Received: 6 February 2023 | Accepted: 26 March 2024

DOI: 10.1002/poi3.391

RESEARCH ARTICLE

P&I Policy & Internet PSO WILEY

Can Facebook's community standards keep up with legal certainty? Content moderation governance under the pressure of the Digital Services Act

Mathieu Fasel | **Sophie Weerts**

Swiss Graduate School of Public Administration, University of Lausanne, Lausanne, Switzerland

Correspondence

Mathieu Fasel, Swiss Graduate School of Public Administration, University of Lausanne, Lausanne, Switzerland.
Email: mathieu.fasel@unil.ch

Abstract

Content moderation by social media companies is a challenge for regulators around the world. The European Union is trying to tackle this challenge with its Digital Services Act (DSA). Notably, Article 14 DSA aims to impose language requirements based on the principle of legal certainty to social media companies' terms and conditions, of which community standards (CS) are a part. The principle of legal certainty is one of the building blocks of international human rights law, and its inclusion in the DSA illustrates the human rights-based approach anchored in this European regulation. Based on a content analysis of Facebook's CS, this paper shows that their standards do not meet European requirements for legal certainty and argues that important changes in content moderation governance would be needed for proper compliance. Such changes could generate a domino effect beyond the European Union and on Facebook's content moderation governance itself. At the same time, the DSA could also generate a boomerang effect on the legal certainty principle as such. From this perspective, the paper contributes to the literature on regulatory governance studies, and on international human rights law in social media studies from a European Union perspective.

KEYWORDS

community standards, content moderation, Digital Services Act, human rights, legal certainty

INTRODUCTION

Social media have an impact on democracy, the protection of minorities and freedom of information and expression (Jørgensen & Pedersen, 2017; Kaye, 2019a; York & Zuckerman, 2019). Events such as the storming of the US Capitol or the persecution of the Rohingyas in Myanmar have been relayed on and exacerbated by social media (Amnesty International, 2022; Zakrzewski et al., 2023). Since the 2010s, public authorities and civil society have called for developing comprehensive regulatory solutions anchored in a human rights-based approach (HR-based approach) to limit the adverse effects on individuals and society without taking away the advantages of such communication networks for dissident voices (Article 19, 2023; Council of Europe, 2021; Kaye, 2019b; La Rue, 2011).

Social media companies have developed content moderation to try to mitigate harm on their platforms (Gillespie, 2018). Content moderation can be defined as 'the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse' (Grimmelmann, 2015, p. 47). From a practical perspective, content moderation occurs through a mixture of algorithmic systems (Gorwa et al., 2020) and human supervision by 'moderators' (Roberts, 2016; York, 2022). The decisions at the core of this process are formalized in written documents usually called community standards (CS) or community guidelines, which form part of social media companies' terms and conditions (Suzor, 2019). Such governance is embedded in the internet's self-regulatory and free market logic that was favoured in the 1990s for online intermediaries (Cohen, 2019). Content moderation aims to guarantee a friendly atmosphere for users, but at the same time serves the business model of companies based on advertisement (Zuboff, 2019). Content moderation documentation contributes to the users' information, echoing the 'notice' logic already applied to users' data (Cate & Mayer-Schonberger, 2013).

With the adoption of the Digital Service Act (DSA) in 2022, the European Union has strengthened its legal framework for social media companies. Among its provisions, Article 14 DSA states that the terms and conditions of companies must be drafted in 'clear, plain, intelligible, user-friendly and unambiguous language'. These legal requirements express the principle of legal certainty, which requires legal texts to be 'clear, foreseeable, coherent, determinate, and predictable' (Ranchordas, 2021, p. 19). With these drafting requirements, the legislator is attempting to respond to the common criticism levelled at companies' practices in terms of the transparency of their general terms and conditions, which play a significant role in shaping users' consent to data processing and moderation decisions (Selbst & Barocas, 2018).

The legal certainty principle embedded in Article 14 DSA is not only a legal principle to be respected in legal relationships between private parties—for example, between a company and its clients (Weerts, 2019)—, it is also a standard from international human rights law (IHRL). The purpose of such a requirement is to enable recipients of a law to determine what behaviour is expected of them (Popelier, 2008). The principle requires human rights (HR) restrictions 'to be formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly' (United Nations, General comment no. 34, 2011, p. 6). Since content moderation has been considered as a practice potentially restricting freedom of expression (Callamard, 2019), the drafting requirements applying to CS, as part of the social media companies' terms and conditions, are then also likely to be analysed through the principle of legal certainty as understood in IHRL.

From a regulatory perspective, the entry into force in 2024 of the DSA does not only initiate a shift from a self-regulatory approach to a co-regulatory approach (Finck, 2018), but it also anchors a HR-based approach in the governance of social media companies. In this way, Article 14 DSA is particularly interesting because it can be analysed from a dual

perspective: on the one hand, that of a strengthening of transparency requirements to ensure users' informed consent with companies' terms and conditions; on the other hand, that of a requirement to ensure that social media companies adopt practices that comply with IHRL. This paper focuses on the second issue by asking the following question: To what extent are CS currently in force at Facebook compliant with the principle of legal certainty as defined by IHRL, and what impact could the implementation of this principle have on CS? The article highlights the current shortcomings of CS' language regarding the principle of legal certainty. It argues that compliance with the legal certainty requirements enshrined in Article 14 is not just a matter of formal drafting, and that, given the ontology of CS and their role in the architecture of content moderation, compliance from an IHRL perspective will not be easy to implement. If successful, the implementation of these requirements could produce a domino effect on the governance of content moderation and a boomerang effect on the very principle of legal certainty.

This paper contributes to several literatures. It contributes to the discussion about the implementation of HR-based approaches from a regulatory governance studies perspective (Donald & Speck, 2020; Murray & Long, 2022). From the perspective of IHRL in social media studies (Dias Oliva, 2020; Jørgensen, 2019; Kaye, 2019a), it enriches the conversation by showing the difficulty of applying an IHRL principle to a digital artefact. The paper can also help practitioners identify challenges of implementation concerning Article 14 DSA.

The following section provides background on content moderation and the HR-based approach in the governance of content moderation. The paper then focuses on the DSA, emphasising Article 14 and its relationship with the legal certainty principle. After description of the methodology, the findings of the content analysis are presented, and the paper discusses the effects of the DSA implementation on the content moderation governance of Facebook. In conclusion, beyond the issue of the effective implementation of the DSA, this paper raises the question of the relationship between law and technology, in light of the fact that the content of the principle of legal certainty may ultimately have to evolve to play a role in the digital world.

COMMUNITY STANDARDS AND THE HUMAN RIGHTS-BASED APPROACH FOR REGULATING CONTENT MODERATION

Community standards in the governance of content moderation

Social media companies offer a service distributing user-generated content among third parties, while actively moderating the content distributed by these users. Users publish all kinds of content on the platforms, which can be unlawful or harmful to individuals or society without being unlawful as such (Macdonald & Vaughan, 2023). Although companies are in principle immune from liability for user-generated content (Husovec, 2020), they still choose to filter it to ensure a pleasant user experience. Therefore, companies carry out content moderation, which ensures that the digital sphere stays safe from content that contravenes companies' values or violates legal obligations (Gillespie, 2018). It also contributes to users spending more time on the platforms, generating more revenue for the companies (Zuboff, 2019).

