Transnational Legal Theory

ISSN: 2041-4005 (Print) 2041-4013 (Online) Journal homepage: www.tandfonline.com/journals/rtlt20

International human rights law in content moderation and the risks of 'misdiagnosing' its limits

Stefania Di Stefano

To cite this article: Stefania Di Stefano (2025) International human rights law in content moderation and the risks of 'misdiagnosing' its limits, Transnational Legal Theory, 16:4, 519-545, DOI: 10.1080/20414005.2025.2576408

To link to this article: https://doi.org/10.1080/20414005.2025.2576408

Published online: 27 Oct 2025.

TRANSNATIONAL LEGAL THEORY
2025, VOL. 16, NO. 4, 519–545
https://doi.org/10.1080/20414005.2025.2576408

Routledge
Taylor & Francis Group

International human rights law in content moderation and the risks of 'misdiagnosing' its limits

Stefania Di Stefano [a,b]

[a]International Law Department, Graduate Institute of International and Development Studies, Geneva, Switzerland; [b]Postdoctoral researcher, LISE (Cnam/CNRS), Paris, France

ABSTRACT

International human rights law (IHRL) has emerged as a dominant discursive framework for articulating and addressing issues raised by digital platforms. Despite its potential to offer a global language to articulate and address the questions raised by digital platforms, the 'IHRL project' has its detractors, who argue that this normative framework is inadequate to address the unique challenges that these new actors and technologies pose. Taking content moderation as a framework of analysis, this article critically engages with the criticisms aimed at IHRL in this sphere and questions whether these critiques are diagnosing an inadequacy of IHRL *in content moderation*. The article argues that the limits of IHRL that have been identified originate from and reflect a *traditional* approach to international law, and offers an alternative diagnosis: it argues that these 'limits' are in fact symptomatic of instances of *change* in international law.

ARTICLE HISTORY Received 17 November 2024; Accepted 31 July 2025

KEYWORDS International human rights law; content moderation; legal change; digital platforms; business and human rights

1. Introduction

International human rights law (IHRL) has recently emerged as a dominant discursive framework for articulating and addressing issues raised by technological developments and, in particular, digital platforms. A *language* originally confined to regulating the relationship between states and individuals, it is now also employed to shape and regulate relationships between individuals and non-state actors, including business enterprises. With the threats that digital platforms pose to human rights being nowadays publicly exposed and put under the spotlight at the international level, IHRL has since become a *lingua franca* in regulatory efforts at the international,

CONTACT Stefania Di Stefano stefania.distefano@graduateinstitute.ch, stefania.di-stefano@lecnam.net

regional and national level to constrain the power of these platforms. Likewise, as the human rights dimension of these issues becomes more pronounced, digital platforms are also increasingly using IHRL in their own operations (at least formally), either through committing to respecting IHRL by adopting corporate human rights policies[1] and/or by creating oversight mechanisms.[2] IHRL is thus employed by a wide variety of actors ranging from policymakers, regulators, civil society, academia, but also digital platforms themselves, to articulate a wide set of issues and challenges raised by these actors.

Yet, despite its potential to offer a *global* language to articulate and address the questions raised by digital platforms, the 'IHRL project'[3] has its detractors, who argue that the adoption of this normative framework was not 'inevitable'[4] and that it is inadequate to address the unique challenges that these new actors and technologies pose.[5] The main criticisms that are aimed at IHRL as a suitable framework for addressing these challenges revolve around themes that are not unique to digital platforms, and yet they are somehow presented, at times, as intrinsically confined to platform governance. These critiques include, for example, the fact that IHRL, a normative framework devised to constrain *governmental* power, is not easily translatable to constrain *corporate* power, that its language could prove to be a legitimising tool for platforms, that its inherent indeterminacy prevents it from offering any real solutions, or that the authoritative body of norms employed in governing platforms is not grounded in binding sources.

Taking content moderation as a framework of analysis, my aim in this article is to critically engage with the criticisms aimed at IHRL in this sphere. In particular, the article questions whether these criticisms are diagnosing a factual inadequacy of IHRL *in content moderation*. The article seeks to make two important contributions to the current debates around platform governance and the adequacy of the human rights framework in governing digital platforms.

[1] Miranda Sissons, 'Our Commitment to Human Rights' (Meta, 16 March 2021) <https://about.fb.com/news/2021/03/our-commitment-to-human-rights/> accessed 8 April 2023.

[2] 'Oversight Board | Independent Judgement. Transparency. Legitimacy'. <https://oversightboard.com/> accessed 9 September 2020. For a detailed account of how the Facebook Oversight Board was created, see Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129 *The Yale Law Journal* 2418.

[3] Brenda Dvoskin., 'Expert Governance of Online Speech' (2023) 64(1) *Harvard International Law Journal* 85.

[4] Evelyn Douek, 'The Meta Oversight Board and the Empty Promise of Legitimacy' (2024) 37(2) *Harvard Journal of Law & Technology* 373, 416.

[5] See for example Evelyn Douek, 'The Limits of International Law in Content Moderation' (2021) 6 1) *UC Irvine Journal of International, Transnational, and Comparative Law* 37; Brenda Dvoskin (n 3), Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: Operationalising a Human Rights-Based Approach to Content Moderation' (2020) 43 *Fordham International Law Journal* 939; Rachel Griffin, 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality' (2023) 2 *European Law Open* 30.

First, I argue that the criticisms aimed at the role and function of IHRL in the context of platform governance are oftentimes overblown and not *unique* to the application of IHRL to content moderation practices, and that they do not sufficiently take into consideration important dimensions of IHRL and its practices.

Second, I contend that these criticisms constitute a misdiagnosis of the limits of international human rights law in this area. This misdiagnosis originates from and reflects a *traditional* understanding of international law, and finds its rationale in the attempt to pigeonhole the phenomena examined in *predetermined* international law categories. I instead suggest that the difficulties linked to the ability to qualify these phenomena through traditional understandings of international law are in fact symptomatic not of the inadequacy of international human rights law to address these issues, but rather of instances of *change* in international law. These considerations are relevant not only for assessing the adequacy of IHRL in addressing the challenges posed by the governance of digital platforms, but, more broadly, for further understanding the evolving nature of human rights governance.

In my analysis, I will proceed as follows: after providing a brief overview of the historical trajectory of the increasing relevance of IHRL in content moderation, I will turn to the four main criticisms that have been identified as evidencing the limits of IHRL in this space to show how these critiques originate from a *traditional* approach to international law. These include (1) the difficulties linked with the application of IHRL to corporate actors; (2) the indeterminacy of IHRL; (3) the risks of co-optation of IHRL by digital platforms; (4) the question of IHRL sources. I conclude by suggesting that the social practices witnessed in the context of the application of IHRL to content moderation, while departing from a traditional understanding of international law, would be better framed and understood as instances of legal change in international law. My analysis in this article will be conducted against the practices of Meta, as it has formally issued a commitment to human rights in its Corporate Human Rights Policy, as well as of the Oversight Board, an oversight mechanism created by Meta which applies IHRL in the decisions it issues.

2. IHRL in content moderation: a brief history

International law is *traditionally* conceived as 'the body of law which will most usually govern the relations of states with each other',[6] and it generally

[6] Rosalyn Higgins, Problems and Process: International Law and How We Use It (Oxford University Press 1995) 39.

sees itself concerned with 'the rights and obligations of states'.[7] By analogy, IHRL, as a branch of international law, is also *traditionally* regarded as a system that considers states as the primary bearers of human rights obligations.[8] The traditional approach to IHRL is therefore founded on the presumption that human rights constitute a contract between the state and individuals,[9] and individuals must be protected from violations committed *by states.* Traditional international law narratives thus depict the state as the 'hero',[10] with the international law stage being accessible only to 'those personae that the legal system, through the medium of the personality doctrine, would allow to appear', while 'all the other entities […] may participate in the production and overall performance of the play [but] would not be granted the privilege of making an appearance on stage'.[11]

Nonetheless, in the last twenty years, academic scholarship has increasingly attempted to open the international law stage to non-state actors, focusing in particular on the relevance of human rights norms for assessing their conduct.[12] These debates have also crossed purely academic boundaries. A milestone in the trajectory for the increased recognition of the role of business enterprises in the human rights field has been the endorsement, by the UN Human Rights Council, of the UN Guiding Principles on Business and Human Rights (UNGPs),[13] which establish that business enterprises have a responsibility to respect human rights throughout their activities and operations. The relevance of the issue of corporate impact on the enjoyment of human rights is further signalled by the ongoing negotiations for an 'international legally binding instrument to regulate, in international human rights law, the activities of transnational corporations and other business enterprises'.[14]

[7] ibid.

