Discrimination on online platforms: legal framework, liability regime and best practices

Laurena Kalaja and Lana Bubalo

in: Matthias C. Kettemann (ed.), How Platforms Respond to Human Rights Conflicts Online. Best Practices in Weighing Rights and Obligations in Hybrid Online Orders (Hamburg: Verlag Hans-Bredow-Institut, 2022)

Discrimination on online platforms: legal framework, liability regime and best practices

Laurena Kalaja and Lana Bubalo
GLOBAL DIGITAL HUMAN RIGHTS NETWORK | UNIVERSITY OF STAVANGER

Introduction

The availability of the internet has had huge importance and significant effects on all aspects of people's lives. It has brought many opportunities, connecting the world as never before.[1] At the same time, it has resulted in new forms of human rights infringements, including the right not to be discriminated against.

Online discrimination is understood as denigrating or excluding individuals or groups on the basis of race, gender, sex orientation, age, disability, religion and beliefs through the use of symbols, voice, video, images, text, and graphic representation or the combination thereof, on the internet.

The discrimination on the internet can be direct - such in case of algorithmic bias or discrimination by design. Algorithmic discrimination includes biases incorporated into algorithms and codes that power machine learning and artificial intelligence systems resulting in systematic disadvantage of certain groups or people.[2] This kind of discrimination is becoming increasingly common and requires regulation. Algorithmic bias can have very serious consequences, leading to unfair exclusion of a specific demographic or otherwise target group, for example in employment or housing. For example, Facebook offered "ethnic affinities" as a category which advertisers could use to target their campaigns. After a lot of negative public attention, Facebook discontinued using this designation.[3]

Discrimination by design means that the way the website is made, enables discrimination. For instance, Airbnb's main service requires each guest to create an online profile with certain information, including a genuine name and phone number. It also encourages inclusion of a real photograph. For Airbnb, the authenticity of this profile information is vital to the operation of the service, as it engenders a sense of trust and connection between hosts and guests. Guests' physical characteristics may contain social cues that instill either familiarity and comfort, on the one hand, or suspicion and distrust, on the other. The sense of authentic connection that Airbnb is adamant about cultivating, however, has dangerous consequences in a market long plagued by discrimination against racial and ethnic minorities.[4] After massive criticism, the Airbnb has taken measures to fight against discrimination. As of October 2018, rather than displaying a potential guest's profile photo before the booking is accepted, hosts now receive a guest's photo only after they've accepted the booking request. Additionally, Airbnb hosts explicitly agree to a standard and to

[1] The Select Committee on Communications, Regulating in a digital world, HL Paper 299, 2019. p. 3.

[2] See Lopez, P, Bias does not equal bias: a socio-technical typology of bias in data-based algorithmic systems, Internet policy review, vol. 10, issue 4, 2021.

[3] Facebook Lets Advertisers Exclude Users by Race — ProPublica, last accesses 8.5.2022.

[4] Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157 (9th Cir. 2008)

adhere to a nondiscrimination policy that goes beyond what is required by law, in most jurisdictions. Additionally, specially trained teams have been brought in to handle discrimination complaints and enforcement.[5]

Another example of discrimination by design is the site Roommates.com which requires subscribers to express preferences in a dropdown menu that lists gender, sexual orientation, and family status as potential options. A participant had to share such a preference to find a match. In the US, the Fair Housing Act forbids advertising housing with any preference for race, sex or family status. The case was brought to trial by Fair housing groups, who accused Roomates.com of facilitating discrimination. A district court agreed, barring the website from soliciting information on users' sex, sexual orientation or family status. However, the appeal court overturned the decision and found that it would be a serious invasion of privacy, autonomy and security "to prevent people from choosing roommates with compatible lifestyles".[6]

Besides the discrimination by the platforms themselves, discrimination can be indirect- when internet users discriminate other internet users online using intermediaries (such as online platforms[7]). Such discrimination can take many different forms. It can occur in social networking sites, chat rooms, discussion boards, through text messaging, web pages, online videos, music, and online games. This issue opens difficult question on the scope of private authority and public regulation. Should the responsibilities of private tech companies derive from human rights law, terms of service, contracts or something else?[8]

Whilst algorithmic discrimination poses a challenge as it is often "hidden"- meaning that the users are often not aware they are being discriminated against- the discrimination users post on online platforms is more prominent and "visible". As platforms such as Facebook, Instagram and YouTube are not traditional media publishers with editorial control, there is uncertainty about whether they should bear liability for the discriminatory conduct and comments their users post online. It is however uncontested that they do have great power to control the information available to the online users.[9]

Online discrimination may resemble traditional discrimination,[10] but it can have more serious consequences, as the internet plays an essential role in our lives, shaping our conception of the world, our opinions and our values.

