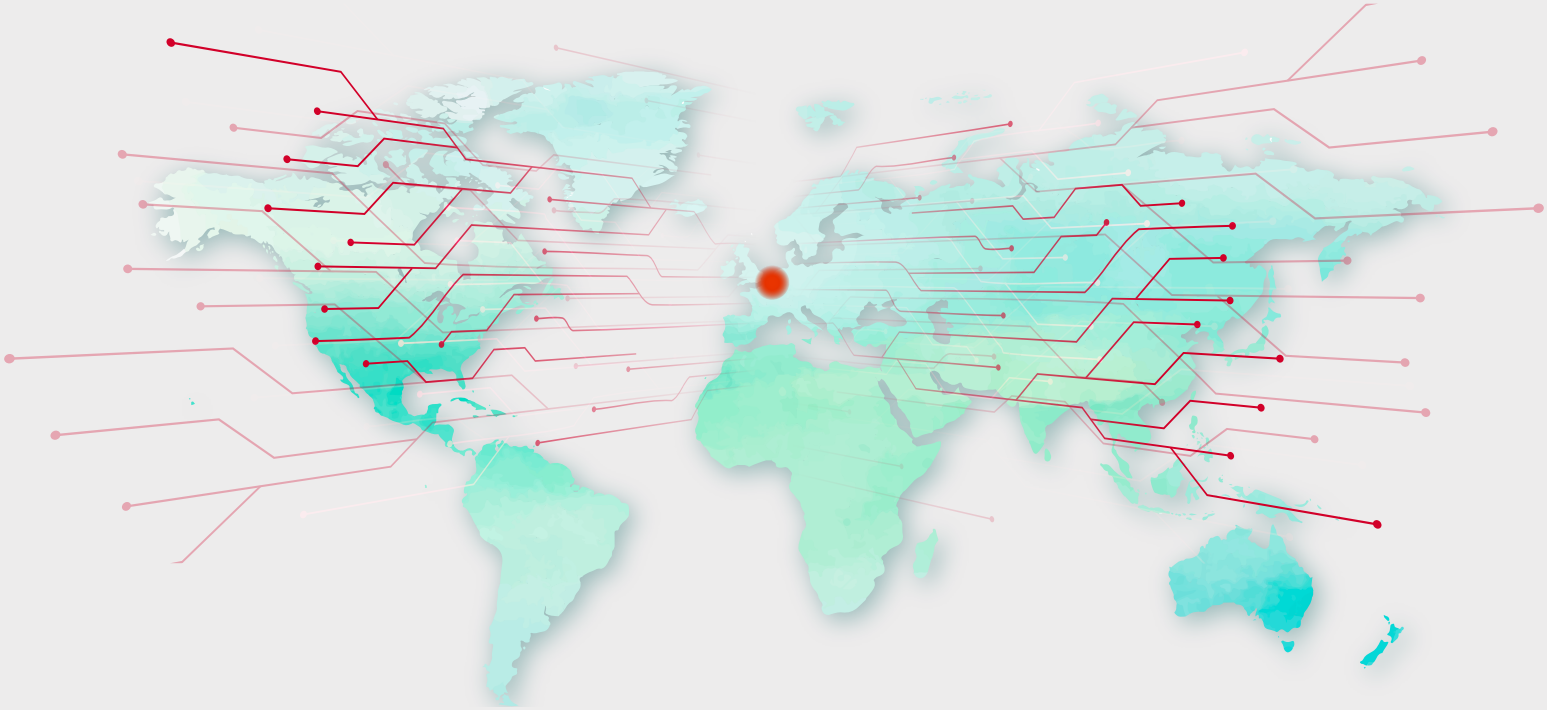
THE FUTURE OF FREE SPEECH

JUSTITIA

REBUILDING THE BULWARK OF LIBERTY



THOUGHTS ON THE DSA:

Challenges, Ideas and the Way Forward through International Human Rights Law

Jacob Mchangama, Natalie Alkiviadou and Raghav Mendiratta

Content

Abstract

National and regional legislative measures/proposals that dramatically enhance platform liability for content developed by users such as the German Network Enforcement Act (NetzDG) and the EU's proposed Digital Services Act (DSA) place free speech at risk and potentially shrink civic space. Such measures render private companies, not bound by International Human Rights Law (IHRL) arbiters of fact and law. To meet obligations and avoid hefty fines, social media platforms (SMPs) are adopting a "better safe than sorry approach," increasingly relying on Artificial Intelligence (AI) to (proactively) remove even contentious areas of speech such as hate speech. Against the backdrop of the current developments in the form of the proposed DSA, this paper provides an overview of the challenges that emanate from the current European approach with a particular emphasis on contested areas of speech such as hate speech and disinformation and puts forth proposals that can be taken into consideration during negotiations and discussions. Whilst cognizant of the fact that the structural composition of the DSA, in particular its platform liability approach will not change (for now), this paper puts forth ideas that could feed into the negotiation process, namely a rights-based approach to content moderation. We make specific reference and recommendations as to the recent proposal by the European Commission of a mechanism for facing disinformation in exceptional circumstances (in reaction to the ongoing Ukrainian crisis and related pro-Russian information manipulation).

Introduction

Social media platforms have "created unprecedented possibilities for widespread cultural participation and interaction".[1] With 4.2 billion active social media users,[2] this has been a great leap forward for humanity, empowering marginalized and traditionally silenced groups, enhancing connectivity and allowing for awareness raising on, *inter alia,* human rights violations. In fact, in today's digital reality, people have "little choice but to participate" on online platforms.[3] Nevertheless, the other side of this coin is marked by phenomena such as hate speech, violent extremism and disinformation, with tragic events including the abuse of platforms during the Rohingya genocide.[4]

Through content moderation policies, platforms define the limits of speech, an issue which has emerged as "one of the most pressing challenges for freedom of expression in the 21st century".[5] Private companies, driven by their own business models are now "content gatekeepers",[6] exerting immense influence over public discourse globally. However, the plot thickens since their role is enhanced by the escalating moderation responsibilities imposed by governments through legislation, making such platforms "even more powerful".[7] As noted by McGowan when discussing the DSA, enhanced platform responsibility and liability "betrays the legislators' goal of limiting corporate power over public discourse by formally assigning companies a role in deciding the legality of our speech".[8]

As in the offline world, freedom of expression on SMPs is not absolute. For example, in May 2020, France passed legislation compelling SMPs to remove "manifestly illicit"[9] hate speech within 24 hours. Companies not complying with this requirement would face fines of up to 1.25 million Euros. In June 2020, France's Constitutional Council ruled that this law limited the freedom of expression in an unnecessary, inappropriate and disproportional manner.[10] This law can be seen as a legislative precedent set by Germany and its 2017 Network Enforcement Act (The NetzDG),[11] which also imposes a legal obligation on social media companies to remove illegal content, including insult, incitement and religious defamation within 24 hours and at risk of a fine of up to 50 million Euros. The NetzDG "alarmed human rights campaigners".[12] Two reports[13] released by Justitia demonstrate how this precedent has spilled over in more than twenty States, including authoritarian regimes such as Russia and Venezuela.