[8] See, for example, Frédéric Mégret, 'Special character' in Daniel Moeckli, Sangeeta Shah and Sandesh Sivakumaran (eds), *International Human Rights Law* (Oxford University Press, 4th edn 2022) 90.

[9] Andrew Clapham, *Human Rights Obligations of Non-State Actors* (Oxford University Press 2006) 58.

[10] Cedric Ryngaert, 'Non-State Actors: Carving out Space in a State-Centred International Legal System' (2016) 63 *Netherlands International Law Review* 183.

[11] Andrea Bianchi, 'The Fight for Inclusion: Non-State Actors and International Law' in Ulrich Fastenrath and others (eds), *From Bilateralism to Community Interest* (1st edn, Oxford University Press 2011) 40.

[12] See, for example, Philip Alston (ed), *Non-State Actors and Human Rights* (Oxford University Press 2005); Anthea Roberts and Sandesh Sivakumaran, 'Lawmaking by Nonstate Actors: Engaging Armed Groups in the Creation of International Humanitarian Law' (2012) 37(1) *Yale Law Journal* 108; Math Noortmann, August Reinisch and Cedric Ryngaert (eds), *Non-State Actors in International Law* (Hart Publishing 2015); Katharine Fortin, *The Accountability of Armed Groups under Human Rights Law* (Oxford University Press 2017); John Gerard Ruggie, *Just Business: Multinational Corporations and Human Rights* (W W Norton & Company 2013); Surya Deva and David Bilchitz (eds), *Human Rights Obligations of Business: Beyond the Corporate Responsibility to Respect?* (Cambridge University Press 2013).

[13] Human Rights Council, 'Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework', A/HRC/17/31 (21 March 2011).

[14] OHCHR, Open-ended intergovernmental working group on transnational corporations and other business enterprises with respect to human rights, https://www.ohchr.org/en/hr-bodies/hrc/wg-trans-corp/igwg-on-tnc, accessed 15 June 2025.

Yet, despite the fact that the UNGPs were unanimously endorsed by the Human Rights Council already in 2011, technology companies, including social media platforms, have been particularly slow in their engagement with the business and human rights field and their corporate responsibility to respect human rights that this instrument establishes and internationally recognises. The reasons for such a slow engagement are to be found in the idea that technology companies were mostly perceived as constituting a tool for promoting human rights rather than a threat to their exercise: this is evidenced, for example, by the fact that UN documents from the early 2010s engaging with the issue focused most of their attention to states' obligations under international law and to governmental restrictions to the exercise of human rights online, with recommendations and guidance directed to requiring states to comply with IHRL when restricting the exercise of human rights in the online sphere.[15] Such an approach is hardly surprising: it is yet another manifestation of the traditional, state-centred approach of international law as a discipline.

Similarly, technology companies strived to present themselves as neutral actors who were facilitating the exercise of human rights, and freedom of expression particularly, and, at best, protecting them from governmental interference.[16] This self-perception is also evidenced by the fact that the Global Network Initiative, a multi-stakeholder initiative for information and technology companies, focuses on helping companies to 'respect freedom of expression and privacy rights when faced with *government* pressure to hand over user data, remove content, or restrict communications'.[17] As technology companies' own activities and operations were generally understood as human rights-friendly more than human rights-threatening, it is not surprising that regulators' attention to these issues was slow to pick up.

[15] For a comparative evolution of the right to freedom of expression at the international level and the Facebook Community Standards, see Konstantinos Stylianou, Nicolo Zingales and Stefania Di Stefano, 'Is Facebook Keeping up with International Standards on Freedom of Expression? A Time-Series Analysis 2005-2020' (11 February 2022) <https://papers.ssrn.com/abstract=4032703> accessed 8 April 2023.; see also Frank La Rue, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/HRC/17/27, 16 May 2011, para. 22; Maud de Boer-Buquicchio, 'Report of the UN Special Rapporteur on the Sale of Children, Child Prostitution and Child Pornography' (2014) A/69/262; Rita Izsák, 'Report of the UN Special Rapporteur on Minority Issues' (2015) A/HRC/28/64.

[16] See, for example, Tarleton Gillespie, 'Platforms Are Not Intermediaries' (2018) 2 *Georgetown Law Technology Review* 198; Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018); Robyn Caplan and Danah Boyd, 'Who Controls the Public Sphere in an Era of Algorithms: Case Studies' (Data & Society 2016); Robert Gorwa, 'What Is Platform Governance?' (2019) 22 *Information, Communication & Society* 854. Rikke Frank Jørgensen, 'Framing Human Rights: Exploring Storytelling within Internet Companies' (2018) 21 *Information, Communication & Society* 340.

[17] 'Our work' *Global Network Initiative* <https://globalnetworkinitiative.org/> accessed 8 April 2023.

It is only from 2016 that a change of course takes place: following the US elections and the Brexit referendum,[18] the Cambridge Analytica scandal,[19] the reports on the role of the Facebook platform in the genocide against the Rohingyas,[20] a domino effect is set off which exposes how technology companies, and social media specifically, shaped and negatively affected the exercise of their users' human rights online.[21] It is the techlash that followed these events, and the loss of users' trust in these platforms, that prompted an interest from policymakers and regulators in the matter[22] and, as a consequence, an interest from the companies themselves and their own turn to IHRL. These incidents, in fact, revealed the extent to which content moderation – a function that platforms had long disavowed[23] – could negatively impact the exercise of the right to freedom of expression and access to information online. The veil of neutrality had somehow shielded social media companies from the scrutiny of the business and human rights field, whose focus was mainly directed at other industries. If until that moment social media platforms had been seen under a positive light and as a tool to promote freedom of expression as exercised online, these scandals exposed their role as drivers for serious human rights violations.

The role of content moderation as the 'definitional part of what platforms do'[24] became evident, and with it the realisation that social media platforms had become 'the architecture for publishing new speech' and 'the architects of the institutional design that governs it'.[25] Importantly, it became clear that social media platforms enforced content moderation policies at a global level, but these standards offered significantly lower protection than the standards offered by international human rights law. As such, content moderation practices often resulted in either over-removal of legitimate content, as

[18] . Tarlach McGonagle. '"Fake news" False Fears or Real Concerns?' (2017) 35 *Netherlands Quarterly of Human Rights* 203. See also Steven Levy, *Facebook: The Inside Story* (Penguin Books, 2020).

[19] Carole Cadwalladr and Emma Graham-Harrison, 'Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach', The Guardian (2018), available at https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

[20] Human Rights Council, 'Report of the Independent International Fact-Finding Mission on Myanmar', A/HRC/39/64 (12 September 2018)

[21] See, for example, Rikke Frank Jørgensen, 'Introduction' in Rikke Frank Jørgensen (ed), *Human rights in the age of platforms* (The MIT Press 2019) xvii.

[22] The EU has recently adopted the Digital Services Act – 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) (Text with EEA relevance) 2022 (OJ L 277). The DSA creates obligations for online intermediaries, aiming at aim creating a safer digital space where the fundamental rights of users are protected. Crucially, it introduces due diligence obligations with respect to content moderation, and risk management. While the DSA is an important framework for the regulation of online content moderation and the protection of fundamental rights of users, its analysis is outside the scope of this article.

[23] Gillespie (n 16)

[24] ibid, 201.