Content moderation is a complex process that implies different actors. In the early days of social media, content moderation was operated through decisions taken at company executives' discretion (Klonick, 2018; Suzor, 2019). Due to the massive increase in the amount of content to be processed, content moderation now mainly relies on the use of automated systems of different types. For example, Gorwa et al. (2020) report that social

media companies use machine-learning approaches based on natural language processing. Such approaches 'generally involve training language classifiers on a large corpus of texts, which have been manually annotated by human reviewers according to some operationalisation of a concept like offence, abuse, or hate speech' (Gorwa et al., 2020, p. 5). The general principles under which moderation operates are decided by company managers who collaborate with various stakeholders (Kettemann & Schulz, 2020). In addition, moderators are hired (or subcontracted) by the companies to provide human oversight and help settle difficult cases (Roberts, 2016). CS emerged from this complex process and interplay of actors, playing the role of the set of internal rules that guide the work of the moderators (York, 2022, pp. 16–20).

As a reaction to critics of opacity regarding content moderation, social media companies have started to make their CS public (Gillespie, 2018). For example, Facebook made their CS public in April 2018, stating that they were 'publishing the internal guidelines used to enforce their standards' (Bickert, 2018). Regarding their content, Klonick notes that CS have become more detailed over time, going from a 'one page of internal 'rules' applied globally with a list of things (moderators) should delete, (…) like Hitler and naked people' (Klonick, 2018, p. 1631), to the development of a more 'intricate system of rules' (p. 1635). Gillespie argues that CS 'constitute a gesture: to users, that the platform will honour and protect online speech and at the same time shield them from offence and abuse; to advertisers, that the platform is an environment friendly to their commercial appeals; and to lawmakers, to assure them of the platform's diligence, such that no further regulation is necessary' (Gillespie, 2018, p. 47). The author underlines that CS must be seen as 'discursive performances', a form of 'statement of principles' that has been rendered public to appease transparency requirements (Gillespie, 2018, p. 45). From a legal point of view however, CS play an additional role to that of guide for moderators, because they are part of companies' terms and conditions. Indeed, when users register on a social media platform, the company states that they must agree to their terms and conditions: CS are thus at the centre of a legal contract between the company and the user (York, 2022, pp. 16–17).

The HR-based approach for regulating content moderation

Content moderation decisions, taken based on CS, illustrate the significant power social media companies hold over online speech. In their early days, the private governance of the Internet and the birth of social media platforms were seen as a promise of an Internet free from state oppression, enabling the advent of a freer and more democratic society (Bietti, 2023; Cohen, 2019). Developments in the last decade have led to the realisation that reality was far from this promise: not only were social media platforms used to convey statements that undermined fundamental values, such as the prohibition on discrimination and hate speech (Castano-Pulgarin et al., 2021); but at the same time, social media companies had become so powerful that the decisions they took through content moderation had the power to reshape social and political discourses around the globe (Taylor, 2021). Such a situation led observers to compare their power to that of the states from which they were supposed to 'free' society, for example 'Facebookistan' (Chander, 2012).

This dual problem has led international organisations and civil society to approach the issue from the angle of IHRL and to propose solutions based on its principles. In particular, David Kaye, former Special Rapporteur to the United Nations for freedom of information and expression, has published several reports stressing how content moderation by social media companies can undermine the freedom of expression of users of their services (Kaye, 2016, 2018, 2019b). Kaye consequently argues that social media companies should

incorporate IHRL standards into their decision-making processes (Kaye, 2019a). The essence of this proposal has had an echo in the work of the Council of Europe (2021).

The idea of anchoring a HR-based approach in the private governance of social media companies is also supported by legal academic literature. The argument is based on the understanding that when companies remove user content, they impact users' ability to express themselves to prevent them from posting harmful or illegal statements. In this context, decisions that social media companies make on users' posts 'have significant effects on (their) ability to generate and share information and expression' (Land, 2019, p. 289). As such, 'powerful companies like Facebook and Google can influence human rights in ways traditionally reserved for governments' (Jørgensen, 2019, p. 163). Callamard argues that 'internet intermediaries can violate human rights', detailing that because it interferes with users' speech, content moderation is at times painted as 'censorship' (Callamard, 2019, pp. 206–207). This opinion is shared by York and Zuckerman who underline that 'how platforms like Facebook control speech critically affects the boundaries for freedom of expression' (York & Zuckerman, 2019, p.137). In this context, scholars designate social media companies as 'Human Rights Arbiters' (Jørgensen & Pedersen, 2017) or the 'New Governors' of online speech (Klonick, 2018). Based on this observation, some scholars have called for a general HR deployment in content moderation governance through the concept of 'digital constitutionalism', calling for social media companies to be subjected to the logic of the rule of law in the same way as states (Celeste, 2019; De Gregorio, 2022; Pollicino, 2021).

Considering that state authorities are the primary subjects of HR obligations at the international and national levels (Donnelly, 2013), guaranteeing that private companies also apply HR has required some changes in the field of IHRL. As the violation of human rights by transnational corporations goes beyond the issue of social media companies, two responses have been formulated in IHRL in general. First, the normative discourse in human rights set by international organisations and legal doctrine has evolved around three principles: the promotion, the protection and the fulfilment of HR. Through their obligation to protect, states must guarantee that they take measures to protect human rights from other private actors (indirect horizontal effect of HR; Cinneide, 2003). Accordingly, Laidlaw points out that the obligation of states to protect human rights 'trickles down to businesses' (Laidlaw, 2015, p. 88). Second, international organisations and the international legal scholars community support the proposal that private companies should voluntarily embrace HR standards from a corporate social responsibility perspective (Huisman, 2021; Laidlaw, 2017). However, regarding this voluntary approach applied to social media companies, several authors highlight the risk of bluewashing (Douek, 2021; Laidlaw, 2017). Sander (2020, p. 1005) underlines that 'given the complexity (of applying IHRL to social media companies), there is an inevitable risk that online platforms will attempt to co-opt the vocabulary of human rights to legitimise minor reforms at the expense of more structural or systemic changes to their moderation processes'. Along the same lines, Griffin argues that critical legal studies have long described how approaches based primarily on individual rights 'are unsuited to addressing systemic inequalities' (Griffin, 2023, p. 37). This same point is taken up by Quintais et al. (2023, p. 907), who point out the limitations of approaches based on fundamental rights because of their individualistic and apolitical character.

In this context, the European Union has chosen to oblige social media companies to incorporate a HR-based approach into their content moderation governance and to go beyond not only the self-regulatory approach but also beyond the voluntary commitments to HR derived from the business and human rights principles. Among the provisions adopted, Article 14 DSA directly echoes one of the elements of the HR framework, the principle of legal certainty.

THE EMBEDDED HR-BASED APPROACH IN THE DSA: ARTICLE 14 AND THE PRINCIPLE OF LEGAL CERTAINTY

With the adoption in 2022 of the DSA, the regulatory regime for content moderation is undergoing radical change in the EU. The original logic of self-regulation is being left behind in favour of a co-regulatory approach (Finck, 2018). Co-regulation describes 'a regulatory regime (…) made up of a complex interaction of general legislation and a self-regulatory body' (Marsden, 2011, p. 46). In adopting a co-regulatory approach, the state can thus consolidate private rules and mechanisms (the self-regulatory regime) in legislation while adding new legal obligations. Examples of co-regulation of social media companies already existed in the EU before the DSA, notably with Germany's NetzDG (Klausa, 2023) or the 2018 Terrorist Content Regulation (Quintais et al., 2023). However, the DSA stands out for its scope, as it aims to regulate the activity of social media companies as a whole and not just particular issues like hate speech or terrorist content. To that end, the DSA imposes transparency and due diligence obligations to social media companies, thus enforcing an approach based on the respect of fundamental rights as stated in its preamble (n3, n9, n22, n40, n47, etc.).