Heightening pressure on private companies to remove content, makes them "even more risk averse in their moderation policies,"[14] thereby shrinking civic space and placing free speech in dire straits. Big platforms already have terms which regulate permissible content beyond that which is illegal, with companies such as Facebook spending "growing amounts of resources to police content and take down illegal, harmful and objectionable content".[15] This is reflected in a 2022 study issued by

Justitia which included a legal analysis of 2.400 Facebook comments that were labelled as "hateful attacks" by a sophisticated algorithm.[16] The comments were a representative sample of over 900,000 hateful attacks found by analyzing 63 million comments on Facebook pages belonging to Danish politicians and media outlets. Justitia found that only 11 comments, or 0.066% of the total could be considered illegal under Danish law. As such, an estimate is that only 1 out of 15.000 comments found on Danish Facebook pages are, in fact, illegal.

Enhanced state pressure to moderate content has also contributed to SMPs' increased use of AI to remove content in order to avoid liability and protect their business models.[17] Relying on AI, even without human supervision, can be supported when it comes to content that could never be ethically or legally justifiable, such as child abuse. However, things are not as straightforward for contentious areas of speech such as hate speech that depend on intent, context and nuance. Whilst technology such as natural language processing and sentiment analysis have been developed to detect harmful text, without having to rely on specific words/phrases, research has shown that they are "still far from being able to grasp context or to detect the intent or motivation of the speaker".[18] As Keller explains, "no reputable experts suggest that filters are good enough to be put in charge of deciding what is illegal in the first place."[19]

Overly broad and restrictive content moderation, both public or private, should be avoided. Several studies suggest that improper and overbroad removals make some users suspicious and may counteractively reinforce falsehoods and violent extremism. For example, a 2022 paper by Bartusevičius et. al analyzed 101 nationally representative samples from three continents and revealed a positive association between perceived levels of repression and intentions to engage in anti-government violence. Additional analyses from three specific countries in the studies characterized by widespread repression and antigovernment violence identified a strong positive association between personal experience with repression and intentions to engage in anti-government violence. These results suggest that political repression of speech, aside from being normatively abhorrent, creates psychological conditions for political violence. Similarly, another recent study published in Harvard's Misinformation Review documented how, once President Trump's election fraud Tweets were labelled as misinformation on Twitter, they gained more traction on other platforms.[20] The study argued that Twitter's labelling of certain Tweets was not only ineffective at preventing the spread of Trump's claims, it might have even backfired at an ecosystem level by drawing additional attention to messages that Twitter deemed problematic. Nevertheless, the pressure on platforms to remove disinformation is steadily increasing.

Further, studies have also shown that extremists who are deplatformed from mainstream social media for violating terms, migrate elsewhere with fewer rules.[21] This may not only defeat law

enforcement but also impede counter-narrative efforts, which could be effective in reducing hate speech. In research conducted by the Royal United Services Institute on the far-right group 'Britain First', scholars found that limiting the accessibility of extremists to Facebook reduced their interaction with others and the dissemination of their ideas. However, their migration to other platforms with less moderation lead to their content becoming more extreme.[22] Ravndal argues that the rise of far-right extremism in Western Europe emanates from a combination of high immigration, low electoral support for radical right political parties and the "extensive public repression of radical right actors and opinions".[23] Although he notes that such repression may discourage people from joining extreme groups, it may also push others to follow more violent paths.[24] As such and as highlighted by Erixon, there is a need to understand the "behavioural consequences that follow from heavy-handed approaches to content regulation".[25]

In light of the above, our position is that the current handling of contentious speech online (such as hate speech and disinformation) through enhanced platform responsibility is deeply problematic. What this paper will attempt to do is put forth formulae that can be taken into account during negotiations on the proposed DSA so as to limit, at least in part, some of the negative effects of enhanced platform control on content. On this premise, it will turn to IHRL as the foundation of content moderation. IHRL provides broader protection to free speech than the European Convention on Human Rights (ECHR)[26] as interpreted by the European Court of Human Rights (ECtHR) and subsequently, the Charter of Fundamental Rights of the European Union, which reaffirms and endorses the obligations of the ECHR.[27] To be compliant with IHRL, a platform's content moderation practices must be legitimate, necessary, and proportional within the framework of Article 19(3) ICCPR (restrictions on freedom of expression), which sets out the grounds for limitation of freedom of expression. In this brief paper, we discuss how IHRL norms could be molded and relied upon to regulate two of the most contentious categories of speech in Europe and the rest of the world, hate speech and disinformation. This is imperative for the future of free speech since the DSA is likely to have global consequences and, as such, the normative framework should emanate from the international rather than the European setting. On the issue of IHRL and content moderation, former Special Rapporteur on the Freedom of Opinion and Expression (SRFOE), David Kaye recommended that platforms "should recognize that the authoritative global standard for ensuring freedom of expression on their platforms is human rights law … [and] re-evaluate their content standards accordingly".[28]

The recognition of IHRL, at least on a theoretical level, has been seen in approaches of major platforms such as Facebook. For example, in March 2021, Facebook launched its Corporate Human Rights Policy,[29] which outlines the human rights standards as defined in international law and sets out how they will be applied to, amongst others, their policies. Individual thought leaders The

Oversight Board (a body of international experts created by Facebook in 2020 to make final decisions regarding content moderation questions, including the evaluation of complaints by users) has also embraced IHRL in judging the appropriateness of Facebook's content moderation decisions. The Oversight Board, in deciding a variety of issues ranging from hate speech to nudity to dangerous individuals and organizations has relied on relevant provisions and principles of IHRL.

The IHRL benchmark is paramount for ensuring legitimacy to the restriction of freedom of expression, one that is emphatically missing from the DSA's very structure and essence. The section on IHRL will thus follow suit on recommendations of the SRFOE and Justitia's extensive report on IHRL as a framework of first reference.[30] In addition to proposing ways in which IHRL can be better embedded in the EU's vision for its digital future, we will also discuss how decentralization of content moderation could boost end user control over data and privacy and simultaneously protect freedom of expression. Whilst we are wary that, at the heart of the DSA lies enhanced intermediary liability, this paper also argues that at least some form and extent of a decentralized approach could be integrated in parallel.