[25] Kate Klonick, 'The New Governors: The People, Rules and Processes Governing Online Speech' (20180 131 *Harvard Law Review* 1598, 1603–4.

was the case, for instance, with the removal of the Napalm girl photograph,[26] or in under-removal of content that was contrary to international human rights law, as was the case for Myanmar and the Rohingya's genocide. The revelation that a few companies had become the main vehicles for exercising the right to freedom of expression while also applying policies that could interfere with the exercise of this right – and negatively impact a wide range of other human rights – resulted in a closer examination of digital platforms. From 2016 onwards, in fact, the number of documents and regulatory efforts aimed at taming the power of platforms sees a rapid and sudden increase.[27]

A key document that consecrated the role of IHRL in content moderation is, without doubt, the 2018 report of the UN Special Rapporteur on Freedom of Expression.[28] The report represents a milestone in the historical trajectory of the increasing relevance of IHRL in content moderation[29] since, exclusively devoted to this issue, it not only clearly confirmed that '[f]ew companies apply human rights principles in their operations, and most that do see them as limited to how they respond to government threats and demands',[30] but it additionally put pen to paper the fact that the UNGPs were a relevant instrument also for technology companies, as they 'establish "global standard[s] of expected conduct" that should apply throughout company operations and wherever they operate'.[31] Indeed, the UN Special Rapporteur did not invent the wheel by underscoring that the UNGPs, as the international authoritative framework outlining the responsibilities of businesses with respect to human rights, were also applicable to social media companies and their content moderation practices. Still, the epistemic authority of the mandate and the lack of a 'better alternative' (at the time at least) contributed to granting the report a milestone status.[32] Building on the UNGPs, the report provides guidance to social media companies on how to moderate content on their platforms in compliance with these standards.

[26] Sam Levin and Julia Carrie Wong Luke Harding in London, 'Facebook Backs down from "napalm Girl" Censorship and Reinstates Photo' *The Guardian* (9 September 2016). <https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo> accessed 30 August 2020.

[27] For an overview of the policy drivers, political dynamics, and institutional characteristics shaping platform regulation in different jurisdictions, see Robert Gorwa, *The Politics of Platform Regulation: How Governments Shape Online Content Moderation* (1st edn, Oxford University Press 2024).

[28] David Kaye, 'Report of the UN Special Rapporteur on Freedom of Opinion and Expression' A/HRC/38/35 (6 April 2018).

[29] Anna Sophia Tiedeke and Martin Fertmann, 'A Love Triangle? Mapping Interactions between International Human Rights Institutions, Meta and Its Oversight Board' (2023) 34 *European Journal of International Law* 907.

[30] Kaye (n 28), at 10.

[31] ibid.

[32] On the relevance of a lack of a better alternative, see, for instance, Marko Milanovic and Sandesh Sivakumaran, 'Assessing the Authority of the ICRC Customary IHL Study' (2022) 104 *International Review of the Red Cross* 1856, 1864; see also Fuad Zarbiyev, 'Cutting off the King's Head: Rethinking Authority in International Law' (2023) 14(3) *Journal of International Dispute Settlement* 285.

Because of the pressure from policymakers and regulators, social media companies started to commit to human rights. Meta, for example, created a Human Rights Policy team in 2019[33] and adopted a Corporate Human Rights Policy in 2021.[34] The company has also published a first Human Rights Report in 2022[35] and a second one in 2023.[36] Through its Corporate Human Rights Policy, Meta commits itself to respecting human rights in accordance with the UNGPs. The UNGPs require companies to respect, *at a minimum,* the International Bill of Human Rights, which includes the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights, as well as the International Labour Organization Declaration on Fundamental Principles and Rights at Work.[37] In the same policy, Meta commits to a supposedly non-exhaustive list of additional international human rights instruments, although it is mentioned that it relies on these instruments '*depending on the circumstances*', which are not further clarified.[38]

Additionally, Meta has created an oversight mechanism, the Oversight Board, which is empowered to (1) review content decisions taken by Meta, and which have been appealed by users and (2) provide policy guidance to Meta either following a case decision or upon a specific request from the company. In reviewing content policy decisions, the Oversight Board must 'interpret [Meta]'s Community Standards and other relevant policies (collectively referred to as 'content policies') in light of [Meta]'s articulated values'[39] as well as pay particular attention to the impact of removing content in light of human rights norms protecting free expression.[40] Despite the fact that, at least strictly speaking, the Charter does not empower the Board to apply IHRL directly in their decisions, but rather to rely on IHRL to inform the interpretation of Meta's Standards and Values, in all the decisions that the Board has delivered so far, it has not only directly applied IHRL,[41] but it

[33] Meta, 'Human Rights Report' (Meta 2022) 77 <https://about.fb.com/wp-content/uploads/2022/07/Meta_Human-Rights-Report-July-2022.pdf> accessed 24 July 2024.
[34] Sissons (n 1).
[35] Meta (n 33).
[36] Meta, 'Human Rights Report' (2023) https://humanrights.fb.com/wp-content/uploads/2023/09/2022-Meta-Human-Rights-Report.pdf.
[37] Sissons (n 1).
[38] ibid. The additional instruments mentioned in the Corporate Human Rights Policy include the International Convention on the Elimination of All Forms of Racial Discrimination, the Convention on the Elimination of All Forms of Discrimination Against Women, the Convention on the Rights of the Child, the Convention on the Rights of Persons with Disabilities, the Charter of Fundamental Rights of the European Union, the American Convention on Human Rights and the UN Declaration on Human Rights Defenders.
[39] 'Oversight Board Charter' (2023) <https://www.oversightboard.com/wp-content/uploads/2023/11/3427086457563794.pdf> art. 1§4.
[40] ibid art. 2§2.
[41] On the Oversight Board's role in applying IHRL to content moderation decisions, see Stefania Di Stefano, 'Translating and Developing International Human Rights Law in the Online Sphere: The Role of Meta's Oversight Board' (9 August 2024) <https://papers.ssrn.com/abstract=4920875> accessed 10 November 2024.

has also relied on a wide number of standards, ranging from human rights treaty provisions, but also general comments of Treaty Bodies, reports from UN Special Procedures, and even UN General Assembly and Human Rights Council resolutions.[42] The Oversight Board justifies the use of and reliance on IHRL on the UNGPs either as a consequence of Meta's own commitment to the framework, or irrespective of Meta's own commitment, since the UNGPs constitute, to date, the only authoritative guidance at the international level that the UN Human Rights Council has issued for states and business enterprises. The Board also heavily relies on the 2018 report of the UN Special Rapporteur on Freedom of Expression mentioned above, which, as discussed, draws recommendations with respect to the application of the UNGPs in the context of online content moderation.

The adoption of IHRL in content moderation was a slow process that was mostly prompted by the contingencies of the time. Despite the existence of a framework that outlined the human rights responsibilities of business enterprises, and the measures to be taken to comply with this responsibility, technology companies and social media platforms started to engage in this area only as a consequence of the techlash that hit them, and which prompted the action and intervention of regulators.

3. International human rights law in content moderation: (mis)diagnosing its limits

The UNGPs, which currently represent the only global framework for corporate responsibility and accountability for business-related human rights harms, have played a crucial role for elevating IHRL as the dominant discursive framework for addressing issues of online content moderation. However, with this increasing relevance of IHRL in this sphere, the academic debate has also started to question to what extent this is a useful framework for addressing the adverse human rights impacts of social media platforms. If, on the one hand, some scholars argue that IHRL can be an effective tool

[42] The list of standards referenced in the Oversight Board's decisions is available in the Oversight Board's Transparency Reports. The breakdown presented in this section relies on the data available in the reports. Oversight Board, 'Oversight Board transparency reports – Q4 2020, Q1 & Q2 2021' (October 2021) https://www.oversightboard.com/news/215139350722703-oversight-board-demands-more-transparency-from-facebook/ (last accessed 5 December 2023); Oversight Board, 'Oversight Board Q3 transparency report' (December 2021). https://www.oversightboard.com/news/640697330273796-oversight-board-publishes-transparency-report-for-third-quarter-of-2021/ (last accessed 5 December 2023); Oversight Board 'Oversight Board Q4 2021 transparency report' (June 2022) https://www.oversightboard.com/news/322324590080612-oversight-board-publishes-first-annual-report/ ; Oversight Board, 'Oversight Board Q1 2022 transparency report' (August 2022) https://www.oversightboard.com/news/572895201133203-oversight-board-publishes-transparency-report-for-first-quarter-of-2022/ (last accessed 5 December 2023); Oversight Board, 'Oversight Board Q2 2022 transparency report' (October 2022).

for constraining the power of these platforms[43] or that oversight mechanisms such as the Oversight Board can contribute to the development of human rights standards and their application in the online sphere,[44] on the other hand, other scholars are more sceptical and contend that IHRL is inadequate for addressing challenges raised by content moderation.[45] If some of the limits identified by these scholars are legitimate and underscore noteworthy issues in regulating platforms through this framework, it is unclear whether these limits that they identify are unique and confined to content moderation or, instead, belong to inherent disciplinary constraints within international human rights law. In particular, some of these criticisms are somehow caricatured and do sufficiently consider important dimensions of IHRL and its practices.

