International Journal of Law and Information Technology, 2023, **31**, 46–74
https://doi.org/10.1093/ijlit/eaad015
Advance access publication 20 July 2023
Article

Content moderation in times of war: testing state and self-regulation, contract and human rights law in search of optimal solutions

Nataliia Filatova-Bilous*

ABSTRACT

The full-scale invasion of Ukraine and crimes against humanity accompanying it have been fuelled by the mass spread of fakes and hatred incitements, forcing the largest online platforms to review and strengthen their content moderation policies. However, the approaches taken by platforms have not been perfect, and some of them could even exacerbate the situation. All in all, this is another evidence of the need to develop mechanisms being able to cope with the challenges to online speech and safety caused by dramatic social events. Basic approaches to address content moderation issues developed by now are self-(co-) and state regulation, on the one hand, and contract and human rights law, on the other. However, neither of them taken separately can ensure the needed level of protection of human rights online. Thus, in this article the ways to combine and improve these approaches are proposed. On the one hand, there is a need to dwell on private law mechanisms allowing to ensure the protection of human rights by virtue of judgements in private disputes. On the other hand, state regulation should be improved by international instruments allowing to provide for a uniform approach to regulation at a global scale.

KEYWORDS: online platforms, content moderation, war in Ukraine, freedom of speech, contract law, human rights law, business-to-consumer relationships

INTRODUCTION

Content moderation has been one of the most widely discussed issues attracting more and more attention from researchers, legislators and IT lawyers during the last decade. In times of infancy of the Internet, content moderation was considered mainly as a kind of bonus for platform users

* PhD, associate professor at civil law department of Yaroslav Mudryi National Law University, Kharkiv, Pushkinska 77, Ukraine. Telephone: +380667677373; E-mail: filatovaukraine@gmail.com. The author expresses acknowledgments to SAR Denmark and to Aalborg University (Denmark) for the support of the safe stay during two months of the war time in Ukraine and for giving all the needed facilities to continue conducting research activity, which allowed to finish this article.

making the digital environment more comfortable, safe and friendly.[1] However, the rapid rise of online platforms[2], together with the widening number of their users, have made platforms the 'gatekeepers of free expression' and 'managers of the world's information'.[3] Moreover, platforms became giant influencers of the most significant social events of recent years: they happened to remedy one of the hugest interferences in the president's election in the USA in 2016[4] and to be the primary means of sharing dangerous conspiracy theories in times of COVID-19 pandemic.[5],[6]

The war initiated by Russian Federation against Ukraine has become another powerful trigger for spreading hate speech, disinformation and extremely harmful content online. On the one hand, manipulative techniques with content made millions of people in Russia and all over the world believe in complete nonsense and justify the unprovoked invasion of Ukraine by Russian armed forces. The news on the invention of 'ethnically oriented' biological weapon[7] and COVID-19 in secret Ukrainian laboratories[8] are the most prominent examples of fakes spread via various online platforms in this regard. On the other hand, the spread of incitements to hostility and violence via social media[9] has contributed to the growth of hatred of Ukrainians, which has become one of the reasons for mass atrocities and other crimes against humanity in Ukraine[10]

The escalation of uncontrolled hatred online forced platforms to urgently strengthen their policies for the sake of online safety. Some online platforms decided to block content of this kind and suspend the accounts of Russian state-owned channels (these were Facebook (Meta) and YouTube) as well as to stop monetization for Russian content creators.[11] Other platforms, like Telegram, continued to apply a *laissez-faire* approach allowing Russian aggressive propaganda

[1] Giovanni De Gregorio and Roxana Radu, 'Digital constitutionalism in the new era of Internet governance' (2022) 30 (1) International Journal of Law and Information Technology, 78; Mark A. Lemley, 'The Contradictions of Platform Regulation' (2021) 1 Journal of Free Speech Law, 306.

[2] In this article for the sake of uniformity I will use the term 'online platforms' to identify the largest social media and content-sharing platforms, like Facebook, Instagram, YouTube, Telegram and others. This term will sometimes be used to identify companies which operate these platforms (platform operators).

[3] Emily Bell, 'The Unintentional Press: How Technology Companies Fail as Publishers' in Lee C. Bolinger and Geoffrey R. Stone, eds., *The Free Speech Century* (New York: Oxford University Press, 2019), 239.

[4] Kyle Langvardt, 'Regulating Online Content Moderation' (2018) 106 Georgetown Law Journal, 1353, 1383; Stephen Macedo, 'Lost in the Marketplace of Ideas: Toward a New Constitution for Free Speech After Trump and Twitter?' (2022) 48 Philosophy & Social Criticism, 496, 510.

[5] Ana Laura Pérez, 'The "Hate Speech" Policies of Majorplatforms during the Covid-19 Pandemic' (UNESDOC, 2021) <https://unesdoc.unesco.org/ark:/48223/pf0000377720_eng/PDF/377720eng.pdf.multi> last time accessed 11 April 2023.

[6] Macedo (n 4) 499.

[7] Svitlana Chorna, 'Combat mosquitos, ethnically oriented birds and Satanists – Kremlin spreads fake news' (Golos, 27 March 2022) <http://www.golos.com.ua/article/357846> last accessed 11 April 2023; 'В США признали наличие своих биологических лабораторий на Украине' [The U.S.A. Confirmed that They Had Their BioLabs in Ukraine] (Тюменское время, 9 March 2022) <https://www.youtube.com/watch?v=bGU0Rnn9zUg&ab_channel=%D0%A2%D1%8E%D0%BC%D0%B5%D0%BD%D1%81%D0%BA%D0%BE%D0%B5%D0%B2%D1%80%D0%B5%D0%BC%D1%8F> last accessed 11 April 2023 [In Russian].

[8] 'Ситуация накаляется: из военного госпиталя для карателей бегут врачи (+ВИДЕО)' [Situation Is Escalating: the Doctors Run Away from the Military Hospital for Ukrainian Punishers] (*Русская Весна [Russian Spring]*, 25 April 2020) <https://rusvesna.su/news/1587817348> last accessed 11 April 2023 (in Russian).

[9] See 'Russian lawmakers advocate freezing and starving Ukrainian civilians, turning them into refugees'. <https://www.youtube.com/watch?v=10AiNAsCnkw&ab_channel=RussianMediaMonitor> last accessed 11 April 2023.

[10] According to the Report of the Independent International Commission of Inquiry on Ukraine transmitted by the U.N. Secretary-General to the General Assembly, 'war crimes and violations of human rights and international humanitarian law have been committed in Ukraine since 24 February 2022'. The Commission documented «indiscriminate attacks using cluster munitions, unguided rockets and air strikes', which 'are highly likely to have indiscriminate effects and cause significant harm to civilians'. The Commission also found that «Russian armed forces had shot at civilians attempting to flee' and documented 'patterns of summary executions, unlawful confinement, torture, ill-treatment, and rape and other sexual violence committed in areas occupied by Russian armed forces' (see Report of the Independent International Commission of Inquiry on Ukraine, U.N. Doc. A/77/533 (October 18, 2022)).

[11] Matt Novak, 'YouTube Stops Monetization for Video Creators in Russia but There's One Exception' (*Gizmodo*, 3 October 2022) <https://gizmodo.com/youtube-stops-monetization-for-video-creators-in-russia-1848633335> last accessed 11 April 2023; 'Meta's Ongoing Efforts Regarding Russia's Invasion of Ukraine' (*Meta*, 26 February 2022) <https://about.fb.com/news/2022/02/metas-ongoing-efforts-regarding-russias-invasion-of-ukraine/> last accessed 11 April 2023.

to coexist with Ukrainian official channels[12] and giving content moderation issues away to the users *per se*.[13] Twitter chose a middle position in this regard. The company decided not to limit itself only to radical practices like blocking the content, but to use various content moderation approaches to cope with harmful content, like labelling, promotion of fact-checked news and content, and making the not fact-checked one less visible.[14] Finally, Tik Tok chose the trickiest approach: the platform decided to block all non-Russian content for Russian users, and all the content generated by Russian state-funded channels—for non-Russian users.[15] As a result, Russian users found themselves in an information bubble, where they mostly saw pro-war content usually full of fake news and hateful expressions.

The analysis of these practices proves that content moderation issues may not be entirely given out to platforms and need external regulation. By this time, several approaches to address these issues have been developed. Initially, content moderation issues were resolved with *self-regulatory mechanisms*, which involved the creation of various codes of conduct or principles (like Santa Clara Principles[16]), the formation of self-regulatory organizations (SROs) (like Global Network Initiative (GNI))[17] and the creation of special entities (like Facebook Oversight Board).[18] During the last decade, states worldwide started developing *special laws* concerning content moderation. One of the first among them was the Chinese Cybersecurity Law of 2016,[19] which has become a prominent example of an authoritarian approach to regulating these issues in the modern world. Democratic countries recently have also made steps towards adopting laws concerning content moderation issues and other practices applied by platforms. In this regard, Digital Services Act (DSA) recently adopted in the European Union[20] and the Online Safety Bill introduced, but not yet adopted in the UK, are worth special attention.[21] Moreover, content moderation issues are subject to contract law since they constitute a part of contractual relationships between platforms and their users. Thus, private disputes between platforms and their users are resolved under contract law remedies. Finally, content moderation is subject to human rights standards, which platforms should observe and guarantee following the U.N. Guiding Principles on Business and Human Rights.[22]

However, neither of these approaches taken separately is flawless and effective enough to combat the risks posed by sharp social conflicts and wars. Hence, this article aims to rethink the mentioned approaches and find how they may be combined and improved. For this purpose, Part II provides an empirical analysis of the content moderation practices used by the largest online platforms to mitigate the risks posed by the full-scale invasion of Ukraine and points out

[12] Vera Bergengruen. 'How Telegram Became the Digital Battlefield in the Russia-Ukraine War' (*Time*, 21 March 2022) <https://time.com/6158437/telegram-russia-ukraine-information-war/> last accessed 11 April 2023.

[13] Pavel Durov <https://t.me/durov_russia/40> last accessed 11 April 2023.

[14] Sinéad McSweeney, 'Our ongoing approach to the war in Ukraine' (BlogTwitter, 16 March 2022) <https://blog.twitter.com/en_us/topics/company/2022/our-ongoing-approach-to-the-war-in-ukraine> last accessed 11 April 2023.

[15] Dan Milmo, 'TikTok users in Russia can see only old Russian-made content' (*The Guardian*, 10 March 2022). <https://www.theguardian.com/technology/2022/mar/10/tiktok-users-in-russia-can-see-only-old-russian-made-content> last accessed 11 April 2023.

[16] 'The Santa Clara Principles on Transparency and Accountability in Content Moderation' *(Santaclara)* <https://santaclara-principles.org/> last accessed 11 April 2023.

[17] Governance Charter (*Globalnetworkinitiative*) <https://globalnetworkinitiative.org/governance-charter/> last accessed 11 April 2023.

[18] Oversight Board Charter *(Oversight Board)* <https://oversightboard.com/governance/> last accessed 11 April 2023.

[19] Cybersecurity Law of the People's Republic of China of 2016 (unofficial translation) (*China Law Translate*, 07 November 2016) <https://www.chinalawtranslate.com/en/2016-cybersecurity-law/> last accessed 11 April 2023.

[20] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) OJ L 277, 27.10.2022, p. 1–102.

[21] Online Safety Bill 121 2022–23 *(UK Parliament,* 18th January 2023) < https://bills.parliament.uk/publications/49376/documents/2822 > last accessed 11 April 2023.

[22] Guiding Principles on Business and Human Rights. Implementing the United Nations 'Protect, Respect and Remedy' Framework (*United Nations*, 2011) <https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf> last accessed 11 April 2023.

the main weaknesses of these practices. Part III focuses on the analysis of regulatory approaches to address various issues faced by platforms and their users considering content moderation, juxtaposing self(co-)-regulation and state regulation proposed in some jurisdictions during the last years (the EU, the UK and China). In Part IV content moderation is regarded from a different angle—through the lens of contract law, on the one hand, and of human rights law, on the other. In Part V I attempt to figure out the way different approaches may be combined and improved. Considering the horizontal approaches, the analysis allows us to resume that content moderation is a point in which contract law and human rights law may meet. To make this tandem enforceable, there is a need to create a judiciary having the authority to resolve disputes between users and platforms on a borderless basis, and here self-regulatory instruments may help. Meanwhile, the role of state regulation should not be downplayed as well. However, to ensure a uniform regulation of content moderation issues on a global scale there need to be international instruments obliging states to implement human rights standards into their inner legislation on content moderation.

