


ARTICLE

# Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation

Niklas Eder\* 

University of Oxford, Centre for Socio-Legal Studies, Oxford, United Kingdom  
Email: [niklaseder@gmail.com](mailto:niklaseder@gmail.com)

(Received 03 July 2023; accepted 15 April 2024)

## Abstract

The European Union’s Digital Services Act (DSA) introduces a new regulatory approach to address the societal harms of online platforms: Systemic risk assessments. While a core component of the DSA, the regulation only outlines the standards and processes governing systemic risk assessments in broad strokes. It remains unclear what these systemic risk assessments will entail in practice. This Article develops a proposal of how systemic risk assessments should be implemented. It situates systemic risk assessments as a critical step toward platform accountability as they address societal harms, while existing approaches, such as remedy mechanisms, only protect user rights. Engaging with intangible harms and regulating speech and public discourse, risk assessments also entail significant challenges. Conventional reference points for content moderation regulation, such as terms and conditions, contractual freedom, fundamental rights and expertise, do not provide practical and legitimate bases to concretize risk assessment obligations. Public actors, such as the European Commission, should refrain from defining substantive standards, too, as they are directly bound by freedom of expression guarantees. Instead, the Article argues, the Commission should foster a procedural framework, a “virtuous loop,” which empowers civil society and allows it to specify and refine the standards governing systemic risks over time. Developing this framework, the Article explains how systemic risk assessment can fix “multistakeholderism,” and “multistakeholderism,” in turn, can help make systemic risk assessments work.

**Keywords:** Digital Services Act; systemic risk assessments; social media; online platforms; platform accountability; platform protection; virtuous loop; multistakeholderism; content moderation

## A. Introduction

Systemic risk assessments have the potential to significantly transform the landscape of content moderation. By requiring platforms to engage in a “proactive and continuous form of governance,” they promise to address content moderation where it really matters.<sup>1</sup> While in the United States, where scholarship on systemic approaches flourishes—and any such efforts depend on the goodwill of platforms to function—the European Union has introduced obligations for

\*Niklas Eder is a Digital Policy Postdoctoral Researcher at the Centre for Socio-Legal Studies at Oxford University; a Visiting Lecturer at the Dickson Poon School of Law at King’s College London; and the Co-Founder of User Rights.

<sup>1</sup>Evelyn Douek, *Content Moderation as Systems Thinking*, 136 HARV. L. REV. 526, 526 (2022) (displaying a critique of Douek and an overview on systemic approaches). See also Kate Klonick, *Of Systems Thinking and Strawmen*, 136 HARV. L. REV. F. 339 (2023) (explaining the importance of systemic perspectives); Rebecca Tushnet, *A Hobgoblin Comes for Internet Regulation*, VERFASSUNGSBLOG (Feb. 19, 2024), <https://verfassungsblog.de/a-hobgoblin-comes-for-internet-regulation/>.

platforms to assess and mitigate systemic risks in its Digital Services Act (DSA).<sup>2</sup> The EU, yet again, passed an ambitious law regulating the digital future, likely with vast extraterritorial effect.<sup>3</sup> It is still uncertain whether the DSA will effectively strengthen platform accountability, or whether the EU merely released yet another European paper tiger. Much depends on how the Commission will implement the new law.<sup>4</sup>

Systemic risk assessments constitute a new regulatory approach to tame the power of platforms.<sup>5</sup> They are supposed to complement individual rights and remedy mechanisms, which are part of what is often described as the “rule of law” approach.<sup>6</sup> Individual rights and remedy mechanisms, which allow users to challenge enforcement decisions, already exist, but are often viewed as ineffective. Many of the most important components of content moderation are out of reach for individual remedy mechanisms, and individual rights fail to capture all relevant societal interests. Articles 34 and 35 of the DSA aim to address these shortcomings by establishing obligations for platforms to assess and mitigate systemic risk.<sup>7</sup> Platforms must analyze the harms of their content moderation practices, including, according to the wording of Article 34, the “amplification of content,” the “design of recommender systems” and generally their “content moderation systems.”<sup>8</sup> They must analyze how their moderation practices systemically effect areas such as civic discourse, electoral processes, public health, and security. The DSA vaguely defines the sources of risk, Article 34 (2) DSA, the kinds of risks, Article 34 (1) DSA, and the sorts of mitigation measures platforms must consider.<sup>9</sup> Beyond that, systemic risk assessments in content moderation are unknown territory for industry, academics, and regulators alike.

<sup>2</sup>Commission Regulation 2022/2065, 2022 O.J. (L 277). See generally *Digital Services Act*, EUR-LEX (Feb. 15, 2023), <https://eur-lex.europa.eu/EN/legal-content/summary/digital-services-act.html> (giving a brief overview on the DSA, including that it amended Council Directive 2000/31, 2000 O.J. (L 178) 1 (EC), which was the EU’s previous regulation on “electronic commerce”); MARTIN HUSOVEC & IRENE ROCHE LAGUNA, *DIGITAL SERVICES ACT: A SHORT PRIMER* (2022) (providing background on the DSA); Lea Katharina Kumkar, *The Digital Services Act: Introduction and Overview*, in 1 *CONTENT REGULATION IN THE EUROPEAN UNION: THE DIGITAL SERVICES ACT 1*, 1 (Antje von Ungern-Sternberg ed., May 23, 2023) (giving a proposal of federal rules in the U.S.); Rory Van Loo, *Federal Rules of Platform Procedure*, 88 *U. CHI. L. REV.* 829 (2020) (providing context for the DSA in the European constitutional order and other regulatory areas); Giovanni De Gregorio & Pietro Dunn, *The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age*, 59 *COMMON MKT. L. REV.* 473 (2022).