Focusing on Article 14 DSA in particular, the EU imposes language requirements for the terms and conditions of social media companies. Article 14 states that:

> Providers of intermediary services shall include information on any restrictions that they impose in relation to the use of their service in respect of information provided by the recipients of the service, in their terms and conditions. That information shall include information on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review, as well as the rules of procedure of their internal complaint handling system. It shall be set out in clear, plain, intelligible, user-friendly and unambiguous language, and shall be publicly available in an easily accessible and machine-readable format. (paragraph 1)

Legal certainty is a key legal principle from an IHRL perspective. Indeed, in IHRL theory, any restrictions on HR need to fulfil three conditions to be considered lawful: a legal basis provides for the limitation of the right, there is a legitimate aim to the restriction, and the restriction is proportionate (Donnelly, 2013). If one of these conditions is not met, courts will consider that the human right has been violated. Among these conditions, the requirement of the legal basis is also called the legality principle, which contains other subprinciples, notably the legal certainty principle (Donnelly, 2013). The legal certainty principle mandates that any law must be 'clear, foreseeable, coherent, determinate, and predictable' (Ranchordas, 2021, p. 19). Such requirements protect people against arbitrary decisions.

Quintais et al. highlight the important role attributed to Article 14 DSA. They emphasise the key importance given to the terms and conditions of social media companies and how Article 14 could allow the application of EU fundamental rights law to content moderation (Quintais et al., 2023). Indeed, with the entry into force of Article 14, the relationship between social media companies and users is changing. Until then, this relationship was a private one, where the terms and conditions formed the basis of the contract. With Article 14 and the co-regulatory framework of the DSA, the pre-existing contractual regime disappears in part and an obligation to comply with HR standards arises. This changes the role of terms and conditions, which evolve from a contractual agreement between the platform and users to a legal basis on which restrictions to freedom of expression must be justified. In this context, content moderation needs to be guided by the 'same standards of legality, necessity and legitimacy that bind State regulation of expression' (Quintais et al., 2023, p. 896). With this

change, the European legislator is materialising, as it is called in IHRL theory, an 'indirect horizontal effect of human rights in the relationship between online platforms and their users' (Quintais et al., 2023, p. 910).

Considering that the requirement for legal certainty has been included in the DSA for social media companies, this paper offers to examine whether such a requirement can be met in the case of Facebook's CS.

METHODOLOGY

To answer the research question, the paper presents a qualitative content analysis of Facebook's CS as a single case study. Case-study research allows intimate knowledge of the properties of a single case (Gerring, 2009). Considering the importance of Facebook in the regulatory landscape of social media companies, we believe its case to be particularly relevant to illustrate the issue of the implementation of the new DSA requirements for CS. In 2023, Facebook had over three billion users, making it the largest social media company worldwide. This confers a particular importance to its CS, which apply to more people than any national law on earth.

The data analysed is the American English version of Facebook's CS. This version was chosen because Facebook indicates on its 'Transparency Center' that 'the US English Version of the CS reflects the most up-to-date set of the policies and should be used as the primary document'. The CS are accessible online.[1] The CS are divided into 'Policies'. These policies are organised under thematic sections, which are themselves divided by subthemes. The six thematic sections are 'Violence and criminal behaviour', 'Safety', 'Objectionable content', 'Integrity and authenticity', 'Respecting intellectual property' and 'Content-related requests and decisions'. In total, there are 24 policies. The policies 'intellectual property', 'memorialisation', 'user-requests' and 'additional protection of minors' were excluded because they did not provide users with commands but were informative policies related to how the company is dealing with special user requests. The analysis focuses on the 20 remaining policies ($n = 20$). Each policy presents a similar structure, showing the 'Policy Rationale' first, which details the theme and reasons underlying it. The policies show a large 'Do not post' sign at the top before listing behaviour examples. The analysis focused on these lists considering that they included the command dimension of the CS. The analysis did not include the 'policy rationale'. This part gives a contextual explanation for the policy and does not issue commands to users. From a legal perspective, policy rationales could be considered similar to the preamble of regulations, which does not have a binding dimension but provides information for the interpretation in the application of the rule (Orgad, 2010). The analysis stops at the version of CS of 10 January 2024.

To assess the language of the CS, the analysis offers an innovative coding scheme based on the principle of legal certainty. Atlas.ti software was used for the coding process. Legal certainty requires the lawmaker to draft its texts so that they are accessible and predictable for the recipients of the law (Popelier, 2008). However, accessibility and predictability are not linguistic characteristics as such. That is why legal scholars agree that when drafting legislation, the two goals of accessibility and predictability translate to three linguistic features: clarity, precision and unambiguity (Kabba, 2011; Majambere, 2011; Xanthaki, 2014). The legislator must make these three characteristics objectives when drafting a law. As a team of two researchers, a first round of coding was made using the legal criteria of clarity, precision and unambiguity as codes. We realised that these three criteria were only superficially informing us on how CS are complying with the requirement of legal certainty. Following an iterative process, two additional rounds of coding were carried out and led to the refinement of these three characteristics into subcodes. The requirement

of clarity was divided into three codes, the requirement of precision in two, and the requirement of unambiguity in three (see Table 1). Disagreements were solved in common. The three characteristics of clarity—precision—unambiguity are not mutually exclusive: the same word or text can fail to comply with several characteristics at the same time.

Like any empirical study, this paper has limitations. First, this paper focused on a specific aspect relating to the principle of legality. Other principles ensuring the restriction regime under IHRL were not analysed. This study would be very interesting to assess Facebook's commitment to HR standards, but other data would have been necessary. Second, this study focuses on the language of CS but does not consider the issue of their enforcement. While the failure of social media companies to enforce their terms and conditions is also an essential issue in studies about content moderation (Zakrzewski et al., 2023), it was beyond the scope of our analysis.

ANALYSIS

The results of our analysis are summarised in Table 2. To illustrate the results, we present selected excerpts with representative examples of issues we encountered.

Clarity

Clarity issues were identified in 10 different policies. The *Adult Sexual Exploitation* policy illustrates a clarity issue among the different cases due to its structure. The repetition of similar problems gives an impression of carelessness. The writing structure still allows the understading of the proscribed behaviour. However, it complicates it, as shown in Figure 1. There is no need to list behaviours with bullet points and to repeat 'or' at the end of the sentence - the bullet points list implies that these are alternative options, making the preposition unnecessary.

A text including several examples systematically completed with exceptions also illustrates a problem of clarity. Such structure renders the reading difficult and thus has consequences for understanding the proscribed behaviour. As illustrated in Figure 2, the repetition of ‘but not limited to’ is not in the right place, and the structure should be entirely changed so that the exceptions make more sense.

Precision

Issues regarding the precision criteria were encountered in 14 different policies. The policy on *Adult Nudity and Sexual Activity* presents good examples of a lack of precision. Instead of using vaguer terms like ‘sexual activity with visible genitalia’ that would signify precisely enough the proscribed behaviour, the policy gives many details about what is forbidden, making it impossible to identify the general rule and assess if behaviours deriving a little from the rule are also covered or not. Figure 3 features many occurrences of precision issues.

Unambiguity

The main issue related to the criteria of unambiguity is the equivocal character of many text passages. Text passages were either contradictory in the worst cases or simply ambiguous. Ambiguity issues were encountered in almost every policy analysed: 16 of the 20 policies.

TABLE 1 Coding scheme.