CONTENT MODERATION PRACTICES AND THE WAR

In the modern digital world, the most significant social problems have always been accompanied by hot debates on various online platforms, which often gave rise to the spreading of disinformation and hate speech online, having harmful effects on the offline world. This was the case with the incitements to ethnic cleansing of Rohingya Muslim groups in Myanmar in 2016–2017 spread via Facebook,[23] sharing of dangerous conspiracy theories in times of COVID-19 pandemic, giving rise to antivaccination and other social protests,[24] and sharing incitement to violence in January of 2021, which led to the storming of the US Capitol.[25]

The full-scale invasion of Ukraine, which started on 24 February 2022, has become another dramatic event that has prominently illustrated that disinformation and hate speech on online platforms multiply hatred in the offline world. Trying to justify the invasion and harmful acts Russian and other media channels and bloggers rapidly increased the production and spreading of various fakes and disinformation. One of the most widespread one was about Ukrainian Nazis in the armed forces and at each level of the Ukrainian state governance.[26] Videos which are said to 'confirm' these 'facts' can still be found on YouTube on Belarusian and some Russian channels.[27] The information about Ukrainian 'Nazi's' has been also widely spread on Twitter

[23] Steve Stecklow, 'Why Facebook is losing the war on hate speech in Myanmar' (*Reuters,* 15 August 2018) < https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/> last accessed 11 April 2023.

[24] Ana Laura Pérez (n 5).

[25] Macedo (n 4) 499.

[26] Timofei Sergeitsev, 'Что Россия должна сделать с Украиной' [What Russia Must Do with Ukraine] (РИА Новости, 03 April 2022) < https://ria.ru/20220403/ukraina-1781469605.html > last accessed 11 April 2023 (only with VPN) [In Russian].

[27] See for example: 'Нацисты Киева, геноцид памяти в ЕС, убийцы стали героями, фамилии и лица вандалов. Понятная политика' [Nazis from Kyiv, Genocide over the Memory in the EU, the Murders have Become Heroes, Last Names and Faces of Vandals] (ATH: новости Беларуси и мира, 9 May 2022) < https://www.youtube.com/watch?v=4zBqJeHGcfo&ab_channel=%D0%90%D0%A2%D0%9D%3A%D0%BD%D0%BE%D0%B2%D0%BE%D1%81%D1%82%D0%B8%D0%91%D0%B5%D0%BB%D0%B0%D1%80%D1%83%D1%81%D0%B8%D0%B8%D0%BC%D0%B8%D1%80%D0%B0> last accessed 11 April 2023; 'Есть ли нацизм на Западе? Преступления под ширмой американской демократии. ЭТО ДРУГОЕ' [Are there Nazis in the West? The Crimes Under the Shelter of American Democracy] (ATH: новости Беларуси и мира, 5 April 2022) < https://www.youtube.com/watch?v=8fyWkuFsMp8&ab_channel=%D0%90%D0%A2%D0%9D%3A%D0%BD%D0%BE%D0%B2%D0%BE%D1%81%D1%82%D0%B8%D0%91%D0%B5%D0%BB%D0%B0%D1%80%D1%83%D1%81%D0%B8%D0%B8%D0%BC%D0%B8%D1%80%D0%B0> last accessed 11 April 2023; «Нацизм на Украине расцветает: нацполк Азов возвращается. Алексей Кочетков» [Nazism in Ukraine Is Rising: Nazi formation 'Azov' is coming Back] (Партия «Справедливая Россия – За правду», 29 December 2022) < https://www.youtube.com/watch?v=zQk4hlEDv6Y&ab_channel=%D0%9F%D0%B0%D1%80%D1%82%D0%B8%D1%8F%D0%A1%D0%9F%D0%A0%D0%90%D0%92%D0%95%D0%94%D0%9B%D0%98%D0%92%D0%90%D0%AF%D0%A0%D0%9E%D0%A1%D0%A1%D0%98%D0%AF%E2%80%93%D0%97%D0%90%D0%9F%D0%A0%D0%90%D0%92%D0%94%D0%A3> last accessed 11 April 2023.

by Russian ambassadors[28] and by chief editors of the main Russian media channels,[29] which have not been blocked and may be easily accessed in various countries. Not less wide-spread information flowing around various social media was about Ukrainian secret bio-laboratories where 'ethnically -oriented' biological weapons (ducks, bats and mosquitoes) had been produced 'purporting to arrange massive attacks on Russian people to infect them'.[30] Moreover, even COVID-19 according to the results of a so-called 'Russian investigation' was invented in Ukraine,[31] which was even said to be 'confirmed' by the US Video-content concerning these issues is still accessible via YouTube.[32] Finally, Russian media and bloggers have been spreading fakes about the creation of nuclear weapon in Ukraine, which 'evidenced the preparation to attack Russia soon'.[33]

Fierce online debates on these and other issues often ended up with hatred of Ukrainians and incitements to violence, which undermined the safe use of social media and fuelled mass atrocities and other crimes against humanity in Ukraine. From the first days of the war various journalists of Russian channels (so-called propagandists), bloggers and influencers started sharing hateful content on their accounts on various platforms. It is still possible to find this content on Telegram or other YouTube channels. Various Russian political leaders, journalists and influencers suggested 'freezing Ukrainian civilians',[34] 'killing everybody who has a blue-yellow chevron'[35] and even 'drowning Ukrainian children in a river' or 'burning them in their houses'.[36] It comes with no surprise that by now there have been created a lot of Telegram channels coordinated by anonymous users which spread mostly hateful content, pictures and videos depicting murdered people, where thousands of followers make fun of someone's death.[37]

The spreading of the content described above forced most of the largest global platforms to correspond to the new challenges without waiting for state bodies' and public officials' decisions and instructions. However, each platform chose its own approach to fight harmful content online.

Western companies operating the largest online platforms (Facebook, Instagram, Google and YouTube) chose rather harsh remedies.

First, they amended their content moderation practices in a way allowing them to identify and directly remove disinformation and hate speech produced by Russia-related information channels. In particular, Meta Platforms Inc. (Meta) established a special operation centre staffed by experts who were native Russian and Ukrainian speakers to monitor the content posted on Facebook and Instagram.[38] It also imposed additional penalties on the accounts and domains

[28] See Mikhail Ulyanov's (Permanent Representative of Russia in Vienna) Twitter account < https://twitter.com/amb_ulyanov/status/1281703911997542409> last accessed 22 February 2023.

[29] See Margarita Simonian's (the chief editor of RT and Sputnic channels) Twitter account <https://twitter.com/M_Simonyan?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor> last accessed 22 February 2023.

[30] 'МО РФ сообщило о разрабатываемом на Украине биологическом оружии при финансировании США' [Ministry of Defence of the RF Informs about the Biological Weapon Being Currently Elaborated in Ukraine with the Financial Support by the USA] (TACC, 6 March 2022) <https://tass.ru/armiya-i-opk/13987899> last accessed 11 April 2023 (only with VPN) [In Russian];

[31] *Русская Весна* (n 8).

[32] See Тюменское время (n 7).

[33] 'Украина была очень близка к созданию ядерного оружия, сообщил источник' [Ukraine Was Extremely Close to Create Nuclear Weapon, Says a Resource] (РИА Новости, 06 March 2022) <https://ria.ru/20220306/ukraina-1776880095.html> last accessed 11 April 2023 (only with VPN) [In Russian].

[34] See 'Russian lawmakers advocate freezing and starving Ukrainian civilians, turning them into refugees' <https://www.youtube.com/watch?v=10AiNAsCnkw&ab_channel=RussianMediaMonitor> last accessed 11 April 2023.

[35] See 'Старше Эдды' post from 4 August 2022 < https://t.me/vysokygovorit/8992> last accessed 11 April 2023.

[36] See 'RT's director of broadcasting Anton Krasovsky suggests drowning or burning Ukrainian children' <https://www.youtube.com/watch?v=8lkshypC2Rk&ab_channel=RussianMediaMonitor> last accessed 11 April 2023.

[37] See for example Telegram Channel 'Повёрнутые на Z войне' [Mad about Z-war]< https://t.me/voenacher> last accessed 11 April 2023; Telegram Channel «Мертвые укропы и хохлы» <https://t.me/+ywk3KdP-uYU4OGJi> last accessed 11 April 2023.

[38] Meta (n 11).

that repeatedly shared false information.[39] The efforts made by the company have targeted even WhatsApp messenger: messages that have not originated with the sender have been labelled, while reposting third-party messages has been limited.[40] Special attention has been paid to Russian state-funded accounts: initially, the company applied labels to content generated by them, but later it restricted access to RT and Sputnik across the EU and the UK and demoted the content from Russian state-controlled media making them harder to find.[41] Alphabet Inc. (Alphabet) made primarily the same steps. The company decided to block YouTube channels connected to RT and Sputnik ahead of the pan-EU sanction on the channels.[42] To resist the spreading of content that denies, minimizes or trivializes 'well-documented violent events' related to the war,[43] YouTube has blocked hundreds of other Russian channels and thousands of videos, including the channels of the most prominent Russian pro-war journalists (so-called 'propagandists').[44] Meanwhile, the company made efforts to promote truthful information by virtue of breaking news and 'top news' sections of its homepage.[45]

Second, the companies took various steps concerning ads restriction. Meta initially prohibited ads from Russian state media and demonetized their accounts. Later it paused ads targeting people in Russia and restricted Russian advertisers from running ads anywhere in the world, including within Russia.[46] Alphabet acted in much the same way. It blocked all ads in Russia, including Search, YouTube and Display ads, which is why people in Russia could not see any ads from Google.[47] This way the company tried to prevent practices of using ads together with posts unreasonably blaming Ukraine for the initiation of the war.[48]

Finally, companies used some novel remedies to mitigate online conflicts. In particular, Meta changed the practice of applying its hate speech policies to Ukrainian users. It derogated from its standard practices and gave Ukrainian users more freedom when expressing their attitude to Russians and Russia.[49] This step was explained by ethical issues: the company stated that restricting people who suffered from unprecedented violence from posting emotional content was too severe and unjustified.[50]

Assessing practices used by Meta and Alphabet, it may be said that generally they have been justified and rather consistent. The companies made many efforts to resist hate speech and disinformation being the basic pillars of Russian information policy online. However, the scrupulous legal analysis shows that these practices lack transparency and consistency. While expanding fact-checking mechanisms, the companies failed to provide the public with any information on the number of fact-checkers they invited and on the level of their expertise.[51] Besides, neither Meta, nor Alphabet has disclosed criteria used by their content moderators and algorithms to downrank, block or delete posts and accounts

[39] Meta (n 11).
[40] Meta (n 11).
[41] Meta (n 11).
[42] Kent Walker, 'Helping Ukraine' (*Bloggoogle*, 04 March 2022) <https://blog.google/inside-google/company-announcements/helping-ukraine/> last accessed 11 April 2023.
[43] Natasha Lomas, 'YouTube is now blocking Russia state-affiliated media globally' (Techcrunch, 11 March 2022) <https://techcrunch.com/2022/03/11/youtube-is-now-blocking-russia-state-affiliated-media-globally/> last accessed 11 April 2023.
[44] 'YouTube removes more than 9,000 channels relating to Ukraine war' (The Guardian, 22 May 2022) < https://www.theguardian.com/technology/2022/may/22/youtube-ukraine-invasion-russia-video-removals> last accessed 11 April 2023.
[45] Walker (n 42).
[46] Meta (n 11).
[47] Natasha Lomas, 'Google pauses its ad sales in Russia, Microsoft pauses sales' (*TechCrunch*, 4 March 2022) <https://techcrunch.com/2022/03/04/google-microsoft-sales-pause-russia/> last accessed 11 April 2023.
[48] Natasha Lomas, 'Alphabet confirms Russia is restricting Google News' (*TechCrunch*, 24 March 2022) <https://techcrunch.com/2022/03/24/russia-blocks-google-news/> last accessed 19 September 2022.
[49] Meta (n 11).
[50] Meta (n 11).
[51] In particular, Meta did not disclose in its 'Ongoing Efforts Regarding Russia's Invasion of Ukraine' any information neither on the number of content moderators hired to cope with new challenges, nor on their expertise. There is only a general information about this fact itself (see Meta (n 11).