<sup>3</sup>Daphne Keller, *The EU’s New Digital Services Act and the Rest of the World*, *VERFASSUNGSBLOG* (Nov. 7, 2022), <https://verfassungsblog.de/dsa-rest-of-world/>. See generally Ioanna Tourkochoriti, *The Digital Services Act and the EU as the Global Regulator of the Internet*, 24 *CHI. J. INT’L L.* 129 (2023); Anupam Chander, *When the Digital Services Act Goes Global*, 38 *CHI. L. REV.* 1067 (2023); Laureline Lemoine & Mathias Vermeulen, *The Extraterritorial Implications of the Digital Services Act*, *DSA OBSERVATORY* (Nov. 1, 2023), <https://dsa-observatory.eu/2023/11/01/the-extraterritorial-implications-of-the-digital-services-act/>; Giancarlo Frosio & Christophe Geiger, *Towards a Digital Constitution: How the Digital Services Act Shapes the Future of Online Governance*, *VERFASSUNGSBLOG* (Feb. 20, 2024), <https://verfassungsblog.de/towards-a-digital-constitution/>.

<sup>4</sup>See Martin Husovec, *Will the DSA Work?: On Money and Effort*, *VERFASSUNGSBLOG* (Nov. 9, 2022), <https://verfassungsblog.de/dsa-money-effort/> (giving an overview on important questions about the implementation of the DSA).

<sup>5</sup>See Gerhard Wagner, Martin Eifert, Axel Metzger & Heike Schweitzer, *Taming the Giants: The DMA/DSA Package*, 58 *COMMON MKT. L. REV.* 987 (2021). See generally Florence G’Sell, *The Digital Services Act: A General Assessment*, in *CONTENT REGULATION IN THE EUROPEAN UNION*, *supra* note 2, at 85, 85 (presenting a proposal of federal rules in the United States); Rory Van Loo, *Federal Rules of Platform Procedure*, 88 *U. CHI. L. REV.* 829 (2020).

<sup>6</sup>See generally Rachel Griffin, *Public and Private Power in Social Media Governance: Multistakeholderism, the Rule of Law and Democratic Accountability*, 14 *TRANSNAT’L LEGAL THEORY* 46 (Aug. 20, 2022) (explaining the DSA’s individual remedy obligations). See also Kuczerawy Aleksandra, *Remedying Overremoval: The Three-Tiered Approach of the DSA*, *VERFASSUNGSBLOG* (Nov. 3, 2022), <https://verfassungsblog.de/remedying-overremoval/>.

<sup>7</sup>See David Sullivan & Jason Pielemeier, *Unpacking “Systemic Risk” Under the EU’s Digital Service Act*, *TECH POLICY PRESS* (July 19, 2023), <https://www.techpolicy.press/unpacking-systemic-risk-under-the-eus-digital-service-act/> (giving an overview of the situation and discussing the implementation of risk assessments). See generally *IMPLEMENTING RISK ASSESSMENTS UNDER THE DIGITAL SERVICES ACT*, *GLOBAL NETWORK INITIATIVE AND DIGITAL TRUST & SAFETY PARTNERSHIP* (2024).

<sup>8</sup>See Commission Regulation 2022/2065, *supra* note 2, at art. 34.

<sup>9</sup>*Id.* 2022/2065, at art. 34(1), 34(2).

A systemic approach to regulate content moderation could be a huge leap forward, allowing to hold platforms accountable for their societal impact. It's also a jump into the unknown and it remains uncertain how the obligations will be implemented. The first round of risk assessments was due end of August 2023, four months after the EU Commission designated platforms as “very large online platforms” (VLOPs).<sup>10</sup> These risk assessments were then sent to auditors, the Commission and the Board for Digital Services, which review the risk assessments, develop best practices and guidelines on specific risks—Article 35 (2) b. and (3) DSA—and will potentially require platforms to take alternative mitigation measures or even fine them.<sup>11</sup> The risk assessment and audit reports have not yet been published and will only be published three months after the Commission has received the audit report.<sup>12</sup> As the audit period is one year, the first reports are not expected until Autumn 2024.<sup>13</sup> Meanwhile, the Commission has announced that it has opened an investigation into Tik Tok. Other than that, little is known about its approach to enforcing systemic risk assessment obligations. As the DSA only outlines the basic parameters of the processes by which risk assessments will be developed and reviewed, much uncertainty remains. The success of systemic risk assessments as a regulatory approach for content moderation will ultimately depend on how the process is implemented, on questions of who takes what role in that process, and on how those involved execute their particular functions.

This Article develops a proposal which contributes to the process of transforming systemic risk assessments into a meaningful mechanism. By providing some foundations and outlining one of many possible visions, it hopes to instigate a debate in which many more proposals for a successful implementation of the risk assessment provisions can be developed.

The Article begins by describing the potential and limitations of individual remedy mechanisms under the DSA. It argues that the DSA significantly strengthens individual remedy by expanding what kind of content moderation decisions can be challenged by users. However, despite this expansion, individual rights and remedy mechanisms remain structurally limited. Responding to these limitations requires systemic approaches, such as the ones outlined in the DSA.

The Article then lays out the challenges of systemic approaches to regulating content moderation. It explains why public actors should refrain from defining the concrete standards governing systemic risk assessments and argues that terms of services and contractual freedom do not provide a legitimate normative basis for assessing systemic risks, either. It describes the challenges of relying on fundamental rights and expertise to specify systemic risk obligations.

Responding to the identified challenges, Section D proposes an approach which focuses on civil society involvement and could contribute to making systemic risk assessments a success. It conceptualizes the processes laid out in the DSA as a “virtuous loop”—a loop involving platforms, the Commission, the Board for Digital Services and auditors—which ultimately aims at empowering civil society organizations. It engages with the shortcomings of conventional “multistakeholderism” and explains how systemic risk assessments provide solutions to these shortcomings. It argues that systemic risk assessments could make civil society involvement mandatory, thus making away with the flaws of extensive discretion of platforms with view to when and how to involve civil society stakeholders and pressuring them to account for positions articulated in civil society engagement. The Article finally proposes a role for social media councils and argues that they could help make civil society involvement fairer and could translate various positions into concrete and implementable solutions.