I. The Digital Services Act: Challenges and Ideas

In January of this year, with 530 votes in favour, 78 against and 80 abstentions, the European Parliament adopted the text of the DSA that will be used in negotiations with Member States. After five trilogues, on the 22nd April, the Council and Parliament reached a provisional political agreement on the DSA. The impact of the DSA on individuals, groups, companies, States, civil society and civic space in Europe and beyond, cannot be stressed enough. These new set of rules, which seek to circumvent the spread of illegal content online and enhance big tech transparency are creating a "new digital world, shaping our society for decades to come".[31] As Facebook whistleblower Frances Haugen told the EU, the DSA could become a "global gold standard"[32] in content moderation. In essence "the DSA will regulate how human rights are exercised or curtailed online".[33] In fact, the DSA has the potential to become an "important tool in order to guarantee a proper protection of fundamental rights by sector specific legislation".[34] It is imperative that the final negotiated text will stay true to the values enshrined in the Charter of Fundamental Rights, including the freedom of expression.

The DSA is significant in terms of imposing transparency and accountability requirements on platforms and new user rights. In relation to transparency reports which will need to be issued by all intermediaries (regardless of size), it is noted that several platforms already issue transparency/enforcement reports, but this content remains inadequate. Ranking Digital Rights' 2020 Corporate Accountability Index highlights that "the most striking takeaway is just how little companies across the board are willing to publicly disclose about how they shape and moderate digital content, enforce their rules, collect and use our data and deploy the underlying algorithms that shape our world".[35] Further, under the DSA, all hosting providers must give reasonings for decisions on content moderation, establish an internal complaint handling mechanisms and partake in out of court dispute settlements. While the DSA introduces some new transparency rules that are "straightforward and desirable"[36] such as transparency reports, other mechanisms are not as simple. For example, by obliging platforms to inform users whose content has been removed of the reasoning does protect freedom of expression and on this level, it is welcomed. However, considering the multitude of obligations under this regulation and the sheer amount of online content, it could be foreseen that platforms will prefer to remove content than maintain and provide reasoning.

In relation to content removal, the text approved by the European Parliament establishes a "notice and action" process. Upon such notices, hosting services should act "without undue delay, taking into account the type of illegal content that is being notified and the urgency of taking action". Rather than endorsing general monitoring obligations, MEPs voted in favor of maintaining conditional liability[37] for online intermediaries, shielding them from liability for user-generated illegal content that is not brought to their attention. This is positive and we urge negotiators to ensure that it stays

this way. Whilst it is indisputable that general monitoring obligations would have led to even more over-removals of legitimate speech than we are used to and despite conditional liability being a preferred mechanism, things are far from perfect. Whilst the text adopted by the European Parliament does not directly impose further liability on platforms (by sticking to conditional liability), the very role endowed to private companies to make decisions on the fundamental right to free speech is problematic. Beyond that, it could be argued that enhanced liability is achieved through alternate means. Specifically, the additional due diligence rules for very large online platforms (VLOPs) in terms of annual risk assessments under the close eye of the Commission as well as the possibility of fines for non-compliance "is the same thing as diluting the liability exemption directly". [38] Further, Barata argues that the mere notification of alleged illegality should not create knowledge or awareness to kick start the notice and action process "unless the notified content reaches a certain threshold of obviousness of illegality".[39] Platform assessments which consider key values of IHRL such as legality, proportionality and necessity should be part and parcel of this process. Also, it does not seem necessary or proportionate (in terms of free speech) that all categories of content/speech "entail the same consequences."[40] On a normative level, as argued by Keller, the DSA's content moderation approach is based on "breaking down human behavior and its governance into rationalized, bureaucratized, calculable components". Whilst this is the approach adopted by large platforms, the DSA, in antithesis to platforms seek to add consistency and foreseeability to "evolving human behavior". [41] So are large platforms' existing content moderation practices.

In tandem with the notice and action process, the DSA stipulates that VLOPs should "assess the systemic risks" stemming from their functioning. This additional obligation will most probably have the same consequence as enhanced platform liability since they may be prone to reducing such risks and subsequently reducing the possibility of a violation of the DSA (and the fines associated therewith).[42] On a practical level, mitigating such risk will probably require the use of AI (with all the problems that come with this route as summed up above). Free expression, freedom from discrimination and due process are all placed at risk when automated mechanisms come into play in the handling of contentious areas of speech. As such, mitigating systemic risks may also impact the exercise of freedom of expression, even within the framework of content which is illegal (albeit loosely defined by the DSA). The DSA recognizes four categories of systemic risks which should be assessed in-depth. The first deals with the amplification of illegal content (such as illegal hate speech) with our comments on this term put forth above. Another category which is of particular interest here concerns "any actual and foreseeable negative effects on the protection of public health…or other serious negative effects to the person's physical, mental, social and financial well-being". Barata notes that the reference to negative effects is "not appropriate in terms of human rights law."[43]At the heart of the functioning of IHRL is the balancing of, at times, competing rights. Blanket bans on

effects to generic areas such as financial well-being could not possibly meet any test of legitimacy, proportionality or necessity.

A major issue that must be highlighted is the broad definition of "illegal content" which is to be removed upon notification. The DSA holds that such content means "any information or activity…which is not in compliance with Union law or the law of a Member State, irrespective of the precise subject matter or nature of that law." The DSA also notes that the general idea that should underpin the concept of illegal content is that "what is illegal offline should also be illegal online" and that this should cover content including "hate speech" but also "unlawful discriminatory content." Three themes are identifiable here. Firstly, that the DSA includes "a lot of constructive ambiguity" working with unclear definition which "would require platforms to take a very cautious approach".[44] Barata argues that the DSA deliberately refrains from providing a sound definitional framework and that this "vagueness and broadness may trigger over-removals of content and affect the right to freedom of expression of users."[45] Secondly, that there is no accepted definition of "hate speech" amongst Member States. In fact, categories such as "hate speech" are given "divergent interpretations across the EU."[46] Thirdly, the fact that illegal content extends also to "unlawful" discriminatory content demonstrates the low threshold attached to what is to be deemed removable by intermediaries, an issue that contributes to the further jeopardization of freedom of expression. Further, mandating the removal of illegal content is achieved through a stringent monitoring system, particularly for "very large" online platforms with more than 45 million users in the EU who are at risk of penalties in cases of non-conformity with the DSA, contributing to the "better safe than sorry" approach discussed above and contributing to the enhanced use of AI, with all the challenges this carries. The DSA also provides for the appointment of a Digital Services Coordinator in each Member State to ensure the application and enforcement of the DSA and who may investigate suspected infringements of certain duties by VLOPs. So, what happens in countries such as Hungary which, in 2021, passed a Russia-inspired "gay propaganda" law banning the promotion of material on LGBTQ rights to minors in schools and the media. [47] President Duda of Poland described LGBT rights as an "ideology even more destructive" than communist ideology which indoctrinated the Polish youth before 1989.[48] This country is following the Russian-Hungarian footsteps with its lower house of its parliament having adopted a similar law. The danger of a catch-all provision prohibiting any illegal or discriminatory content is also reflected in the fact that several EU countries continue to maintain blasphemy and religious insult laws (although rarely implemented). For example, Article 525 of Spain's Penal Code punishes, amongst others, the "vilification of religious feelings". Or, for example, Article 283 of the Austrian Criminal Code that punishes "publicly incite[ing] to commit a hostile act" against a church or religious community. So, how do we ensure that coordinators will uphold the values of the Charter of Fundamental Rights such as free speech and equality. To make matters worse "trusted flaggers" whose reports must be processed by intermediaries "expeditiously" must be