Online discrimination, just like any other discrimination is often motivated by hate or prejudice, and it is sometimes not possible to distinguish (online) discrimination and (online) hate speech.[11] Hate speech covers

[5] Airbnb Works To Clean Up Its Reputation For Racial Discrimination In New 3-Year Report - Essence last accessed 6.5.2022.

[6] Roommate-matching site does not violate housing laws, court | Reuters last accessed 6.5.2022.

[7] OECD (2019), What is an "online platform?, in An Introduction to Online Platforms and Their Role in the Digital Transformation, OECD Publishing, Paris, 2019, https://doi.org/10.1787/19e6a0f0-en ,last accessed 28.04.2022.

[8] Human Rights Council, Report of the special Rapporteur on the promotion and protection of the freedom of opinion and expression, 2016, p .3.

[9] Levy, K, Barocas, S, Designing Against Discrimination in Online Markets Berkeley Technology Law Journal, vol. 32:1183, 2017.

[10] Gaylord-Harden NK, Cunningham JA, The impact of racial discrimination and coping strategies on internalizing symptoms in African American youth. Journal of Youth and Adolescence. 2009;38(4):532–543. doi: 10.1007/s10964-008-9377-5; Umana-Taylor AJ, Wong JJ, Gonzales NA, Dumka LE. Ethnic identity and gender as moderators of the association between discrimination and academic adjustment among Mexican-origin adolescents. Journal of Adolescence. 2012;35(4): 773–786. doi: 10.1016/j.adolescence.2011.11.003 .

[11] The Equality and Anti-Discrimination Ombud's Report: Hate Speech and Hate Crime, 2015, available at: Hatytringer og hatkriminalitet_Engelsk.indd (ldo.no), last accessed 30.3.2022, p. 6.

many forms of expressions which advocate, incite, promote, or justify hatred, violence and discrimination against a person or group of persons for a variety of reasons.[12]

Internet has no borders, and discrimination online is a global issue that needs to be addressed at international, regional, and national level. The question is whether the current legislation is sufficient to provide for protection against the discrimination in the online setting. This study therefore aims, *inter alia*, at mapping out the existing legal solutions, accompanying policy measures, community standards and good practice to address and redress discrimination on online platforms.

By analysing existing regulations and community standards, we intend to evaluate which measures are best suited for preventing and redressing online discrimination, and we will eventually provide examples of best practices.

The paper starts with an overview of principle of non-discrimination in international and regional (European) documents. It then looks at the question of liability of social media platforms for the content posted by third parties. Outsourcing the task of defining the infringement and balancing human rights to private companies (privatization of justice) results in several complex legal issues.

The analysis ends with the conclusion on best practices which can contribute to achieve a goal of making the internet a better, more respectful environment where the infringements of human rights are minimized, and vulnerable groups are not exposed to discriminatory practices.

Regulation of online discrimination

There is no doubt the internet is going from the space of freedom to becoming more regulated. The result is that there is more accountability for the users and providers of content, as well as online platforms. The speed of technological change and its transnational character however make the regulation of digital world a challenging task.[13]

Any interference by the state with freedom of expression and information must comply with the rule of law and meet the strict criteria laid down in international human rights law; it must be prescribed by law, pursue a legitimate aim and be proportionate.[14] However, as argued by Jørgensen and Anja Møller Pedersen, the current regulatory schemes are insufficient to provide the standards and compliance mechanisms required to meet these standards.[15] In addition, as it will be shown in the following, most of the sources of law deal with discrimination in general, and not particularly online discrimination, making the legal framework quite fragmented and complex, as the rules on intermediary liability come into play. This eventually leads to legal uncertainty for victims of online discrimination, as well as for the social media platforms which need to comply with the rules. Complex and fragmented laws can increase operational costs, potentially leading them to simplify by being too restrictive.

[12] European Commission against Racism and Intolerance, Hate speech and violence (coe.int), last accessed 30.3.2022.

[13] The Equality and Anti-Discrimination Ombud's Report: Hate Speech and Hate Crime, Hatytringer og hatkriminalitet_Engelsk.indd (ldo.no), last accessed 31.3.2021, p. 7.

[14] Council of Europe, Recommendation CM/Rec(2014)6 of the Committeee of Ministers to member States on a guide to human rights for Internet users - Explanatory Memorandum. Strasbourg: Council of Europe, para. 47.

[15] Jørgensen and Pedersen, p. 2.

International legal framework

The UN Framework recognizes that States have the duty under international human rights law to protect everyone within their territory and/or jurisdiction from human rights abuses. This includes both the positive and negative obligations of the state. Besides not infringing the citizens' rights themselves, States have a duty to have effective laws and regulations in place to prevent and address human rights abuses and ensure access to effective remedy for those whose rights have been abused.[16] Several international documents, such as the UN Charter[17] and the Universal Declaration of Human rights,[18] prohibit discrimination. According to these documents, human rights are universal – to be enjoyed by all people, no matter who they are or where they live.