Legal requirements	Linguistic requirements	Codes	Explanation	Example
Accessibility & predictability	Clarity	Accessible language	The text must not be too complex for its recipients. A text is considered too complex when it requires the knowledge of a field specialist.	'Ponzi scheme' requires specific knowledge, whereas terms like fraud or scam are more accessible (Policy on *Fraud and Deception*)
		Clear structure	The different parts of a text fit together and do not make the reading difficult. It includes logical spacings and linking of paragraphs.	-
		Spelling and grammar	Spelling and grammar errors make a text sound less official and would not be expected in a legal text	-
	Precision	Not too vague	Too much vagueness does not allow individuals to adapt their behaviours to the rule as they do not know what is covered or not	The sentence 'statements of intent to commit high-severity violence' was assessed as too vague (Policy on *Violence and Incitement*)
		Not too precise	Too much precision does not allow individuals to assess the lawfulness of behaviours that are not precisely described but are still similar	The text passage 'Other activities, except in cases of medical or health context, advertisements, and recognized fictional images or with indicators of fiction, including but not limited to: Squeezing female breasts, defined as a grabbing motion with curved fingers that shows both marks and clear shape change of the breasts' was considered too precise (Policy on *Adult Nudity and Sexual Activity*)
	Unambiguity	No contradictions in the text	The text analysed and its immediate context (sentence, paragraph, section) must not contradict each other and must be coherent with each other	The text passage 'Do not post: Sadistic remarks towards the following content which includes a label so that people are aware it may be sensitive' is contradictory. Should the user not post the content, or can the user post the content and a label will be added? (Policy on *Violent and Graphic Content*)

(Continues)

TABLE 1 (Continued)

Legal requirements	Linguistic requirements	Codes	Explanation	Example
		Logical meaning	The behaviour described must make sense regarding the aim of the rule and/or regarding usual logic applicable to the field considered	The sentence 'Comparisons to animals or insects that are not culturally perceived as intellectually or physically inferior ("tiger," "lion")' was assessed as lacking logical meaning (Policy on *Bullying and Harassment*)
		Conciseness	The text must avoid synonyms or details that can lead to confusion regarding the sense of the described behaviour. This code differs from the analysis of precision in that it does not focus on a word or group of words, but analyses what is around a word to see what might trigger confusion.	The text passage 'Do not post content that promotes, encourages, coordinates, or provides instruction for' was assessed as lacking concision because of the use of numerous synonyms—Even though every one of the words taken individually would be precise enough (Policy on *Suicide and Self Injury*)

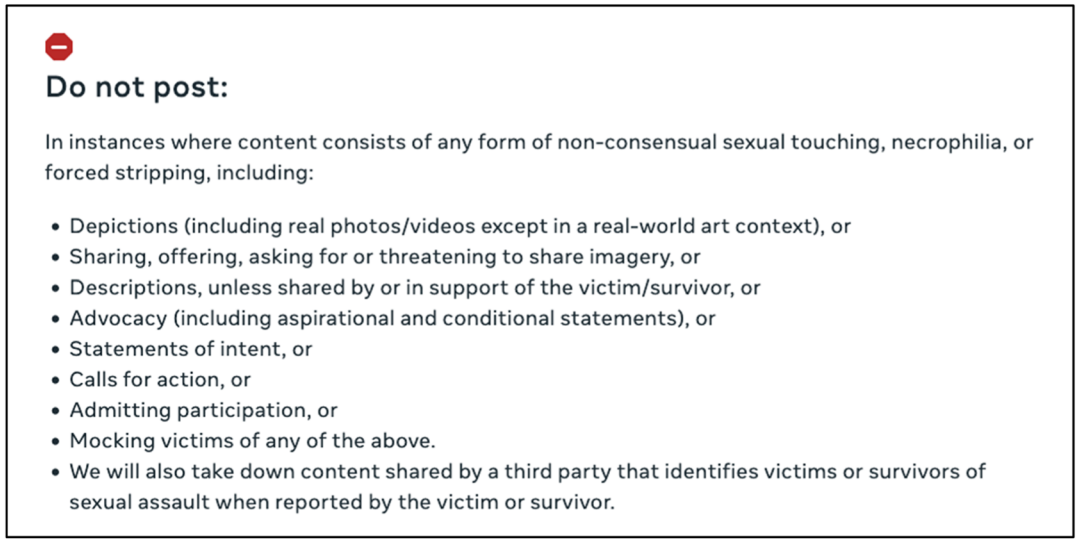
TABLE 2 Results.

Community standards		CRITERIA RESPECTED / CRITERIA NOT RESPECTED		
Thematic section	Policies	*Clarity*	*Precision*	*Unambiguity*
Violence and criminal behaviour	Violence and incitement			
	Dangerous individuals and organisations			
	Coordinating harm and promoting crime			
	Restricted goods and services			
	Fraud and deception			
Safety	Suicide and self-injury			
	Child sexual exploitation, abuse and nudity			
	Adult sexual exploitation			
	Bullying and harassment			
	Human exploitation			
	Privacy violations			
Objectionable content	Hate speech			
	Violent and graphic content			
	Adult nudity and sexual activity			
	Sexual solicitation			
Integrity and authenticity	Account integrity and authentic identity			
	Spam			
	Cybersecurity			
	Inauthentic behaviour			
	Misinformation			

The policy on *Bullying and Harassment* contains examples of ambiguity. As shown in Figure 4, some parts of the text are rather positive and unharmful ('positive physical descriptions'), which does not seem logical regarding the behaviours that the rule intends to forbid.

Results

In the end, the analysis shows that Facebook's CS lack the 'clear, precise and unambiguous' character that legal certainty requires according to IHRL and Article 14 DSA. Out of the 20 policies analysed, only two did not present an issue. Of the three criteria, the one assessing unambiguity was mainly not respected. In almost every policy, ambiguous elements distort the understandability of the text. This element is particularly problematic because unambiguity is important for understanding a rule. From the users' point of view, if a rule has several potential meanings, it is impossible to know which behaviour to adopt. The criterion of precision was also not often respected. The drafting of CS is characterised by long lists of prohibited behaviour, supplemented by exceptions. It makes it difficult to identify what behaviours are allowed. Some formulations are so precise that CS only cover very specific behaviours, and it is impossible to deduct from the text whether similar behaviours are prohibited. The clarity issues identified did not, for the most part, hinder the understandability of the proscribed behaviour. Yet, numerous times, spelling errors, deficient structures, hasty formatting and inconsistencies were present.