(channels) of their users: in their policies one can find only general terms like 'Russian state-controlled media', 'violent content', 'malicious activity',[52] 'illegal content', 'channels engaging in coordinated deceptive practices'[53] etc. The decisions to impose ads restrictions for Russian users, although generally justified, have deprived Russian anti-war influencers from the possibility of getting financing from ads and enlarging their audiences. Finally, the decision to extend the borders of permissible fury expressions of Ukrainian users, although being justified from an ethical perspective, does not comply with the standard of clarity and foreseeability, and was not duly communicated to the platform users. Unsurprisingly, these vulnerabilities were used by Russian state bodies as an occasion to justify severe limitations applied to Meta's and Alphabet's services to limit free speech online. Meta's platforms Facebook and Instagram were overally blocked in Russia by Russian state body Roskomnadzor.[54] The restrictions were accompanied by the recognition of Meta an extremist organization according to the judgement of Tverskoy district court of Moscow.[55] Alphabet's services have not been blocked yet, however, there are a lot of voices in the Russian parliament and government claiming for it.[56]

Unlike Alphabet and Meta, Twitter during the first months of the war acted more carefully. The keynote of its practices considering the Russian-Ukrainian war has been that 'content moderation must extend beyond the leave-up-take-down binary'.[57] The company has been dwelling on promoting fact-checked information rather than removing the false one. For this reason, it created special rubrics where fact-checked news on the latest events concerning the war was gathered and spread.[58] However, the platform has not suspended or blocked the Russian-funded media accounts (twits) worldwide: it blocked them only in the EU in response to the EU sanctions, while in the rest of the world these accounts were merely withdrawn from recommendations and marked with special labels.[59] Moreover, the company decided not to remove Russian governmental accounts and the accounts of the most well-known Russian propagandists and pro-war bloggers.[60] Twitter mainly labelled these accounts with a special sign, however, they remain accessible from every corner of the Earth.[61] Finally, in the Spring of 2023 the platform started acting in a very surprising manner: it started featuring Russian government accounts (including Vladimir Putin's one) at the top of certain search results and showing them up in suggestions, which presumably means that Twitter removed restrictions announced in the beginning of the war.[62]

[52] See Meta (n 11).

[53] Walker (n 42).

[54] 'Приняты ответные меры на ограничение доступа к российским СМИ' [The Remedies in Response to the Limitation of Access to Russian Media Have Been Taken] (*RKN*, 4 March 2022) <https://rkn.gov.ru/news/rsoc/news74156.htm?utm_source=google.com&utm_medium=organic&utm_campaign=google.com&utm_referrer=google.com> last accessed 11 April 2023; 'Об ограничении доступа к социальной сети Instagram' [On the Limitation of Access to Instagram] (*RKN*, 11 March 2022) <https://rkn.gov.ru/news/rsoc/news74180.htm> last accessed 11 April 2023.

[55] Фаина Ваулина, 'Страна-агрессор считает, что деятельность компании направлена против ее вооруженных сил' [The Aggressor State Considers the Company's Activity Being Against Its Armed Forces] (ZN, 21 March 2022) < https://zn.ua/TECHNOLOGIES/v-rossii-sud-priznal-kompaniju-meta-ekstremistskoj-orhanizatsiej.html> last accessed 11 April 2023.

[56] Reuters, 'Russia takes steps to punish Google over YouTube "fakes"' (Euronews, 08 April 2022) <https://www.euronews.com/2022/04/07/ukraine-crisis-russia-google> last accessed 11 April 2023.

[57] McSweeney (n 14).

[58] McSweeney (n 14).

[59] McSweeney (n 14).

[60] McSweeney (n 14).

[61] For example, Margarita Simonian's account (Chief Editor of RT and Sputnik) < https://twitter.com/M_Simonyan?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor> and Vladimir Solovyov's account (Russian journalist being under personal sanctions provided by the EU, the UK and the USA) < https://twitter.com/VRSoloviev> are accessible and are not blocked. Twitter accounts created and held by Ministry of Foreign Affairs of Russian Federation < https://twitter.com/MID_RF> is also accessible and posts content rather often.

[62] James Titcomb, 'Putin's Twitter account resurfaces as Russia comes in from the cold' (The Telegraph, 7 April 2023) < https://www.telegraph.co.uk/technology/2023/04/07/elon-musk-twitter-lifts-restrictions-putin-kremlin-russia/> last time accessed 11 April 2023,

Twitter has also changed its ads policy in light of the war. Unlike Meta and Alphabet banning monetization only for Russian users, Twitter has been pausing advertising both for Ukrainian and Russian users 'in order to ensure that critical public safety information is elevated, and ads don't detract from [it]'.[63] Moreover, the platform decided to stop monetization for content focussing on the war or content considered misleading.[64]

However, this approach has also shown obvious vulnerabilities. The platform's policy of mitigating disinformation turned out to be not enough transparent and reliable: it led to the surprising blocking of some accounts sharing details of the invasion during the first days of the war, although the accounts had not posted anything harmful or misleading.[65] The platform's 'polite' approach to Russian state media and government accounts has also shown its weaknesses: a lot of Russian government accounts continue spreading unchecked information and aggressive propaganda with no limitations.[66] Moreover, it led to extremely contradictory decisions in Spring 2023: the platform decided not to remove[67] a twit created by Dmitry Medvedev (Russian ex-president) where he calls Ukraine a 'country 404' and its government 'a Nazi regime'.[68] However, this 'polite' approach did not save the platform from the restrictions inside Russia: Russian state bodies responded to it with the 'polite' decision to limit access of Russian users to Twitter, although did not block the platform completely.[69]

Alongside platforms putting at least some restrictions in response to the invasion, some platforms took a completely libertarian approach. The most prominent among them is Telegram—a kind of a 'hybrid' of a private messenger and a social media, which allows its users to type private messages and create large channels and groups joined by millions of users.[70] Because Telegram combines the features both of a private messenger and a social media, it is sometimes called 'messenger-as-social media'.[71] This platform architecture causes other peculiarities distinguishing it from 'classical' social media. First, Telegram does not use algorithmic suggestions for its users: everyone can freely choose which channel to join or leave and refrain from joining any.[72] Second, Telegram does not censure speech on channels or in groups giving this issue away to their administrators. Since its foundation in 2013, the platform has been standing for a 'hands-off' approach to content moderation. It has not been interfering with any content, unless it falls under the definition of terrorist content, content violating copyrights or other illegal content.[73] The company does not moderate content posted in private groups and chats and does not remove content or channels which merely express an 'alternative view contradicting the official position of a particular state'.[74]

[63] McSweeney (n 14).
[64] McSweeney (n 14).
[65] Minhaj Adnan, 'Twitter 'mistakenly' blocked accounts sharing info on Russian-Ukraine war' (The Siasat Daily, 24 February 2022) <https://www.siasat.com/twitter-mistakenly-blocked-accounts-sharing-info-on-russian-ukraine-war-2281203/> last accessed 11 April 2023.
[66] Timothy Graham, Jay Daniel Thompson, 'Russian government accounts are using a Twitter loophole to spread disinformation' (*The Conversation*, 15 March 2022) <https://theconversation.com/russian-government-accounts-are-using-a-twitter-loophole-to-spread-disinformation-178001> last accessed 11 April 2023.
[67] Elon Mask, 'In response to @TwitterDaily' (Twitter, 10 April 2023) < https://twitter.com/elonmusk/status/1645177202961534977> last accessed 11 April 2023.
[68] Dmitry Medvedev, 'Why Will Ukraine Disappear? Because Nobody Needs It' (Twitter, 8 April 2023) https://twitter.com/MedvedevRussiaE/status/1644669039095037953 last accessed 11 April 2023.
[69] 'Роскомнадзор ограничил доступ к Twitter' (Habr, 5 March 2022) <https://habr.com/ru/news/t/654473/> last accessed 11 April 2023.
[70] 'Telegram FAQ' (Telegram) <https://telegram.org/faq?setln=uk#groups-and-channels> last accessed 11 April 2023.
[71] William Marks, David Nemer, 'Telegram's Embrace of Contradiction' (Lawfare, April 6, 2022) < https://www.lawfareblog.com/telegrams-embrace-contradiction> last accessed 11 April 2023.
[72] William Marks (n 71).
[73] Mariëlle Wijermars and others, 'Is Telegram a "harbinger of freedom"? The performance, practices, and perception of platforms as political actors in authoritarian states' (2022) 38:1–2 Post-Soviet Affairs, 125, 133.
[74] Telegram FAQ (n 70).

According to 'The Times' investigation Telegram has become a 'digital battlefield in the Russian-Ukrainian war'[75] and the most popular platform used by Russian and Ukrainian users to communicate via private chats and share various information concerning the war. The hands-off approach of the platform allowed both Ukrainian[76] and Russian[77] state officials and ultra-right forces on both sides of the frontline to create their own channels having millions of followers. In the end, a lot of such channels have filled the platform with the severest hate expressions, terrible photos and videos, incitements to violence etc. The platform has not taken any step to mitigate these facts and conflicts. It only obeyed the European sanctions and 'barred Kremlin-backed media outlets from using its platform within the EU'. However, these measures did not restrict these media from creating other accounts with fake names, which the platform is not going to resist anyhow[78].

The approach to content moderation used by Telegram is very controversial. On the one hand, the platform makes it clear in its policies that it does not interfere with discussions between users and reserves the right not to respond to local governments' requests in any country. On the other hand, it allows various extremists and aggressive users to incite millions of people to mass violence. As a result, Telegram may become the largest breeding ground for hate speech and disinformation in history.

Finally, TikTok has been taking the most extravagant approach to content moderation in light of the invasion. Among its first steps the platform announced combatting misinformation by partnering with independent fact-checking organizations and labelling the content from state-controlled media.[79] The company also blocked access to Russian state media accounts under the respective EU sanctions.[80] The platform's next step was much more radical: on 6 March 2022 it decided to suspend the possibility of livestreaming and posting new content for Russian users and to block Russian users from seeing any content from elsewhere except Russia.[81] As the platform explained, this step was needed to protect Russian users from severe legal consequences brought about by adopting the 'fake news' law in Russia.[82]

However, in practice the measures applied by TikTok led to very controversial consequences. Concerning disinformation, the remedies used by the platform were considered far from being enough: expert organizations found out that there were loopholes in TikTok algorithms which allowed the platform to promote Russian content to Russian and European users even after the EU sanctions and updates of internal policies.[83] Moreover, new pro-war content recorded

[75] Bergengruen (n 12).

[76] For example, President V. Zelensky has a channel on Telegram, which he uses for sharing various information and his opinion on various events (see Zelenskiy/Official (Telegram) <https://t.me/V_Zelenskiy_official> last accessed 19 September 2022).

[77] Russian ex-president Dmitrii Medvediev has his channel on Telegram where he also shares his opinion on various issues, and this opinion is radically anti-Ukrainian (see Дмитрий Медведев (Telegram) <https://t.me/medvedev_telegram> last accessed 19 September 2022).

[78] Mark Scott, 'Telegram bans Russian state media after pressure from Europe' (*Politico*, 4 March 2022) <https://www.politico.eu/article/russia-rt-media-telegram-ukraine/> last accessed 19 September 2022.

[79] 'Bringing more context to content on TikTok' (Newsroom TikTok, 4 March 2022) <https://newsroom.tiktok.com/en-us/bringing-more-context-to-content-on-tiktok> last accessed 19 September 2022.

[80] Ciarán O'Connor, 'TikTok is on the precipice of a disinformation scandal that's going unchecked' (Business Insider, 21 March 2022) <https://www.businessinsider.com/tiktok-is-on-the-precipice-of-a-disinformation-scandal-2022-3?r=US&IR=T> last accessed 19 September 2022.

[81] Newsroom TikTok (n 53).

[82] On 4 March 2022 Russian parliament adopted the Law amending the Criminal Code of RF and articles 31 and 151 of the Criminal Procedure Code of RF (see Федеральный закон 'О внесении изменений в Уголовный кодекс Российской Федерации и статьи 31 и 151 Уголовно-процессуального кодекса Российской Федерации' от 04.03.2022 N 32-ФЗ < http://publication.pravo.gov.ru/Document/View/0001202203040007> last accessed 19 September 2022). The law sets criminal liability for the public sharing of misleading information and for public acts aiming to discredit the use of Armed Forces of the Russian Federation to defend the interests of the Russian Federation, its citizens and maintain global peace and safety.

[83] The Cube, 'TikTok is still promoting banned Russian content to users, says report' (*Euronews*, 11 August 2022) <https://www.euronews.com/my-europe/2022/08/10/tiktok-is-still-promoting-banned-russian-content-to-users-says-report> last accessed 11 April 2023.

by Russian state media far after March 2022 turned out to be displayed in TikTok recommendations for users worldwide, including EU users.[84] The decision to cut Russian users from the platform happened to be even more questionable. Russian TikTok users found themselves in an information bulb where they could only reach old videos mainly created by other Russian users and could not reach any new content from abroad. Meanwhile, the blockage of access to a new content turned out to be rather tricky: by virtue of VPN and other technologies active Russian TikTok users have continued creating and posting pro-war content, while users being usually passive could watch only content of this kind and still could not access any content created by users from other countries.[85]

Hence, neither of practices used by platforms to combat risks brought about by the war were flawless. Considering the scale of the conflict and the number of people involved, these flaws have already caused dramatic effects not only on protection of individual human rights, but on global safety in whole. Thus, the full-scale Russian invasion and informational policy facilitating it is another proof of the need to have some instruments and mechanisms capable of addressing content moderation issues and ensuring online safety.