In addition, two international covenants from 1966 - Covenant on Civil and Political Rights (ICCPR), and the Covenant on Economic, Social, and Cultural Rights (ICESCR) contain general and specific non-discrimination clauses. The principal clause on non-discrimination is found in Article 26 of the ICCPR. It provides an autonomous right of equality and prohibits discrimination in law or in fact in any field regulated and protected by public authorities.[19] This Convention also on art 19(3) allows for restrictions on the freedom of expression, when these are useful, reasonable and desirable.[20]

Other conventions are more specific in terms of discrimination grounds, or they apply only to certain vulnerable groups. The Convention on the Elimination of All Forms of Racial Discrimination (ICERD, 1965) and Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW, 1979), are legally binding universal instruments containing implementation mechanisms. The Committee on the Elimination of Racial Discrimination (CERD) bases its practice on the 'living instrument' doctrine - a key vehicle for evolution and innovation within the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD). This ensures the treaty can respond to contemporary manifestations of racial discrimination while staying within the scope of its provisions.[21] This approach is important for the purpose of extending the scope of these conventions to online discrimination.

We should also mention the UN Guiding Principles on Business and Human Rights (United Nations Human Rights Council, 2011) as a set of guidelines for States and companies to prevent and address human rights abuses committed in business operations. They are a prevailing soft law standard for the human rights responsibility of private actors.[22]The Guiding principles reaffirm that states must ensure that not only State organs, but also businesses under their jurisdiction respect human rights. [23] These Principles

---

[16] The UN Working Group On Business And Human Rights, The UN Guiding Principles On Business And Human Rights An Introduction, Intro_Guiding_PrinciplesBusinessHR.pdf (ohchr.org), last accessed 30.3.2022.

[17] § 26 of the Charter of United Nations, available at: Charter of the United Nations.pdf (undp.org), last accessed 28.3.2022.

[18] Of the thirty articles, some are in one way or another explicitly concerned with equality, and the rest implicitly refer to it by emphasizing the all-inclusive scope of the Universal Declaration of Human Rights.

[19] § 26 of ICCPR reads: "All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status."

[20] European Court of Human Rights, App. No. 6538/74, The Sunday Times v The United Kingdom, 26 April 1979, para 59.

[21] Keane, David, Mapping the International Convention on the Elimination of All Forms of Racial Discrimination as a Living Instrument, Human Rights Law Review, Vol. 20, Issue 2, 2020, p. 236.

[22] Jørgensen, Rikke Frank and Møller Pedersen, Anja, Online service providers as human rights arbiters, The Responsibilities of Online Service Providers. Mariarosaria Taddeo; Luciano Floridi (eds.) Springer, Law, Governance and Technology Series, Vol. 31).2017. p. 179-199.

[23] Guiding Principles on Business and Human Rights, Ch. 1 (A) (1).

assert a global responsibility for businesses to avoid causing or contributing to adverse human rights impacts through their own activities, and to address such impacts when they occur and seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products of services, even if they have not contributed to those impacts. [24] As a matter of transparency, the Guiding Principles state those businesses should be prepared to communicate how they address their human rights impacts externally, particularly when the concerns are raised by or on behalf of affected stakeholders. [25]

None of these documents mentioned above are specific for online discrimination but are used as general sources of law. There does not exist any international conventions on intermediary liability either.

Regional documents

Council of Europe (CoE)

The most important document on human rights in Europe - European Convention on Human Rights (hereafter Convention or ECHR) from 1950 explicitly forbids discrimination in Art 14.

According to the Court's case-law, the principle of non-discrimination is of a "fundamental" nature and underlies the Convention together with the rule of law, and the values of tolerance and social peace.[26]

The expression "direct discrimination" describes a "difference in treatment of persons in analogous, or relevantly similar situations" and "based on an identifiable characteristic, or 'status'[27] protected by Article 14 of the Convention.[28] Article 14 however is an accessory right, and it applies only in conjunction with other Convention rights.

In addition to article 14, in Article 1 of Protocol 12 to the Convention,[29] it is stated:

1. The enjoyment of any right set forth by law shall be secured without discrimination on any ground such as sex, race, color, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.
2. No one shall be discriminated against by any public authority on any ground such as those mentioned in paragraph 1.

Article 1 of the Protocol 12 hence represents a general prohibition of discrimination[30] and an independent right not to be discriminated against. It confirms that the state has both positive and negative obligations - to secure protection of individuals against discrimination and not to actively discriminate.

---

[24] Ibid. Ch. II (A) 11-13.

[25] Ibid. Ch. II (B) (21).

[26] S.A.S. v. France [GC], Application no. 43835/11 of July 1st 2014, § 149; Străin and Others v. Romania, Application no. 57001/00 of July 25 2005, § 59.