<sup>10</sup>See *Digital Services Act: Commission Designates First Set of Very Large Online Platforms and Search Engines*, EUR. COMM'N (Apr. 25, 2023), [https://ec.europa.eu/commission/presscorner/detail/en/IP\\_23\\_2413](https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413) (naming seventeen VLOPs and two “very large online search engines” (VLOSEs)). See also Commission Regulation 2022/2065, *supra* note 2, at art. 33(6).

<sup>11</sup>See Commission Regulation 2022/2065, *supra* note 2, at art. 35(2), 35(3), 74.

<sup>12</sup>See *id.* at art. 42(4).

<sup>13</sup>*Delegated Regulation on Independent Audits Under the Digital Services Act*, EUR. COMM'N (Oct. 20, 2023) <https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act> (explaining rules on the performance of audits for VLOPs and VLOSEs).

## B. The Future of Content Moderation Regulation is Systemic

This Section describes the idea underlying existing individual remedy mechanisms. It explains why individual remedy might become significantly more impactful because of the DSA and why it still remains a structurally limited form of platform accountability.

### 1. The Idea Behind Individual Remedy

Individual remedy mechanisms allow users to contest enforcement decisions that platforms impose on users for violating the policies platforms establish for speech. The range of punishments levied on a user can range from the removal of a post to temporary or even permanent account suspensions. Individual remedy can be conceived as part of a broader approach to hold platforms accountable, which has been characterized as the rule of law approach.<sup>14</sup> Individual remedy has been celebrated as an important step in protecting users' rights to free expression, constituting an important component of what has been described as "digital constitutionalism."<sup>15</sup> Conceptually, the justification for requiring platforms to provide individual remedy mechanisms is straightforward: As platforms increase in size and power, their decisions to sanction users become comparable to decisions a state imposes on its citizens, and their policies resemble laws,<sup>16</sup> making them the "new governors."<sup>17</sup> Therefore, similar to citizens having rights to contest actions of the state, users should have rights to contest actions of platforms.

From this analogy follows the idea that, although fundamental rights are originally intended to constrain the power of the state and conventionally only apply in the *vertical* relation between a citizen and the state, the extraordinary power of platforms justifies applying them in the *horizontal* relation between a user and platforms. More concretely, the idea is to apply *the negative dimension* of fundamental rights in a horizontal relationship, meaning the dimension which allows users to contest sanctions which platforms impose on them for the alleged violation of a rule. This idea has shaped self-regulatory approaches of platforms, which often provide appeal mechanisms to users for decisions to remove their content. The idea that fundamental rights may have a horizontal dimension is not new and a common subject of debate in European constitutional law.<sup>18</sup> National Courts, such as the German Federal Court of Justice, have begun relying on a horizontal dimension of national fundamental rights to address the overwhelming power of platforms with view to their users, too.<sup>19</sup> The DSA appears to enshrine this practice in European secondary

<sup>14</sup>Griffin, *supra* note 6, at 54–58.

<sup>15</sup>See, e.g., Nicolas Suzor, *Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms*, 4 SOC. MEDIA & SOC'Y (2017) (explaining the notion of digital constitutionalism); Giovanni De Gregorio & Roxana Radu, *Digital Constitutionalism in the New Era of Internet Governance*, 30 INT'L J.L. & INFO. TECH. 68, 69 (2022). See generally Edoardo Celeste, *Digital Constitutionalism: A New Systematic Theorization*, 33 INT'L REV. L. COMPUTS. & TECH. 76 (2019); Giovanni De Gregorio, *The Rise of Digital Constitutionalism in the European Union*, 19 INT. J. CONST. L. 41 (2020).

<sup>16</sup>See generally Molly K. Land, *The Problem of Platform Law: Pluralistic Legal Ordering on Social Media*, in OXFORD HANDBOOK OF GLOBAL LEGAL PLURALISM 975, (Paul Schiff Berman ed., 2019). See also Orly Lobel, *The Law of the Platform*, 101 MINN. L. REV. 87 (Mar.7, 2016).

<sup>17</sup>See generally Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2017); TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018); Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27 (2019).

<sup>18</sup>See, e.g., Eleni Frantziou, *The Horizontal Effect of Fundamental Rights in the European Union: A Constitutional Analysis* (2019); Eleni Frantziou, *The Horizontal Effect of the Charter: Towards an Understanding of Horizontality as a Structural Constitutional Principle*, 22 CAMBRIDGE Y.B. EUROP. LEGAL STUD. 208 (2020); Hans D. Jarass, *Die Bedeutung der Unionsgrundrechte unter Privaten*, 2 ZEITSCHRIFT FÜR EUROPÄISCHES PRIVATRECHT 310, 311 (2017); CHRISTOPHER UNSELD, ZUR BEDEUTUNG DER HORIZONTALWIRKUNG VON EU-GRUNDRECHTEN (2018).

<sup>19</sup>See Tobias Lutz, *Plattformregulierung durch AGB-Kontrolle?: Der Beitrag des Zivilrechts zum Grundrechtsschutz auf Online-Plattformen*, VERFASSUNGSBLOG (July 30, 2021), <https://verfassungsblog.de/facebook-agb-kontrolle/>, (giving a brief overview on the case law of the German Federal Court). See also Ruth Janal, *Impacts of the Digital Services Act on the Facebook*

law.<sup>20</sup> It is still largely unclear if and how secondary law can command the application of European fundamental rights, whose scope is defined in the Charter of Fundamental Rights—Article 51 (1) Charter of Fundamental Rights<sup>21</sup>—and what follows from a potential application with view to the substance of decisions taken by platforms.<sup>22</sup> Nonetheless, the DSA in any case creates clear procedural obligations: Platforms will have to provide a textual foundation for interventions against users in their terms and conditions, per Article 14 DSA; platforms will need to notify users with a statement of reasons when intervening, per Article 17 DSA; platforms will need to provide internal remedy, per Article 20 DSA: and platforms will be required to cooperate with external remedy mechanisms, per Article 21 DSA, meaning that users can appeal decisions to sanction them for violative content to independent bodies.<sup>23</sup>

