*, 2017, 2. For conducting this study the researchers selected "the drag queens' Twitter profiles among former participants of the drag reality show 'RuPaul's Drag Race'". T.D. Oliva, D.M. Antonialli, A. Gomes, *op. cit*., 2021, 707.

[72] "The white nationalist accounts analyzed in the research were selected from a list published by HuffPost USA that included 62 Twitter users who identify themselves as white nationalists and share content supporting white supremacy". T.D. Oliva, D.M. Antonialli, A. Gomes, *op. cit.*, 2021, 708.

[73] T.D. Oliva, D.M. Antonialli, A. Gomes, *op. cit.*, 2021, 703.

[74] *Ivi*, 712.

[75] *Ivi*, 716.

[76] Therefore, words such as 'fag', 'bitch', and 'sissy' detached from a context analysis, have toxicity levels of 91%, 98%, and 83%, respectively. *Ivi,* 716-717.

[77] *Ivi*, 729.

Downloaded from www.biodiritto.org.
ISSN 2284-4503

In 2019, five youtubers filed a class action against YouTube and Google[78] alleging that YouTube 'restricted mode policy'[79] limited and banned a variety of LGBTQ+ videoclips and accounts, on the basis of explicitly queer content and on the sexual orientation and gender identity of the content creators.[80] In particular, according to the plaintiffs, videos containing, in the title or caption, keyword such as 'gay', 'lesbian', 'transgender', 'bisexual' or more generally 'LGBT+' were hidden or demonetized after being automatically flagged as 'adult content' by platform's algorithms.[81] The charges were dismissed by the United States District Court Northern District of California, San Jose Division, on the grounds that "tech businesses are not state actors susceptible to judicial examination under the First Amendment".[82]

However, besides the Court's decision, YouTube publicly admitted to the risk of restriction and demonetization of queer content. Indeed, while stating that YouTube does not have a list of LGBTQ+-related words that trigger demonetization, a YouTube spokesperson confirmed that the platform "uses machine learning to evaluate content against guidelines" and that "sometimes the systems get it wrong".[83]

Similar reports have been released about other social media, such as Instagram and Twitter. In 2018, the American queer activist Eli Erlick complained that Instagram shadow-banned[84] her account after posting a picture using the hashtag #lesbian.[85] In 2017, Twitter blocked the possibility to search for words like 'gay', 'bisexual', 'trans'. As later explained by the platform, the company's new policy attempting to restrict sensitive content used "a list of terms that frequently appear alongside adult content". That list was "out of date and incorrectly included terms that are primarily used in non-sensitive contexts".[86]

[78] A. ROMANO, *A group of YouTubers is trying to prove the site systematically demonetizes queer content*, in *Vox*, Oct. 10, 2019, available at: https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report.

[79] 'Restricted mode' is a YouTube option setting that restrict access to content that is flagged as 'mature' or that contains profanity, nudity, and violence.

[80] *Class Action Complaint for Damages, Injunctive Relief, Restitution, and Declaratory Judgment*, United States District Court Northern District of California, San Jose Division, August 13, 2019, available at: https://www.docdroid.net/g7RXXi1/youtube-lgbtq-lawsuit-pdf

[81] *Ibidem*.

[82] N. LANG, *This Lawsuit Alleging YouTube Discriminates Against LGBTQ+ Users Was Just Tossed Out*, in *Them*, January 8, 2021, available at: https://www.them.us/story/lawsuit-alleging-youtube-discriminates-against-lgbtq-users-tossed-out

[83] A. ROMANO, *op. cit.*, 2019.

[84] The 'shadow-ban' is a feature that Instagram uses to filter out potential harmful content by not removing an account but by making it invisible to the other users for a certain period.

[85] E. ERLICK, *How Instagram May Be Unwittingly Censoring the Queer Community*, in *Them*, January 30, 2018, available at: https://www.them.us/story/instagram-may-be-unwittingly-censoring-the-queer-community. See also, S.K. KATYAL, J.Y. Jung, *The Gender Panopticon: AI, Gender, and Design Justice*, in *U.C.L.A. Law Review*, 68, 2021, 738.