Four of the main criticisms that are aimed at IHRL and that I will be addressing in the following subsections revolve around the following issues: (1) the inherent inadequacy of IHRL to regulate corporate actors, which includes the claim that, even if IHRL applies to corporate actors by virtue of the UNGPs, its current application to content moderation issues relies on a misinterpretation of the UNGPs themselves; (2) the idea that IHRL is inherently indeterminate and, as such, cannot provide clear standards that would meaningfully address the human rights issues raised by content moderation; (3) the fact that IHRL is a language that is being co-opted by platforms to further legitimise their conduct; (4) the question of the IHRL sources that are being relied upon in the context of content moderation, which are either 'soft law' sources or, if they are 'hard law', they are not binding on corporations and, as a consequence, on platforms. As will be shown, all these criticisms share a common, underlying (and often silent) thread: they reflect, at a more fundamental level, a traditional understanding of international law and, in taking such an approach, they could be overlooking important symptoms of legal change.

3.1. Applying international human rights law to corporate actors: the role of the UNGPs

The application of and reliance on IHRL in content moderation finds its rationale in the UNGPs. Meta has committed to respecting this framework through its Corporate Human Rights Policy. The Oversight Board relies

[43] See, for example, Giovanni De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (1st edn, Cambridge University Press 2022); Kate Jones, 'Online Disinformation and Political Discourse: Applying a Human Rights Framework'. <https://www.chathamhouse.org/publication/online-disinformation-and-political-discourse-applying-human-rights-framework> accessed 3 January 2020.

[44] See, for example, Laurence R Helfer and Molly K Land, 'The Facebook Oversight Board's Human Rights Future' (2023) 44 *Cardozo Law Review* 2233.

[45] See, for example, Douek (n 4); Douek (n 5); Dvoskin (n 5); Sander (n 5); Griffin (n 5).

upon the UNGPs when applying IHRL in their decisions. The UNGPs set out the responsibility of business enterprises to respect human rights when conducting their business activities and operations, meaning that they 'should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved'.[46] They also make a vital distinction between the state *duty* to protect human rights, which emanates from states' international *legal* obligations, and the companies' *responsibility* to respect human rights, which stems from a *social* norm.[47] Indeed, the UNGPs are not a binding document, although, as also pointed out by the former UN Special Rapporteur on Freedom of Expression, 'the companies' overwhelming role in public life globally argues strongly for their adoption and implementation'.[48]

Nevertheless, it has been argued that IHRL is inadequate for addressing issues that arise in content moderation. Dvoskin, for instance, has argued that '[t]he IHRL project for content moderation interprets the UNGPs in a new fashion'.[49] She states that if '[u]nder its original meaning, social media companies are expected not to infringe on individuals' rights to freedom of expression (and any other rights) to the extent that international law already recognises those rights',[50] the new interpretation advanced by the IHRL project 'proposes that companies ensure a new right to freedom of expression on privately-owned social media companies'.[51] It is therefore contented that the interpretation of the UNGPs in the content moderation context 'collapses the substance of states' duties and businesses' responsibilities'[52] and that it 'asks from corporations what international law does not require states to protect'.[53]

Under the UNGPs, the corporate responsibility to respect human rights requires that business enterprises '(a) avoid causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur' and '(b) seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts'.[54] Content moderation being an essential part of social media's business model – having been defined as 'the

[46] Human Rights Council, 'Guiding Principles on Business and Human Rights – Implementing the United Nations "Protect, Respect and Remedy" Framework' HR/PUB/11/04 14, 13.
[47] John Gerard Ruggie, 'The Social Construction of the UN Guiding Principles on Business and Human Rights', *Research Handbook on Human Rights and Business* (Edward Elgar Publishing 2020) 75–76.
[48] Kaye (n 28), at 10.
[49] Dvoskin (n 5) 102.
[50] ibid.
[51] ibid.
[52] ibid.
[53] ibid 103.
[54] Human Rights Council (n 46) 14.

essence of platforms'[55] – it falls squarely within the activities through which the company should avoid causing or contributing to adverse human rights impacts. As such, if the content moderation policies a company adopts, or their implementation and enforcement, lead to adverse human rights impacts, it is within the company's responsibility to ensure that these adverse impacts are correctly identified and addressed. Importantly, an adverse human rights impact is defined as occurring 'when an action removes or reduces the ability of an individual to enjoy his or her human rights'.[56] It follows that limiting the corporate responsibility to respect, as defined by the UNGPs, to the idea 'that companies ought to take action regarding speech that international law prohibits' significantly misinterprets the normative content of the UNGPs. IHRL explicitly prohibits only a handful of speech categories, namely child pornography, direct and public incitement to genocide, advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, and incitement to terrorism.[57] However, this does not translate in the idea that any kind of speech that would not fall within these categories would be legitimate speech, since it could still cause or contribute to adverse human rights impacts. For instance, disinformation is not an explicitly prohibited content category in international law: on the contrary, international law protects expression even when false or misleading.[58] Yet, disinformation can have significant adverse human rights impacts, such as during elections or during public health crises. If companies were to interpret the UNGPs as merely requiring them to take action only with respect to content that is prohibited under international law, this would translate in their failure to comply with their responsibility to respect human rights.

The UNGPs also explicitly mention that the corporate responsibility to respect human rights 'refers to internationally recognised human rights – understood, at a minimum, as those expressed in the International Bill of Human Rights and the principles concerning fundamental rights set out in the International Labour Organization's Declaration on Fundamental Principles and Rights at Work'.[59] Provided that the UNGPs do not limit themselves to require companies to comply with international law only with respect to what international law prohibits, it is therefore unclear why it would be unusual or even contrary to the UNGPs to 'call on companies to look at the text of international treaties and the interpretations that the

[55] Gillespie (n 16) 201.
[56] OHCHR, 'The Corporate Responsibility to Respect Human Rights: An Interpretive Guide' (2012) HR/PUB/12/02 5.
[57] Frank La Rue, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression 2011 [A/66/290].
[58] Irene Khan, Disinformation and freedom of opinion and expression – Report of the UN Special Rapporteur on Freedom of Opinion and Expression 2021 [A/HRC/47/25] at 38.
[59] Human Rights Council (n 46) 13.

U.N. Human Rights Committee and U.N. Special Rapporteurs have made of them as a set of rules that companies should apply on their platforms'.[60] The text of international treaties is the starting point for defining the rights companies should respect, and the interpretations provided by human rights bodies further clarify the contours of those rights. Indeed, in Ruggie's vision, international treaties are to be considered by companies as an enumeration of rights that they should respect when conducting their business activities.[61] Even so, a company would still be required to engage in an interpretive exercise to devise a course of action that would respect these rights, and it is by looking at the text of these treaties and their authoritative interpretations that social media companies can identify the instances whereby speech, albeit not entirely prohibited, can be limited in a manner that is still protective of rights.

A related argument that has been advanced with respect to the impossibility of relying on international treaties and applying them directly to the activities of social media companies relates to the idea that the normative framework of these treaties is designed for states and therefore unsuitable for regulating corporate conduct.[62] This argument is not new in the business and human rights field.[63] Clapham refers to this objection as the 'legal impossibility argument', which is based on the idea that private non-state actors cannot incur responsibilities under international law.[64] He explains that '[p]artisans of this thesis point to a lack of evidence that international law accepts such a general development' and 'argue that treaties are negotiated and entered into by states and [they] cannot bind those who are not a party to them'.[65] Nonetheless, he contends that the language of human rights 'has generated meanings and significance beyond the realm of international legal obligations owed by states' and therefore 'excluding any obligations for non-state actors through appeals to the "definition", "essence", or "original sense" of the term "human rights" are unconvincing'.[66]

Proponents of the legal impossibility arguments often advance this objection with respect to corporate criminal liability: it is argued that international law cannot impose criminal liability on legal persons. The argument essentially relies on the idea that since international law treaties do not attach criminal responsibility to corporations, but only to individuals,[67] it is not possible for corporations to be liable for international crimes. However,

[60] Dvoskin (n 5) 102–3.
[61] Ruggie (n 47) 76.
[62] Douek (n 4) 44.
[63] See, for example, Clapham (n 9) 35–41.
[64] ibid 35.
[65] ibid 35–36.
[66] ibid 41.
[67] This is precisely the argument advanced by the majority in Kiobel v. Royal Dutch Petrol. Corp., 621 F.3d 111, 120 (2d Cir. 2010).