## II. The New Potential of Individual Remedy: Demotion of Content

The DSA obliging platforms to provide individual remedy mechanisms is a step forward, for two reasons: One, providing remedy is no longer a courtesy of platforms but a legal requirement. Two, the DSA defines very, perhaps shockingly broadly, what counts as a sanction which triggers the individual remedy framework—meaning that it requires a foundation in the terms and conditions, a statement of reasons and triggers internal and external remedy mechanisms. Any “restrictions that they impose in relation to the use of their service,”<sup>24</sup> including the demotion of content<sup>25</sup> and shadow-banning<sup>26</sup> count as sanctions.<sup>27</sup> The DSA’s definition of what is eligible as a “sanction,” and subsequently able to be challenged, is broad by design. This approach creates grounds for contestation for all the measures, explicit or covert, employed by platforms to enforce against content violating their policies.<sup>28</sup> The DSA basically codifies an idea that courts often took years or decades to develop: Measures that can be considered sanctions need not be as formal as an intervention or with the intent to sanction an individual. Instead, courts have championed an understanding of sanctions to include measures that practically and detrimentally affect individuals.

This is the “Dassonville” moment for platforms, comparable to the pathbreaking decision of the European Court of Justice declaring that any measure which is “capable of hindering directly or indirectly, actually or potentially intra-community trade are to be considered as measures having an effect equivalent to quantitative restrictions,” meaning that these measures constitute encroachments on trade rights and require a justification.<sup>29</sup> The ECJ’s broad understanding of what counts as an encroachment of individuals’ free trade rights paved the path for the ECJ to

---

“Hate Speech” German Federal Court of Justice, in *CONTENT REGULATION IN THE EUROPEAN UNION*, *supra* note 2, at 119 (discussing how the DSA might impact the approach of the German Federal Court).

<sup>20</sup>See Janal, *supra* note 19, at 124–31 (giving a more detailed discussion of the obligations under the DSA and how they relate to the German Federal Court’s case law). See also Commission Regulation 2022/2065, *supra* note 2, at arts. 14, 15, 17, 20, 21.

<sup>21</sup>See Mattias Wendel, *Taking or Escaping Legislative Responsibility? EU Fundamental Rights and Content Regulation under the DSA*, in *CONTENT REGULATION IN THE EUROPEAN UNION*, *supra* note 2, at 59. See also Thomas Wischmeyer & Peter Meißner, *Horizontalwirkung der Unionsgrundrechte – Folgen für den Digital Services Act*, 2023 NJW 2673 (2023).

<sup>22</sup>See generally João Pedro Quintais, Naomi Appelman & Ronan Ó Fathaigh, *Using Terms and Conditions to Apply Fundamental Rights to Content Moderation*, 24 GERMAN L.J. 881 (2022) (providing extensive discussions of that question). See also Janal, *supra* note 19, at 134–135 (giving an analysis of if and how the Unfair Contract Terms Directive might apply).

<sup>23</sup>See Commission Regulation 2022/2065, *supra* note 2, at arts. 14, 17, 20, 21.

<sup>24</sup>See *id.* at art. 14(1).

<sup>25</sup>See *id.* at art. 3(t) (providing the definition of “content moderation”).

<sup>26</sup>See *id.* at recital 55 (giving the definition of “shadow banning”).

<sup>27</sup>See Janal, *supra* note 19, at 126–27 (showing the broad scope of DSA’s individual remedy framework). See also G’Sell, *supra* note 5, at 94–97.

<sup>28</sup>See Eric Goldman, *Content Moderation Remedies*, 28 MICH. TECH. L. REV. 1 (2021) (giving an overview on the broad spectrum of enforcement measures platforms use).

<sup>29</sup>Case 8-74, *Procureur du Roi v. Dassonville*, 1974 E.C.R. 00837. See generally Robert Schutze, *Re-Reading ‘Dassonville’: Meaning and Understanding in the History of European Law*, 24 EUR. L.J. 376 (2018) (analyzing the *Dassonville* decision and its meaning for the EU legal order).



become the EU's "engine of integration," as it could declare inapplicable any national measure that it found to fall under this extremely broad definition.<sup>30</sup> Any such measure would need to serve a legitimate interest and satisfy necessity and proportionality requirements, and the ECJ is the final authority to assess whether such measures are indeed justified. Applying a broad definition as to what counts as an encroachment of individual rights changed the European legal order, and it will now change content moderation.

The DSA's definition opens demotion -related content to user challenge, providing an urgently needed remedy: Demotion without justification or notification is a powerful and opaque form of content moderation that has lacked effective accountability mechanisms.<sup>31</sup> For example, Elon Musk explained that as CEO of Twitter, he would no longer remove content but "max deboost[] & [sic] demonetize[]" it instead, likely to avoid all the scrutiny that comes with removal.<sup>32</sup> The EU proposes its own answer to the question of whether "freedom of speech" is also "freedom of reach,"<sup>33</sup> introducing for the first-time requirements for platforms to justify reducing the reach of content. This means that in the EU, demotion is treated as a minus to removal: The normative framework is identical, only that the threshold to justify demotion will likely be lower than justifying removal. Besides strengthening individual remedy, this also raises difficult questions, such as how to delimit demotion from amplification.<sup>34</sup>

### III. Persistent Shortcomings

While individual remedy will become significantly more meaningful due to the broad definition of what counts as a restriction, its contribution to rectifying the societal impact of content moderation remains structurally limited.<sup>35</sup> The structural deficits of individual remedy consist, one, in their still limited scope: Crucial components of content moderation, such as amplification of content, the design of recommender systems and the newsfeed, clearly of significant importance for social media,<sup>36</sup> remain out of reach for individual remedy. Individual remedy does not address the algorithmic infrastructure, which ultimately makes a platform what it is. Two, many societal harms do not manifest themselves in the violation of individual rights, such as impacts on civic discourse, electoral processes or public health and security. Individual remedy only empowers users of platforms, although non-users equally suffer the detrimental consequences of content moderation. Three, content moderation at scale happens through automated means, and the availability of individual remedy, understood as an additional review through humans, cannot match the scale of automated decisions. Four, a regulatory approach which solely builds on individual remedy ultimately puts the burden of holding platforms accountable on the individual.