[86] Z. FORD, *Twitter Offers Incoherent Explanation for Anti-LGBTQ Censorship*, in *ThinkProgress*, November 7, 2017, available at: https://archive.thinkprogress.org/twitter-lgbtq-censorship-719192dbb0fd/.

Indeed, algorithms of content moderation often conflate LGBTQ+ content and terms with sexual imagery,[87] leading to an overinclusive blocking of LGBTQ+ speech, whether the content is sexual or not.[88]
In March 2023, Twitter has removed thousands of tweets related to the "Trans Day of Vengeance" a transgender rights protest taking place in Washington, D.C., banning from the platform both tweets in support and critics of the trans rights movement.[89] According to Ella Irwin, Twitter's Trust & Safety Lead, the ban is the result of an automated process that remove tweets that violate the platform's content policies, including prohibitions on content that incites violence, since the word "vengeance" implies a non-peaceful protest.

On the same days, LGBTQ+ activists, such as the Trans Safety Network reported that tweets shared via Direct Message on Twitter were no longer directly visible with the usual preview if they included certain words such as 'trans', 'gay', 'lesbian', 'queer', and 'sex'.[90]

What has been discussed shall allow to conclude that the employment of AI technologies for dealing with harmful content online without human monitoring may endanger freedom of expression of LGBTQ+ members, as well as members of other socially vulnerable groups. In other words, queer people may face restrictions when sharing anything online based on their LGBTQ+ identity, sexual orientation, or gender identity. The risk of over-blocking LGBTQ+ content, as well as biased implementation of platform ToS, jeopardizes queer individuals' freedom to express themselves.

In conclusion, the self-regulation of harmful speech by private tech companies sparks an essential debate about how to reach a balance between online content moderation and the need to protect users' freedom of expression.[91] As noted by Balkin, freedom of expression in the digital age is a triangle.[92]

Indeed, the nineteenth-century conception of free speech, focused on protecting citizens against public authorities' interferences with their freedom to manifest their own thought, is becoming "outmoded and inadequate to protect free expression today".[93] In the early twenty-first century, freedom of expression is increasingly dependent on a third group of players: "the privately owned infrastructure of digital communication composed of firms that support and govern the digital public sphere that people use to communicate".[94]

Although a distinction is often made between legal traditions that inform the concept of freedom of expression by giving relevance to the culture of dignity (European Union) or the culture of freedom

[87] See, for a general overview of case related to ban from platforms, S.K. KATYAL, J.Y. Jung, *The Gender Panopticon: AI, Gender, and Design Justice*, in *U.C.L.A. Law Review*, 68, 2021, 738.

[88] C. SOUTHERTON, D. MARSHALL, P. AGGLETON, M. RASMUSSEN, R. COVER, *Restricted Modes, Social Media Classification and LGBTQ Sexual Citizenship*, in *New Media & Society*, 23, 2020, 927.

[89] CBSNews, *Twitter removes tweets about 'Trans Day of Vengeance'*, Marc 30, 2023, available at: https://www.cbsnews.com/atlanta/news/twitter-removes-tweets-about-trans-day-of-vengeance/.

[90] H. WILLIAMSON, *Twitter appears to censor LGBTQ+ terms including 'trans' in DM previews*, in *www.thepinknews.com*, April 2, 2023, available at: https://www.thepinknews.com/2023/04/01/twitter-appears-to-censor-lgbtq-terms-including-trans-in-dm-previews/.

[91] See, among others, K. KLONICK, *The New Governors: The People, Rules, and Processes Governing Online Speech*, in *Harvard Law Review*, 131, 2018, 1598; K. LANGVARDT, *Regulating Online Content Moderation*, in *The Georgetown Law Journal*, 106, 2018, 1353.