this exercise seeks to deduct substantive law, ie, that international law does not or cannot impose corporate criminal responsibility, from the constitutive documents of international criminal tribunals, whereas these treaties should be taken for what they are: documents 'expressly grant[ing] jurisdiction to try only natural persons, not legal persons, ie, corporations'.[68] For instance, several countries, as underscored by Kolieb, have incorporated the Rome Statute in their respective domestic system and have omitted the jurisdictional distinction between natural and legal persons, thereby prosecuting also corporations for the crimes enshrined in the Rome Statute.[69] Similarly, despite the fact that the International Criminal Tribunal for Rwanda did not have jurisdiction to prosecute legal persons, it still 'successfully prosecuted corporate leaders for utilising the resources of their corporations and their positions of authority to commit war crimes and genocide, as well as allowing their employees to engage in such crimes'.[70]

In a similar fashion, the argument made with respect to the impossibility to apply international human rights treaties to corporations relies on the fact that IHRL treaties are designed for states and apply only to those states that are parties to it; moreover, because those treaties do not attach any obligations to corporate entities, it is understood that international law does not impose any human rights obligations on these actors. However, as Clapham points out, 'a case can be made that, as a rule, human rights are to be respected by all persons, groups, and states, and that exceptional additional duties for the state have been explicitly articulated'.[71] The UNGPs can therefore be conceived as an articulation of additional responsibilities for companies. It would follow that, in principle, there is nothing that prevents actors other than states to rely on those treaties and even apply the rules contained therein. Indeed, the question as to whether these rules, as articulated in these treaties, are effective for addressing these issues is entirely legitimate, and it is not excluded that these rules might not answer *all* the questions that content moderation raises. But to say that IHRL does not answer all the questions that content moderation raises is significantly different than saying that IHRL is inherently inadequate for answering the questions that content moderation raises, especially because, as will be discussed below, IHRL does not offer clearcut answers, but a framework for articulating and devising answers to this sort of questions that respect human rights. As underscored by Douek, 'the key question is *how such rules differ in the context of a private company versus when they are being*

[68] Jonathan Kolieb, 'Through the Looking-Glass: Nuremberg's Confusing Legacy On Corporate Accountability Under International Law' (2015) 32 American University International Law Review 598.
[69] ibid 600.
[70] ibid; *Prosecutor v Nahimana* [2007] International Criminal Tribunal for Rwanda Case No. ICTR-99-52-A, Judgement, 2; *Prosecutor v Musema* [2000] International Criminal Tribunal for Rwanda Case No. ICTR-96-13-A, Judgement, Sentence 250.
[71] Clapham (n 9) 35.

applied to a state'.[72] This key question has not been overlooked: in the 2018 Report of the UN Special Rapporteur on Freedom of Expression, it has been underscored that social media platforms 'do not have the obligations of Governments, [but] their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression'.[73] Similarly, in the decisions delivered by the Oversight Board, it has been mentioned that one of the challenges facing the Board is precisely the identification of these differences and of 'whether Facebook's answers fall within the zone of what the UN Guiding Principles require'.[74] Perhaps these differences have not yet been outlined and developed sufficiently, but this does not necessarily translate in the fact that IHRL is an unsuitable framework to do so or that corporate duties cannot be further articulated. It also does not necessarily mean that, in some contexts, corporate duties will have to significantly differ from state duties.

A final aspect relating to the application of the UNGPs in this context that has been described as distinctive and deviating from the original meaning of the UNGPs pertains to the idea that, in order to comply with IHRL, a platform has to constantly balance the rights of its users against the rights of other users or society's interests.[75] It is not clear, however, why would this balancing exercise manifest itself only in the context of content moderation. Indeed, the UNGPs were drafted against a different backdrop of adverse human rights issues in mind.[76] Yet, content moderation is not the only context in which the content of corporate human rights responsibilities would be nebulous or would not involve balancing different interests and rights against each other. An example that would come to mind relates to business operations in the context of conflict-affected areas and, for instance, the conditions that would ensure a responsible exit when conducting business operations is deemed to be no longer viable. As underscored by the UN Working Group on Business and Human Rights,[77] the UNGPs themselves require that 'at all times, enterprises need to be aware of any risks that a particular course of action may pose to affected stakeholders and take these into account in their decisions'.[78] As such,

[72] Douek (n 4) 44.
[73] David Kaye, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression 2019 [A/74/486] para 41.
[74] See, for example, *Armenians in Azerbaijan* (2021) 2020-003-FB-UA (Oversight Board).
[75] Douek (n 5) 58.
[76] See, for example, Justine Nolan, 'Business and Human Rights in Context' in Dorothée Baumann-Pauly and Justine Nolan (eds), *Business and Human Rights* (Routledge 2016); Florian Wettstein, 'The History of Business and Human Rights and Its Relationship with Corporate Social Responsibility', *Research Handbook on Human Rights and Business* (Edward Elgar Publishing 2020).
[77] UN Working Group on the issue of human rights and transnational corporations and other business enterprises, 'Business, Human Rights and Conflict-Affected Regions: Towards Heightened Action' (2020) A/75/212.
[78] OHCHR (n 56) 79.

the idea of balancing rights is already taken into account in the framework presented by the UNGPs: it might well be that this exercise presents itself more often or in a different guise in the context of content moderation, but this issue is not *unique* to content moderation. Similarly, the UNGPs failing to provide specific guidance with respect to business conduct in conflict-affected areas,[79] it falls on the companies to determine which steps, which would be extremely context-dependent, would need to be taken while still complying with their responsibility to respect human rights.[80]

The criticisms that have been aimed at the current application of the UNGPs in the context of content moderation constitute a misdiagnosis of the issue. As has been shown, the interpretation of the UNGPs in the context of content moderation does not deviate from the intended scope of the instrument. The UNGPs, while differentiating between the state duty to protect and the corporate responsibility to respect, do not limit the responsibility to respect to a mere compliance with what international law prohibits: such a stance would fundamentally counter the concept of the corporate responsibility to respect, which encompasses both the responsibility to avoid causing or contributing to adverse human rights impacts through business activities as well as the responsibility to seek to prevent and mitigate adverse human rights impacts. Importantly, adverse human rights impacts are not limited to human rights violations, but they also include any action that *reduces* the ability of an individual to enjoy his or her human rights. Indeed, companies are not directly bound by IHRL treaties, and the rights contained therein are to be considered as an enumeration of the rights to be respected. Yet, relying on such a differentiation does not necessarily equate with the idea that corporate actors are prevented from relying on or even applying those rules. Finally, the balancing exercise required in content moderation is not alien to the UNGPs: on the contrary, the UNGPs themselves expects companies to engage in this exercise. By misinterpreting and misdiagnosing the limits of the UNGPs in content moderation, the risk is losing sight of the larger picture, which should prompt us not to immediately dismiss the UNGPs and the IHRL framework, but to investigate how the framework can be further improved to address its possible shortcomings.

[79] See, among others, Andreas Graf and Andrea Iff, 'Respecting Human Rights in Conflict Regions: How to Avoid the "Conflict Spiral"' (2017) 2 *Business and Human Rights Journal* 109; Jonathan Kolieb, 'Don't Forget the Geneva Conventions: Achieving Responsible Business Conduct in Conflict-Affected Areas through Adherence to International Humanitarian Law' (2020) 26 *Australian Journal of Human Rights* 142.

[80] See, for example, UN Working Group on the issue of human rights and transnational corporations and other business enterprises (n 77).