approved by Digital Services Coordinators. Trusted flaggers must meet certain conditions such as being objective and transparent in terms of funding. However, their very integration into the DSA, beyond the complexities of illiberal states is cause for concern since this creates a two-path system of removal whereby private (non-judicial non-state) companies (not bound by IHRL) are directed by non-judicial and possibly non-state but potentially state influenced entities to remove speech, with the former having to give priority to such requests.

Further, the DSA states that due diligence obligations are adapted to the "type, nature and size" of the intermediary. The short and long-term effect of this distinction can be a migration of content and users as described in the introductory section of this paper. Users who are easily caught up by algorithms or humans for the undefined notion of hate speech may migrate to smaller platforms (as they are already doing) which are not under such stringent control as is anyhow already the case.[49] Enhanced demands and liability imposed by the EU on VLOPs will "accelerate this development".[50]

In light of the above problematic aspects of the DSA when it comes to freedom of expression in particular, as noted by Keller, lawmakers …. still have an opportunity to resist provisions that will be to little good to users. [51]

In order to have a well-rounded understanding of the treatment of contentious areas of speech by the DSA, a few words should be said about disinformation which is not included in the list of illegal content (as is hate speech). Instead, the DSA's approach to the handling of disinformation is to transform the existing Code of Practice on Disinformation[52] into a co-regulatory instrument which essentially means more involvement and monitoring by the EU. Therefore, although disinformation is not (yet) tackled through enhanced platform liability as is hate speech, the ropes around platforms are definitely tightened through co-regulation. The voluntary Code which, for the moment, adopts self-regulatory standards was signed by companies such as Facebook, Google, Twitter and Mozilla and advertisers in 2018. Others such as TikTok joined later. It includes provisions on issues such as diluting the visibility of disinformation by "improving the findability of trustworthy content". The Code specifically states that 'signatories should not be compelled by governments, nor should they adopt voluntary policies, to delete or prevent access to otherwise lawful content or messages solely on the basis that they are thought to be false". In 2021 and based on the DSA's outlook on the handling of disinformation, the Commission issued a guidance to strengthen to the Code.[53] Measures include increased fact checking across all Member States and the establishment of an effective monitoring framework. Revision of the Code has been extended to March 2022.[54] On the 1st March 2022, in a drastic move and against the backdrop of the Russian invasion of Ukraine, the European Union adopted Regulation 2022/350 which suspended the broadcasting activities of Russian outlets "Russia Today' and 'Sputnik".[55] The justification of this measures is that Russia has entered into a

"systematic, international campaign of media manipulation and distortion of facts to enhance its strategy of destabilization of its neighboring countries and of the Union and its Member States". In an email released on the Lumen database addressed to Google, the Commission informally stated that search engines must delist Russia Today and Sputnik, accounts of their affiliates must be suspended on social media and posts reproducing content from the two outlets must be deleted. This general monitoring obligation is a deviation from the conditional liability provided for in the E-Commerce Directive[56] and the text adopted by the European Parliament for its negotiations on the DSA. Such overbroad content removals are problematic in two ways: firstly, they target and chill free speech for individual users who might wish to discuss and engage with such content in the furtherance of political discourse, and, secondly, they set a negative precedent for social media platforms by signaling those blanket removals of content such as these may be necessary and proportional. Such moves could ultimately lead to inappropriate censorship and stifle discussion of issues of public interest as well as criticism of governments. An interesting historical comparison could be made here to the broadcasting of the Reich's radio from 1939 to 1945 in Britain when the country was at war with Germany. On the other hand, Nazi Germany prohibited the listening of enemy broadcasts, arresting Germans for listening to the BBC.[57]

II. The way ahead: IHRL as a framework for first reference, and adopting a human rights approach to content moderation

The DSA is seeking to enhance obligations on platforms of all sizes and increase accountability on platforms under the threat of penalties. Thus, it could be foreseen that platforms will increasingly rely on automated content moderation tools and will be incentivized to remove more content rather than allowing content to remain up and provide explanations for doing so. In this context, the legal standards on which platforms' content policies are formed and enforced, as well as the threshold for removing content becomes even more crucial.

We are turning to IHRL and seeking to feed into the negotiation process that international human rights norms must become a framework of first reference for content moderation on platforms. As private entities, social media platforms are not signatories to or bound by international human rights instruments. However, there appears to be increasing consensus among international thought leaders such as former SRFOE David Kaye and members of the Oversight Board, that IHRL could be a reliable means to facilitate a more rights-compliant and transparent model of content moderation. Several leading academics echo this sentiment. Evelyn Aswad argues that international law is the most suitable framework for protecting freedom of expression. [58]Similarly, Susan Benesch suggests that, even though IHRL cannot be used "right off the shelf", it can be the framework for content moderation. [59]Hilary Hurd underlines that, while Article 19 of the International Covenant on Civil and Political Rights (ICCPR) only applies to states, there have been "renewed calls to apply Article 19 to technology companies".[60]

From the lens of social media platforms that operate across jurisdictions, the global nature of IHRL may also prove useful in dealing with the differences in national perception and legislation that characterize the global ecosystem of online expression. By adopting an IHRL approach to content moderation, private platforms would also accommodate user demand since many remain deeply skeptical about state regulation of social media. Justitia's 2021 global survey on attitudes towards free speech showed that people in two-thirds of the 33 countries surveyed prefer the regulation of social media content to be carried out solely by the companies themselves. While a plurality in the rest prefers the regulation of content to be carried out by social media companies along with national governments.[61]

Yet, applying IHRL to private companies is a difficult task involving a plethora of challenges and dilemmas. Scholars such as Danielle Citron argue that IHRL is just too flexible to provide the level of clarity that social media platforms need.[62] Evelyn Douek underlines the problem of IHRL's flexibility and argues that there is little that actually compels such platforms to adhere to IHRL. [63]Moreover, an IHRL approach to content moderation will necessarily have to be adapted to the specific

circumstances of social media rather than copied wholesale for entities that are very different from states for which IHRL was developed to constrain and guide. Further, IHRL Thus, IHRL will not resolve all thorny issues and dilemmas related to content moderation and it is unrealistic to expect that all content moderation decisions will be compliant with IHRL. We also acknowledge that an IHRL approach to content moderation will result in a significantly more speech-protective social media environment, leaving in place much content that is likely to be false and misleading and/or cause offense and be deemed unacceptable/hateful/harmful by various states and constituencies across the globe. Accordingly, an IHRL approach should be seen as an imperfect improvement rather than a perfect solution. Nevertheless, we believe that the adaption is possible and would be beneficial for moderating the hate speech and disinformation found on platforms in today´s centralized social media environment.