[92] J.M. BALKIN, *Free Speech is a Triangle*, in *Columbia Law Review*, 118, 2018, 2012-13.

[93] *Ibidem*.

[94] *Ibidem.*

(United States),[95] this distinction today gives way to the global trend toward self-regulation by digital intermediaries and their algorithms.[96]

This form of self-regulation bears problems of biases, opacity, lack of transparency and accountability,[97] and leaves private actors free to balance and enforce fundamental rights without any public guarantees, *de facto* delegating to the major Internet companies the manifestations of the freedom of expression on their platforms,[98] and eventually leading to a much-needed call for a legislative intervention.

However, while the debate in the United States is almost entirely focused on the issue of platforms' liability for content posted by users,[99] which clashes with the *totem* of Section 230 of the Communication Decency Act (CDA),[100] the European Union has decided to reclaim its normative power over digital intermediaries with the enactment of the Digital Services Act (DSA).[101]

4. The European Union attempt to regulate the online content moderation: the Digital Services Act

The Digital Services Act (DSA) (Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC) came into effect in November 2022 and will be applicable in the European Union starting from March 2024. Although it includes specific obligations regarding the protection of users' personal data and transparency obligations related to online advertising, the core regulatory purpose of the DSA is to establish rules for online content moderation. In fact, the EU legislator intends the DSA to ensure user safety online, establishing governance mechanisms for online content while balancing these mechanisms with the necessary protection of fundamental rights, particularly freedom of expression.[102]

Generally speaking, the DSA includes rules for all online intermediary services, a broad notion that comprises hosting service providers, and therefore, for what matters here, social media platforms as

[95] M. Nogueira de Brito, *Hate Speech and social media*, in C. Blanco de Morais, G. Ferreira Mendes, T. Vesting (eds), *The Rule of Law in Cyberspace*, Cham, 2022, 283-307.

[96] *Ibidem*.

[97] G. De Gregorio, *Democratising Online Content Moderation: A Constitutional Framework*, in *Computer Law and Security Review*, 36, 2020, 4: "The high degree of opacity and inconsistency of content moderation frustrates democratic values".

[98] G. Cerrina Feroni, A. Gatti, *Online Hate Speech and the Role of Digital Platforms: What Are the Prospects for Freedom of Expression?*, in C. Blanco de Morais, G. Ferreira Mendes, T. Vesting (eds), *The Rule of Law in Cyberspace*, Cham, 2022, 261-282.

[99] D.K. Citron, M.A. Franks, T*he Internet as a Speech Machine and Other Myths: Confounding Section 230 Reform*, in *University of Chicago Legal Forum*, 2020, 45.

[100] Section 230 of the Communication Decency Act states that "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider" (47 U.S.C. § 230).

[101] The Digital Services Act proposal was first presented by the European Commission in December 2020 and approved by the European Parliament in July 2022: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0825.

[102] A. Turillazzi, M. Taddeo, L. Floridi, F. Casolari, *The digital services act: an analysis of its ethical, legal, and social implications*, in *Law, Innovation and Technology*, 2023.

well. Specific obligations are then directed at a particular category of online platforms, defined as 'Very Large Online Platforms', which includes social media that reaches more than 10% of the 450 million European consumers.[103]

The EU's mechanism for governing online content and related moderation, including algorithmic moderation, is based on three main aspects: the liability system, provided in articles 6 and followings, the removal order (article 9), and the notice and action mechanism (article 16).[104]

It should be noted that the DSA provides for a dual-track action, on the one hand for illegal content, and on the other hand for "dangerous" content on the contractual basis of the ToS (article 14).[105] In this context, space is also dedicated to the use of automated moderation tools, with particular reference to transparency obligations (article 15).[106]