REGULATORY APPROACHES TO ADDRESS CONTENT MODERATION ISSUES

Although the war initiated by Russia started in 2022, the risks posed by scarcely controlled information flow online have been one of the most serious concerns of modern society during the last decade.[86] Thus, it actualized a need for external mechanisms to regulate platforms' practices and led to the development of different approaches, which may be grouped into two main ones: a bottom-up or a self(co-)-regulatory and a top-down (a state) one.

In this part, I will analyse both of them to find out whether they can combat illegal and harmful content and speech online in times of sharp and mass social conflicts.

Self (co-)-regulatory approach

Self-regulation is said to be the 'dominant form of regulation in the online environment', especially in early days of Internet.[87] Platforms started using self-regulatory mechanisms to improve their practices when they faced social pressure and responsibility for users' security[88] for the first time. There are various forms of self-regulation, but basically they all come down to the following.

The first form is a creation of an independent or half-independent body granted a right to review the decisions of a platform concerning content moderation. The most prominent example in this context is the Oversight Board created by Meta. The purpose of the Board is to 'protect free expression by making principled and independent decisions about important pieces of content and by issuing policy advisory opinions on Facebook's content policies'.[89] For these purposes, experienced and skilled persons have been chosen to be its

[84] The Cube (n 83).

[85] Will Oremus, 'TikTok created an alternate universe just for Russia' (*The Washington Post*, 13 April 2022) <https://www.washingtonpost.com/technology/2022/04/13/tiktok-russia-censorship-propaganda-tracking-exposed/> last accessed 11 April 2023.

[86] Examples illustrated in the previous Part prove this statement t.

[87] Teresa Quintel and others, 'Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond', Bilyana Petkova and Tuomas Ojanen (ed), *Fundamental Rights Protection Online: The Future Regulation Of Intermediaries* (Edward Elgar Publishing 2020) 200.

[88] 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 Int'l L.J. 939, p. 952.

[89] Charter (n 18).

members and to deliver judgements in various cases concerning content moderation.[90] The Board has the authority to review Facebook's decisions concerning some pieces of content on request of an 'original poster of the content', 'a person who previously submitted the content to Facebook for review' or on Facebook's request to deliver judgement in a particular case concerning content moderation or to deliver a policy recommendation.[91] Noticeably, the Board has already delivered many judgements in very resonating cases, including the suspension of Donald's Trump account.[92] The Board's decisions in cases concerning the removal of content or other content moderation issues are binding for Facebook. However, 'this obligation is strictly limited to the precise content decided upon'.[93] Thus, the decisions do not have a power of precedent in other similar cases.

The second form of self-regulation is establishing non-governmental organizations with various powers over its members (so-called SROs). SROs regulate their members' activity by creating codes of conduct and observing compliance with them. Moreover, these organizations seek to protect their members' interests and improve their business activity in other ways. The most prominent example of organization of this kind is a GNI— a 'multistakeholder platform' comprising leading ICT companies, civil society organizations, academics, investors, and others[94] to protect and advance freedom of expression and privacy rights in the ICT industry.[95] A core of this organization is its Principles on Freedom of Expression and Privacy based on internationally recognized laws and standards of human rights.[96] Companies joining GNI declare to be committed to these principles and to implement them into their practices. To observe compliance with principles, the organization ensures the assessment of each participating company's activity by independent assessors.[97] If the assessors conclude that a company does not comply with the principles, the GNI Board may 'place that company under special review'.[98] Meanwhile, the outcome of each assessment is also reported to the public.

Finally, the last form of self-regulation worth mentioning is a joint elaboration of principles and codes of conduct by various platforms and stakeholders. One of the prominent examples in this regard is Manila Principles on Intermediary Liability. The document was developed by a broad coalition of civil society groups and experts from around the world and subsequently signed by various ICT companies and other organizations.[99] The Principles provide for a nuanced and balanced approach to issues of content moderation, seeking to protect the users of ICT services from harmful and fake information and to protect ICT companies from unjustified interference of states with their activity.[100] A not less prominent example is Santa Clara Principles on Transparency and Accountability in Content Moderation. They also were developed by academics and human rights organizations and signed by the largest ICT companies (like Meta, Google and Twitter). Just like Manila Principles, Santa Clara Principles are based

[90] See the detailed overview of the process of selection of members and the initial problems of this selections here: Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression (2020) 129 (8 The Yale Law Journal 2418, p. 2458).

[91] Article 2 of Charter (n 18).

[92] Case decision 2021-001-FB-FBR (Oversight Board, 5 May 2021) <https://oversightboard.com/decision/FB-691QAMHJ/> last accessed 11 April 2023.

[93] Kate Klonick (n 90) p. 2464.

[94] About GNI <https://globalnetworkinitiative.org/about-gni/> last accessed 11 April 2023.

[95] Our mission <https://globalnetworkinitiative.org/team/our-mission/> last accessed 11 April 2023.

[96] The GNI Principles <https://globalnetworkinitiative.org/gni-principles/> last accessed 11 April 2023.

[97] Accountability, Policy & Learning Framework. URL: https://globalnetworkinitiative.org/accountability-policy-learning/.

[98] Accountability (n 97).

[99] The Manila Principles on Intermediary Liability Background Paper of 30 May 2015 <https://www.eff.org/files/2015/07/08/manila_principles_background_paper.pdf> last accessed 11 April 2023.

[100] The Principles stand for legality, legitimacy and proportionality of any state interference and contain rather high thresholds for such interference into the practices of ICT companies. They also shield intermediaries from strict liability for third-party content (See Manila Principles (n 99)).

on human rights standards and seek to provide basic rules on how content moderation should protect Internet users and balance their rights and interests in the online environment.[101]

Self-regulation has undoubtful advantages, being a balanced and borderless type of regulation. It attempts to provide the most flexible mechanisms for addressing various issues arising in practice and caused by the largest and the hardest social conflicts. For instance, GNI was the first among all organizations to warn platforms and governments about the possible dramatic consequences of the Russian invasion on the freedom of expression and privacy online[102] and the first to address to the UN Special Rapporteur the Submission on Freedom of Expression of armed conflict and other disturbances.[103]

However, it also has undeniable weaknesses. All described forms of self-regulation risk to lack transparency and independence. In particular, the first form of self-regulation shows that entities like the Oversight Board in fact remain dependent on the platform which created them.[104] It also demonstrates that procedures used by the Board while hearing cases lack transparency.[105] SROs also risk lacking transparency when exercising their regulatory powers: being associations of competing ICT companies, these organizations may start promoting the interests of some members to the detriment of others.[106] Moreover, being less publicly observed than state regulators, SROs can hide some violations committed by their members from the public.[107] However, the most serious weakness of self-regulation is its lack of enforceability and generally voluntary nature. The example of the Oversight Board shows that it lacks the authority to oblige Facebook to do something.[108] SROs' powers are also insufficient to make platforms liable for their violations. Finally, self-regulatory codes of conduct and principles have the weakest mechanisms of enforcement since there is no entity entitled to observe compliance with them.

Trying to fix these vulnerabilities, some states attempted to support self-regulation with the mechanisms of state enforcement, which gave rise to co-regulation.[109] It may have various forms. The simplest is the form of an agreement between a state (a local community) and a platform.[110] Another form is a joint preparation of codes of conduct by state bodies and platforms

[101] Santa Clara (n 16).

[102] GNI Statement on Protecting and Respecting Freedom of Expression and Privacy in the Context of Russia's Invasion of Ukraine <https://globalnetworkinitiative.org/wp-content/uploads/2022/03/GNI-Statement-on-Protecting-and-Respecting-Freedom-of-Expression-and-Privacy-in-the-Context-of-Russias-Invasion-of-Ukraine.pdf> last accessed 11 April 2023.

[103] 'GNI Submission to UN Special Rapporteur on Freedom of Expression on Armed Conflict' (*Globalnetworkinitiative*, 19 July 2022) <https://globalnetworkinitiative.org/gni-submission-un-sr-foe-2022/> last accessed 11 April 2023.

[104] According to the Charter of Oversight Board its members are initially elected by a group of co-chairs selected by Facebook and formally appointed by trustees, who in their turn are also appointed by Facebook (article 1). Facebook is also the source of the Board's funding, even though formally it is provided via the trust (article 5) (see Charter (n18)). These issues pose risks for the transparency of the Oversights Board's structure and the observation of due process by the Board (See Kate Klonick (n 90) p. 2460).

[105] The Charter of the Board says that it 'has the discretion to choose which requests it will review and decide upon'. Moreover, the Board establishes its own set of procedures that its staff will use to select a pool of cases from which the board can choose (article 2) (see Charter (n 18)).

[106] That is why SROs are sometimes called 'fox in a henhouse' (see Jodi Lynne Short, *From command-and-control to corporate self-regulation: How legal discourse and practice shape regulatory reform* (University of California, Berkeley, 2008) 22).

[107] In this context a scandal concerning GNI is worth mentioning. The organization turned out to veil the human rights violations by Google and Yahoo! in China (see Evelyn Douek, 'The Limits of International Law in Content Moderation' (2021) 6 UCI J. INT'L TRAN'L & COMP. L. 37, 58).

[108] The Charter of the Board outlines its major powers, however, it does not oblige Facebook or other entities to respond to what the Board has ordered (see Charter (n 18)). In the Trump's case the Board asked Facebook 46 questions, and Facebook refused to answer part of them simply because in its opinion 'this information was not reasonably required for decision-making' (see Case decision (n 92)).

[109] There is no uniform definition of this term. However, in the EU Interinstitutional agreement on Better Lawmaking it was defined as a mechanism whereby a Community legislative act entrusts the attainment of the objectives defined by the legislative authority to parties which are recognized in the field (see Interinstitutional agreement on better law-making, *Official Journal* C 321, 31/12/2003 P. 0001–0005).

[110] For example, Airbnb and Amsterdam signed a memorandum of understanding that imposes the duty on Airbnb to prevent rentals for periods extending local rules. The same memorandums have been signed with other local communities (see Michèle Finck, 'Digital Co-Regulation: Designing a Supranational Legal Framework for the Platform Economy' (2017) 43 (1) European Law Review, 60).

and joint compliance monitoring. The most prominent examples in this context are Code of Conduct on Countering Illegal Hate Speech Online 2016[111] and Code of Practice on Online Disinformation 2018[112] (and the renewed Code of 2021[113]). The Codes were developed and agreed on by the leading tech companies and players in the advertising industry, which was initiated by the European Commission. The monitoring and assessment of compliance are also given to the signatories of the Codes, nevertheless the Commission retains the right to prepare its own assessments.[114] The regulatory potential of the Codes has been further strengthened by the recently adopted DSA, which states that refusal without proper explanations by an online platform to participate in the application of such a code of conduct could be taken into account when determining whether the online platform has infringed the obligations laid down by the Regulation.[115]

However, the example of the mentioned EU codes of conduct shows that these co-regulatory instruments still are not deprived of weaknesses typical for self-regulation. Like the latter, the Codes are generally non-binding and not dully enforceable against platforms.[116] There is no effective monitoring mechanism to ensure compliance with these Codes: they shift from monitoring to reporting systems obliging platforms to make periodical public reports, but leaving the issue of monitoring aside. Noticeably, even the DSA does not contain any provision on the consequences of non-compliance with the Codes—it only mentions consequences for refusing to sign the Codes, but not for compliance with them in the routine practice of content moderation. All in all, these vulnerabilities of self- and co-regulation make them weak in countering harmful content and disinformation online, especially in times of mass social conflicts like wars and armed conflicts.

State (top-down) regulation

For a long time the Internet has been considered an area free from state regulation.[117] However, the last decade has shown a kind of race towards so-called tech nationalism[118]: there are more and more voices claiming that 'rules for what people can say online are too important to leave entirely to private actors'.[119] As a result, many states worldwide have started to adopt their laws on content moderation and platforms' liability. Due to the lack of space below, I will provide only a short analysis of the main provisions of these laws. For this purpose, I have chosen the latest bills developed in China, the EU and the UK. This choice is not a random one. First, these countries have been among the first and the few to introduce their approaches regulating content moderation issues. Second, the comparison of the laws (bills) developed in these countries can outline significant differences between authoritarian (Chinese) and democratic (European and British) views on the way content moderation should be regulated. Meanwhile, the comparing the Regulation recently adopted in the EU and the Bill introduced in the UK may help

[111] Code of conduct on countering illegal hate speech online of 30 June 2016 <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> last accessed 11 April 2023.

[112] 2018 Code of Practice on Disinformation <https://ec.europa.eu/newsroom/dae/redirection/document/87534> last accessed 11 April 2023.

[113] The 2022 Code of Practice on Disinformation <https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation> last accessed 11 April 2023.

[114] In particular the Commission published its assessment of 2018 Code of Practice on Disinformation in 2020 (see Assessment of the Code of Practice on Disinformation—Achievements and areas for further improvement SWD (2020) 180 final).