[27] Biao v. Denmark [GC], Application no. 38590/10 of May 24th 2016, § 89; Carson and Others v. the United Kingdom [GC], Application no. 42184/05 of March 16th 2010, § 61; D.H. and Others v. the Czech Republic [GC], Application no. 57325/00 of November 13th 2007, § 175; Burden v. the United Kingdom [GC], 2008, § 60.

[28] Varnas v. Lithuania, Application no. 42615/06 of December 9th 2013, § 106; Hoogendijk v. the Netherlands, Application no. 58641/99 of January 6th 2005.

[29] 20 states have ratification Protocol 12 as of November 2021.

[30] Savez crkava "Riječ života" and Others v. Croatia, Application no. 7798/08, December 9th 2010, § 103; Sejdić and Finci v. Bosnia and Herzegovina [GC], application nos. 27996/06 and 34836/06, December 22nd 2009, § 53.

Article 10 ECHR contains the right to freedom of expression, and in paragraph 2 of Article 10 the Convention acknowledges that the exercise of these freedoms carries with it duties and responsibilities, and may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others (...).

In addition to ECHR, CoE countries have adopted several other conventions with regard to specific protected grounds. The Framework Convention for the Protection of National Minorities of 1994 (FCNM) in its Article 6, encourages Parties to intercultural dialogue and to take appropriate measures to protect persons who are subject to threats or acts of discrimination, hostility or violence due to their ethnic, cultural, linguistic or religious identity. This convention does however not have any specific rules on online discrimination.

Other relevant Council of Europe documents

The Council of Europe Commission against Racism and Intolerance (ECRI) established in 1993 is the CoE's independent human rights monitoring body.[31] Its mandate is combating racism, discrimination, (on grounds of "race", ethnic/national origin, color, citizenship, religion, language, sexual orientation and gender identity), xenophobia, antisemitism and intolerance in Europe.

ECRI issues General Policy Recommendations (GPRs) addressed to the governments of all member states. These recommendations provide guidelines which policymakers are invited to use when drawing up national strategies and policies. One recommendation relevant for this study is ECRI General Policy Recommendation No. 6 (2000) - Combating the dissemination of racist, xenophobic and antisemitic material via the Internet.

Recommendation CM/Rec (2014) 6 of the Committee of Ministers to member states on a Guide to human rights for Internet users states that restrictions to online freedom of expression may apply to expressions which incite discrimination, hatred or violence. These restrictions must be lawful, narrowly tailored and executed with court oversight. So even though the freedom of expression is highly valued right, it is not absolute, and can be restricted in order to protect other interests, such as the right not to be discriminated against.

A distant reference to social media is contined in Recommendation CM/Rec (2011) 7 on a new notion of media. According to this document, the actors operating collective online, shared spaces which are designed to facilitate interactive mass communication should be attentive to the use of, and editorial response to, expressions motivated by racist, xenophobic, anti-Semitic, misogynist, sexist (including as regards LGBT people) or other bias. These actors may be required (by law) to report to the competent authorities' criminal threats of violence based on racial, ethnic, religious, gender or other grounds that come to their attention.[32] The threshold for the platforms to act is high, as it requires the discrimination to reach the level of criminal threats.

Some of the CoE recommendations contain the measures which are meant to help remedy the violations of the human rights online. In appendix to Recommendation Rec (2001) 8 of the Committee of Ministers to

[31] European Commission against Racism and Intolerance (ECRI) - Homepage (coe.int) last accessed 31.3.2022.

[32] § 91.

member states on self-regulation concerning cyber content (self-regulation and user protection against illegal or harmful content on new communications and information services), the member states are encouraged to establish content complaints systems, such as hotlines, which are provided by Internet service providers, content providers, user associations or other institutions. Such content complaints systems should, where necessary for ensuring an adequate response against presumed illegal content, be complemented by hotlines provided by public authorities.[33]

Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries requires the states to guarantee accessible and effective judicial and non-judicial procedures that ensure the impartial review, in compliance with Article 6 of the ECHR, of all claims of violations of Convention rights in the digital environment, including the right not to be discriminated against in the enjoyment of all the rights and freedoms set forth in the ECHR. They should furthermore ensure that intermediaries provide users or affected parties with access to prompt, transparent and effective reviews for their grievances and alleged terms of service violations, and provide for effective remedies, such as the restoration of content, apology, rectification or compensation for damages. Judicial review should remain available, when internal and alternative dispute settlement mechanisms prove insufficient or when the affected parties opt for judicial redress or appeal. [34]

In the most recent Recommendation CM/Rec(2022) 13 of the Committee of Ministers to member States on the impacts of digital technologies on freedom of expression, on 6 April 2022, the issue of algorithmic discrimination is expressly addressed:

"When there are legitimate concerns that their policies may lead to discrimination, internet intermediaries should provide information that allows independent third parties to evaluate whether their policies are implemented in a non-discriminatory way, including by disclosing the datasets upon which automated systems are trained in order to identify and correct sources of algorithmic bias."[35]

Some of the CoE recommendations relate to protection against discrimination of particular groups. For instance, Recommendation CM/Rec (2010) 5 on measures to combat discrimination on grounds of sexual orientation or gender identity includes the obligation to combat inciting hatred or other forms of discrimination against LGBTI+ persons. It covers all forms of discrimination, including online discrimination.