Firstly, it should be noted that article 6 of the DSA does not modify the liability system already provided for in the E-Commerce Directive.[107] In fact, the new intermediary liability system still provides a *safe harbor regime* if the platforms are not aware of the presence of illegal content (article 6.1 a) and if, having obtained this knowledge, they have acted promptly to remove the illegal content (article 6.1 b). Article 9 then provides for an obligation to remove content following notification by an authority, judicial or administrative, of a Member state, to which effect must be given "without undue delay" (article 9.1). The Regulation also reiterates, as already provided for in the E-Commerce Directive, that "no general obligation to monitor the information which providers of intermediary services transmit or store, nor actively to seek facts or circumstances indicating illegal activity shall be imposed on those providers" and establishes a protection regime for so-called 'Good Samaritans', serving as a liability shield for good-faith efforts to remove illegal content proactively,[108] an effort that could be entrusted to the identification capabilities of algorithms for content moderation of illegal and dangerous content.

The DSA also introduces general rules on the mechanisms that allow each user to report content deemed illegal with a 'notice and action' mechanism that obliges platforms to act on the content once they receive notice. If providers decide to remove or block the content, they must inform the user who posted it, establishing an obligation to provide a reason for the decision and to inform them in case the content was processed through an algorithmic tool (article 16).

In addition to these moderation mechanisms related to the potential illegality of content, the DSA imposes specific obligations related to the ToS that serve as a basis for the spontaneous and proactive removal of content by platforms. In this sense, article 14 states that Providers of intermediary services shall include information on any restrictions that they impose in relation to the use of their service in respect of information provided by the recipients of the service, in their terms and conditions. That

[103] A list of the designated VLOPs can be found on https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops.

[104] A.P. HELDT, *EU Digital Services Act: The White Hope of Intermediary Regulation*, in T. FLEW, F.R. MARTIN (eds.), *Digital Platform Regulation*, 2022.

[105] See P. LEERSSEN, *An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation*, in *Computer Law & Security Review*, 48, 2023, 6-7.

[106] *Ibidem.*

[107] 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').

[108] A. SAVIN, *The EU Digital Services Act: Towards a More Responsible Internet*, Copenhagen Business School, CBS LAW Research Paper No. 21-04, 2021.

information shall include information on any policies, procedures, measures, and tools used for the purpose of content moderation, including algorithmic decision-making and human review, as well as the rules of procedure of their internal complaint handling system.

The DSA requires important transparency obligations, referred also to the use of algorithms in content moderation activities. Article 15 set that providers of intermediary services shall make publicly available, at least once a year, comprehensible reports on any content moderation that they engaged in during the relevant period. In particular, these reports must contain an account of the moderation activities carried out by platforms' own initiative, including the number and type of measures implemented (Art. 15 (c) DSA).[109] With specific reference to the use of algorithms, the report must account "any use made of automated means for the purpose of content moderation, including a qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error of the automated means used in fulfilling those purposes, and any safeguards applied (art. 15 (d)). Moreover, specifically for "Very Large Online Platoform", the DSA establishes that the Commission shall have access to information about the algorithms used in content moderation, for the purpose of ensuring the effective implementation of and compliance with the obligations laid down in this Regulation, throughout the Union (artt. 69 ff.).

Clearly, the Digital Services Act (DSA) is an important piece of legislation that seeks to balance the need for effective content moderation with the need to protect freedom of expression online. The DSA represents a significant step towards creating a safer and more transparent online environment for all users. The DSA acknowledges the importance of protecting freedom of expression as a fundamental human right, while also recognizing the need to address the proliferation of harmful and illegal content online. However, while the DSA offers a framework for regulating digital services that shall be both effective and respectful of individual rights, it is not without limitations.[110] For instance, concerns over potential over-censorship by digital service providers in an attempt to avoid liability under the DSA have been raised. Additionally, there may be challenges in balancing the diverse and often conflicting perspectives of different stakeholders in the digital ecosystem. Nevertheless, the DSA is an important example of how legislation can attempt to balance competing interests, and as the digital landscape continues to evolve, it will be important to reassess such legislation to ensure that it effectively addresses the challenges of moderating content online while upholding the right to freedom of expression of marginalized groups. To this extent, in the next session, both merits and pitfalls of the DSA shall be analyzed.