<sup>30</sup>See Eric Stein, *Lawyers, Judges, and the Making of a Transnational Constitution*, 75 AM. J. INT'L L. 1, 1 (1981) (explaining the role of the European Court of Justice in European integration); HJALTE RASMUSSEN, ON LAW AND POLICY IN THE EUROPEAN COURT OF JUSTICE: A COMPARATIVE STUDY IN JUDICIAL POLICYMAKING (1986). See generally J.H.H. Weiler, *The Transformation of Europe*, 100 YALE L. J. 2405 (1991); JUDICIAL ACTIVISM AT THE EUROPEAN COURT OF JUSTICE (Mark Dawson, Bruno De Witte & Elise Muir eds., 2013) (giving a perspective from a more recent work).

<sup>31</sup>Tarleton Gillespie, *Do Not Recommend? Reduction as a Form of Content Moderation*, 8 SOC. MEDIA & SOC'Y 1 (2022) (talking about demotion and non-recommendation as content moderation measures).

<sup>32</sup>See @Elonmusk, X, (Nov. 18, 2022, 1:31 PM) <https://x.com/elonmusk/status/1593673339826212864>.

<sup>33</sup>See Renée DiResta, *Free Speech Is Not the Same As Free Reach*, WIRED (Aug. 30, 2018), <https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/>.

<sup>34</sup>See Paddy Leerssen, *An End to Shadow Banning? Transparency Rights in the Digital Services Act Between Content Moderation and Curation*, 48 COMPUT. L. & SEC. REV. 105790, 1 (2022).

<sup>35</sup>See, e.g., Rachel Griffin, *Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality*, 2 EUR. L. OPEN 30, 30 (2022) (discussing these shortcomings more extensively); Griffin, *supra* note 6, at 55–76; Douek, *supra* note 1, at 43.

<sup>36</sup>See Tarleton Gillespie, *The Relevance of Algorithms*, in MEDIA TECHNOLOGIES: ESSAYS ON COMMUNICATION, MATERIALITY, AND SOCIETY 167, 167 (Tarleton Gillespie, Pablo J. Boczkowski & Kirsten A. Foot eds., 2013).

The transition from individual rights to systemic risks assessments entails a paradigm change, it introduces a very different approach to regulating tech. Previous legislation, such as the General Data Protection Regulation (GDPR), has focused heavily on individual rights and remedies, building on a neoliberal construction of individual freedom and responsibility.<sup>37</sup> The privacy theory that underpins the individual-centered regulatory approach that characterizes the GDPR is based on the idea of the “liberal self” that possesses “the capacity for rational deliberation and choice” and is “capable of exercising its capacities.”<sup>38</sup> This approach finds its continuation in proposals to respond to machine learning–based decision making with transparency, explainability and due process requirements<sup>39</sup>—the effectiveness of which can be questioned.<sup>40</sup> Systemic risk assessments can be distinguished from regulatory approaches that assume that individuals will read terms and conditions, make meaningful choices when clicking through pop-up windows, and that this will keep large corporations in check. They shift responsibility from the individual to the state. They take the burden of holding platforms to account off the shoulders of individuals. In much the same way that the state is expected to ensure that the products we use and the food we eat are safe, the medicines we take are tested, and the toys our children play with are free of toxic substances, the welfare state is now expected to police the harmful effects of social media.

This Section argued that the DSA significantly strengthens individual remedy by expanding what kind of content moderation decisions can be challenged by users. Despite this expansion, individual remedy has its limitations. Responding to these limitations requires an understanding of systemic harms and regulatory approaches to address these harms. The following Section develops such an understanding.

### C. The Challenges of Assessing Systemic Risks

To develop an adequate procedural and normative framework for risk assessments, we first need to understand the challenges that a systemic approach to content moderation entails. This Section outlines some of these challenges. It first explains why public actors should refrain from defining the concrete standards governing systemic risk assessment themselves. It then argues that terms of services and contractual freedom do not provide a legitimate normative basis for assessing systemic risks. Finally, it describes the challenges of relying on fundamental rights to substantiate systemic risk obligations.

<sup>37</sup>See Niklas Eder, *Beyond Automation: Machine Learning-Based Systems and Human Behavior in the Personalization Economy*, 25 STAN. TECH. L. REV. 1, 37–40 (2021).

<sup>38</sup>See Julie E. Cohen, *What Privacy Is For*, 126 HARV. L. REV. 1904, 1907 (2013).

<sup>39</sup>See generally JESSICA FJELD, NELE ACHTEN, HANNAH HILLIGOSS, ADAM CHRISTOPHER NAGY, MADHULIKA SRIKUMAR, BERKMAN KLEIN CENTER, *PRINCIPLED ARTIFICIAL INTELLIGENCE: MAPPING CONSENSUS IN ETHICAL AND RIGHTS-BASED APPROACHES TO PRINCIPLES FOR AI* (2020) (giving an overview on rights based responses to automated decision-making). See also Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503 (2013); Sandra Wachter, Brent Mittelstadt, Luciano Floridi, *Transparent, Explainable, and Accountable AI for Robotics*, 2 SCI. ROBOTICS 1 (2017); Andrew Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018); Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1 (2019); John Zerilli, Alistair Knott, James Maclaurin, Colin Gavaghan, *Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?*, 32 PHIL. & TECH. 661 (2019).