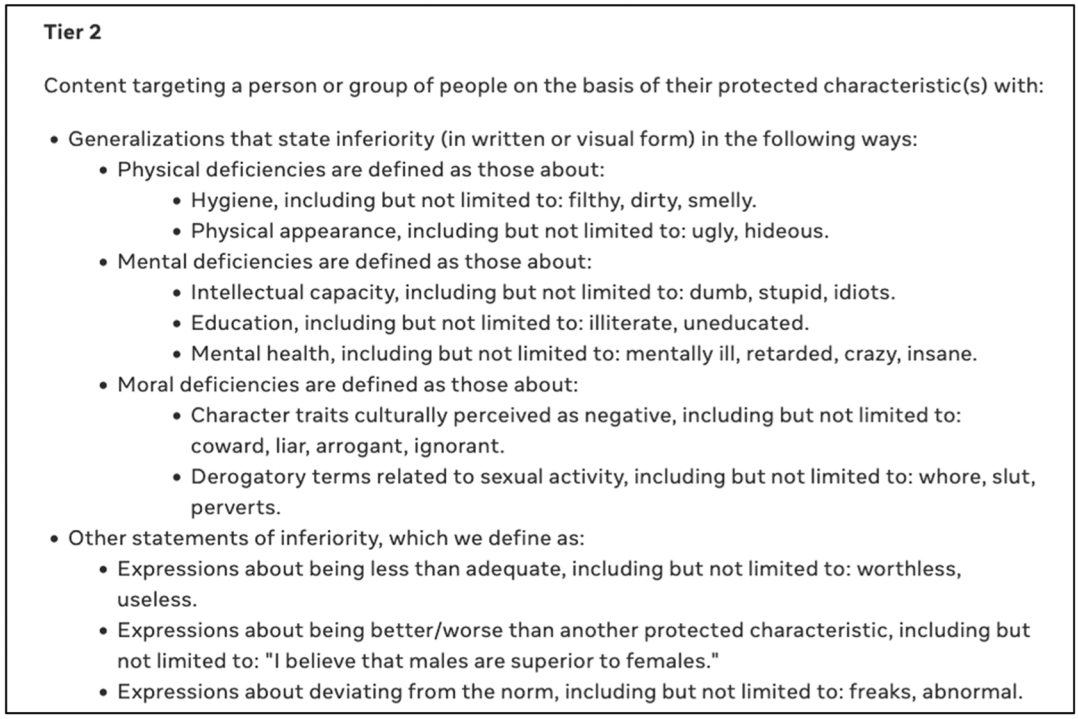


Do not post:

In instances where content consists of any form of non-consensual sexual touching, necrophilia, or forced stripping, including:

- Depictions (including real photos/videos except in a real-world art context), or
- Sharing, offering, asking for or threatening to share imagery, or
- Descriptions, unless shared by or in support of the victim/survivor, or
- Advocacy (including aspirational and conditional statements), or
- Statements of intent, or
- Calls for action, or
- Admitting participation, or
- Mocking victims of any of the above.
- We will also take down content shared by a third party that identifies victims or survivors of sexual assault when reported by the victim or survivor.

FIGURE 1 Excerpt from the policy adult sexual exploitation. Retrieved on 10 January 2024, at: https://transparency.fb.com/policies/community-standards/sexual-exploitation-adults.

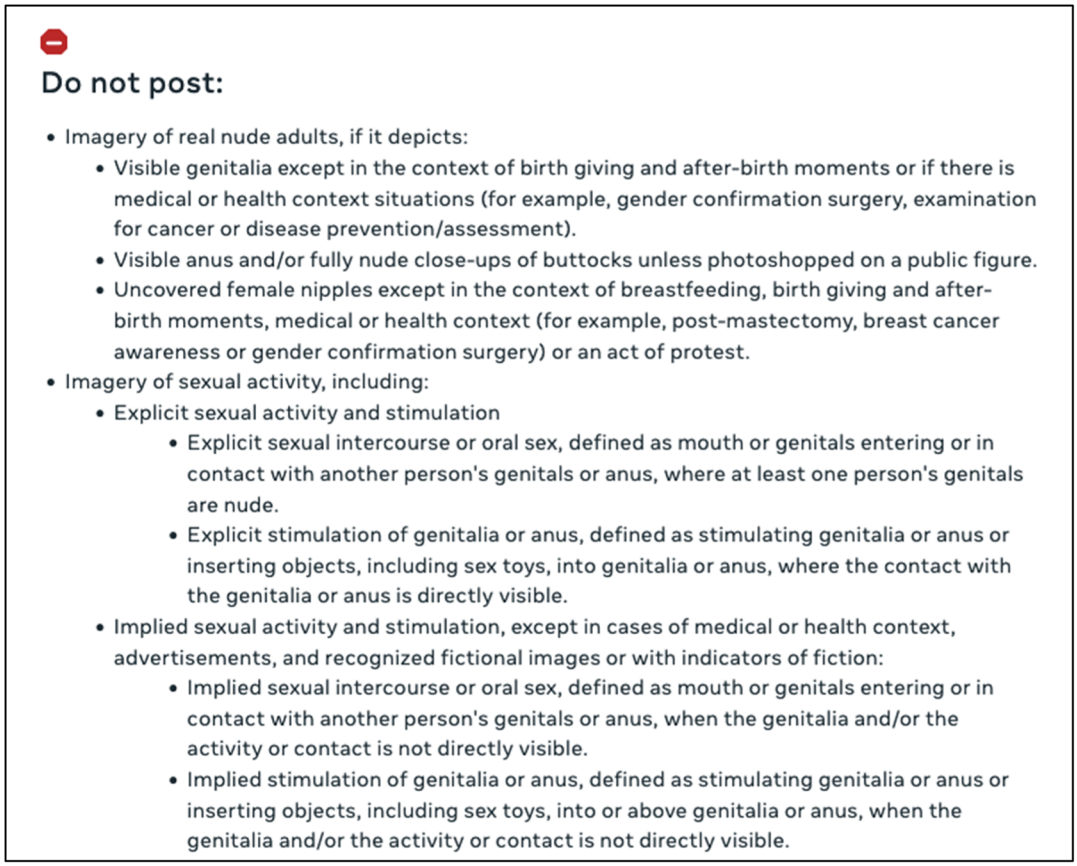


Tier 2

Content targeting a person or group of people on the basis of their protected characteristic(s) with:

- Generalizations that state inferiority (in written or visual form) in the following ways:
  - Physical deficiencies are defined as those about:
    - Hygiene, including but not limited to: filthy, dirty, smelly.
    - Physical appearance, including but not limited to: ugly, hideous.
  - Mental deficiencies are defined as those about:
    - Intellectual capacity, including but not limited to: dumb, stupid, idiots.
    - Education, including but not limited to: illiterate, uneducated.
    - Mental health, including but not limited to: mentally ill, retarded, crazy, insane.
  - Moral deficiencies are defined as those about:
    - Character traits culturally perceived as negative, including but not limited to: coward, liar, arrogant, ignorant.
    - Derogatory terms related to sexual activity, including but not limited to: whore, slut, perverts.
- Other statements of inferiority, which we define as:
  - Expressions about being less than adequate, including but not limited to: worthless, useless.
  - Expressions about being better/worse than another protected characteristic, including but not limited to: "I believe that males are superior to females."
  - Expressions about deviating from the norm, including but not limited to: freaks, abnormal.

FIGURE 2 Excerpt from the policy hate speech. Retrieved on 10 January 2024, at: https://transparency.fb.com/policies/community-standards/hate-speech.



Do not post:

- Imagery of real nude adults, if it depicts:
  - Visible genitalia except in the context of birth giving and after-birth moments or if there is medical or health context situations (for example, gender confirmation surgery, examination for cancer or disease prevention/assessment).
  - Visible anus and/or fully nude close-ups of buttocks unless photoshopped on a public figure.
  - Uncovered female nipples except in the context of breastfeeding, birth giving and after-birth moments, medical or health context (for example, post-mastectomy, breast cancer awareness or gender confirmation surgery) or an act of protest.
- Imagery of sexual activity, including:
  - Explicit sexual activity and stimulation
    - Explicit sexual intercourse or oral sex, defined as mouth or genitals entering or in contact with another person's genitals or anus, where at least one person's genitals are nude.
    - Explicit stimulation of genitalia or anus, defined as stimulating genitalia or anus or inserting objects, including sex toys, into genitalia or anus, where the contact with the genitalia or anus is directly visible.
  - Implied sexual activity and stimulation, except in cases of medical or health context, advertisements, and recognized fictional images or with indicators of fiction:
    - Implied sexual intercourse or oral sex, defined as mouth or genitals entering or in contact with another person's genitals or anus, when the genitalia and/or the activity or contact is not directly visible.
    - Implied stimulation of genitalia or anus, defined as stimulating genitalia or anus or inserting objects, including sex toys, into or above genitalia or anus, when the genitalia and/or the activity or contact is not directly visible.