[115] Recital 104 of the Preamble of the Digital Services Act (n 20).

[116] Quintel (n 87) 205.

[117] Giovanni De Gregorio (n 1) 78.

[118] Terry Flew, 'The Challenge of Media Platform Regulation for Small and Medium-Sized Nations' (SSRN, 28 October 2021) 1, 3 <https://ssrn.com/abstract=3951610> last accessed 19 September 2022.

[119] Douek (n 107) 41.

to check whether democratic regimes (like the ones existing in the UK and in the EU) share the same approaches to these issues or look at them from quite different angles. Moreover, such a comparison may be interesting in light of Brexit as it may show whether the UK previously being a member of the EU still follows the same path as the EU in terms of regulatory approaches.

It comes with no surprise that China was among the first countries which started developing their own legislation regarding the online activity. Although there is no official translation of its laws, much may be learned about Chinese policy from the U.N Special Rapporteur's Reports on freedom of expression.[120] Chinese legislation in the analysed field is built on the concept of 'Internet sovereignty'[121] meaning that the communist party has absolute power in determining state policy concerning Internet use. The main legislative act concerning freedom of expression online is the Cybersecurity Law of the People's Republic of China of 2016. The law imposes a very wide range of rather vague duties on Internet users restricting them, in particular, from using networks for 'the overturn of socialist order', 'inciting ethnic hatred and discrimination', 'creating and disseminating false information to disrupt economic or social order' etc. (article 12).[122] Meanwhile, for network operators[123] the law also sets rather broad duties, like a duty 'to be honest and credible', 'to perform obligations to protect network security', 'to accept supervision from the government and public', and 'to bear social responsibility' (article 9)[124]. Network operators shall also 'strengthen the management of information published by users, immediately stop transmission of information prohibited by the law or administrative regulations and report it to the relevant competent departments' (article 47).[125] For violation of these provisions, platforms hold strict liability including sanctions provided for by the criminal law.[126]

Besides the Cybersecurity Law, there are other laws imposing obligations on network operators. According to the U.N. Special Rapporteur's report, there are numerous laws and regulations which impose general obligation of monitoring, surveillance and censorship on online platforms, requiring them to institute regular 'patrols' of users' posts online.[127] Compliance with these rules is ensured with criminal sanctions.[128] All in all, the Chinese approach to regulating content moderation is one of the strictest and reflects an authoritarian policy of the state concerning freedom of speech.

Although for quite a long time jurisdictions representing well-developed democracies refrained from adopting special laws concerning content moderation, during the last three years their attitude has changed dramatically. Having originated in some European countries,[129] the approach towards strengthening the regulation in this field has become pan-European. In the end of 2020 European Commission published Proposal for a Regulation on a Single Market for DSA, which was adopted by the European Parliament in July 2022[130] and came into force

[120] Submission to U.N. Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye, By Human Rights in China Regarding the Telecommunications and Internet Services Sectors in China (*OHCHR*, November 16, 2016) <https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/Telecommunications/HRIC.pdf> last accessed 11April 2023; Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression 70, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018).

[121] Submission (n 120) 9.

[122] Cybersecurity Law (n 19).

[123] A vague term which is not defined in the law, but due to its broadness is considered to cover online platforms.

[124] Cybersecurity Law (n 19).

[125] Cybersecurity Law (n 19).

[126] In particular, for failing to stop the transmission of information prohibited by law or administrative orders competent state bodies may impose fines up to 500 000 RMB, temporary suspend operations or even cancel business licenses (article 68). See Cybersecurity (n 19).

[127] Submission (n 120) p. 17.

[128] Submission (n 120) 23–24.

[129] The most prominent example in this context is the German network enforcement law (*NetzDG*) of 2017, tackling online hate speech and other harmful content (see Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (*NetzDG*)).

[130] Digital Services Act (n 20).

on 16 November 2022.[131] The Act aims ensure 'safe, predictable, and trustworthy online environment in the EU' and 'for allowing Union citizens and other persons to exercise their fundamental rights […], in particular, the freedom of expression and of information […], the right to non-discrimination and the attainment of a high level of consumer protection'.[132] To reach this aim the Act creates two main pillars. First, it stands for a broad definition of 'illegal content' covering both information which is illegal itself and information which is illegal by its reference to illegal activity.[133] Second, the Act extends its scope to all providers of intermediary services whose activity has a 'substantial connection with the EU', which covers situations when a provider has a significant number of users in one or more Member States, when it uses a language or a currency generally used in a Member State, or when a provider's application is available in the relevant national application store etc.[134] Thus, the Act is applied by virtue of the 'Brussels effect'.[135]

Considering providers of intermediary services and, in particular, online platforms, the Act follows the general line established in E-commerce Directive 2000/31/EU: it does not impose a general monitoring obligation on them and exempts them from any liability for third-party content considering its substance.[136] However, the DSA obliges platforms to comply with some procedural and reporting obligations concerning content moderation and imposes significant sanctions for the breach of these obligations, thus shifting from substance to procedure. In particular, all platforms are obliged to set notice and action mechanisms allowing their users to make complaints on some pieces of content (article 16); to annually report on any content moderation engaged in a relevant period (article 15); to provide users with access to the internal complaint-handling systems (article 20) and others. On so-called 'very large online platforms'[137] the Act additionally imposes obligations to conduct the assessment of risks stemming from the functioning and use of their services (article 34); to be subject to audits to assess compliance (article 37) and others. The Act also imposes on platforms obligations towards their users in horizontal (contract) relations: platforms shall provide their users with a 'clear and specific statement of reasons' to remove certain content (article 17); clarify in their terms of use any restrictions on the content provided by their users (article 14) and others.

The authority to observe compliance with the DSA requirements is granted to special state bodies, Digital Services Coordinators, which shall be created in each Member State (articles 49–55). Meanwhile, concerning very large online platforms the power of observation is given to the European Commission (articles 64–76). These bodies have the authority to order a cessation of infringements, impose periodic penalties to ensure the cessation, and to impose fines for failure to comply with the DSA requirements (article 51 and article 74). The amount of fines generally is to be determined by Member States considering the maximum amounts laid down in the DSA (article 52), except for very large online platforms for which the fines are determined in the DSA directly (article 74).

Another piece of legislation purporting to regulate content moderation practices has recently been introduced in the UK—the Online Safety Bill, aiming to protect users from illegal content. Just like the European DSA, the UK Bill outlines its scope very broadly: it may be applied to

[131] Digital Services Act: EU's landmark rules for online platforms enter into force (An official website of the European Union, 16 November 2022) < https://ec.europa.eu/commission/presscorner/detail/en/IP_22_6906> last time accessed 22 February 2023.
[132] Recital 3 of Digital Services Act (n 20).
[133] Article 3(h) of Digital Services Act (n 20).
[134] Article 3(e) of Digital Services Act (n 20).
[135] Anu Bradford, 'The Brussels Effect' (2012) 107 (1) Northwestern University Law Review 1.
[136] See article 4–6 and article 8 of Digital Services Act (n 20).
[137] Platforms having 45 million and more active users each month (see article 33 of Digital Services Act (n 20)).

user-to-user and search services[138] 'having links with the UK', that is having a significant number of UK users, being capable of being used in the UK by individuals or being capable of causing material harm to individuals in the UK (section 3).[139] Thus, it might represent a kind of 'Brussels effect' in the UK. In its purposes to provide online safety, the Bill initially purported to fight not only against illegal content, but also against legal, but harmful content to adults, which was supposed to encompass disinformation, fakes and other types of content which do not fit within the notion of illegal content.[140] However, in the latest version of the Bill introduced in January 2023 the notion of this content and duties of platforms to moderate it have been excluded. Thus, for now the Bill directly obliges platforms to moderate illegal content and content harmful to children. Illegal content is defined as the one which amounts to a relevant offence, where the latter is defined through listing the categories of criminal offences (like terrorism, threats to kill, fraud, inchoate offences etc.).[141] Content harmful to children is defined as either primary priority content, priority content or content that presents a material risk of significant harm to an appreciable number of children.[142] Considering legal but harmful content to adults, the Bill in its current version does not put any direct obligations on platforms concerning combating content of this kind. However, it suggests that platforms should decide how to moderate it themselves in their Terms of Services (ToS)[143] and should be consistent in applying these Terms.[144] Thus, the issue of legal but harmful content is given away to platforms who may voluntarily moderate it under the rules involved in their ToS. Besides, the Bill suggests giving more power to filter this content to the users themselves: for this purpose, it obliges platforms to dwell on user empowerment, that is to give users the tools they may need to control the content they wish or wish not to encounter.[145]

The Bill is generally based on the idea of a statutory duty of care imposed on platforms.[146] In particular, clause 9 of the Bill obliges platforms to operate a service using proportionate systems and processes designed to prevent individuals from encountering illegal content; minimize the length of time for which any illegal content is present and swiftly take down such content upon a notice.[147] Besides the duty of care, the Bill specifies the duty of risk assessment (clause 8), content reporting (clause 16), complaints procedures (clause 17), record-keeping and review (clause 19), and other duties. Considering the number and the scope of obligations the Bill distinguishes three categories of service providers (platforms): Category 1, 2A and 2B. The threshold conditions for each category have to be specified by the Secretary of State in its regulations.[148] For the Category 1 service providers the Bill imposes extra duties: user empowerment duty (described in short above), duty to protect content of democratic importance, duty to protect journalistic content and some other duties listed in clause 6 (5). In particular, considering the duty to protect the content of democratic importance the Bill obliges platforms to take into account the importance of freedom of expression when designing proportionate systems and

[138] According to the Bill 'user-to-user service' is a term used to determine online content-sharing platforms, while 'search service' is a term identifying online platforms used as search tools (like Google) (see Part 2 of the Bill). For the sake of terminological unity throughout the article I will use the term 'online platform' instead of both 'user-to-user service' and 'search service'.

[139] Online Safety Bill (n 21).

[140] The definition of harmful content used to be provided in clause 54 (3) (see Online Safety Bill 121 2022–23 [As Amended in Public Bill Committee] *(UK Parliament, 28 June 2022)* <https://publications.parliament.uk/pa/bills/cbill/58-03/0121/220121.pdf> last accessed 11 April 2023.

[141] See clause 53 and Schedules 5, 6, 7 of Online Safety Bill (n 21).

[142] See clause 54 of Online Safety Bill (n 21).

[143] See clause 9 (8) of Online Safety Bill (n 21).

[144] See clause 64 of Online Safety Bill (n 21).

[145] See clause 12 of Online Safety Bill (n 21).

[146] Online Harms White Paper—Executive summary (*Gov.uk*, 30 April 2019) <https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper-executive-summary--2#contents> 11 April 2023.

[147] Online Safety Bill (n 21).

[148] See Schedule 11 of Online Safety Bill (n 21).

processes for making decisions about this content or about users who post it (clause 13 (1)). For this purpose, the platforms should set out policies concerning this kind of content in their ToS, which should be clear and consistently applied (clause 13 (5)). The content of democratic importance is defined as news publisher content or regulated user-generated content, which is, or appears to be, specifically intended to contribute to democratic political debate (clause 13 (6)).[149] Examples of such content would be content promoting or opposing government policy and content promoting or opposing a political party.[150] Much the same obligations are put on platforms regarding journalistic content, which is defined as news publisher content and regulated user-generated content that is generated for the purposes of journalism, and which is 'UK-linked'.[151] This includes, but is not limited to, content generated by news publishers, freelance journalists and citizen journalists.[152]

The Bill sets a mechanism for state observation of compliance with its rules by platforms. The authority to monitor platforms' activity is given to a state body, OFCOM. If the body determines a failure to comply with the Bill, it may issue a so-called 'confirmation decision' obliging the platform in breach to take certain steps to fix the situation, to pay a certain penalty or to do both of this (section 120). Noticeably, the Bill introduces rather high penalties (up to £18 million or 10% of the person's qualifying worldwide revenue for the most recent complete accounting period) (schedule 13).[153]

Having analysed the most recent legislative acts, one may conclude, that a move towards state regulation of content moderation becomes a global and apparently an inevitable trend. However, it cannot be said to be the best possible option. Each of the mentioned bills and acts has gained serious criticism for giving ground for excessive restrictions on freedom of speech, privacy, and access to information. Naturally, Chinese legislation has been considered one of the most dangerous for containing too vague provisions and prohibitions giving ground for biased governmental decisions and for imposing a general monitoring obligations on Chinese companies encouraging them to over-remove content that merely seems to be illegal or harmful.[154]

Bills elaborated in democratic states contain more deliberate and balanced provisions, although they also may threaten freedom of expression and other human rights. First, both the European DSA and British Online Safety Bill give full authority to observe platforms' compliance with their requirements to administrative bodies. Meanwhile, the U.N. Special Rapporteur on Freedom of Expression expressed his concerns with respect to this regulatory solution and stressed that 'states should refrain from adopting models of regulation where government agencies, rather than judicial authorities, become the arbiters of lawful expression'.[155] Second, both legislative acts contain rather vague definitions of basic terms like 'illegal content' which creates uncertainty about the way the acts will be applied and the way fundamental human rights will be guaranteed.[156] Finally, some provisions of the Bills provide for at least atypical if not to say unjustified state interference into contractual platform-user relationships. In particular, DSA obliges platforms to suspend the provision of their services to users that frequently provide manifestly illegal content (article 22), while Online Safety Bill prescribes the way the largest platforms should treat illegal content, content of democratic importance, journalistic content etc.