<sup>40</sup>Eder, *supra* note 37, at 37 (explaining limitations). See also Lilian Edwards & Michael Veale, *Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18 (2017); Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1 (2017); Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT’L DATA PRIV. L. 76 (2017); Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC’Y 973 (2018).

### 1. Regulating Speech and Public Discourse

A comparison can help illustrate the challenges of regulating content moderation and the role that public actors should and should not play.

The starting point for such a comparison is that there is nothing new about regulating the operation of private companies to mitigate social harm. Environmental regulation, for example, has the goal of limiting the detrimental impact of production processes on our climate. How is regulating industrial processes generating CO<sub>2</sub> different from regulating content moderation?<sup>41</sup>

The first point this comparison helps us clarify is the reasons justifying regulators intervening in platforms' content moderation practices. When a company runs a factory and emits CO<sub>2</sub>, it cannot argue that the way it conducts its business is a purely private matter. Emissions affect everyone, which makes the processes creating them a public matter and which justifies regulating them. The regulation of content moderation and systemic risks is based on the same idea. Because social media, and more particularly the content on social media including its curation, affects public discourse, elections and public health, and produces-real world violence, content moderation is not a private matter. The impact of content moderation on our society justifies regulating content moderation.

Secondly, the comparison illustrates the particularities of public institutions regulating speech and public discourse.<sup>42</sup> We may take no issue with the state defining concrete standards on how much CO<sub>2</sub> a company may emit. When the state regulates content moderation, however, it indirectly regulates speech and it shapes public discourse as it takes place online. By defining concrete standards for content moderation, governments would, to limit the power of platform, unduly expand their own power. This is the conundrum of regulating content moderation.<sup>43</sup> Regulating content moderation should not be a back door to regulating speech. In a democracy, public discourse shapes elections which ultimately determine who holds power. Democratic governments and more broadly public institutions should not be able to manipulate public discourse, and content moderation shaping that public discourse, in their favor. There is, therefore, a natural and healthy skepticism to endowing the EU Commission with a mandate to dictate platforms how they should moderate and curate content.<sup>44</sup>

A third aspect, which the comparison allows us to see more clearly, is why regulating content moderation is so hard—and why we need a complex process to succeed in this task. The comparison helps us specify the challenges of assessing and mitigating the systemic risks of content moderation. When regulating CO<sub>2</sub> emissions, emissions can be measured in an objective manner and their effects can be proven in an empiric, scientific process. Regulation can establish clear, objective and easily measurable thresholds for what is allowed and what is not allowed, how much CO<sub>2</sub> can and cannot be emitted. Regulating content moderation is very different from regulating emissions. The societal harms of content moderation are not easily quantifiable. We lack objective standards and empirical measures to assess these harms; thus hard questions arise: What are the emissions or public harms produced by social media? How to measure these emissions and harms? What causes them? What conditions must content moderation fulfil to minimize emissions and mitigate harm? These questions are hard to answer, and there are no objectively correct solutions. And as we deal with speech and public discourse, public institutions

<sup>41</sup>See Douek, *supra* note 1, at 71 (discussing this comparison).

<sup>42</sup>See Christian Fuchs, *Social Media and the Public Sphere*, 12 TRIPLEC: COMM'N CAPITALISM & CRITIQUE 57 (Feb. 19, 2014) (discussing this point in-depth).

<sup>43</sup>See Evelyn Douek, *What Kind of Oversight Board Have You Given Us?*, U. CHI. L. REV. ONLINE (May 11, 2020) <https://lawreview.uchicago.edu/online-archive> (referring to it as “Gordian knot” as well).

<sup>44</sup>See MARTIN HUSOVEC, THE DIGITAL SERVICE ACT'S RED LINE: WHAT THE COMMISSION CAN AND CANNOT DO ABOUT DISINFORMATION (2024) (displaying a critical discussion of what powers the EU Commission has and does not have under the DSA in the context of disinformation). See also Suzanne Vergnolle, *A New European Enforcer?: Why the European Commission Should Not Stand alone in the Enforcement of the Digital Services Act*, VERFASSUNGSBLOG (May 23, 2023), <https://verfassungsblog.de/a-new-european-enforcer/>; Chander, *supra* note 3.



should restrain from dictating standards in a similar way as they might dictate emission limitations. This is why the processes around risk assessments are complex and why we need a procedural framework which allows us to define and refine standards over time.

## II. The Limited Legitimizing Power of Terms of Service

Another point, which the comparison illustrates, is that the reference point for systemic risk assessments is not the user base of a platform, but the public more broadly. One important idea underlying systemic risk assessments is that the way platforms moderate content affects not only the users of platforms, but also all other people and society at large. As the impact of content moderation is not only internal but also external, producing real-world harms affecting fundamental rights and the functioning of democracy, problems pertaining to content moderation cannot simply be resolved between a company and its users. It is for that very reason, that the idea that user choice pertaining to the content that they see on their newsfeed solves the problem of content moderation is flawed, and why relying on user polls to decide hard questions is not convincing either. Anyone involved in systemic risk assessments, including social media councils, must focus on the impact of content moderation on the general public, rather than only on the platforms' users.

Based on this observation, we can identify important differences in the normative framework which underlies individual remedy and a normative framework we might find suitable for systemic risk assessments. Individual remedy allows users to contest platforms' enforcement decisions. The starting point of any claim brought under the individual remedy framework is that an enforcement decision did not align with the terms and conditions. Platforms, in turn, invoke their terms and conditions to justify decisions that affect their users. The underlying assumption here is that basing enforcement decisions on terms and conditions is legitimate because the user agreed to these terms and conditions. Being bound by them is a consequence of users exercising their contractual freedom. One can contest this assumption, arguing that contractual freedom only legitimizes the relation between two parties if both contracting parties are actually on an equal footing and that, in fact, platforms are often so big and market-dominating that this balance does not exist at all, leaving the user no real choice. However, one does not have to accept this argument, or engage in that argument at all to see why terms and conditions are largely irrelevant when it comes to justifying systemic risks. Systemic risks affect everyone, not just the users, including all people who never agreed to the terms and conditions. In the normative framework for systemic risk assessments, terms and services are not the solution to the problem but very much a part of it.