FIGURE 3 Excerpt from the policy adult nudity and sexual activity. Retrieved on 10 January 2024, at: https://transparency.fb.com/policies/community-standards/adult-nudity-sexual-activity.

DISCUSSION

The result of the analysis of Facebook's CS shows that their semantics do not correspond to those expected of a legal text compliant with the principle of legal certainty. It illustrates the gap existing between the CS' drafting in the self-regulatory approach of content moderation and the legal requirements under the DSA. It comes as no surprise considering that CS were not designed as external communication tools aimed at users but derive from the internal dimension of content moderation where their purpose is that of a guide for company moderators (Klonick, 2018) and where they play a role with automated processes, notably through 'Community standards classifiers' (Gorwa et al., 2020). From this perspective, the analysis of CS highlights the casuistic and iterative logic of content moderation. This does not mean that CS do not have a prescribing dimension: their formulation remains like that of 'command-and-control' instruments prohibiting or prescribing behaviours (Sinclair, 2002). However, with the poor linguistic quality that the analysis has revealed, the language of CS does not meet the IHRL requirements according to which users should be able to assess what they can and cannot say.

In this context, the adoption of the DSA could drive significant changes if Facebook decides to be compliant with the new European regulation, at least among European Member States. Social media companies are no longer just being asked to respect HR voluntarily but must comply with legal obligations to act by adapting their governance of

Tier 3: Additional protections for Private Minors, Private Adults, and Minor Involuntary Public Figures:

- In addition to all the protections listed above, all private minors, private adults (who must self-report), and minor involuntary public figures are protected from:
  - Targeted cursing.
  - Claims about romantic involvement, sexual orientation or gender identity.
  - Calls for action, statements of intent, aspirational or conditional statements, or statements advocating or supporting exclusion.
  - Negative character or ability claims, except in the context of criminal allegations and business reviews against adults.
  - Expressions of contempt, disgust, or content rejecting the existence of an individual, except in the context of criminal allegations against adults.
- When self-reported, private minors, private adults, and minor involuntary public figures are protected from the following:
  - First-person voice bullying.
  - Unwanted manipulated imagery.
  - Comparison to other public, fictional or private individuals on the basis of physical appearance.
  - Claims about religious identity or blasphemy
  - Comparisons to animals or insects that are not culturally perceived as intellectually or physically inferior ("tiger," "lion").
  - Neutral or positive physical descriptions.
  - Non-negative character or ability claims.
  - Attacks through derogatory terms related to a lack of sexual activity.

FIGURE 4 Excerpt from the policy bullying and harassment. Retrieved on 10 January 2024, at: https://transparency.fb.com/policies/community-standards/bullying-harassment.

content moderation. In this respect, European institutions should not be the only ones trying to make sure that these obligations are fulfilled. Indeed, Article 14 also puts the conditions for guaranteeing the indirect horizontal effect for substantive rights in place, as Quintais et al. (2023) underline. National courts could therefore examine whether a European Member State that admits CS for restricting freedom of expression complies with its obligations to respect, protect and fulfil fundamental rights. In this context, each European Member State should be highly motivated to ensure the comprehensibility of the CS since it may be held responsible. From this perspective, the HR-based approach implemented through the DSA will offer more guarantees for protecting freedom of expression than in states that only opt for self-regulatory approaches.

In the case that Facebook decides to be compliant with the linguistic requirements, the enforcement of Article 14 DSA could generate a 'domino effect' on the scope of the application of CS. We focused on the American English version of the CS. This choice was justified, although Facebook publishes CS in several European Union official languages. Indeed, the company states that the American English version is the primary document, meaning that this version prevails over other versions, emphasizing the American mindset of Facebook's content moderation governance. However, as we have shown, CS—in all the official European languages—must now meet the requirements of Article 14 DSA. There are, therefore, three options for Facebook: first, the company decides to modify each version of its CS in the official European languages and breaks with the idea that the American English version takes precedence over the other language versions; second, it adapts the

UK English version (as Ireland is an EU Member State) and makes it the primary interpretative document in terms of contestation in other European languages; third, the company decides to change the American English version of its CS, keeping its principle of the unity of interpretation based on this version. The company's choice will have the effect either of contributing to the fragmentation of content moderation (Ahn et al., 2022) or of reinforcing the Brussels effect of European regulation (Bradford, 2020).

Depending on Facebook compliance's decision, a subtle transformation of the initial self-regulatory approach of content moderation implemented in the United States could take place, since the American company would 'voluntarily' incorporate the HR-based approach that is integrated into the European Union's co-regulatory approach beyond the territorial jurisdiction of the EU and its Member States. Several reasons could bring the company to implement Article 14 DSA globally. The first is the desire to avoid an administrative burden by favouring the least demanding option. The second is to avoid undermining other elements of content moderation governance such as the work of its Oversight Board, which applies United Nations HR recommendations, notably those that require the respect of the principle of legal certainty (e.g., Oversight Board, Nazi Quote, 2021). A last and more contextual element can be found in the discussions on digital regulations. Since Cambridge Analytica, a succession of events showed that the activities of social media companies generate systemic risks for democracies and vulnerable groups. The US Congress and US Supreme Court have not yet found an alternative to the Section 230 immunity regime (Liptak, 2023). This leaves large companies free on the grounds that the market will naturally balance out. In this context, the indirect effect of the DSA on the American English version of the CS could therefore be appreciated by American policymakers advocating for legislative changes, and Facebook could be encouraged in this direction.

However, the implementation of Article 14 DSA could also produce another domino effect, on content moderation governance itself this time. Indeed, as mentioned earlier, CS primarily play an internal role, between their function as a guide for moderator and their interrelationship with automated processes in algorithmic content moderation. Only later were they published to help users understand moderation decisions, as a 'gesture of transparency' (Gillespie, 2018, p. 47). The DSA is bringing external constraints on CS in making them an essential piece of information for users and requiring them to fulfil the formal standards of the legal certainty principle. Such a shift from an internal or bottom–up to an external or top–down perspective could have impacts on the internal 'community standards-making chain' amongst company executives, employees involved in automation processes (labellers, coders, engineers, etc.), the moderators and the community. It would require the company to find a way to include the legal requirements of Article 14 in this 'community standards-making chain' while ensuring that CS continue to play their 'internal' role for companies' employees engaged in content moderation. Given the initial self-regulatory tradition in which CS are anchored, it is questionable whether such an evolution is even possible. Regarding such a challenge, the possible failure to be compliant with Article 14 would show a limit of the HR-based regulatory strategy for social media.

CONCLUSION

This paper has shown that Facebook's CS do not comply with Article 14 DSA, which requires the respect of the legal certainty principle, one of the conditions for restricting freedom of expression. From this perspective, this paper opens the discussion about the effectiveness of one of the legal solutions anchored in the DSA to guarantee a safer internet and contributes to the debate on the operationalisation of a HR-based approach to the regulation of social media companies.