We argue that to be compliant with IHRL, a platform's content moderation practices must be legitimate, necessary, and proportional within the framework of Article 19(3) ICCPR (restrictions on freedom of expression), which sets out the grounds for limitation of freedom of expression. In this brief paper, we discuss how IHRL norms could be molded and relied upon to regulate two of the most contentious categories of speech in Europe and the rest of the world, hate speech and disinformation.

A. Hate speech

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. For example, LGBT groups and anti-racism activists have reported that Facebook's algorithms have flagged words such as 'tramp' and 'fag' that activists have reclaimed to 'cope with hostility' and have censored posts talking about racial discrimination.[64]

To regulate hate speech on their platforms in a rights-compliant manner, platforms should frame terms and conditions based on a threshold established by Article 20(2) ICCPR (prohibition of advocacy of hatred) and the Rabat Plan of Action's (RPA) six-part threshold test for context, speaker, intent, content and form, extent of dissemination, and likelihood of imminent harm before taking any enforcement action. Article 20(2) of the ICCPR states that "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law." General Comment 34 notes that Articles 19 and 20 are compatible with and complement each other. The acts that are addressed in Article 20 are all subject to restriction pursuant to Article 19, paragraph 3. Thus, we argue that the better approach to regulating hate speech should emanate

from Articles 19 and Article 20(2) of the ICCPR and our references in this paper in relying on IHRL for content moderation refer to relying upon Article 19 and Article 20(2) ICCPR.

In terms of the threshold of what constitutes as 'hate speech', the 2012 SRFOE report explained that "the threshold of the types of expression that would fall under the provisions of Article 20(2) should be high and solid". [65]The RPA aims at clarifying the threshold of Article 20(2) ICCPR and sets a high bar for legitimate restriction of expression. The RPA understands that to assess the severity of the hatred and, therefore, determine whether the high threshold is met, potential issues to be considered are "the cruelty of what is said or of the harm advocated and the frequency, amount and extent of the communications". The RPA includes a six-part threshold test to be used when applying Article 20(2), which incorporates a consideration of the social and political context, the status of the speaker, the intent to incite the audience against a target group; the content and form of the speech; the extent of its dissemination and the likelihood of harm, including imminence.[66]

Further, any restrictions of freedom of expression on the grounds of hate speech under Article 19 must be necessary, legitimate, and proportional and must occur within the framework of one of the grounds set forth in Article 19. The restriction must be the least intrusive measure and it is crucial to point out that the right to freedom of expression under IHRL allows "even expression that may be regarded as deeply offensive" as well as hate speech-adjacent categories such as denial of historical facts and religious insult/blasphemy.

Thus, when dealing with hate speech through the negotiation process, policymakers and social media companies should look at the nature of hate speech as conceptualized in Articles 19 and 20(2), HRC documents, reports of the SRFOE and the RPA. The use of the RPA allows for a practical examination of the disputed content by looking at the six-part test on issues such as context and likelihood of harm to ensure that the determination that content should be removed matches the benchmarks laid out by IHRL. Accordingly, platforms should make sure that their Terms and content moderation guidelines properly reflect the standards set out by Articles 19 and 20 and train their AI and human moderators accordingly. Adopting an IHRL approach to tackling online hate will significantly narrow the applicable definition of hate speech that currently exist on social media platforms and raise the threshold for determining when deeply offensive and hurtful speech can be removed. The negative consequences of strengthening the protection of controversial speech could be mitigated if platforms provide users with better means to adopting their own filters or use filters developed by third parties so that users can protect themselves from content they may deem offensive but falls short of the high threshold required to constitute hate speech under ICCPR Article 20(2).

B. Disinformation

This topic is particularly relevant to the current negotiation given that the European Commission introduced the idea of having a mechanism for facing disinformation in exceptional circumstances, in reaction to the ongoing Ukrainian crisis and related pro-Russian information manipulation.

We hope that such a protocol incorporates IHRL as the backbone. For example, Empirical data increasingly suggests that the perception of social media being awash in misinformation is exaggerated.[67] At the same time the protocol would have to bear in mind counter narratives and their efficiency. In relation to IHRL and given that this recommendation was a reaction to the invasion of Ukraine, we note the following. Article 20(1) of the International Covenant on Civil and Political Rights (ICCPR) prohibits any propaganda for war. However, General Comment 34 of the Human Rights Committee highlights that any legal prohibitions arising from Article 20 must be justified and be in strict conformity with Article 19 which provides for freedom of expression. Restrictions under Article 19 can only be legitimate if they meet the strict tests of necessity and proportionality which entail an immediate and direct connection between the speech and the threat whilst measures must be "the least intrusive instruments" to achieve the legitimate aim pursued. These could include labelling or downranking or tech-oriented solutions to prevent virality.

Further, the report of the European Commission´s independent High-Level Group on fake news and online disinformation ("EU HLEG") in had concluded "that the best responses to disinformation are multidimensional, with stakeholders collaborating in a manner that protects and promotes freedom of expression, media freedom, and media pluralism". [68]The Commission recommended "to disregard simplistic solutions" and that "any form of censorship either public or private should clearly be avoided". The UK's Royal Academy also cautions against censorship, particularly of scientific misinformation online, noting that such measures may even "exacerbate feelings of distrust in authorities" and push it towards "harder-to address corners of the internet".[69] Research also demonstrates that removals may counteractively reinforce falsehoods.[70]

On principle, the right to impart information and ideas is not limited to "correct" statements, but, at the same time, this does not justify the dissemination of knowingly or recklessly false statements by official or state actors. [71]That said, IHRL does not generally justify the "dissemination of knowingly or recklessly false statements, especially by official or State Actors."[72] It also does not protect disinformation that rises to the level of incitement to violence, hate speech or fraud. [73]Nevertheless, legislation that seeks to tackle disinformation and platforms' terms on disinformation must not be overbroad and must be narrowly tailored. Disinformation impacts a variety of rights under IHRL. However, unlike hate speech, disinformation does not constitute a specific category of speech exempted from the protection of Article 19 or subject to specific prohibition such as, for example,

under Article 20(2). As a result, there is scope for discussion about what kinds of disinformation are permitted under IHRL.