EU law

European union has for decades worked to establish the values of inclusion, non-discrimination, multilingualism and cultural diversity, which are epitomized in EU's motto: "United in diversity", and it is crucial for the EU as a project that these values also are reflected in the online setting. The prevention and response to discrimination are values enshrined in Article 2 of the Treaty of the European Union (TEU).[36] This provision states that:

[33] Chapter IV, 12.

[34] Art. 1.5.1 and 1.5.2.

[35] § 3.6.

[36] Official Journal of the European Union C 326/15 of 26.10.2012.

"The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail."

Prohibition of discrimination is contained in EU's Charter of fundamental rights[37] Article 21:

1. "Any discrimination based on any ground such as sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.
2. Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited."

On the level of secondary legislation, European antidiscrimination legislation is contained in directives,[38] such as Directive 2000/43/EC against discrimination on grounds of race and ethnic origin,[39] Directive 2000/78/EG establishing a general framework for equal treatment in employment and occupation[40], Directive 2006/54/EG, on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast)[41]and the E-Commerce Directive.[42]

The EU regulatory framework on content moderation online is increasingly complex and has been differentiated over the years according to the category of the online platform and the type of content reflecting a risk-based approach. For instance, Audio-Visual Media Services Directive,[43] imposes particular obligations to one category of online platforms, the VideoSharing Platforms. They should take appropriate and proportionate measures, preferably through co-regulation, in order to protect the general public from illegal content. Those measures must be appropriate in the light of the nature of the content, the category of persons to be protected and the rights and legitimate interests at stake and be proportionate taking into account the size of the platforms and the nature of the provided service.

The Counter-Racism Framework Decision (CRFD), which was adopted by the Council alone, seeks to combat particularly serious forms of racism and xenophobia through criminal law but does not define racism and xenophobia nor use the terms racist and xenophobic hate speech.[44]

Instead, the CRFD criminalizes two types of speech - publicly inciting to violence or hatred and publicly condoning, denying or grossly trivializing crimes of genocide, crimes against humanity and war crimes - when they are directed against a group of persons or a member of such a group defined by reference to race,

[37] Charter of Fundamental Rights of the European Union, OJ C 326, 26.10.2012, p. 391–407.

[38] Directives are a method of harmonization of the law, i.e. legal acts that are binding as to the result to be achieved, but that leaves member states discretion as to how to achieve the result.

[39] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. OJ L 180, 19.7.2000, p. 22–26.

[40] OJ L 303, 02.12.2000, p. 16-22.

[41] OJ L 204, 26.7.2006, p. 23–36.

[42] OJ L 178, 17.7.2000, p. 1–16.

[43] Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive), OJ L 95, 15.4.2010, p. 1–24.

[44] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, O.J. [2008] L 328/55.

color, religion, descent or national or ethnic origin. The list of protected grounds is limited to these five characteristics.

While only hate speech which is racial and xenophobic has been made illegal by the CRFD, Member States may go beyond the EU minimum and criminalize other types of hate speech, by referring to a broader list of protected characteristics (a list including e.g., religion, disability, sexual orientation).

In 2017 European commission issued a Communication on tackling illegal content online[45] where it expressed the need for the online platforms to, 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.[46] In 2018 a Recommendation on measures to effectively tackle illegal content online[47] was adopted calling the online platforms to act in a diligent and proportionate manner towards the content they host, especially when processing notices and counter-notices and deciding on the possible removal of or disabling of access to content considered to be illegal. 'Illegal content' arguably encompasses a large variety of content categories that are not compliant with EU and national legislation, including discrimination.

According to the E-Commerce Directive**,** which is considered to be foundational legal framework for online services in the EU , the intermediaries benefit from "safe harbors"- which means that they cannot be subject to a general obligation to monitor users' online content, and they are exempt from liability unless they are aware of the illegality and are not acting adequately to stop it. [48] As the social media platforms are considered to be passive and neutral, they are exempted from liability for illicit content posted by their users. However, they cannot rely on the exemptions from liability if they were aware of the facts or circumstances on the basics of which a diligent economic operator should have realized that the publication was unlawful and failed to act expeditiously. [49]

National laws - examples from Australia, Germany and France

On September 8th 2021 the High Court of Australia found that media companies may be liable for the defamatory comments of third parties on social media platforms.[50]

By this, the stronger protection is given to right to reputation vis-a vis right to freedom of expression, as the liability for damages is not limited to the author of the defamatory comments but extends to publishers who allow the defamatory content to be posted on their social media platforms and thereby encourage or facilitate the defamatory publication. Australian law has thereby abandoned the safe harbor principle and opted for a stricter liability of intermediaries with regards to illegal content.