The regime established in Articles 34 and 35 DSA requires to assess whether terms and conditions cause systemic risks.<sup>45</sup> While the DSA does not directly interfere with platforms' ability to write their own terms and conditions, systemic risk assessments constrain their discretion.<sup>46</sup> Cementing an everything-goes approach in X's terms and conditions would not shelter the platform from obligations under the DSA. Platforms will invoke their right to conduct a business to insist that they can design their terms and services as they please. But this does not prevent EU institutions to address terms and conditions as a source of risks, as explicitly named in Article 34(2) DSA and presents us with the challenge of balancing a platform's right to conduct a business with the objective of mitigating societal harms.

<sup>45</sup>See Commission Regulation 2022/2065, *supra* note 2, at arts. 34–35.

<sup>46</sup>See Quintais, Appelman & Ó Fathaigh, *supra* note 22 (discussing in depth how Article 14(4) of the DSA, which requires platforms to apply and enforce restrictions "with due regard to fundamental" rights, might create obligations pertaining to terms of service).

### III. The Challenges of Applying Human and Fundamental Rights

We can look at individual remedy processes to reflect on if and how human or fundamental rights could provide foundations for systemic risk assessments. Over the last few years, the idea that international human rights law should guide content moderation practices has been discussed by academics, civil society organizations and UN rapporteurs alike.<sup>47</sup> Social media councils involved in individual remedy mechanisms have developed approaches to review enforcement decisions based on human rights.<sup>48</sup> Exploring how international human rights law evolves, scholars analyze if and how it imposes obligations on intermediaries, including platforms.<sup>49</sup> Relatedly, the DSA requires platforms to account for fundamental rights in their enforcement processes—Article 14 DSA—and requires them to assess negative effects on fundamental rights on a systemic level—Article 34 DSA.<sup>50</sup> With secondary law aiming to define the scope of application of what has the status of EU primary law, the EU Charter, it is still largely unclear what the application of European fundamental rights to content moderation will look like.<sup>51</sup>

While relying on human or fundamental rights in the individual remedy framework typically means applying them *horizontally in a negative dimension*, applying them to systemic risks would mean to apply them *horizontally in a positive dimension*. That means that platforms would not only be expected to not violate the freedom of expression of their users by removing their content, but that they would be expected to take active steps in order to assure that their content moderation practices do not have detrimental effects of fundamental rights generally, including effects on fundamental rights of people who are not their users, but who can still be affected by online content, such as hate speech. This entails even more uncertainties. It is difficult to assess whether a post should be removed based on fundamental rights, and it is even more difficult to assess whether a particular way to amplify content “*impacts fundamental rights*.” Examining international human rights law, scholars have assessed whether states have positive obligations, meaning that they should pass laws that protect human rights.<sup>52</sup> The question the DSA raises is whether similar obligations should be extended to private actors, such as social media platforms. In the world of international human rights law, the idea that businesses have human rights responsibilities is already well established, and addressed under the notion of human rights responsibilities of businesses, which are defined in the non-binding UN Guiding Principles on Business and Human Rights. Civil society organizations, human rights consultancies and UN agencies and officials have worked to specify human rights due diligence obligations in the context of freedom of expression and social media for years.<sup>53</sup> To better grasp what the obligations

<sup>47</sup>See, e.g., Barrie Sander, *Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation*, 43 FORDHAM INT'L L.J. 939 (2019). See also Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26 (2018); Susan Benesch, *But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies*, YALE J. REG. ONLINE BULL. (2020); Evelyn Mary Aswad, *To Protect Freedom of Expression, Why Not Steal Victory from the Jaws of Defeat?*, 77 WASH. & LEE L. REV. 609 (2019); David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Rep. on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018), para. 41–43.

<sup>48</sup>See generally Laurence Helfer & Molly K. Land, *The Facebook Oversight Board's Human Rights Future*, 44 CARDOZO L. REV. 2234 (2022).

<sup>49</sup>See, e.g., Dafna Dror-Shpoliansky & Yuval Shany, *It's the End of the (Offline) World as We Know It: From Human Rights to Digital Human Rights – A Proposed Typology*, 32 EUR. J. INT'L L. 1249, 1269–1270 (2020). See also Yuval Shany, *Digital Rights and the Outer Limits of International Human Rights Law*, 24 GERMAN L.J. 461, 466 (2022); Javier Pallero & Carolyn Tackett, *What the Facebook Oversight Board Means for Human Rights, and Where We Go From Here*, ACCESSNOW (June 1, 2020), <https://www.accessnow.org/facebook-oversight-board-where-we-go-from-here/>.

<sup>50</sup>See Commission Regulation 2022/2065, *supra* note 2, at arts. 14, 34.

<sup>51</sup>See Quintais, Appelman & Ó Fathaigh, *supra* note 22 (discussing that question extensively). See also Wendel, *supra* note 21; Wischmeyer & Meißner, *supra* note 21.

<sup>52</sup>See, e.g., Dror-Shpoliansky & Shany, *supra* note 49, at 1269–70. See also Shany, *supra* note 49, at 466.

<sup>53</sup>With a focus on AI systems, see, for example EVALUATING THE RISK OF AI SYSTEMS TO HUMAN RIGHTS FROM A TIER-BASED APPROACH, EUR. CTR. FOR NOT-FOR-PROFIT L. (2021) (discussing guidance developed by civil society organizations).

pertaining to systemic risks could entail for platforms, we can thus draw a comparison between due diligence obligations under the UN Guiding Principles on Business and Human Rights and the DSA's obligations to assess systemic risks. The EU can learn from the efforts originating in the human rights and corporate social responsibility world, and the global struggle can leverage hard legal obligations originating in the EU to hold platforms accountable to human rights responsibilities.