[109] Article 13 (d) states that those reports shall include: "The content moderation engaged in at the providers' own initiative, including the number and type of measures taken that affect the availability, visibility and accessibility of information provided by the recipients of the service and the recipients' ability to provide information, categorized by the type of reason and basis for taking those measures".

[110] ARTICLE19.ORG, *At a glance: Does the EU Digital Services Act protect freedom of expression?*, in *www.article19.org*, February 19, 2021, available at: https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/.

5. The DSA's merits and pitfalls in the protection of queer people's freedom of expression

Social media platforms play a crucial role in supporting freedom of expression in today's digital societies, especially for those voices that are frequently silenced in the offline world. Platforms, on the other hand, contain harmful content, typically directed at minorities, while, at the same time, LGBTQ+ content is likely to being arbitrarily censored by algorithmically biased moderation systems.[111]

However, regulations intended to improve the digital environment by imposing and incentivizing large platforms to adopt moderation policies based only on algorithmic mechanism could end up exacerbating the problem of unjustified restrictions on the LGBTQ+ community's freedom of expression, due to the algorithms' lack of transparency, accountability, and technological capacity to read the context of a content, as demonstrated by studies and content creators' exposures.[112]

Hence, the EU regulatory attempts to force large platforms to be more transparent go in the right direction.

The DSA's is a praiseworthy legislative effort to regulate the conduct of digital companies in the governance of the digital environment, with an emphasis on respect for fundamental rights protected by EU law.[113] In particular, the Digital Services Act would strive, in the European legislator's intentions, to reduce the risk of unlawful content dissemination on the one hand, and potentially unreasonable limitations on users' freedom of expression on the other.[114]

Moving in this direction are those measures specifically aimed at "Online Platform"[115] on the transparency of their performance in content moderation activities. Indeed, the DSA imposes to social media platforms significant transparency and accountability standards, including the publication of an exhaustive report on moderation activity and methods at least once a year (Art. 15 DSA).[116]

Obliging platforms to provide detailed reports on moderation activity would first and foremost incentivize them to work on continuously improving their ability to remove illicit content, and to mitigate the risks of errors in removing lawful content,[117] such as LGBTQ+ content that is unfairly deemed to be

[111] C. Dinar, *The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act*, Washington D.C., 2021.

[112] *Supra* section 3.

[113] See European Commission, *The Digital Services Act: ensuring a safe and accountable online environment*, 2020, available at: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en

[114] V. Golunova, *The Digital Services Act and freedom of expression: triumph or failure*?, Maastricht University, March 8, 2021, available at: https://www.maastrichtuniversity.nl/blog/2021/03/digital-services-act-and-freedom-expression-triumph-or-failure.

[115] The DSA imposes asymmetric due diligence obligations, with various duties assigned to intermediaries according on their size. Only very large online platforms (VLOPs), that are those platforms with at least 45 million active users in the EU, are subject to the Regulation's full extent.

[116] Article 15, Transparency reporting obligations for providers of intermediary services: "Providers of intermediary services shall publish, at least once a year, clear, easily comprehensible and detailed reports on any content moderation they engaged in during the relevant period".

[117] M. MacCarthy, *How online platform transparency can improve content moderation and algorithmic performance*, in *Brooking*, February 17, 2021, available at: https://www.brookings.edu/blog/techtank/2021/02/17/how-online-platform-transparency-can-improve-content-moderation-and-algorithmic-performance/.

illicit (for example, as seen, because wrongly flagged as sexual content or unduly restricted as adult material).