[45] Communication from the Commission to the European parliament, the Council, the European economic and social committee and the Committee of the regions tackling illegal content online -towards an enhanced responsibility of online platforms, COM/2017/0555 final

[46] § 10 of the Communication

[47] Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online. C/2018/1177

[48] See Directive 2000/31 EC (e-commerce directive), art. 12.

[49] C-324/09 of 12 July 2011, § 119.

[50] High Court of Australia in Fairfax Media Publications Pty Ltd v Voller; Nationwide News Pty Ltd v Voller; Australian News Channel Pty Ltd v Voller, S236/2020 S237/2020 S238/2020.

In Europe, Germany's Network Enforcement Act, or NetzDG law[51] represents a key test for combatting hate speech on the internet. Under the law, which came into effect on January 1, 2018, online platforms face fines of up to €50 million for systemic failure to delete illegal content. Supporters see the legislation as a necessary and efficient response to the threat of online hatred and extremism. Critics view it as an attempt to privatize a new 'draconian' censorship regime, forcing social media platforms to respond to this new painful liability with unnecessary takedowns.[52]

In France, recent case law has imposed proactive monitor obligations on intermediaries for copyright infringement even in cases where their liability is not engaged.[53]

This shows that the national legislators are moving towards strengthening the protection of rights online and the position of individuals in relation to big tech companies, and these rules can be applied *mutatis mutandis* to online discrimination as a form of illegal content.

Tech companies' internal regulations and policies

Over the past decade, tech giants such as Google (Alphabet) and Facebook (Meta) have become the biggest companies in the world.[54]

Despite the fact the companies such as Facebook have insisted they are only neutral platforms and have no editorial responsibilities, they have recently been under pressure to filter communication that appears on their platforms, including discriminatory content. As a result, they have introduced more rules to govern speech and participation and instituted several mechanisms for the removal of people and content that transgress these rules.

Terms of service (TOS) which individuals typically must accept as a condition to access the platform, often contain restrictions on content that may be shared. The inconsistent enforcement of terms of service however has also attracted public scrutiny. Some have argued that the world's most popular platforms do not adequately address the needs and interests of vulnerable groups, for example there have been accusations of reluctance to "engage directly with technology related violence against women, until it becomes a public relation issue". [55]

Voluntary codes of conduct for internet service providers have been adopted in several European states (e.g. the Netherlands, UK). In May 2016, European Commission and several platforms[56] agreed on the common "Code of conduct on countering illegal hate speech online"[57] to prevent and counter the spread of

---

[51] » Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG) German Law Archive (iuscomp.org), last accessed 31.3.2021.

[52] The Impact of the German NetzdG law – CEPS

[53] See APC v. Google, Tribunal de Grande Instance [TGI] [ordinary court of original jurisdiction] Paris, Nov. 28, 2013 (Fr.), https://www.legalis.net/jurisprudences/tribunal-de-grande-instance-de-paris-ordonnance-de-refere-28-novembre-2013/ [https://perma.cc/7JA2-37PB]; Nord-Ouest Prod. v. S.A. Daily Motion, Tribunal de grade instance [TGI] [ordinary court of original jurisdiction] Paris, July 13, 2007 (Fr.), https://www.legifrance.gouv.fr/affichJuriJudi.do?oldAction=rechJuriJudi&idTexte= JURITEXT000018861366&fastReqId=728956270&fastPos=2 [https://perma.cc/JX6L-45GZ]

[54] Top 20 Biggest Tech Companies in The World in 2021 - The Teal Mango last accessed 31.3.2021.

[55] See: Human Rights Council, Report of the Special Rapporteur on promotion and Protection to the Right of freedom of opinion and Expression, p. 14.

[56] Code of conduct is joined by Facebook, Microsoft, Twitter and YouTube, Instagram, Snapchat, Dailymotion Jeuxvideo.com. TikTok joined in September 2020. On 25 June 2021, LinkedIn also announced its participation to the Code of Conduct.

[57] The EU Code of conduct on countering illegal hate speech online | European Commission (europa.eu), last accessed 30.4.2022.

illegal hate speech online. The last evaluation shows that on average the companies are now assessing 81% of flagged content within 24 hours and 62.5% of the content deemed illegal hate speech is removed.[58]

The biggest disadvantage of such Codes of conduct are that they are self-regulatory mechanisms and are joined by platforms on voluntary bases. In other words, they are not enforceable.