[149] Online Safety Bill (n 21).
[150] Online Safety Bill. Explanatory Notes (2023) (Parliament.uk, 18 January 2023) < https://bills.parliament.uk/publications/49377/documents/2735> last time accessed 11 April 2023.
[151] See clause 15 (9) of Online Safety Bill (n 21).
[152] Explanatory Notes (n 150).
[153] Online Safety Bill (n 21).
[154] Report (n 120) recital 15.
[155] Report (n 120) recital 68.
[156] Markus Trengove and others, 'A critical review of the Online Safety Bill' (2022) 3 (8) Patterns.

However, even if there were ideal pieces of legislation elaborated by some states, overally the tendency towards state regulation of platforms functioning globally is rather endangering: it leads to the emergence of a so-called 'splinternet' where each country will have its own piece of 'network property'.[157]

Thus, online platforms may either avoid targeting their services to certain jurisdictions, which poses the risk for freedom of speech in these jurisdictions, or implement the strictest requirements introduced by certain states so as to simplify and unify their content moderation practices globally. These tendencies will facilitate digital ostracism of users from particular countries or social (national) groups in times of war and thus may contribute to the rise of enmity between people and peoples being on the different sides of the frontline.

HORIZONTAL APPROACHES TO CONTENT MODERATION

Besides the mentioned regulatory approaches, there are other approaches to address content moderation issues. They are horizontal by nature, and during the last years they have been gaining more and more attention from scholars. In this part, I will analyse and compare these approaches and test them for the ability to counter harmful content and speech online in times of war and other mass conflicts.

Contract law approach

This approach is based on a presumption that content moderation issues may be addressed by contract law remedies. Online platforms are usually characterized as 'contractual architectures'[158] where all the relationships between them and millions of users are based on myriads of contracts. Content moderation policies are also parts of contracts (so-called 'Terms of Use') drafted by platforms and adhered to by their users.[159] Therefore, as UNCITRAL has recently stated, it is contract law and its principles of good faith and fair dealing, as well as the terms of particular contracts agreed by the parties under the principle of freedom of contract, which constitute a primary source of the rights and obligations among the various actors involved in an online platform.[160]

Applying this approach to particular cases where, for example, a user's post or account has been removed from the platform or, on the contrary, a user was insulted by a piece of content which a platform failed to remove or block, the aggrieved user may claim for redress in a private dispute against the platform. Primarily a court will check the grounds of the claim and redress mechanisms in the platform's ToS. If the latter is silent or is too vague to find a violation or a redress mechanism, the court should refer to contract law provisions and principles. For instance, if the ToS expressly allow the platform to remove content or even an account without any notification and explanation of reasons to the user, the respective ToS provision may be considered a 'surprising term',[161] an 'unfair term'[162] or an 'unconscionable clause'[163] and thus

[157] Giovanni De Gregorio (n 1) 82.

[158] T. R. de las Heras Ballell, 'The Legal Anatomy of Electronic Platforms: A Prior Study to Assess the Need of a Law of Platforms in the EU', 3(1) Italian Law Journal, 149, 150.

[159] Niva Elkin-Korenand and others, 'Social Media as Contractual Networks: A Bottom-Up Check on Content Moderation' (2022) 107 Iowa L. Rev. 987, 1000.

[160] Recital 17 of Legal issues related to the digital economy (including dispute resolution)—progress report, U.N. Doc. A/CN.9/1064/Add.3 (29 June–16 July 2021).

[161] In particular, this concept is determined in UNIDROIT Principles of commercial contracts (PICC) in article 2.1.20.: if a term is of such a character that the other party could not reasonably have expected it, it is not effective unless it has been expressly accepted by that party (see UNIDROIT Principles 2016, art. 2.1.20).

[162] As defined, for instance, in Principle No. IV.3.5 of Principles of European Contract Law (PECL) or in Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts, *Official Journal*, 95, 21.4.1993, p. 29–34.

[163] A concept typical for common law contract law (see Ingeborg Schwenzer, Pascal Hachem and Christopher Kee, *Global Sales and Contract Law* (OUP 2012) 263).

unenforceable against the user. If the ToS do not specify the remedies of redress for users whose content or accounts have been unjustifiably removed, these remedies may be deduced from general provisions of contract law: the platform should compensate damages (general remedy typical for common law contract law) or conduct a specific performance, which may come down to restoration of a content or an account (remedy typical for civil law legal tradition).[164] Moreover, since users are consumers and platforms are businesses, disputes concerning content moderation may be resolved under consumer protection law. In particular, a lack in a platform's ToS of adequate and precise provisions addressing harmful speech may be considered an unfair practice, whereas failure to disclose all the content moderation rules to users may constitute a misleading or deceptive practice.[165]

An undeniable advantage of the contract law approach is that it is not attributed to a particular state-imposed regulation and may be applied to solve content moderation issues on a borderless basis. The main source of regulation within this approach is a contract between a platform and a user and a contract law chosen by contracting parties or identified under international private law rules. Thus, this approach is a more open-ended and flexible in combating various content moderation problems on a global scale.

However, it has some weaknesses as well.

First, it does not have remedies needed to address all the problems arising in practice of content moderation. In particular, a considerable number of such problems stem from the vagueness of ToS and the lack of concrete provisions allowing users, on the one hand, to identify which content may not be posted online and, on the other hand, which content moderation practices are used to combat harmful content online and how and when they are applied.[166] Contract law has some instruments helping to fill in gaps in contractual obligations, like the doctrine of implied obligations allowing to deduce some lacking obligations from other contract terms, practices, and usages.[167] However, these instruments cannot help in case of gaps in ToS concerning content moderation. Platform users need not only to restore *some* obligations presumably implied by a platform, but also to have a very precise, concrete, and foreseeable framework of how their rights may be limited to balance them with the rights of other users and vice versa. ToS lacking this precision cause various risks for their users, beginning with the risk of unjustified limitation of users' rights to freedom of expression and ending with the risk of being a victim of bullying, hatred and disinformation.

Second, contract law does not provide adequate remedies of redress for users. Contract law remedies typically come down to the liability of one party towards the other, and this liability usually constitutes the obligation to pay damages or penalties provided for by the contract. However, in most jurisdictions, platforms enjoy a so-called 'safe harbour' regime and are released from any liability (be it contractual, tort or other). On the one hand, platforms generally may not be held liable for failing to do something concerning harmful content (in particular, for failing to remove or block content or account). Thus, users suffering from such content have few opportunities to seek for a redress. On the other hand, platforms are not held liable for any actions voluntarily taken in good faith to restrict access

[164] Ingeborg Schwenzer (n 163) 541–542.

[165] Mark MacCarthy, 'A Consumer Protection Approach to Platform Content Moderation' (2019 Bilyana Petkova and Tuomas Ojanen (ed), *Fundamental Rights Protection Online: The Future Regulation Of Intermediaries* (Edward Elgar Publishing 2020) 115.

[166] This was underlined in a number of judgements of Facebook Oversight Board (See Case Decision (n 92), Case 2020-003-FB-UA (*Oversight Board*, 28 January 2021) <https://www.oversightboard.com/decision/FB-QBJDASCV/> last accessed 11 April 2023; Case decision 2021-004-FB-UA (Oversight Board, 26 May 2021) <https://www.oversightboard.com/decision/FB-6YHRXHZR/> last accessed 11 April 2023).

[167] Provisions on implied obligations may be found in PICC (see article 5.1.2) and in PECL (see article 6:102).

to or availability of material that the platform or the user considers illegal or harmful (a so-called 'Good Samaritan' principle),[168] which deprives users whose content or account has been removed from the opportunity to ask for redress. Therefore, generally, contract law remedies are helpless in unfair content moderation practices applied by platforms. Noticeably, there are many cases where platforms were released from liability by courts due to 'Good Samaritan' rule despite their breach of their own ToS provisions.[169]

Finally, contract law remedies are designed for private disputes arising usually between two parties. Thus, they may help to address content moderation issues only in relationships between a platform and its particular user. They cannot address content moderation problems concerning groups of users and, more so—whole societies. Therefore, pure contract law remedies cannot counter the mass spreading of disinformation and hate speech giving rise to social conflicts and wars.

Human rights approach

This approach is based on the idea of a so-called 'horizontal effect' of human rights which is a part of academic discussion on the responsibilities of businesses concerning respect for human rights.[170] This idea has become a part of the international agenda and gave birth to the U.N. Guiding Principles on Business and Human Rights, which require businesses to protect and respect human rights and to mitigate adverse human rights impacts.[171]

In line with these Principles, platforms have been considered primarily responsible for respect for and protection of human rights while conducting their content moderation practices. For the first time at the international level, this argument was expressed by David Kaye, a U.N. special rapporteur on the promotion and protection of rights to freedom of expression in 2018.[172] The need for platforms to comply with international human rights standards was also underlined in various self-regulatory principles[173] and emphasized in Facebook Oversight Board's judgements.[174] The human rights approach provides platforms and other parties concerned (their users, states, etc.) with standards allowing them to determine when the users' rights may be limited and to what extent. On the other hand, they help to determine when the limitation is needed for the sake of maintaining online safety and preventing harmful behaviour online.

Following the provisions of international law concerning freedom of speech (like article 19 of the International Covenant on Civil and Political Rights (ICCPR)), when deciding to restrict users' right to freedom of expression, platforms should satisfy tripartite test: legality, legitimacy and necessity (proportionality) of limitation.[175] Considering the realms of platforms' activity, *legality* here means that all the limitations of the users' rights, the grounds and procedures of their application shall be directly mentioned in platforms' rules

[168] This rule is expressly provided in a well-known Section 230 Section 230 of the Communications Decency Act of the U.S.A. (see 47 U.S. Code § 230—Protection for private blocking and screening of offensive material) and in article 6 of the EU Digital Services Act (n 20).

[169] These cases are analysed by Niva Elkin-Koren et al (see Niva Elkin-Koren (n 159)). These are: Mishiyev v. Alphabet, Inc., 444 F. Supp. 3d 1154, 1158 (N.D. Cal. 2020); Schneider v. YouTube, LLC, No.5:20-cv-4423 (N.D. Cal. July 2, 2020), ECF No. 1; Johnson v. Twitter, Inc., No. 18CECG00078 (Cal. Sup. Ct. June 6, 2018).

[170] See Olena Uvarova, 'Business and Human Rights in Times of Global Emergencies: Comparative Perspective' (2020) 26 Comparative Law Review 225; Б. П. Карнаух, 'Захист власності Європейським судом з прав людини і горизонтальний ефект' (2021) 5 Право України 149.

[171] Guiding Principles (n 22).

[172] Report (n 120) recital 68.

[173] Manila Principles (n 99); The GNI Principles (n 96).

[174] In one of its judgements, the Oversight Board directly cited D. Kaye's report and mentioned that 'although companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right' (see Case decision 2020-003-FB-UA (n 166)).