On the contrary, platforms should not be incentivized by government to take down content that may be illegal but that actually are not, just because regulations instruct them to minimize risk[118] – this risk has been called by scholars the "better safe than sorry" effect.[119]

Indeed, other DSA's measures are likely to aggravate the concerns raised in relation to minorities' freedom of expression, particularly those groups that face disproportionate biases in moderation activities via algorithms, such as the LGBTQ+ community. This could occur despite the forward-looking decision not to impose a general monitoring requirement on digital intermediaries, which would have had the effect of incentivizing the use of moderation through artificial intelligence.

As a matter of fact, measure such as the "notice and action" mechanism (Art. 16 DSA) could achieve an unintended restrictive effect on freedom of expression by leading to the risk of removal of genuinely legal content.[120] Indeed, the DSA establishes a "notice and action" mechanism for the removal of illegal content online: providers of hosting services should act on receipt of such a notice "without undue delay", considering the type of illegal content that is being notified and "the urgency of taking action" and remove the content. By imposing a mandatory short-term takedown limit, with the threat of economic sanction in case of noncompliance, the DSA risks incentivizing platforms to engage in restrictive moderation, only possible by relying on artificial intelligence tools.

As noted by human rights activists, this aspect of the DSA is likely to produce collateral damage towards freedom of expression, affecting the ability of ordinary people to debate issues such as immigration, gender, religion, and identity and potentially the very minority groups that hate speech bans are supposed to benefit.[121] The effect of mandatory 'notice and action' policy is that incentives platforms to proactively enforce their own Terms of Service, and to extend automatic content moderation to identify potentially illegal content and remove it before receiving a notice by the authorities, in order to avoid potential fines and economic sanctions.

Therefore, as explicitly stated by Guido Scorza, member of the Italian Data Protection Autority, "the [DSA's] ambition is noble and clear: limiting the circulation of harmful and illicit content online. But there is a risk that, while pursuing it, we will end up immolating a significant portion of free speech in the digital dimension on the altar of the principle "the end justifies the means", and that what is technologically possible must also be considered legally legitimate and democratically sustainable".[122]

[118] F. ERIXON, *"Too Big to Care" or "Too Big to Share": The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules*, in *European Center for International Political Economy*, Policy Brief n. 5, 2021.

[119] A. TURILLAZZI, M. TADDEO, L. FLORIDI, F. CASOLARI, *The digital services act: an analysis of its ethical, legal, and social implications*, in *Law, Innovation and Technology*, 2023.

[120] J. BARATA, *The Digital Services Act and its impact on the right to freedom of expression: special focus on risk mitigation obligations*, PLI, 2021.

[121] J. MCHANGAMA, *The Real Threat to Social Media Is Europe*, in *Foreign Policy,* Apr. 25, 2022, available at: https://foreignpolicy.com/2022/04/25/the-real-threat-to-social-media-is-europe/. See also: *At a glance: Does the EU Digital Services Act protect freedom of expression?*, in *Article19*, Feb. 11, 2021, available at: https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/.

[122] G. SCORZA, *Digital Services Act, "Le luci e le poche ma gravi ombre delle nuove regole Ue", Intervento di Guido Scorza, Componente del Garante per la protezione dei dati personali*, in *AgendaDigitale*, April 28, 2022 [translated

6. Conclusions

Algorithmic content moderation poses a significant threat to the freedom of expression of the LGBTQ+ community on social media platforms. The queer community's presence on digital platforms has enabled them to share their experiences and perspectives with a wider audience,[123] but eventually this narrative has been challenged by the use of automated moderation tools. Indeed, these tools have been criticized for perpetuating biases and silencing queer voices, further marginalizing those minorities that are already marginalized. Indeed, as shown, the implementation of algorithmic content moderation has brought about a significant impact on the freedom of expression of the LGBTQ+ community in digital spaces, due to lack of transparency, discrimination, and the technical incapability to contextualize the discourse.

As explained, the EU's Digital Services Act (DSA) aims to regulate online content moderation and promote transparency and accountability on digital platforms. The DSA's provisions include requirements for platforms to implement clear and transparent moderation policies and provide users with the opportunity to challenge moderation decisions.