Facebook has been under fire for giving advertisers the ability to exclude people from their targeting based on race, religion, sexual orientation. Now, it needs advertisers to comply with their updated non-discrimination policy. [59] In addition, Meta Platform Terms and Developer Policies state that one of the prohibited practices is:

"Processing Platform Data to discriminate or encourage discrimination against people based on personal attributes including race, ethnicity, color, national origin, religion, age, sex, sexual orientation, gender identity, family status, disability, medical or genetic condition, or any other categories prohibited by applicable law, regulation, or Meta policy."[60]

Tik Tok's content moderation policy contains unusual measures to protect supposedly vulnerable users. The platform instructed its moderators to mark videos of people with disabilities and limit their reach, as these users are „susceptible to harassment or cyberbullying based on their physical or mental condition".[61]

Online platforms as gatekeepers

Online platforms as private actors are powerful forces in facilitating freedom of expression online.[62] They are also seen as gatekeepers[63] and the first line of defence for protection of users' human rights online. EU regulators for example have outsourced the "first line" protection of human rights to intermediaries who are given responsibilities and tasks to disable or remove alleged illegal content on the internet. The companies have notice and take down procedures and are supposed to react timely and transparently to complaints of discriminatory content on their sites. These procedures should be simple and clear, making it easier for the victims of discrimination to get remedied. Here, the American Digital Millennium Copyright Act (DMCA) can be used as an inspiration. It explicitly regulates who should issue the notification of (copyright) infringement, to whom and what it should contain.

According to the E-Commerce Directive, intermediaries are exempted from liability for third party content (so called Safe harbors).[64] This has been confirmed in the ECtHR practice in the case of *MTE and*

[58] The EU Code of conduct on countering illegal hate speech online | European Commission (europa.eu), last accessed 30.04.2022.

[59] Review compliance for Facebook's Non-discrimination Policy | Facebook Business Help Centre

[60] 3 a (i), available at: Platform Terms - Facebook for Developers

[61] Discrimination: TikTok curbed reach for people with disabilities (netzpolitik.org), last accessed 31.3.2021.

[62] Laidlaw, Emily, Regulating Speech in Cyberspace: Gatekeepers, Human Rights, and Corporate Responsibility, Cambridge University Press, 2015, p. i.

[63] Ibid.

[64] 12-14 of 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)

Index v. Hungary[65] where the ECtHR had to decide whether a non-profit self-regulatory body of Internet content providers (MTE) and an Internet news portal (Index) were liable for offensive comments posted on their websites. The Court found that the intermediaries had no liability for the infringement of rights, and that "the notice-and-take-down-system could function in many cases as an appropriate tool for balancing the rights and interests of all those involved".[66]

However, it was pointed out in literature that this system allows platforms to tolerate discrimination, and this can perpetuate inequalities in our society. [67] Given the current state of technology, it should not be impossible for the tech companies to police user content take as a part of their due diligence.

It has been suggested to introduce the "systemic duty of care"[68] model- a legal standard for assessing a platforms overall system for handling online content. The idea is that platforms should improve their systems for reducing online harms, including detecting and removing illegal content. The systemic duty of care could be based on one of two models: A "prescriptive" model which defines precisely the measures a platform must take, and "flexible model in which the measures are left undefined.[69]

It is believed that the legal basis for this could be the E-Commerce Directive, which references to potential duties to "detect and prevent" illegal activities.[70]

Conclusion - suggestions and best practices

The online discrimination is often only a manifestation of the existing disparity and discrimination that exists in the "real world". In other words, digital rights violations often affect those who are already marginalized.

Online discrimination can be seen as a more serious form of discrimination, as the internet can make discrimination more visible and hurtful. As the issue of digital discrimination is likely to persist in the future, the platforms have the responsibility to prevent and minimize the effects of this form of human rights infringement, under the control of the state. Even though the automated systems for content detection or notice and take down systems enable prevention and quick reaction to infringements, these measures suffer from the lack of procedural safeguards and judicial review.

Online discrimination requires effective responses on international, regional, national levels, setting standards for the tech companies who are the gatekeepers and first lines of defense against this practice. The current situation in Europe regarding the liability of platforms is very diffuse and unclear.

Outsourcing the difficult task to protect against discrimination and balance different interests online is problematic as the procedural safeguards cannot be ensured. The platforms are not judicial organs and cannot be said to have a legal expertise to evaluate requests against general legal criteria. The Interamerican

[65] Magyar Tartalomszolgáltatók Egyesülete and Index.hu zrt v. Hungary, Application no. 22947/13, February 2nd 2016.

[66] Magyar Tartalomszolgáltatók Egyesülete and Index.hu zrt v. Hungary, Application no. 22947/13, February 2nd § 9.

[67] Sylvain, O., Discriminatory Designs on User Data, Emerging threats, 2018, available at: Discriminatory Designs on User Data | Knight First Amendment Institute (knightcolumbia.org) last accessed 01.05.2022.