[175] Barrie (n 88) 970.

and policies embedded in contracts with their users. These rules should be sufficiently clear, detailed and transparent.[176] *Legitimacy* means that any restriction should have a legitimate aim outlined in international conventions and documents. In particular, according to ICCPR, it may be the need to respect the rights or reputation of others, the protection of national security, the protection of public order, public health or public morals,[177] and according to European Convention on Human Rights (ECHR) also to prevent disorder or crime, to prevent the disclosure of information received in confidence, or to maintain the authority and impartiality of the judiciary.[178] Finally, *necessity (proportionality)* means that the restriction 'must be necessary and the least restrictive to achieve the purported aim'.[179] In particular, to make the right decision and to choose the best remedy platforms should consider the context of the content and various circumstances surrounding it, the characteristics of their audience (children, elderlies, etc.), the impact of remedies which may be applied, the difference between them and the effectiveness of the remedies at hand.[180]

Besides the mentioned substantive rules, human rights law provides for several *procedural rules* which platforms should observe to prevent or to remedy the infringement of their users' rights, namely due process requirements. They have been mentioned in various documents and sources and currently are not systematized. One of the most prominent among them is a notification requirement: platforms should notify their users of measures applied to their content or account and the reasons for this application since users should know why their content or account has been remedied anyhow.[181] Another one often described in various documents is a requirement to provide users with mechanisms to lodge appeals and to get redress in case of violation of their rights, in particular, using external out-of-court procedures.[182] Due process also requires using only reliable and accurate automated processes to identify or remove content, suspend accounts and combine algorithmic and human evaluation when deciding whether content complies with platform policies.[183]

Human rights approach has essential advantages which other described approaches lack. Unlike state regulation, human rights law allows to deploy a universal mechanism of balanced content moderation functioning equally globally.[184] Moreover, this approach can also limit the abusive acts of states purporting to put restrictions on the freedom of speech for particular users giving platforms legal grounds to delimit the borders of state interference.[185] Compared to contract law, the human rights approach can solve not only particular private disputes between a platform and a user, but also address content moderation issues and provide remedies for various groups of users. It also helps to find balanced solutions in disputes between platforms and their users when ToS fail to provide them and contract law is helpless either: from the human rights perspective, the lack of precision of ToS and the failure to ensure due process while applying content moderation rules constitutes the breach of human rights, which requires platforms to restore them. All in all, due to its balanced, flexible, and globally applicable nature,

[176] Report (n 120) recital 46; Barrie (n 88) 971.
[177] See article 19 (3) of the International Covenant on Civil and Political Rights.
[178] See article 10 of the European Convention on Human Rights.
[179] Douek (n 107) 44.
[180] A detailed analysis of these characteristics of necessity is provided by Sander Barrie (see Barrie (n 88) 979–988).
[181] This requirement was specifically mentioned in the UN Special Rapporteur's report (Report (n 120) recital 37) and in Manila principles (Manila principles (n 99) Principle V.a. and V.c). This requirement was also mentioned in several Facebook Oversight Board Decisions (see Case decision (n 92)).
[182] See Report (n 120) recital 38; Santa Clara (n 16) Principle 1; Manila Principles (n 99) Principle V.B).
[183] See Report (n 120) recital 36; Santa Clara (n 16) Principle 1.
[184] Douek (n 107) 44.
[185] Douek (n 107) 44.

human rights approach seems to be rather effective in times of war, when it could also be applied together with standards of international humanitarian law.[186]

However, the human rights-based approach has serious weaknesses as well. A lot of them have been mentioned by scholars particularly the vagueness of criteria and standards used in international conventions on human rights,[187] the lack of uniformity in the way the rights and their limitations are formulated in international and in national legal sources.[188] As a result, it is difficult to implement the human rights approach equally and similarly on a global level and to translate it into concrete platform policies and algorithms.[189] Meanwhile, the biggest problem is the lack of enforceability of this approach taken in its pure form.[190] Modern international sources do not impose strict obligations and responsibilities concerning respect for human rights on businesses. Moreover, no instruments hold businesses liable for failing to ensure the respect and protection of human rights in their business activity. Thus, however balanced and nuanced this approach may be, its implementation and application wholly depend on the will and decisions of particular platforms.

HOW TO ADDRESS CONTENT MODERATION ISSUES: IN SEARCH OF A GOOD COMBINATION OF METHODS

The analysis provided in the previous parts has shown that there are a lot of approaches to tackling content moderation issues. However, neither of them is effective enough to address them globally and entirely, more so in times of harsh social conflicts and wars. While self-regulation, contract law and human rights law lack enforceability and thus remain mostly volunteer solutions, state regulation, on the contrary, has a tendency to over-regulate content moderation practices, which poses risks for the balance and freedom in the online space. However, it does not mean that these approaches will still be ineffective when combined and improved in some way. What we need to do is to find the optimal combination and the ways of improvement.

Platform-user relationships and content moderation are rather sensitive issues where private and public interests coexist and are very close to each other. Although initially they are regulated by private law sources (contracts), the scale and the number of users joining the platforms make them 'in essence, a system of administrative law … [where] the administrative agency is a private company'.[191] Moreover, these issues also are subjected to state interference, which poses additional challenges for their regulation. Thus, there needs to be a twofold approach to address these issues. On the one hand, horizontal and self-regulatory approaches should be combined and strengthened to provide for effective horizontal solutions to disputes between platforms and their users. On the other hand, the role of state regulation should not be downplayed as well. However, it should be improved in a way allowing it to address its main problems (ie fragmentation of the Internet). In this part, I will outline the way both approaches may be improved.

Contract law, human rights and self-regulation: in search of combination and a way to enforce

Contract law and human rights law represent a horizontal approach and provide substantive rules for content moderation. Meanwhile, self-regulation is a remedy to consolidate their rules

[186] Arturo J. Carrillo, 'Between a Rock and a Hard Place? ICT Companies, Armed Conflict, and International Law' *(Global Network Initiative*, 1 July 2022) <https://medium.com/global-network-initiative-collection/between-a-rock-and-a-hard-place-41f1ac3e62dc> last accessed 11 April 2023.
[187] Douek (n 107) 50.
[188] Douek (n 107) 50.
[189] Barrie (n 88) 969.
[190] Barrie (n 88) 969.
[191] Jack M. Balkin, 'Free Speech is a Triangle' (2018) 118 COLUM. L. REV. 2011, 2028–29.

in the codes of conduct or principles and to enforce them by virtue of creating special procedures and entities (like the Oversight Board or SROs).

Thus, first, it should be decided whether contract law and human rights law may work together to solve disputes between platforms and their users and thus help to improve content moderation practices in whole. The idea that human rights may be embedded in contract law is not new and stems from a more general academic debate on the constitutionalization of private law, which reflects the increasing influence of fundamental rights in relationships between private parties.[192] Although some scholars have expressed sceptical views emphasizing that 'fundamental rights control State power only'[193] and that '[human rights law] does not inject any values that [private] law already has',[194] a lot of scholars are of the different view and underline that '[human rights law] finds application in private law and contributes to its constitutionalization'.[195] Considering contract law, it has also been stated that 'fundamental rights do not merely influence contract law as a conceptually distinct and autonomous category, [but also] govern contract law, thereby enjoying priority over its internal principles of justice'.[196]

Although the applicability of this approach to platforms has not been specifically discovered yet, some studies in this area seem rather interesting and persuasive. In particular, in the research by Niva Elkin-Koren et al.,[197] platforms are viewed through the lens of contractual network theory. However, using the contractual approach, the authors come to conclusions being unexpectedly close to the ones made by scholars discovering platforms from the human rights perspective. Based on contractual network theory, the authors outline the main principles which should be governing for platforms' content moderation practices: these are predictability of content moderation rules, facilitation of the due process, fairness and equal treatment and provision of effective remedies of redress for users.[198] Doesn't it remind the standards of legality, legitimacy, proportionality and procedural guarantees stemming from the human rights-based approach? Surely, it does. Thus, looking at content moderation issues from seemingly opposite standpoints (human rights-based and contract law) may lead to the same conclusions. This finding proves that contract law and human rights-based approaches do not contradict each other and, more so, are fully compatible.

Having made this conclusion, it is now important to determine how contract law and human rights align regarding content moderation. Researchers of the horizontal effect of human rights distinguish several models of this alignment. However, the most widespread among them is the one where human rights fill in contract law through the open-ended concepts of the latter, like good faith, good morals and public policy.[199]

To identify the exact concept or provision of contract law which opens the door to human rights standards, we need to find an issue in contract regulation of content moderation which is the weakest in terms of enforcement, and which needs to be backed by human rights more than others. This is the issue of liability of platforms for the application

[192] Jan M. Smits, 'Private law and fundamental rights: a sceptical view' in Tom Barkhuysen & Siewert Lindenbergh (ed) *Constitutionalisation of Private Law* (Leiden/Boston, Martinus Nijhoff Publishers, 2006) 9, 10. This debate originated in the concept of positive obligations of states and attempted to find out whether human rights standards could 'affect [private] relations even though no government body is involved yet' (see T.Barkhuysen & M.L. Emmerik, 'Constitutionalisation of Private Law: The European Convention on Human Rights Perspective' in Tom Barkhuysen & Siewert Lindenbergh (ed) *Constitutionalisation of Private Law* (Leiden/Boston, Martinus Nijhoff Publishers, 2006) 43, 54).

[193] Smits (n 192) 20.

[194] Francois du Bois, 'Human Rights and English Contract Law: Parallel Worlds?' In: Siliquini-Cinelli, L., Hutchison, A. (ed) *The Constitutional Dimension of Contract Law* (Springer, Cham, 2017) 32.

[195] Barkhuysen (n 192) 57.

[196] Olha O Cherednychenko, 'Fundamental Rights, Contract Law and Transactional Justice' (2021) 17 (2) European Review of Contract Law, 130, 133.

[197] Niva Elkin-Koren (n 159) 1041.

[198] Niva Elkin-Koren (n 159) 1041–1048.

[199] Smits (n 192) 13.

of content moderation practices or for the failure to apply the ones when facing harmful content. While liability is the primary contract law remedy which helps to punish the party in breach and provides the aggrieved party with the redress, in most countries platforms are exempted from any liability. However, in most of jurisdictions this exemption applies only if a platform uses such practices *in good faith*. In particular, according to section 230 (2) of the Communication Decency Act of the USA, no provider or user of an interactive computer service shall be held liable on account of any action voluntarily taken *in good faith* to restrict access to or availability of material that the provider or user considers obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected.[200] European DSA also mentions the good faith principles in its Article 7 which contains the 'Good Samaritan' rule having much in common with the one in the Communication Decency Act.[201]

Despite its universality and flexibility, the principle of good faith here leaves a lot of questions opened and may be difficult to apply to particular circumstances and disputes. In this regard, human rights standards can serve as an instruction providing precise criteria for distinguishing acts done in good faith from those that do not satisfy this principle. In particular, a platform may be said to comply with the principle of good faith if its actions done to block some content or account have been legal (the platform's ToS precisely identified grounds to impose restrictions and restrictive measures), legitimate (restrictions have had sufficient grounds), and proportional (there have been no softer measures capable of combating harmful consequences), and its users have been provided with due process guarantees and remedies to object its decision. On the contrary, if a platform refrains from applying any restrictions on harmful content or a user, it should be considered acting in good faith if its ToS clearly determined that a platform could refrain from any interference in such cases (legality), by refraining from imposing restrictions the platform managed to avoid severer consequences for other users (legitimacy and proportionality), the platform enacted feedback mechanisms allowing users to notify it about harmful content or activity, and by obtaining such knowledge acted expeditiously to remove (block) the content.

Hence, under this combined approach, it becomes possible to gather all the needed puzzles within a horizontal perspective on content moderation together. While contract law helps to provide redress for the aggrieved party and to impose sanctions on the party in breach, human rights standards allow finding the most balanced approach in each case and clarifying the way contract law remedies should be applied.

The combined approach may already be used in practice and ensured. In particular, users whose right to free speech has been infringed or, on the contrary, whose dignity and honour have been offended by a platform using too strict or too liberal content moderation practices may lodge their suits in state courts or use out-of-court dispute settlement bodies, like the ones mentioned in the EU DSA (article 18).[202] However, these systems obviously will not ensure the effective and full-fledged protection of users' rights. State litigation will take a lot of time, split up content moderation problems commonly faced by users on a platform into the myriads of individual disputes, and all in all, lead to the formation of different approaches to the same content moderation problems in different countries. Meanwhile, out-of-court settlements given out to independent bodies with their own procedures and rules may lack transparency, independence and coercion of their decisions. Moreover, it will not facilitate the formation of precedential legal positions or allow for collective claims.

[200] 47 U.S. Code § 230.
[201] Digital Services Act (n 20).
[202] Digital Services Act (n 20).

Hence, it is time to attempt to design a new system of dispute resolution considering the main features it should have and the main functions it should fulfil.

First and foremost, the system of dispute resolution concerning content moderation issues should be of a borderless nature, and its jurisdiction should cover disputes between users and platforms regardless of their place of residence or incorporation. Online platforms provide their services globally, and most of relationships between platforms and their users are of a cross-border nature where each party is subject to different jurisdictions. To let disputes between them be resolved on a regional basis means to split practices used by online platforms up and to create 'national' solutions to issues emerging globally.

Second, the system should reach out to all or at least most of the platforms which provide services concerning sharing of content among users. This means that the designed system should not be a creature of some or several platforms (like Meta Oversight Board). On the contrary, it should be the one whose jurisdiction covers most (ideally all) social media platforms and whose decisions are mandatory for them.

Third, this system should allow lodging claims not only individually, but also collectively. Individual claims, although helpful in protecting individual rights and interests in particular cases, are helpless in the face of mass violation of rights by applying the same practices. Meanwhile, for social media platforms the latter is more dangerous and topical. On the one hand, platforms can violate the right to free speech of the large groups of users by using unreliable algorithms or by the erroneous application of content moderation practices. On the other hand, platforms may enable mass offence of human honour and dignity because of the use of algorithms and practices being too liberal to fight against hate speech, disinformation and bullying.