The requirements of Article 14 could impose a profound change in CS' language. In the EU, digital service providers could quickly come under pressure, given that Article 14 gives rise to an indirect horizontal obligation requiring the European Member States to verify that the conditions have been put in place by companies to respect fundamental rights. In the context of Facebook, the decision to comply with Article 14 could also have an impact on its global content moderation policy and HR. Moreover, redrafting CS in compliance with the legal requirements of Article 14 DSA could have unintended effects on content moderation governance considering how the current drafting of CS is ontologically embedded in content moderation processes. The changes that are needed regarding CS and their potential impacts on content moderation governance will certainly question the company's motivation to comply with its European legal obligations.

Beyond the discussion about the implementation effects of the HR-based approach anchored in the DSA, this regulatory approach could also produce a 'boomerang effect' of technology on the law. Indeed, the analysis highlights the casuistic dimension of CS. The phenomenal amount of content that social media companies must process makes automation indispensable to identify and filter problematic content. Casuistic drafting—listing what is permitted and prohibited—can be easily translated into code. From this perspective, the requirements of Article 14 DSA illustrate the difficulty of merging the legal rationale with the algorithmic rationale. In mandating social media companies to not only have 'clear, plain, intelligible', but also 'user-friendly and unambiguous language' in their policies, the European legislator extended the classic criteria of the legal certainty principle. The mention of 'user-friendly' is a feature of the digital environment and remains unknown from a legal point of view. Article 14 also requires that social media companies draft their terms and conditions in a 'machine-readable' way, contributing to the movement in favour of the automation of legal texts (Huggins et al., 2022). From this perspective, social media companies will have to find a way to conciliate legal and computing perspectives. The study of these different elements drawn from the DSA and how they will be implemented by social media companies will contribute to the philosophical debate about the impact of technology on the law. Here again, further empirical research is needed to inform the necessary theoretical reflection on possible adaptations and the requirements for the resilience of legal principles in the context of digital transformation.

ACKNOWLEDGMENTS

The idea for this paper was first discussed at the 2022 conference of the International Society for Public Law in Wrocklaw, Poland. We would like to thank Professor Oreste Pollicino and Thomas Streinz for their inspiring initial inputs. We also warmly thank Alicia Pastor y Camarasa for her valuable feedback throughout the process. Open access funding provided by Universite de Lausanne.

ORCID

Mathieu Fasel http://orcid.org/0000-0002-3755-9989
Sophie Weerts http://orcid.org/0000-0002-8579-8124

ENDNOTE

[1] https://transparency.fb.com/policies/community-standards/

REFERENCES

Ahn, S., Baik, J., & Krause, C. S. (2022). Splintering and centralizing platform governance: How Facebook adapted its content moderation practices to the political and legal contexts in the United States, Germany and South Korea. *Information, Communication & Society*, *26*(14), 2843–2862. https://doi.org/10.1080/1369118X.2022.2113817

Amnesty International. (2022). *Myanmar: Facebook's systems promoted violence against Rohingya*. https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

Article 19. (2023). *Content moderation and freedom of expression handbook*. https://www.article19.org/wp-content/uploads/2023/08/SM4P-Content-moderation-handbook-9-Aug-final.pdf

Bickert, M. (2018). *Publishing our internal enforcement guidelines and expanding our appeals process*. https://about.fb.com/news/2018/04/comprehensive-community-standards/

Bietti, E. (2023). A genealogy of digital platform regulation. *Georgetown Law Review*, *7*(1), 1–68.

Bradford, A. (2020). *The Brussels effect: How the European Union rules the world*. Oxford University Press.

Callamard, A. (2019). The human rights obligations of non-state actors. In R. K. Jørgensen (Ed.), *Human rights in the age of platforms* (pp. 191–225). MIT Press. https://doi.org/10.7551/mitpress/11304.003.0015

Castano-Pulgarin, S. A., Suarez-Betancur, N., Tilano Vega, L. M., & Herrera Lopez, H. M. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 1–7. https://doi.org/10.1016/j.avb.2021.101608

Cate, F. H., & Mayer-Schonberger, V. (2013). Notice and consent in a world of Big Data. *International Data Privacy Law*, *3*(2), 67–73. https://doi.org/10.1093/idpl/ipt005

Celeste, E. (2019). Digital constitutionalism: A new systemic theorisation. *International Review of Law, Computers & Technology*, *33*(1), 76–99. https://doi.org/10.1080/13600869.2019.1562604

Chander, A. (2012). Facebookistan. *North Carolina Law Review*, *90*, 1807–1842.

Cinneide, C. (2003). Taking horizontal effect seriously: Private law, constitutional rights and the European Convention on human rights. *Hibernian Law Journal*, *4*, 77–108.

Cohen, J. (2019). *Between truth and power—The legal constructions of information capitalism*. Oxford University Press.

Council of Europe. (2021). *Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation*. https://rm.coe.int/content-moderation-en/1680a2cc18

Dias Oliva, T. (2020). Content moderation technologies: Applying human rights standards to protect freedom of expression. *Human Rights Law Review*, *20*(4), 607–640. https://doi.org/10.1093/hrlr/ngaa032

Donald, A., & Speck, A.-K. (2020). The dynamics of domestic human rights implementation: Lessons from qualitative research in Europe. *Journal of Human Rights Practice*, *12*(1), 48–70. https://doi.org/10.1093/jhuman/huaa007

Donnelly, J. (2013). *Universal human rights in theory and practice*. Cornell University Press. https://www.jstor.org/stable/10.7591/j.ctt1xx5q2

Douek, E. (2021). The limits of International Law in content moderation. *UC Irvine Journal of International, Transnational, and Comparative Law*, *37*(6), 37–76. https://scholarship.law.uci.edu/ucijil/vol6/iss1/4

Finck, M. (2018). Digital co-regulation: Designing a supranational legal framework for the platform economy. *European Law Review*, *41*, 33–67.

Gerring, J. (2009). The case study: What it is and what it does. In C. Boix & S. Stokes (Eds.), *The Oxford handbook of comparative politics* (pp. 1133–1165). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199566020.003.0004

Gillespie, T. (2018). *Custodians of the internet*. Yale University Press. https://yalebooks.yale.edu/book/9780300261431/custodians-of-the-internet/

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 1–15. https://doi.org/10.1177/2053951719897945

De Gregorio, G. (2022). *Digital constitutionalism in Europe—Reframing rights and powers in the algorithmic society*. Cambridge University Press. https://doi.org/10.1017/9781009071215

Griffin, R. (2023). Rethinking rights in social media governance: Human rights, ideology and inequality. *European Law Open*, *2*, 30–56. https://doi.org/10.1017/elo.2023.7

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law & Technology*, *17*(42), 42–109.

Huggins, A., Burdon, M., Witt, A., & Suzor, N. (2022). Digitising legislation: Connecting regulatory mind-sets and constitutional values. *Law, Innovation and Technology*, *14*(2), 325–354. https://doi.org/10.1080/17579961.2022.2113670

Huisman, W. (2021). Corporations, human rights and compliance. In B. Van Rooij & D. Sokol (Eds.), *The Cambridge handbook of compliance* (pp. 989–1009). Cambridge University Press. https://www.cambridge.org/core/books/abs/cambridge-handbook-of-compliance/corporations-human-rights-and-compliance/F8266DB6779FC395EB7FFD8F9A4170D6

Husovec, M. (2020). Remedies first, liability second: or why we fail to agree on optimal design of intermediary liability. In G. Frosio (Ed.), *The Oxford handbook of online intermediary liability*, (pp. 90–103). Oxford University Press.