[68] Keller, D, Systemic duties of care and intermediary liability, Center for internet and society, 2020, available at: Systemic Duties of Care and Intermediary Liability – Daphne Keller – Inforrm's Blog, last accessed 8.5.2022.

[69] Ibid.

[70] Recital 48.

Commission on Human Rights has observed that private actors "lack the ability to weigh rights and interpret the law in accordance with freedom of speech and other human rights standards."[71] In addition, there is problem of lack of transparency and to a large extent different attitudes and approaches by the various providers.[72] In addition, intermediaries that operate in diverse range of markets inevitably face "complex value judgements", issues with cultural sensitivity and diversity and "difficult decisions about the conflict of laws".[73]

In other words, even though the alleged infringement of right not to be discriminated against handled by the platform is a quicker and more efficient way to justice, it lowers legal certainty. Therefore, controlling the access to content and services has to be subject to judicial review.

The online platforms should, as part of their social responsibly due diligence, contribute to prevention and protection of the right not to be discriminated against in order to secure the equality of all citizens. They should be incentivized to uncover the illegal content, rather then be punished for not preventing it. This way, the potential "over-removing" in order to avoid liability would be avoided. The applications should be designed to elicit and ensure that certain kinds of content do not even occur in the first place.

Even though the safe harbor principle still applies in Europe and the USA, platforms can hardly be said to be neutral and passive in the modern world. Intermediaries are not mere conduits that purport to provide free and uninhibited forum for social interaction. They are implicated in every user utterance or act, even if they do not moderate posts.[74] They have the ability to control over the information published on the platform, and they have financial profits form the information they convey.[75] In addition, intermediaries design their platforms in ways that shape the form and substance of their users' content.[76] Their role is no longer to transmit or store material on behalf of the users-rather it fulfils an active role in the organization and functioning of the websites.[77] It is in control of what users see, much like a newspaper editor - and like editors should have some form of liability.[78]

Given all said, what could be effective practices and policies that address, mitigate and/or prevent online discrimination?

- Online discrimination can have grave consequences for public safety and social inclusion and should be expressly addressed in international, legal and national regulations, and these sources of law should be harmonized
- States, tech companies and NGOs should work together on raising awareness of the problem of discrimination online, so people can recognize discriminatory practices and know their rights

[71] Inter-American Commission in Human Rights, Freedom of Expression and the Internet, pp. 47-48.

[72] Van der Sloot, p. 224.

[73] Taylor, E, The Privatization of Human Rights: Illusions of Consent, Automatisation and Neutrality, GCIG Paper No. 24 (2016).

[74] Sylvain, O, Intermediary Design Duties, Connecticut Law Review, vol 50, nr. 1 2018, p. 226.

[75] Sylvain, O., Discriminatory Designs on User Data, Emerging threats, 2018, available at: Discriminatory Designs on User Data | Knight First Amendment Institute (knightcolumbia.org) last accessed 01.05.2022.

[77] Van der Sloot, B, Welcome to the Jungle:the Liability of Internet Intermediaries for Privacy Violations in Europe, JIPITEC, 6(2015), p. 212.

[78] Keller, D, op.cit.

- More research about online discrimination is needed so this practice can be recognized and better addressed
- Tech companies ought to share best practices in detecting and avoiding discriminatory practices
- Tech companies ought to cooperate on developing the automated systems of content control instead of developing parallel systems, which would be more cost efficient and result in more harmonized systems
- Filtering algorithms would require human review to prevent human rights violations and discrimination.
- The existing mechanisms for reputational and copyright protection such as notice and take down procedures and the right to be forgotten can analogously be applied in case of online discrimination
- Online platforms should have independent bodies consisting of law experts evaluating the reported cases of discrimination in order to achieve better balancing of rights
- There is a need for additional transparency measures for online platforms, including on the algorithms used. Platforms that feature user-generated content should offer users a clear explanation of their approach to evaluating and resolving reports of hateful and discriminatory content, highlighting their relevant terms of service
- Greater ease for reporting cases of online discrimination (user-friendly mechanisms and procedures)
- Platforms should enforce sanctions of their terms of service in a consistent, timely and fair manner
- Platforms should abide by duty of care, going beyond notice-and-takedown based legal models
- Internet intermediaries can no longer be considered passive and neutral transmitters of information, and there should not be exempt from liability for online discrimination
- Legislative framework for handling of requests to take down discriminatory content should be put in place
- Procedural protections should be built into platforms notice-and-takedown systems
- Rules should incentivize intermediaries and users to detect illegality, while minimizing the risks and the costs of errors and safeguarding a balance between the different human rights at stake
- Tech companies need to ensure algorithm transparency and neutrality
- A balance between citizens and tech companies must be struck in designing the liability rules
- Setting up a detailed and harmonized European notice and take down procedure would provide more legal certainty