Finally, the system should provide the timely resolution of claims lodged by users. Although disputes concerning content moderation issues are usually rather sophisticated and need scrupulous analysis of various circumstances to take a balanced decision, the process of resolution should not be too long. In the modern reality of informational overload and quick changes in the world (especially in the online world), resolutions on content moderation issues should be adopted without any delays. Thus, the structure of the system and the procedure underlying dispute resolution should ensure minimal time expenditures.

Considering these features and functions, one can design a basic model of this system. This is where a third element, self-regulation, turns out to be important.

To reach the purposes outlined in this section, ideally there should be a single judicial body whose authority is acknowledged by most of the platforms on a global scale. Thus, this body should be created by virtue of consensus between platforms based on their initiative. Just like national states when signing international treaties on the protection of human rights agree to limit their sovereignty and to create judicial bodies having supranational power to resolve disputes concerning human rights, platforms today should realize the need to create a judiciary having the authority to observe their compliance with human rights standards by virtue of resolving disputes initiated by their users. The judicial body should be granted the authority to resolve individual and collective disputes, require parties to the dispute to present certain evidence and explanations, and oblige the parties to exercise its decisions. To ensure its independence, the body should involve an equal number of representatives from each platform acknowledging its authority as well as from stakeholders, like organizations of human rights protection, consumer protection, academic institutions, etc. All the members should be high-level specialists in the area of human rights law, content moderation and contract law.

There may be several ways to create the judiciary and to provide it with authority.

First, platforms may, jointly with stakeholders, develop a kind of 'Code of practices' (or 'Code of procedures') where all the detailed rules concerning the formation of the

judiciary, procedural requirements, stages of dispute resolution and powers of the judiciary over parties to the dispute will be provided. In this case, the Code is a kind of agreement between platforms, and thus all the platforms that adhered to it are bound by the decisions of the judicial body and shall exercise them during the period mentioned in the Code or in the decision.

Second, the judicial body may be created at a self-regulatory organization involving most of the online platforms. Among all existing organizations of this kind nowadays GNI is the largest one. Although it makes many efforts to improve content moderation practices used by its members and to assess the compliance with human rights standards by platforms, it lacks a judiciary having the authority to resolve disputes between platforms and their users and to ensure compliance of platforms with the GNI Principles and the human rights standards in general. To create the judiciary and to facilitate its work, the organization should first develop the rules regulating issues of formation of this body and all the procedural issues. The legitimacy of the judicial body over the platforms will stem from the membership in the organization: platforms which have joined GNI and have adhered to its codes and principles should also be considered to have acknowledged the authority of the judicial body created at GNI and to exercise its decisions. If they fail to, the organization may develop a system of sanctions (including reputational ones) and apply them in each case where the obligation to exercise the judiciary's decisions is violated.

The enforceability of the judiciary's decisions may be strengthened by virtue of their recognition by nation-states. Here the model used in international commercial arbitration may be used. Despite the independence of arbitrators and arbitral bodies from state judicial systems worldwide, their awards are generally recognized by states which ratified the United Nations Convention on the Recognition and Enforcement of Foreign Arbitral Awards.[203] Decisions taken by the judicial body in disputes concerning content moderation may also be additionally enforced this way, however, for this purpose an international convention, like the New York one, shall be drafted and ratified by states. This purpose seems to be rather ambitious, however, it is not impossible.

State regulation: in search of ways to improve and harmonize

Although relationships on platforms are private by nature, private law remedies cannot be effective enough to cope with the risks brought about in modern realm especially in times of harsh social conflicts. State regulation proves to be more effective in this regard since it goes beyond particular private disputes and creates uniform approaches capable of addressing common issues, which is particularly important when the rights and interests of large social groups are at stake. Hence the role of state regulation should not be downplayed. However, to facilitate a comprehensive and secure functioning of state regulation, its main weaknesses should be eliminated or at least minimized. As mentioned in Part III, these are over-interference of states into the platform-users relationships and the creation of artificial state-dependent borders in addressing the issues being common globally.

To address these issues, we need first to answer the question of what state regulation should be like and how it should be improved. It is impossible to suggest concrete provisions or amendments which should find their place in the respective legal acts, but it is possible to outline the basic approach to develop or improve state regulation in whole. In this context, nothing seems fit better than the approach based on human rights law and the principle of the rule of law. Herein the recommendations expressed by David Kaye in his well-known report should be followed: states should 'repeal laws that criminalize

[203] United Nations Convention on the Recognition and Enforcement of Foreign Arbitral Awards (New York, 10 June 1958).

or unduly restricts expression', 'refrain from imposing disproportionate sanctions … on internet intermediaries', and 'avoid delegating responsibility to companies as adjudicators of content'.[204]

Besides these general recommendations, states should implement human rights standards in their laws in more concrete ways.

First, state regulation should be based on the principle of legality, which should be reflected in two aspects. On the one hand, this principle should be formative for the provisions regulating platform-user relationships: the state laws should require platforms to formulate their ToS in a clear and comprehensive manner, to include and to clearly express the provisions on any restrictions concerning the expressions online. Noticeably, Online Safety Bill[205] and DSA[206] already contain this requirement. On the other hand, the state should also follow the legality principle in its relations with platforms. In this regard, the state should avoid using in its laws vague and ambiguous terms and clearly outline the procedures and sanctions applicable to platforms in breach.

Second, state laws should reflect the standard of legitimacy, which again should find its place in the regulation of platform-user and platform-state relationships. Considering platform-user relationships, state laws should impose on platforms an obligation to apply the restrictions provided for by their ToS when there is a justification for their application, namely, when there is a legitimate aim. For platform-state relationships the state should limit its discretion to interfere with platform-user relationships to the cases when such interference is needed to protect human rights, public interests and other critically essential values. Moreover, legitimacy should be the primary limit for the state's discretion to impose sanctions on platforms or block access to platforms.

Thirdly, the state should implement the principle of proportionality (necessity). Regarding platform-user relationships, state laws should require platforms to apply specific remedies only if there are no softer ones or if the softer ones cannot ensure the needed balance or protection for the aggrieved parties. Considering platform-state relationships, this principle should guarantee that the sanctions applied by the state to the platforms are justified and that no other sanctions or remedies can restore the balance in the online environment and guarantee online and offline security.

Finally, state laws should contain procedural guarantees for users and platforms, namely, due process guarantees. In this respect, platforms should be encouraged to create mechanisms allowing their users to notify the platform about harmful content or a suspicious user or his/her activity, to lodge complaints on decisions taken by a platform, and to get an explanation from a platform on the reasons of deciding this or that way in a particular situation. Meanwhile, the states should also be obliged to guarantee due process for platforms. For this purpose, state laws should provide detailed and clear procedures allowing respective state bodies to interfere with platform-user relationships, initiate an investigation or apply sanctions to platforms.

Having outlined the approach which should form the basis for state regulation, it is now vital to answer the other question—how should it be ensured in every country and globally in the end?

Undoubtedly this issue is even more challenging since it involves too many factors. State regulation generally is the subject of the states' discretion. Thus, the question at hand reaches out to the state sovereignty and the will of particular states to limit it to a certain extent when drafting the provisions concerning information flow online. However, it does not mean that finding

[204] Report (n 120) recital 68.
[205] Online Safety Bill (n 21).
[206] Digital Services Act (n 20).

answers to it is impossible: for these purposes, international instruments may be referred to and used.

The first and the most obvious option in this regard is the development of an international treaty (convention), which should outline the basic obligations of states in promoting human rights standards in their laws on content moderation and platform-user relationships. This treaty may remind UN treaties on the protection of human rights[207] with the difference that its provisions should not only mention the rights and freedoms of some categories of persons, which the states should protect, but also outline specific obligations which the states should exercise when developing their inner legislation in the field. To ensure its effectiveness, the treaty should consist of two parts—substantive and procedural.[208] The treaty's substantive part may be formulated with the use of principles which the adhering parties undertake to follow while drafting their national laws. The principles should reflect the human rights standards in the way described above. Generally, they may remind the Manila Principles with the difference that they have an obliging effect on the adhering states. The procedural part should dwell on creating some institution responsible for ensuring compliance with its provisions or putting this responsibility on some institution with similar functions in some other field (eg the U.N. Human Rights Committee[209]). The institution may function as an arbiter having the authority to hear cases initiated by persons whose rights have been violated because of improper performance by the state of positive obligations imposed by the treaty. To make the institution's decisions more effective, they may be ensured with some sanctions (in particular, peculiar) applied to the states who failed to perform their obligations duly.[210]

The proposed option obviously requires serious political consensus, and its importance should be admitted by most states all over the world. Although this option may be very effective, it seems to be rather utopian at this stage.

Another option that may help promote common regulatory approaches in different countries is drafting a declaration outlining basic principles which states should follow when developing their inner legislation on content moderation. Just like the Universal Declaration of Human Rights, this declaration may only have a recommendatory character. However, considering the role of the Universal Declaration and its influence on the development of national legislation[211] worldwide, the suggested declaration may also become an important source of soft law.

Finally, the harmonizing effect may be reached by virtue of a model law serving as a pattern for state legislators. This option is mainly used by UNCITRAL to harmonize international commercial issues.[212] However, it does not seem inapplicable to the issues of content moderation and platform-user relationships. The main advantage of this option is that it allows not only to outline basic principles of state regulation of the issues at hand, but also to introduce a comprehensive legal act which has many chances to be taken as a pattern by states to avoid spending

[207] Like the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights.

[208] This structure is typical for international declarations and treaties in the field of human rights protection. See Aaron Tucek, 'The Missing Middle: Procedural Rights in the Human Rights System' (Human Rights, 25 February 2019) < https://human-rights.uchicago.edu/blog/2019/2/the-missing-middle-procedural-rights-in-the-human-rights-system-by-aaron-tucek> last time accessed 11 April 2023.

[209] Introduction to the Committee Human Rights Committee (OCHR) <https://www.ohchr.org/en/treaty-bodies/ccpr/introduction-committee> last time accessed 11 April 2023.

[210] The system close to the one provided for by European Convention on Human Rights.

[211] The Declaration was translated into more than 500 languages and has inspired the constitutions of many newly independent States and many new democracies (see Universal Declaration of Human Rights (United Nations) <https://www.un.org/en/global-issues/human-rights#:~:text=They%20include%20the%20Convention%20on,the%20Rights%20of%20the%20Child%20()> last accessed 19 September 2022.

[212] See UNCITRAL Model Law on International Commercial Arbitration (1985), with amendments as adopted in 2006; UNCITRAL Model Law on Electronic Transferable Records (2017); UNCITRAL Model Law on Electronic Signatures (2001) and many others.

extra time and putting excessive efforts to elaborate the one. However, its main disadvantage is that it remains the weakest option among others since it neither obliges states to implement it, nor constitutes an especially authoritative legal source as declarations do.

CONCLUSION

The war in Ukraine, which began in 2022, triggered the massive spread of disinformation and hate speech on various online platforms. Although many platforms have made many efforts to moderate dangerous content, these efforts were insufficient to combat it.

These problems have actualized a need to regulate platforms' practices in content moderation and their relationships with their users. Today there are four main instruments that regulate them: self- and co-regulation, contract law, human rights law and the laws of the particular states.

However, taken separately, these instruments have only limited success in combating harmful and illegal content. Meanwhile, their combination may help to find effective solutions.

On the one hand, there may be an effective combination of contract and human rights law. This combination helps to determine in which cases platforms violate the rights of their users and thus should bear a liability towards their users and in which cases they are not. The key point in this question is whether a platform acts in good faith when combating illegal and harmful content. Although the concept of good faith is a contract law one, human rights law can give precise criteria to provide a correct application of this concept. In particular, the standards of legality, legitimacy, proportionality, and due process may be an instruction when assessing whether the platform acted in good faith.

To provide the combination of contract and human rights law approaches with the reliable mechanism of enforcement, it may be strengthened with self-regulatory instruments. By signing a 'Code of practices' or creating a self-regulatory organization, platforms may create independent judiciaries which will be able to resolve individual and collective disputes between platforms and their users and make the decisions mandatory for platforms.

On the other hand, the human rights law approach should be combined with state regulation. The standards of legality, legitimacy, proportionality and due process should be embedded in the laws of particular states to ensure a balanced and comprehensive approach and to avoid different regulations of the same issues depending on the country and its political regime. To ensure this uniform solution in state laws, it is time to think about the development of international documents (international treaties, declarations, principles, or model laws), which will oblige states to embed human rights standards in their legislation on content moderation or at least serve as a pattern when drafting national laws in this field.

FUNDING DECLARATION

Funded by the European Union. This research is part of the Jean Monnet Center of Excellence project "European Fundamental Values in Digital Era", 101085385 – EFVDE – ERASMUS-JMO-2022-HEI-TCH-RSCH. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.