However, before we move to discussing the threshold of what constitutes disinformation warranting removal, we must acknowledge the fundamental question of whether anybody should have the authority to determine definitively whether content is true or not. It is dangerous to vest this authority in the hands of governments because this means that governments become the arbiters of truth for online speech and governments often get it wrong. In her 2021 Report on Disinformation, SRFOE Irene Khan called for multidimensional responses to disinformation that are well grounded in an IHRL framework. She urged platforms to adopt clear, narrowly defined content and advertising policies on disinformation and misinformation with a special emphasis on adopting clear policies relating to public figures that are consistent with IHRL standards, applying them consistently across geographical areas.

In determining the limits of disinformation from an IHRL perspective, social media companies may focus on 1) content, 2) context, 3) intent, and 4) impact in their assessment of cases involving interference with the right to freedom of expression. Although this test is broadly based on the RPA, it has slight variations. The RPA isn't fully relevant for misinformation in the way that it is for hate speech. For example, the RPA lays emphasis on the status of the speaker but this test does not do so. This is because if misinformation leads to imminent physical harm (for example, medical misinformation encouraging users to administer harmful substances), it warrants removal under the public health restriction under Article 19(3) even if it is posted by a user with only 20 friends.

In removing content based on its "impact" or harm, IHRL clarifies that free expression may only be restricted on the grounds provided in Article 19(3). The European Commission's 'Communication on Disinformation' explained 'public harm' to mean "threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security".[74] However, the relevant threat that disinformation causes to such interests must be serious and immediate in nature. From the outset, only limited and qualified instances of intentional or "bad faith" disinformation entailing immediate real-world harm should be subject to the most intrusive restrictive measures such as content removal. Immediate real-world in this scenario means physical or mental injury or substantial impairment of fundamental rights. Other forms of misinformation likely to result in less serious harm may be subject to less restrictive measures such as labelling or downranking.

Thus, there is a strong argument to be made that policymakers and platforms should adopt a cautious approach in dealing with disinformation. If policymakers or platforms were to adopt overbroad regulation or policies, it might lead to the systematic removal of legitimate information

and opinion and unreasonably restrain the right to expression under Article 19 ICCPR. Lastly, any approach by platforms to deal with disinformation must be accompanied by maximum transparency with respect to company policies, and platforms must engage in ongoing due diligence to determine their policies' impact on freedom of expression.[75]

Conclusion

The risks brought about to free speech by the very structure of the DSA is an issue not to be taken lightly. The DSA essentially empowers States (including increasingly illiberal ones) to monitor private companies (not bound by IHRL) to quickly remove content (even discriminatory content) at risk of fines. Robust free speech guarantees are of particular significance in a text which binds 27 European democracies at a time where censorship and authoritarianism is on the march, and we are living through a "free speech recession".[76] This becomes even more significant given the fact that countries like Poland and Hungary are systematically deviating from democratic and rule of law norms, proposing and passing laws censorship laws on basic human rights. As such, the EU must take care that it does not give a carte blanche to such states to play their own mandate whilst at the same time realizing the global effect the DSA will have. Negotiators and decision makers must not undermine the Brussels Effect[77] that this piece of legislation will have, even in countries where regimes will jump at any opportunity to further censor online speech.

Further, the adopted text came with other failings. For example, MEPs refused to grant people the right to choose the ranking and recommendation algorithms they prefer. This seems to be a missed opportunity as doing so would have meant that individual users could exercise greater control over the content they consume and thus have more democratized control over information streams.[78] On a normative level, platforms are not static, evolving over time to develop new uses, sources of revenue, and communities of users.[79] As noted above, Keller argues that human communication and behaviour is difficult to rationalize and standardize in such an organized manner as the DSA seeks to achieve. To add to that, there are no provisions set in stone to highlight that actions put forth by the DSA, whether those are risk assessments or content removal are to be founded and bound by doctrines and principles of IHRL, a position tirelessly advocated for by the former SRFOE. Mandating private companies to decipher on the legitimacy of speech quickly and efficiently (i.e to do so and respond to users) or face fines despite the fact that they will only need to remove speech which has been reported to them is troublesome. Whilst this cannot change at this stage, incorporating IHRL provisions to underly not only content moderation but also other ambits of the DSA such as risk assessment will give the regulation but also its implementation legitimacy. This will be a step in the right direction for free speech and a healthy civic space. Moreover the "unduly rigid" approach is problematic, unfitting with the organic nature not only of platforms but also of human behaviour and communication. [80]

To this end, we recommend that the following are taken into account:

- Coherently and substantially placing IHRL and particularly Article 19 and Article 20(2) of the International Covenant on Civil and Political Rights at the heart of provisions involving

content moderation of hate speech, taking inspiration from the recommendations of the former SRFOE. To achieve this, a platform's content moderation practices must be legitimate, necessary, and proportional within the framework of Article 19(3) ICCPR (restrictions on freedom of expression), which sets out the grounds for limitation of freedom of expression. For hate speech, platforms should frame terms and conditions based on a threshold established by and take strictly into consideration the Rabat Plan of Action's six-part threshold test for context, speaker, intent, content and form, extent of dissemination, and likelihood of imminent harm before taking any enforcement action. For disinformation, a platform's terms and conditions should be tailored to protect the grounds in Article 19(3) ICCPR and Article 25 ICCPR (right to participate in voting and elections). In addition, platforms must refrain from adopting vague blanket policies for removal. Only disinformation promoting real and immediate harm should be subject to the most intrusive restrictive measures such as content removal. In determining the limits of disinformation, platforms should focus on the post's content, its context, its impact, its likelihood of causing imminent harm, and the speaker's intent. The approach to disinformation should be a mix of IHRL and measures suggested in the Code of Practice on Disinformation such as labelling and downranking.

- Provide a sustainable definition of what is deemed to be "illegal content" for purposes of content removal and limit all actions provided for by the DSA in the ambit of such content without moving to lower thresholds such as unlawful discriminatory speech.

- Re-conceptualise the manner in which the risk assessments are quantified and qualified with "negative effects" on, inter alia, 'health' not constituting a legitimate ground for the limitation of freedom of expression.

- Ensure that rule of law safeguards are incorporated adequately in all requirements, particularly those pertaining to VLOPs. This is significant not only for liberal democracies in the EU but also illiberal countries therein and countries across the globe which may replicate the DSA model. One of the issues that should be looked at is the role of the Digital Services Coordinator and method of appointment as well as the role of trusted flaggers and their methods of appointment.