3.2. Indeterminacy and lack of precedent

Another criticism that is often aimed at IHRL in this area is the idea that the framework is too indeterminate to be able to offer any concrete answers to the challenges presented by content moderation[81] and that the lack of precedent addressing issues relating content moderation further exacerbates these challenges.[82] Moreover, it is argued that, due to the inherent indeterminacy of IHRL, the answers that the framework might offer with respect to content moderation might be different than those offered to governmental restrictions and that these answers, when different, are contrary to IHRL itself.[83]

The preoccupation of international lawyers with indeterminacy is not new.[84] That IHRL does not offer clearcut answers to content moderation questions is not something *unique* to content moderation. As has been argued by Venzke, interpretation makes international law:[85] as such, any development of normative standards that would be applicable to content moderation (or any other issue for that matter) will be the result of interpretive acts made by participants in international law making. What IHRL does is offer a language and a framework to articulate issues and design answers that would respect fundamental human rights.[86]

The criticisms that are aimed at IHRL as an indeterminate framework are hard to reconcile with the criticisms moved to the inadequacy of the UNGPs and their purported misinterpretation that supposedly conflates the state duty to protect with the corporate responsibility to respect. Dvoskin, for example, if on the one hand considers that the interpretation of the UNGPs deviates from the original meaning of the instrument by imposing the same standards to states and corporations, on the other hand claims that any standard or interpretation of IHRL that moves away from the

[81] Douek (n 5) 56–58.
[82] Douek (n 4) 42–46.
[83] Dvoskin (n 5) 119–22.
[84] The concept of indeterminacy in international law is mostly associated with the work of David Kennedy and Martti Koskenniemi, who have focused on the concept of structural indeterminacy. David Kennedy, *International Legal Structures* (Nomos 1987); Martti Koskenniemi, 'The Politics of International Law' [1990] *European Journal of International Law* 4; Martti Koskenniemi, *From Apology to Utopia: The Structure of International Legal Argument* (Cambridge University Press 2006).
[85] Ingo Venzke, *How Interpretation Makes International Law: On Semantic Change and Normative Twists* (1st edn, Oxford University Press 2012).
[86] 'Amidst growing debate about whether companies exercise a combination of intermediary and editorial functions, human rights law expresses a promise to users that they can rely on fundamental norms to protect their expression over and above what national law might curtail. Yet human rights law is not so inflexible or dogmatic that it requires companies to permit expression that would undermine the rights of others or the ability of States to protect legitimate national security or public order interests. Across a range of ills that may have more pronounced impact in digital space than they might offline — such as misogynist or homophobic harassment designed to silence women and sexual minorities, or incitement to violence of all sorts — human rights law would not deprive companies of tools. To the contrary, it would offer a globally recognized framework for designing those tools and a common vocabulary for explaining their nature, purpose and application to users and States' Kaye (n 73) at 43.

standards imposed on states deviates from IHRL itself. She attributes such deviation to the fact that IHRL is too indeterminate[87] and not to the fact that the deviation can be justified precisely by the idea that the corporate responsibility to respect might differ from the state duty to protect. The picture that emerges from her analysis is thus one depicting international law as a 'system [that] is considered as unitary and [where] problems are supposed to have *one* correct legal answer',[88] which, as will be discussed below, is one of the main features of a traditional approach to the discipline.

Indeed, it is true that IHRL does not offer many decisions on issues related to content moderation that would hold precedential value.[89] But this does not mean that new standards cannot be developed. The meaning of rights changes constantly in light of societal developments. In fact, there are several examples of IHRL evolving through time, with the meaning of human rights being expanded through interpretation and following societal changes, and these new challenges are constantly addressed by human rights tribunals.

The European Court of Human Rights, through the interpretation of the European Convention of Human Rights as a 'living instrument', has precisely done so. The most obvious example is the lowering of the threshold of conduct amounting to torture and inhuman and degrading treatment.[90] However, the Court has also expanded the protection of LGBT rights.[91] In one of the most discussed cases of 2024, *Klimaseniorinnen v Switzerland*, the Court has also determined that states have an obligation to mitigate

[87] See, for example, Dvoskin (n 5) 121: 'The Board had two options to avoid that tension. First, given the open texture of IHRL, the Board could have argued that IHRL allows prohibitions on blackface, at least in the online context. However, that would have meant committing to the view that states can also issue these regulations. What seems to animate the Board's reasoning is the belief that states and companies should actually govern speech differently. If that is the case, the Board also had a second option. It could have stated that IHRL was not the appropriate framework in this case because states and corporations are different. However, that would have meant losing the legitimizing force of IHRL. Ultimately, the Board adopted a third choice: states and corporations ought to regulate speech differently and both options are compatible with IHRL. In order to preserve the claim to objectivity, the Board framed this third option as determined by technical facts'.

[88] Andrea Bianchi, *International Law Theories: An Inquiry into Different Ways of Thinking* (Oxford University Press 2016) 21. Emphasis added.

[89] Douek (n 4) 42–46.

[90] See, for example, *Selmouni v France* [2000] 29 EHRR 403 [101]. 'The Court has previously examined cases in which it concluded that there had been treatment which could only be described as torture (see the Aksoy judgment cited above, p. 2279, § 64, and the Aydın judgment cited above, pp. 1891–2, §§ 83–84 and 86). However, having regard to the fact that the Convention is a "living instrument which must be interpreted in the light of present-day conditions" (see, among other authorities, the following judgments: Tyrer v. the United Kingdom, 25 April 1978, Series A no. 26, pp. 15–16, § 31; Soering cited above, p. 40, § 102; and Loizidou v. Turkey, 23 March 1995, Series A no. 310, pp. 26–27, § 71), the Court considers that certain acts which were classified in the past as "inhuman and degrading treatment" as opposed to "torture" could be classified differently in future. It takes the view that the increasingly high standard being required in the area of the protection of human rights and fundamental liberties correspondingly and inevitably requires greater firmness in assessing breaches of the fundamental values of democratic societies'.

[91] Laurence R Helfer and Erik Voeten, 'International Courts as Agents of Legal Change: Evidence from LGBT Rights in Europe' (2014) 68 *International Organization* 77.

climate change under Article 8 of the Convention, which protects the right to respect for private and family life.[92]

Other examples include the 2020 decision by Human Rights Committee that countries may not deport individuals who face climate change-induced conditions that violate the right to life:[93] the decision constitutes a landmark ruling precisely because it addressed, for the first time, negative human rights impacts stemming from climate change. Still with respect to environmental law, the recent recognition of the right to a clean, healthy, and sustainable environment[94] as a human right constitutes a significant development, although the contours of this right and what it actually entails are still to be properly defined.

What is referred to as indeterminacy is *the* feature that allows human rights to adapt to changes, be they societal, cultural, or technological. In the same way as the meaning of 'family life' or 'marriage' has been reinterpreted to recognise the right of LGBT couples to marry,[95] the meaning of the right to freedom of expression is evolving and being reinterpreted in light of technological changes and, in the case of digital platforms, as a consequence of the fact that, today, the right of freedom of expression is being exercised in a new configuration that has been aptly named 'the free speech triangle'.[96] Perhaps one of the issues underlying the 'indeterminacy' criticism hides another criticism, which is more aimed at questioning the legitimacy of the actors relying on IHRL to make this sort of decisions. Indeed, the examples I brought with respect to evolving interpretations of IHRL belong to institutions with a clear mandate from states. However, as I will further discuss below, restricting our field of analysis to formal mechanisms for legal change[97] might not allow us to capture in full the developments of IHRL in general, and of IHRL in content moderation in particular.

To sum up, the preoccupations with the inherent indeterminacy of IHRL are not confined to content moderation and do not constitute a unique 'limit' of this framework in this area. On the contrary, indeterminacy is a feature that allows IHRL to respond to new challenges, whose solution, by definition, would not necessarily draw from precedent decisions. Several examples from human rights tribunals demonstrate how IHRL and the framework it offers have been utilised to operate legal change, expanding the meaning of rights following societal developments. The arguments in support of the inadequacy of IHRL because of its

[92] *Verein Klimaseniorinnen Schweiz and Others v Switzerland* [2024] ECtHR [GC] 53600/20.
[93] *Teitiota v New Zealand* [2019] CCPR/C/127/D/2728/2016 (Human Rights Committee).
[94] Human Rights Council, The human right to a clean, healthy and sustainable environment 2021 A/HRC/RES/48/13.
[95] *Christine Goodwin v the United Kingdom* [2002] App No 28957/95 (European Court of Human Rights).
[96] Jack M Balkin, 'Free Speech Is a Triangle' (2018) 118 *Columbia Law Review*.
[97] Nico Krisch, 'The Dynamics of International Law Redux' (2021) 74 *Current Legal Problems* 269, 271.

inherent indeterminacy are also hard to reconcile with the arguments made in support of the inadequacy and misinterpretation of the UNGPs. The underlying element of these criticism perhaps finds its rationale not in the lack of clear rules addressing content moderation matters, but rather on the legitimacy of the actors who participate in these processes.