The EU's Digital Services Act represents a significant step towards regulating online content moderation and promoting transparency and accountability on digital platforms. Nevertheless, it remains essential to address concerns about the potential over-censorship of queer content, and the need to ensure that the act's provisions adequately protect the rights of marginalized communities.

Indeed, concerns have been raised that the DSA's implementation may not adequately protect the freedom of expression of queer people, as it incentivizes social media platforms to rely solely on algorithms to moderate online content.

Although, as discussed, it is increasingly necessary and efficient to employ algorithms in online content moderation, as much for the identification and flag of dangerous or illicit content as for removal, it is appropriate to take some measures that could mitigate the risks of discrimination and restriction of freedom of expression of marginalized groups, such as the LGBTQ+ community.

First of all, it must be underlined the importance to foster a culture of transparency and accountability by imposing to social media platform reporting obligations. As it has been widely acknowledged, if corporation are required to disclose substantial information regarding their operation and the use of algorithms both to the public and regulators, they are more likely to conform to societal values and expectations, thus resulting in an enhanced level of public trust.[124]

by the Author], available at: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9765212

[123] O. JENZEN, *LGBTQ Youth Cultures and Social Media*, in I. WEST (ed.), *The Oxford Encyclopedia of Queer Studies and Communication*, Oxford, 56 ff.

[124] M. MACCARTHY, *How online platform transparency can improve content moderation and algorithmic performance*, in *www.brookings.edu*, February 17, 2021, available at: https://www.brookings.edu/blog/techtank/2021/02/17/how-online-platform-transparency-can-improve-content-moderation-and-algorithmic-performance/.
See also N.P SUZOR, *What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation*, in *International Journal of Communication*, 13, 2019, 1526.

Secondly, it is necessary to emphasize the importance of respecting the so-called principle of human in the loop, both in designing algorithms for content moderation and in using them.[125] However, it is worth noticing that ensuring that content moderation algorithms are designed with human oversight in mind does not alone provide prevention of bias and discrimination against LGBTQ+ individuals.[126]

Indeed, it is also important that social media platform provide education and training for content moderators teams on issues related to diversity bias and freedom of expression. Incorporate diversity of perspective by ensuring that both algorithm designers but also moderator teams and appeal boards for content removal decisions are diverse and represent a range of background and experiences.[127]

Finally, it could be useful to regularly evaluate and refine moderation processes to ensure that they are effective, equitable, and transparent.[128] This can include soliciting feedback from users and moderators, as well as conducting regular audits and assessments of moderation decisions by involving and empowering marginalized communities, such as the LGBTQ+ community, in content moderation policies and decision-making process.[129]

In conclusion, it is vital to develop more nuanced approaches to algorithmic content moderation that consider the diversity of human experiences and potential impacts on marginalized communities.[130] Only then can we ensure that the internet remains a safe and inclusive space for all, including the LGBTQ+ community.

[125] See, for a better understanding of the technical challenges of the human-in-the-loop principle, R. GORWA, R. BINNS, C. KATZENBACH, *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*, in *Big Data & Society*, 7, 1, 2020.

[126] B. GREEN, A. KAK, *The False Comfort of Human Oversight as an Antidote to A.I. Harm*, *in* www.slate.com, June 15, 2021, available at: https://slate.com/technology/2021/06/human-oversight-artificial-intelligence-laws.html.

[127] O.L. HAIMSON, D.DELMONACO, P.NIE, A. WEGNER. *Disproportionate Removals and Difering Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas*, in *Proc. ACM Hum.-Comput. Interact*., 5, 2021, 466:27.

[128] Institut Montaigne, *Algorithms: Please Mind the Bias!*, Executive Report, March 2020.

[129] C. DINAR, *The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act*, Washington D.C., 2021, 16.

[130] See European Union Agency for Fundamental Rights, *Bias in Algorithms – Artificial Intelligence and Discrimination*, 2022, 14.