We guide you to our report on IHRL as a Framework of First Reference for further analysis on how the above could be achieved.

Ultimately, the future of free speech online may be best served by a more distributed media environment and/or through enhanced user control over content. However, until such

decentralization is achieved, we believe that IHRL as a "framework of first reference" for major social media platforms may cultivate a more transparent, legitimate and speech-protective approach to handling online hate speech and disinformation.

Endnotes

[1] Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human-Rights Based Approach to Content Moderation' (2020) 43 *Fordham International Law Journal* 4, 941

[2] Digital 2021: Global Overview Report https://datareportal.com/reports/digital-2021-global-overview-report#:~:text=Internet%3A%204.66%20billion%20people%20around,now%20stands%20at%2059.5%20percent.

[3] Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human-Rights Based Approach to Content Moderation' (2020) 43 *Fordham International Law Journal* 4, 939

[4] 'Genocide Incited on Facebook, with Posts from Myanmar's Military' (2018) *New York Times* <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html>

[5] Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human-Rights Based Approach to Content Moderation' (2020) 43 *Fordham International Law Journal* 4, 939

[6] Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human-Rights Based Approach to Content Moderation' (2020) 43 *Fordham International Law Journal* 4, 941

[7] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 2

[8] Iverna McGowan, 'European Plans to Regulate Internet Will have Major Impacts on Civic Space at Home and Abroad' (17 May 2021) *Centre for Democracy and Technology* < https://cdt.org/insights/european-plans-to-regulate-internet-will-have-major-impacts-on-civic-space-at-home-and-abroad/>

[9] Proposition de Loi visant *à* Lutter Contre Les Contenus Haineux Sur Internet <http://www.assemblee-nationale.fr/dyn/15/textes/l15t0388_texte-adopte-seance>

[10] 'French Law on Illegal Content Online Ruled Unconstitutional for the EU to Learn' (2020) <https://www.patrick-breyer.de/?p=593729&lang=en>

[11] Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG) <https://germanlawarchive.iuscomp.org/?p=1245>

[12] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 2-3

[13] Jacob Mchangama & Joelle Fiss, 'The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship' (2019) *Justitia* < https://futurefreespeech.com/wp-content/uploads/2020/06/analyse-the-digital-berlin-wall-how-germany-accidentally-created-a-prototype-for-global-online-censorship.pdf>; Jacob Mchangama & Natalie Alkiviadou, 'The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship – Act Two' (2020) *Justitia* < https://justitia-int.org/wp-content/uploads/2020/09/Analyse_Cross-fertilizing-Online-Censorship-The-Global-Impact-of-Germanys-Network-Enforcement-Act-Part-two_Final-1.pdf>

[14] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 2-3

[15] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 3

[16] Jacob Mchangama & Lukas Callesen, 'The Wild West? Illegal Comments on Facebook' <https://justitia-int.org/en/the-wild-west/>

[17] Thiago Oliva Dias, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression' (2020) 20 *Human Rights Law Review* 4, 609

[18] Thiago Dias Oliva, Dennys Marcelo Antonialli & Alessandra Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2020) 25 *Sexuality and Culture,* 702

[19] Daphne Keller, 'Internet Platforms: Observations on Speech, Danger and Money' (2018) *Hoover Institute,* 7

[20] 'Twitter Flagged Donald Trump's Tweets with Election Misinformation: They Continued to Spread Both On and Off the Platform' (2029), 2 Harvard Kennedy School Misinformation Review 4.

[21] Aleksandra Urman & Stefan Katz, 'What They Do in the Shadows: Examining the Far-Right Networks on Telegram' (2020) Information, Communication & Society

[22] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 7

[23] Jacob Aasland Ravndal, 'Explaining Right-Wing Terrorism and violence in Western Europe: Grievances, Opportunities, and Polarization' (2017) 57 European Journal of Political Research 4

[24] Jacob Aasland Ravndal, 'Explaining Right-Wing Terrorism and violence in Western Europe: Grievances, Opportunities, and Polarization' (2017) 57 European Journal of Political Research 4

[25] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 3

[26] Jacob Mchangama & Natalie Alkiviadou, 'Hate Speech and the European Court of Human Rights: Whatever Happened to the Right to Offend, Shock or Disturb?' (2021) 21 *Human Rights Law Review* 4

[27] Charter of Fundamental Rights of the European Union, 2012/C 326/02

[28] Report of the Special Rapporteur on the Promotion and Protection of the Right to

Freedom of Opinion and Expression, UN Doc A/HRC/38/25, 6 Apr. 2018 [hereinafter Kaye

Content Moderation Report], para 43

[29] Facebook, 'Corporate Human Rights Policy' <https://about.fb.com/wp-content/uploads/2021/04/Facebooks-Corporate-Human-Rights-Policy.pdf>

[30] Jacob Mchangama, Natalie Alkiviadou & Raghav Mendiratta, 'A Framework of First Reference – Decoding a Human Rights Approach to Content Moderation on Social Media' (2021) < http://justitia-int.org/report-a-framework-of-first-reference-decoding-a-human-rights-approach-to-content-moderation-on-social-media/>

[31] Sebastian Becker, 'Framing the Future of the Internet' (2022) *Social Europe* <https://socialeurope.eu/framing-the-future-of-the-internet>

[32] Oliver Noyan, 'Content Moderation Policies Continue to Face Core Dillemas' (2021) *Euractiv* < https://www.euractiv.com/section/politics/news/content-moderation-policies-continue-to-face-core-dilemmas/>

[33] Sebastian Becker, 'Framing the Future of the Internet' (2022) *Social Europe* <https://socialeurope.eu/framing-the-future-of-the-internet>

[34] Joan Barata, 'The Digital Services Act and its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' *Platforma por la Libertad de Infoacion,* 12

[35] Anne Brouillette, 'Key Findings: Companies are Improving in Principle but Failing in Practice' *2020 Ranking Digital Rights Corporate Accountability Index* <https://rankingdigitalrights.org/index2020/key-findings>

[36] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 1

[37] 'European Parliament Approves Rights-Respecting DSA & Proposes Ban on Use of Sensitive Personal Data for Online Ads' (2022) < https://edri.org/our-work/european-parliament-approves-rights-respecting-dsa-proposes-ban-on-use-of-sensitive-personal-data-for-online-ads/>

[38] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 5

[39] Joan Barata, 'The Digital Services Act and its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' *Platforma por la Libertad de Infoacion,* 16

[40] Joan Barata, 'The Digital Services Act and its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' *Platforma por la Libertad de Infoacion,* 16

[41] Daphne Keller, 'The DSA's Industrial Model for Content Moderation' (2022) *Verfassungsblog*