3.3. The question of legitimacy

The legitimacy criticism that is addressed to IHRL in content moderation has two dimensions: as mentioned, if on the one hand, critics question the legitimacy of actors beyond states to rely on IHRL and interpret this body of law, which is designed for states, on the other hand they also question the legitimacy dividends that these actors earn by committing to and relying on IHRL as a justification for their own decisions and actions.[98] In this section, I will focus on the second dimension of the legitimacy criticism: the risk of co-optation of IHRL by digital platforms.

The risk that digital platforms are co-opting IHRL 'to legitimize minor reforms at the expense of undertaking more structural or systemic changes to their moderation processes'[99] is very real and should not be underestimated. However, these concerns might be overblown or caricatured, and, similarly to other critiques made to IHRL, might be overlooking important dimensions of the human rights movement.[100]

Co-optation of IHRL is not an exclusive prerogative of digital platforms. Not only is this a widespread concern in business and human rights more generally,[101] but states also co-opt the language of IHRL: besides the fact that it should be recognised that a state's ratification of a human rights treaty does not necessarily translate into compliance, international human rights law is also used by states to legitimise human rights violations. An obvious example would be NATO's 'Operation Allied Force' against the Federal Republic of Yugoslavia, which, despite the lack of authorisation by the UN Security Council, was still depicted as consistent with international law precisely because of its aim to prevent massive breaches of human

[98] Douek (n 5) 56–58; 63–64. See also Griffin (n 5).
[99] Sander (n 5) 1005.
[100] Gráinne de Búrca, *Reframing Human Rights in a Turbulent Era* (Oxford University Press 2021) 4.
[101] Eg., the 2023 Corporate Human Rights Benchmark 2023 Report by the World Benchmarking Alliance, focusing on companies in the extractives and apparel sectors, shows how 'while most companies (70%) are making progress towards fulfilling their responsibility to respect human rights, the pace of improvement remains too slow to deliver the change that rightsholders so urgently need. There are still 47% of extractives and 62% of apparel companies that score below 20 out of 100 points – demonstrating that a large group of companies is not keeping up with stakeholder expectations on human rights'. World Benchmarking Alliance, '2023 Corporate Human Rights Benchmark Insights Report' (2023). <https://assets.worldbenchmarkingalliance.org/app/uploads/2024/03/2023_Corporate_Human_Rights_Benchmark__Insights_Report_13Mar2024.pdf> accessed 29 April 2024.

rights.[102] Other examples would include the deployment of IHRL by Israel as a tool to subjugate Palestinians and legitimise domination and dispossession,[103]or what UN Special Rapporteur on the situation of human rights in the Palestinian Territory occupied since 1967, Francesca Albanese, has named 'humanitarian camouflage', meaning the use of 'international humanitarian law terminology to justify its systematic use of lethal violence against Palestinian civilians as a group and the extensive destruction of life-sustaining infrastructures'.[104]

Yet, what seems to be somehow underestimated is the fact that adopting the language of IHRL is not an endeavour without consequences. Despite the fact that the international human rights system lacks legally binding enforcement mechanisms, which is generally considered as 'an impediment to state compliance',[105] research show that there is cause for optimism.[106] The commitment to human rights, be it via the ratification of a treaty or a formal corporate commitment via a policy, provides a yardstick against which to assess compliance. As demonstrated by Beth Simmons, the commitment by states to IHRL eventually translates into improved outcomes because it allows for the mobilisation of a variety of actors at the domestic level who can leverage those commitments. As also underscored by de Búrca, Simmons' work demonstrates that 'the response of domestic constituencies provides the crucial ingredient that enables international human rights law to operate to produce positive change'.[107]

Similarly, when a company commits to human rights, despite its voluntary nature, the commitment still sets up a clear bar against which to assess their conduct. Human rights can be mobilised by a variety of different actors, including states, international organisations, civil society, or academia. It is not enough to commit to human rights without taking action that follows through: while the steps that Meta has taken or other similar initiatives have been qualified as 'window-dressing',[108] these accounts do not take into consideration the interactions that the use of human rights language may generate. Meta's commitment to human rights or the decision to create the Oversight Board, even conceding that they might have

[102] See, eg., Hilary Charlesworth, 'International Law: A Discipline of Crisis' (2002) 65 *The Modern Law Review* 377.
[103] See Nicola Perugini and Neve Gordon, *The Human Right to Dominate* (Oxford University Press 2015).
[104] Francesca Albanese, Report of the Special Rapporteur on the situation of human rights in the Palestinian territories occupied since 1967 2024 [A/HRC/55/73].
[105] Andreas Johannes Ullmann, 'Compliance with UN Treaty Body Decisions: A Glass One-Third Full or Two-Thirds Empty?' (*OpenGlobalRights*, 5 September 2023). <https://www.openglobalrights.org/compliance-UN-treaty-body-decisions/> accessed 29 April 2024.
[106] ibid. See also Valentina Carraro, 'Promoting Compliance with Human Rights: The Performance of the United Nations' Universal Periodic Review and Treaty Bodies' (2019) 63 *International Studies Quarterly* 1079.
[107] Búrca (n 100) 18.
[108] Douek (n 5) 59.

amounted to window-dressing when the company took these steps, are translating in improved outcomes with respect to the company's responsibility to respect human rights. The company's commitment to human rights is being leveraged by the Oversight Board and other actors to hold the company accountable. The Oversight Board in particular did not turn out to be as toothless as initially predicted: in relying on a wide variety of human rights norms, the Board is effectively mobilising IHRL to promote change in the company's content moderation practices.

Moreover, voluntary commitments, which are usually defined as ineffective, can actually translate in binding standards for a company: the recent decision by the United Kingdom's Supreme Court in *Vedanta Resources PLC and another (Appellants) v Lungowe and others (Respondents)*[109] found that Vedanta's group-wide policies on environmental control and sustainability standards can give rise to the parent company's duty of care to third parties when 'the parent does not merely proclaim [the policies], but takes active steps, by training, supervision and enforcement, to see that they are implemented by relevant subsidiaries' and 'if, in published materials, it holds itself out as exercising that degree of supervision and control of its subsidiaries, even if it does not in fact do so'.[110] Importantly, the Supreme Court stated that '[i]n such circumstances its very omission may constitute the abdication of a responsibility which it has publicly undertaken'.[111]

While the risk of co-optation exists, it still is a risk that is not confined to the application of IHRL to content moderation issues, but reflects, once again, inherent disciplinary constraints that manifest themselves in other areas and with respect to other actors, including states. Nonetheless, the reliance on the language of IHRL is not as meaningless as it is portrayed by commentators who are sceptical of the human rights project. Moreover, as research shows, monitoring mechanisms or tribunals still have some munitions to stimulate compliance. A question that could be asked is how to stimulate corporate compliance, including digital platforms' compliance.

3.4. The question of legality

A final criticism that is aimed at IHRL relates to the use of sources in the context of content moderation. This criticism too has two dimensions: if on the one hand the reliance on IHRL treaties is contested because these instruments do not directly bind corporations, on the other hand the reliance on other standards emanating from human rights bodies is contested because these documents are not binding but constitute only persuasive

[109] *Vedanta Resources PLC and another (Appellants) v Lungowe and others (Respondents)* [2019] UKSC 20.
[110] ibid 53.
[111] ibid.

authority.[112] What this criticism reveals is the longstanding concern with the 'bindingness' of law: in one case, although the standards are considered 'hard law', they do not bind the relevant actor; in the other case, despite the standards being relevant for non-state actors, or even tailored to them,[113] their legitimacy and relevance are questioned due to falling within the realm of 'soft law'. I have already addressed the first dimension of this criticism with respect to the use and direct application of IHRL treaties. In this subsection, I wish to take a step back and address the wider question that this criticism ultimately conceals, which is a question of legality or normativity.

Taking an approach that seeks to neatly divide law in the binary binding/non-binding distinction is symptomatic of a *traditional* approach to international law: as underscored by Bianchi, '[o]ne of the most noticeable features of the traditional approaches is their tendency to draw boundaries', where '[t]he main dividing line or boundary […] is the one drawn between "law" and "non-law"'.[114] Indeed, a traditional approach to international law ultimately also underlies the other critiques that I have addressed in the previous sections: the issue of boundaries (between state and non-state actors, binding and non-binding law, law and non-law) characterises the foundations of each critique.