Jørgensen, R. K. (2019). Rights talk: In the kingdom of online giants. In R. K. Jørgensen (Ed.), *Human rights in the age of platforms*, (pp. 163–187). MIT Press. https://doi.org/10.7551/mitpress/11304.003.0013
Jørgensen, R. K., & Pedersen, A. M. (2017). Online service providers as human rights arbiters. In M. Taddeo & L. Floridi (Eds.), *The responsibilities of online service providers*, (pp. 179–199). Springer.
Kabba, K. (2011). Gender-neutral language: An essential language tool to serve precision, clarity and unambiguity. *Commonwealth Law Bulletin*, *37*(3), 427–434. https://doi.org/10.1080/03050718.2011.595141
Kaye, D. (2016). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. Human Rights Council, Thirty-second session. https://www.ohchr.org/en/documents/thematic-reports/ahrc3238-report-freedom-expression-states-and-private-sector-digital-age
Kaye, D. (2018). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. Human Rights Council, Thirty-eighth session. https://www.ohchr.org/en/documents/thematic-reports/ahrc3835-report-special-rapporteur-promotion-andprotection-right-freedom
Kaye, D. (2019a). *Speech police—The global struggle to govern the internet*. Columbia Global Reports. https://www.jstor.org/stable/j.ctv1fx4h8v
Kaye, D. (2019b). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. UN General Assembly, Seventy-fourth session. https://www.ohchr.org/en/documents/thematic-reports/a74486-report-online-hate-speech
Kettemann, M., & Schulz, W. (2020). *Setting rules for 2.7 Billion. A (first) look into Facebook's norm-making system*. Working Papers of the Hans-Bredow-Institut. https://hans-bredow-institut.de/en/publications/setting-rules-for-2-7-billion-a-first-look-into-facebook-s-norm-making-system
Klausa, T. (2023). Graduating from 'new-school'—Germany's procedural approach to regulating online discourse. *Information Communication & Society*, *26*(1), 54–69. https://doi.org/10.1080/1369118X.2021.2020321
Klonick, K. (2018). The new governors—The people, rules and processes governing online speech. *Harvard Law Review*, *131*, 1598–1670. https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/
Laidlaw, E. B. (2015). *Regulating speech in cyberspace—Gatekeepers, human rights and corporate responsibility*. Cambridge University Press. https://doi.org/10.1017/CBO9781107278721
Laidlaw, E. B. (2017). Myth or promise? The corporate social responsibilities of online service providers for human rights. In M. Taddeo & L. Floridi (Eds.), *The responsibilities of online service providers*. (pp. 135–156). Springer.
Land, M. K. (2019). Regulating private harms online: Content regulation under human rights law. In R. F. Jørgensen (Ed.), *Human rights in the age of platforms*, (pp. 285–316). MIT Press. https://doi.org/10.7551/mitpress/11304.003.0018
Liptak, A. (2023). Supreme court won't hold tech companies liable for user posts. *The New York Times*. https://www.nytimes.com/2023/05/18/us/politics/supreme-court-google-twitter-230.html
Macdonald, S., & Vaughan, K. (2023). Moderating borderline content while respecting fundamental values. *Policy & Internet*, 1–15. https://doi.org/10.1002/poi3.376
Majambere, E. (2011). Clarity, precision and unambiguity: Aspects for effective legislative drafting. *Commonwealth Law Bulletin*, *37*(3), 417–426. https://doi.org/10.1080/03050718.2011.595140
Marsden, C. (2011). Internet co-regulation and constitutionalism. In C. Marsden (Ed.), *Internet co-regulation: European law, regulatory governance and legitimacy in cyberspace*, (pp. 46–70). Cambridge University Press. https://www.cambridge.org/core/books/internet-coregulation/7179CDF556745BA2313666AEE0A60E
Murray, R., & Long, D. (2022). *Research handbook on implementation of human rights in practice*. Edward Elgar Publishing.
Orgad, L. (2010). The preamble in constitutional interpretation. *International Journal of Constitutional Law*, *8*(4), 714–738. https://doi.org/10.1093/icon/mor010
Oversight Board. (2021). Nazi Quote. https://www.oversightboard.com/decision/fb-2rdrcavq/
Pollicino, O. (2021). *Judicial protection of fundamental rights in internet. towards digital constitutionalism?* Hart Publishing.
Popelier, P. (2008). Five paradoxes on legal certainty and the lawmaker. *Legisprudence*, *2*(1), 47–66. https://doi.org/10.1080/17521467.2008.11424673
Quintais, J. P., Appelman, N., & Ó Fathaigh, R. (2023). Using terms and conditions to apply fundamental rights to content moderation. *German Law Journal*, *24*, 881–911. https://doi.org/10.1017/glj.2023.53
Ranchordas, S. (2021). Experimental regulations and regulatory sandboxes: Law without order? *Law & Method, 1*, 1–23. https://doi.org/10.5553/REM/.000064
Roberts, S. (2016). Commercial content moderation: Digital laborers' dirty work. In S. U. Noble & B. Tynes (Eds.), *The intersectional internet: Race, sex, class and culture online*, (pp. 147–159). New York: Peter Lang US. https://ir.lib.uwo.ca/commpub/12/
La Rue, F. (2011). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. Human Rights Council, Seventeenth Session. https://digitallibrary.un.org/record/706331

Sander, B. (2020). Freedom of expression in the age of online platforms—The promise and pitfalls of a human rights-bsed approach to content moderation. *Fordham International Law Journal*, *43*(4), 939–1006. https://ir.lawnet.fordham.edu/ilj/vol43/iss4/3

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, *87*(3), 1085–1140. https://heinonline.org/HOL/P?h=hein.journals/flr87&i=1118

Sinclair, D. (2002). Self-regulation versus command and control? Beyond false dichotomies. *Law & Policy*, *19*(4), 529–559. https://doi.org/10.1111/1467-9930.00037

Suzor, N. (2019). *Lawless—The secret rules that govern our digital lives*. Cambridge University Press. https://osf.io/preprints/socarxiv/ack26/

Taylor, L. (2021). Public actors without public values: Legitimacy, domination and the regulation of the technology sector. *Philosophy & Technology*, *34*, 897–922. https://doi.org/10.1007/s13347-020-00441-4

United Nations. (2011). Human rights committee, general comment No. 34. CCPR/C/GC/34. https://www2.ohchr.org/english/bodies/hrc/docs/GC34.pdf

Weerts, S. (2019). The law and the principle of legality. In A. Ladner, N. Soguel, Y. Emery, S. Weerts, & S. Nahrath (Eds.), *Swiss public administration* (pp. 69–86). Springer Nature.

Xanthaki, H. (2014). *Drafting legislation—Art and technology of rules for regulation*. Hart Publishing.

York, J. (2022). *Silicon values—The future of free speech under surveillance capitalism*. Verso.

York, J. C., & Zuckerman, E. (2019). Moderating the public sphere. In R. K. Jørgensen (Ed.), *Human rights in the age of platforms*, (pp. 137–161). MIT Press. https://doi.org/10.7551/mitpress/11304.003.0012

Zakrzewski, C., Lima, C., & Harwell, D. (2023). What the Jan. 6 probe found out about social media, but didn't report. *Washington Post*. https://www.washingtonpost.com/technology/2023/01/17/jan6-committee-report-social-media/

Zuboff, S. (2019). *The age of surveillance capitalism*. Profile Books.

How to cite this article: Fasel, M., & Weerts, S. (2024). Can Facebook's community standards keep up with legal certainty? Content moderation governance under the pressure of the Digital Services Act. *Policy & Internet*, 1–19. https://doi.org/10.1002/poi3.391