<https://verfassungsblog.de/dsa-industrial-model/>

[42] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 1

[43] Joan Barata, 'The Digital Services Act and its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' *Platforma por la Libertad de Infoacion,* 18

[44] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 5

[45] Joan Barata, 'The Digital Services Act and its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' *Platforma por la Libertad de Infoacion,* 15

[46] Joan Barata, 'The Digital Services Act and its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' *Platforma por la Libertad de Infoacion,* 18

[47] Hungary's Controversial Anti-LGBT law goes into Effect despite EU Warnings' (2021) *France 24* <https://www.france24.com/en/europe/20210707-hungary-s-controversial-anti-lgbt-law-goes-into-effect-despite-eu-warnings>

[48] Andrzej Duda's speech during his 2020-presidential campaign: <https://www.youtube.com/watch?v=8VRRWuryb4k>

[49] Aleksandra Urman & Stefan Katz, 'What They Do in the Shadows: Examining the Far-Right Networks on Telegram' (2020) Information, Communication & Society

[50] Frederik Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) *European Centre for International Political Economy,* 8

[51] Daphne Keller, 'The DSA's Industrial Model for Content Moderation' (2022) *Verfassungsblog*

< https://verfassungsblog.de/dsa-industrial-model/>

[52] EU Code of Practice on Disinformation <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54454>

[53] Guidance on Strengthening the Code of Practice on Disinformation <https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation>

[54] Molly Killen, 'Code of Practice on Disinformation Revision Extended into 2022' (2022) *Euractiv* <https://www.euractiv.com/section/digital/news/code-of-practice-on-disinformation-revision-extended-into-2022/>

[55] Regulation 2022/350 amending Regulation No.833/2014 concerning restrictive measures in view of Russia's actions destabilising the situation in Ukraine.

[56] 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')

[57] Mary Kenny, "When Speech Becomes Treason" (2006) *Index on Censorship*

<https://journals.sagepub.com/doi/pdf/10.1080/03064220500532560>

[58] Evelyn Mary Aswad, 'The Future of Freedom of Expression Online' (2018) 17 Duke Law & Technology Review 1, 52-53.

[59] Susan Benesch, 'But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies' (2020) 38 Yale Journal on Regulation Online Bulletin, 90.

[60] Hilary Hurd, 'How Facebook Can Use International Law in Content Moderation' (2019) Lawfare ,

[61] Svend-Erik Skaaning & Suthan Krishnarajan, 'Who Cares about Free Speech? Findings from a Global Survey of Support for Free Speech (2021) Justitia.

[62] Danielle Keats Citron, 'What to Do about the Emerging Threat of Censorship Creep on the Internet' (2017), Cato Institute.

[63] Evelyn Douek, 'U.N. Special Rapporteur's Latest Report on Online Content Regulation Calls for Human Rights by Default' (2018).

[64] Thiago Dias Oliva, Dennys Marcelo Antonialli & Allesandra Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) 25 Sexuality & Culture, 702; Facebook's Hate Speech Policies Censor Marginalized Users' Wired.

[65] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/67/357 (7 September 2012), para. 45.

[66] Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that constitutes Incitement to Discrimination, Hostility or Violence (2002) para. 22.

[67] See, for example, Yavgeniy Golovchenko 'Measuring the Scope of Pro-Kremlin Disinformation on Twitter' (2020) *Humanities and Social Sciences Communications,* Olga Boichak, Jeff Hemsley, Sam Jackson, Rebekah Tromble, Sikana Taunparburngsun. 'Not the Bots you are Looking For: Patters and Effects of Orchestrated Interventions in the US and German Elections.' (2021) 15 *International Journal of Communication.*

[68] European Commission, 'Final Report of the High Level Expert Group on Fake News and Online Disinformation' (2018).

[69] Royal Society, 'The Online Information Environment: Understanding how the Internet Shapes People's Engagement with Scientific Information.' (2022)

[70] Twitter Flagged Donald Trump's Tweets with Election Misinformation: They Continued to Spread Both On and Off the Platform' (2029), 2 Harvard Kennedy School Misinformation Review 4.

[71] Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda.

[72] Joint Declaration on Freedom of Expression and Fake News, Disinformation and Propaganda' (2017).

[73] Sarah Shirazyan, Allen Weiner, Yvonne Lee & Madeline Magnuson et al., 'How to Reconcile International Human Rights Law and Criminalization of Online Speech: Violent Extremism, Misinformation, Defamation, and Cyberharassment' (2020), Stanford Law School Law and Policy Lab.

[74] EU Code of Practice on Disinformation.

[75] Catharine Christie, Edison Lanza & Michael Camilleri, 'Covid-19 and Freedom of Expression in the Americas' (2020), The Dialogue.

[76] Jacob Mchangama, 'The War on Free Speech – Censorship's Global Rise' (2022) *Foreign Affairs* <https://www.foreignaffairs.com/articles/world/2022-02-09/war-free-speech-censorship>

[77] Anu Bradford, 'The Brussels Effect: How the European Union Rules the World' (2020) *Oxford Scholarship Online* <https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190088583.001.0001/oso-9780190088583>

[78] Digital Services Act (text adopted by European Parliament) <https://www.europarl.europa.eu/doceo/document/TA-9-2022-0014_EN.pdf>

[79] Sasha Desmaris et al., 'Creating a French Framework to Make Social Media Platforms More Accountable: Acting in France with a European Vision' Mission Report submitted to the French Secretary of State for Digital Affairs, May 2019) <https://perma.cc/C9TA-L77A>

[80] Daphne Keller, 'The DSA's Industrial Model for Content Moderation' (2022) *Verfassungsblog*

<https://verfassungsblog.de/dsa-industrial-model/>

Thoughts on the DSA: Challenges, Ideas and the Way Forward through International Human Rights Law

WHO WE ARE

Justitia

Founded in August 2014, Justitia is Denmark's first judicial think tank. Justitia aims to promote the rule of law and fundamental human rights and freedom rights both within Denmark and abroad by educating and influencing policy experts, decision-makers, and the public. In so doing, Justitia offers legal insight and analysis on a range of contemporary issues.

The Future of Free Speech Project

The Future of Free Speech is a collaboration between Justitia, Columbia University's Global Freedom of Expression and Aarhus University's Department of Political Science. At the Future of Free Speech, we believe that a robust and resilient culture of free speech must be the foundation for the future of any free, democratic society. To understand better and counter the decline of free speech, "The Future of Free Speech" project will seek to answer three big questions: Why is freedom of speech in global decline? How can we better understand and conceptualize the benefits and harms of free speech? And how can we create a resilient global culture of free speech that benefits everyone?

The project is sponsored by:

THE FUTURE OF FREE SPEECH

REBUILDING THE BULWARK OF LIBERTY