However, such an understanding of law, ie, 'the idea that international law is a system of objective principles and neutral rules that emanate from States' will, either directly through treaty or indirectly through custom, and the notion that States are the primary subjects of the international legal order'[115] does not reflect law as we see it in action today. As underscored by Alston already in 2005, 'the result of recent developments has been to highlight and/or expand the *de facto* roles played by non-state actors in national and international affairs'.[116] Similarly, Krisch notes how 'in recent years, this neat, unitary picture of law has come under pressure from denationalisation, privatisation and globalisation', adding that 'multiplicity has increasingly become accepted as a condition of law'.[117] He underscores that, in many different contexts, including human rights, 'an account of law that focuses solely on one legal system, be it national, subnational or

[112] Douek (n 4) 43. See also Andreas Kulick, 'Meta's Oversight Board and Beyond – Corporations as Interpreters and Adjudicators of International Human Rights' (2022) 22 The *Law & Practice of International Courts and Tribunals* 161.

[113] Eg., UN Special Procedures are tasked with making recommendations to states generally but also to the UN or other actors within the 'international community', and enjoy more flexibility in defining the scope of their mandate, thus allowing them to be more responsive to emerging issues and to easily enlarge their audience, which could include, as relevant, also corporate actors.

[114] Bianchi (n 88) 24.

[115] ibid 21.

[116] Philip Alston, 'The "Not-a-Cat" Syndrome: Can the International Human Rights Regime Accommodate Non-State Actors?' in Philip Alston (ed), *Non-state actors and human rights* (Oxford University Press 2005) 19.

[117] Nico Krisch, 'Entangled Legalities in the Postnational Space' (2022) 20 *International Journal of Constitutional Law* 476, 477.

international, would today appear as utterly deficient' and that 'such an account would not properly reflect the rules that govern it – the rules that matter for actors in the field'.[118] He also adds that not only, in these contexts, 'the different layers of law have not amalgamated into one legal order', but that '[t]heir relations are often not fully defined, they do not have a common source of validity, and they are often created, monitored and practiced by different institutions and actors'.[119]

A *traditional* approach to international law, drawing neat boundaries, aimed at including and excluding actors or norms in the international law realm, does not allow for a full capture of the phenomena that are being observed. With respect to legality, the adoption of a binary understanding of law prevents a thorough account of the 'practices of all kinds of actors related to law and legal norms', since '[i]n many normative orders inside and outside the state, actors other than judges – regulators, informal dispute settlers, addressees – play a large role'.[120] By understanding legality as a matter of degree rather than as a binary,[121] it is possible to 'take seriously the ways in which this broader range of actors construe law and legal relations through their social practices'.[122]

The criticism on the use of sources in content moderation ultimately conceals a longstanding preoccupation in international law more generally that stems from 'the way in which international law is *traditionally* thought about and taught',[123] but which does not necessarily reflect what international law *is*. If international law is conceived as a social construct that responds to (and, in my opinion, has to respond to) a set of particular urgencies to address impending human needs,[124] and 'not a set of neutral rules, elaborated independently of context and historical background, [but where] the human condition remains central',[125] the concerns about boundaries must be overcome to depict and analyse the social practices and legal relations being witnessed.

4. An alternative diagnosis: instances of legal change?

Some the criticisms that have been aimed at IHRL in the context of content moderation are not symptomatic of *unique* limits of this regulatory framework *in content moderation*. The limits of IHRL that have been identified in content moderation, in fact, do not specifically arise in this particular operational context, but they either reflect a misinterpretation of the

[118] ibid.
[119] ibid.
[120] ibid 488–9.
[121] ibid 489.
[122] ibid.
[123] Bianchi (n 88) 21.
[124] ibid 303.
[125] ibid 310.

regulatory framework, or they caricature some issues while removing them from the larger dimensions of IHRL practices. More generally, however, they also originate from and reflect a *traditional* approach to international law: they diagnose, at a more fundamental level, deviations from international law as *traditionally* understood.

The misdiagnosis of the limits of IHRL in content moderation therefore finds its rationale in the attempt to pigeonhole these phenomena in predetermined international law categories. The criticism relating to the application of IHRL to corporate actors, for instance, emanates from the traditional understanding that states are the primary subjects of international law. Yet, drawing such a neat boundary between state and non-state actors neglects the fact that '[t]he traditional doctrine of subjects is no longer able to offer a satisfactory explanation of the complexities of international legal relations'.[126] It is precisely this acknowledgment that has prompted a growing attention towards the relevance of human rights norms for assessing the conduct of non-state actors. In framing the use of human rights norms by corporate actors through a traditional lens, their interaction with these norms is instead understood and depicted as a *deviation* from international law. Such an approach also misinterprets the normative content of the UNGPs, since an interpretation of this framework as merely requiring companies to take action only with respect to content that international law prohibits would result in their abdication of their corporate human rights responsibilities. If indeed corporate responsibilities may differ from state duties, this does not equate with the fact that, in some contexts, they may not also overlap: the fact that the UNGPs operate a distinction between these actors does not necessarily mean that the substantive measures to be implemented by one or the other *must* be different.

Similarly, the criticisms relating to the issues of indeterminacy and legitimacy still reflect a concern about the participants in these processes. In the context of indeterminacy, the criticism seems more directed to the legitimacy of the actors interpreting and developing IHRL rather than the lack of precedent for content moderation challenges. Such an approach reveals, once again, a traditional understanding of the doctrine of subjects, recognising only states (or state-mandated institutions) as legitimate actors for the development and application of IHRL. In the same vein, the issue of legality is rooted in the longstanding concern with the 'bindingness' of the law. However, a binary understanding of legality hides a variety of social practices through which actors other than states construe law and legal relations.[127]

[126] Bianchi (n 11) 41.
[127] Krisch (n 117) 489.

There are significant risks in misdiagnosing the limits of IHRL in content moderation: the difficulties linked to the ability to qualify these phenomena through traditional understandings of international law can in fact be symptomatic not of the inadequacy of IHRL to address these issues, but rather of instances of *change* in international law.

As underscored by Krisch, '[i]n many instances, international law is changing much more flexibly than the traditional picture leads us to believe – states are not always at the centre, legality is not necessarily treated as binary, and the ways in which international law changes also vary significantly across issue areas and institutional contexts'.[128] Change should not be equated with 'the replacement of one norm with another', but 'there are different degrees of change, some more limited, others more radical, and [...] no clear line can be drawn between a mere change in interpretation and the appearance of a new rule'.[129] According to Krisch and Yildiz, not only '[c]hange occurs when, at a second point in time, the scope of possible interpretations or the weight of particular positions in legal discourse has shifted' but it 'may or may not correspond with doctrinal reconstructions of what the law is – at times, the law practised by actors in a given field will differ from what [...] an application of the doctrinal requirements for new customary law would result in'.[130]

An alternative diagnosis of the symptoms identified with respect to the application of IHRL in content moderation could instead indicate that the deviations from the traditional understanding of international law constitute the seeds of legal change: they could be a manifestation of 'a paradigm shift in our understanding of the power and utility of human rights',[131] an attempt to 'turn human rights on their heads and [the] realiz[ation] that while they have protected private power, they also contain the seeds for action against private power'.[132]

If this diagnosis is correct, then the treatment to be adopted is not an entire dismissal of IHRL as a framework to address issues raised by content moderation, but rather a deeper investigation and assessment of these manifestations of change and their processes.

Acknowledgements

This article was initially presented at the workshop 'The Promise and Perils of Human Rights for Governing Digital Platforms', hosted by Leiden University from

[128] Krisch (n 97) 271.
[129] Nico Krisch and Ezgi Yildiz, 'The Many Paths of Change in International Law: A Frame' in Nico Krisch and Ezgi Yildiz (eds), The Many Paths of Change in International Law (Oxford University Press 2023) 10.
[130] ibid.
[131] Clapham (n 9) 56.
[132] ibid.

18 to 19 January 2024. I am grateful to the workshop organisers, Rachel Griffin, Barrie Sander, Henning Lahmann, Matthew Canfield, Jelena Belic, for their generous engagement with the article and for their feedback. I am also grateful to Andrew Clapham for his precious feedback and comments on earlier drafts of the article, as well as to Molly K. Land for her thoughtful comments. Lastly, I thank the two anonymous reviewers for their thorough engagement with this work and their detailed feedback and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID


Stefania Di Stefano http://orcid.org/0009-0003-2304-388X