However, while holding platforms accountable based on human rights responsibilities is seen as an important pathway to constrain platforms' power, there is also persistent skepticism that human rights actually provide an adequate basis to assess platforms' decisions,<sup>54</sup> and questions around how this development, in turn, affects international law.<sup>55</sup> There are, indeed, many good reasons to doubt that, as the criticism focusing on individual remedy already illustrates: Human rights responsibilities are rather general and do not say anything specific about how to moderate content; perhaps framing a discussion on what should happen to content in terms of human rights disguises more than it reveals; and perhaps it is impossible to resolve individual cases based on human rights responsibilities through legal reasoning, and social media councils should not purport that the outcomes they suggest are somehow determined by international law. The skepticism on human rights providing an adequate framework for the existing individual remedy framework is likely to become even more pronounced with view to relying on human rights, or, rather, in the context of the DSA, fundamental rights, to assess systemic risks. We do not have to discuss the merits of this skepticism in detail and need not share a skeptical view of the role of human rights in content moderation, to understand that human or fundamental rights will not provide objective detailed standards based on which systemic risks could be measured and mitigated.

This also means that human rights experts can only play a limited role in defining the standards that should govern systemic risks. Platforms can, and already do, work with human rights experts such as BSR or Article 1, who can develop human rights impact assessments.<sup>56</sup> They also work with Trust & Safety experts, such as the Trust & Safety Professional Association or the Integrity

---

See also *A Human Rights-Based Approach to Content Governance*, BUS. FOR SOC. RESP. (2021), [https://www.bsr.org/reports/A\\_Human\\_Rights-Based\\_Approach\\_to\\_Content\\_Governance.pdf](https://www.bsr.org/reports/A_Human_Rights-Based_Approach_to_Content_Governance.pdf) (explaining approaches developed by human rights consulting firms); *UN Guiding Principles and Accompanying Documents* (explaining work by UN agencies); *Corporate Human Rights Due Diligence Identifying and Leveraging Emerging Practices*, OHCHR, <https://www.ohchr.org/en/special-procedures/wg-business/corporate-human-rights-due-diligence-identifying-and-leveraging-emerging-practices>; *Report on freedom of expression, states and the private sector in the digital age*, A/HRC/32/38, OHCHR, (May 11, 2016), <https://www.ohchr.org/en/documents/thematic-reports/ahrc3238-report-freedom-expression-states-and-private-sector-digital-age> (showing the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression); See also *Report on online hate speech*, A/74/486, OHCHR, (Oct. 9, 2019), <https://www.ohchr.org/en/documents/thematic-reports/a74486-report-online-hate-speech> (showing the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression); See also *Disinformation and freedom of opinion and expression - Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, A/HRC/47/25 (Apr. 13, 2021), <https://www.ohchr.org/en/documents/thematic-reports/ahrc4725-disinformation-and-freedom-opinion-and-expression-report>.

<sup>54</sup>See Evelyn Douek, *The Limits of International Law in Content Moderation*, 6 U. CAL. IRVINE J. INT'L TRANSNAT'L. & COMPAR. L. 37 (2020). See also Brenda Dvoskin, *Expert Governance of Online Speech*, 64 HARV. INT'L L.J. 85 (2022).

<sup>55</sup>See Shany, *supra* note 49. See also Rishi Gulati, *Meta's Oversight Board and Transnational Hybrid Adjudication—What Consequences for International Law?*, 24 GERMAN L.J. 473 (2022); Dana Burchardt, 24 GERMAN L.J. 438 (2023).

<sup>56</sup>See, e.g., "Assessing the Human Rights Impact of Meta's Platform in the Philippines," Art. 1 (Dec. 2020), [https://about.fb.com/wp-content/uploads/2021/12/Meta-Philippines\\_HRIA\\_Executive-Summary\\_Dec-2021.pdf](https://about.fb.com/wp-content/uploads/2021/12/Meta-Philippines_HRIA_Executive-Summary_Dec-2021.pdf); See also "Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021," BSR, (Sept. 22, 2022), <https://www.bsr.org/en/blog/human-rights-due-diligence-of-meta-impacts-in-israel-and-palestine-may-2021>; See also "Towards Meaningful Fundamental Rights Impact Assessments under the DSA," Policy Paper by the European Centre for Not-for-Profit Law," ACCESSNOW, (Oct. 30, 2023), <https://www.accessnow.org/wp-content/uploads/2023/09/DSA-FRIA-joint-policy-paper-September-2023.pdf> (discussing a proposal of how fundamental rights impact assessments should be conducted under the DSA and play into systemic risk assessments).

Institute.<sup>57</sup> These organizations aim at developing and publicly sharing knowledge, developing processes and standards which can be applied by platforms. While strengthening the role of outside experts may be one way to substantiate the systemic risk assessment process, and we should develop proposals for them to do so in an impactful, independent and more transparent way, it has its inherent limitations.

Human rights experts alone, cannot decide some of the hardest questions pertaining to systemic risks, and the tradeoffs involved in choosing between mitigation measures. As acknowledged above, they cannot claim that their particular assessment of systemic risks directly follows from international human rights, or European fundamental rights.<sup>58</sup> They, as well as Trust & Safety and any other involved expert, ultimately make choices that, while based on expertise, are not neutral. Platforms may attempt to frame the work with outside experts "as ostensibly neutral acts of consultations or advice," but doing so "obfuscate(s) the fact that each entity has its own goal and agenda, further complicated by the fact that many of these entities offer their 'expertise' for a fee."<sup>59</sup> Surely, different experts will come to different solutions when it comes to assessing and mitigating systemic risks and answering hard questions on the impact of content moderation practices.

As we lack a clear normative basis which could govern systemic risks, and regulators should not define standards in detail, we need to develop a procedural approach which allow to concretize systemic risk obligations over time. The following Section proposes such an approach.

#### D. A Way Forward: A Virtuous Loop to Assess Systemic Risks

This Section explains the process as outlined in the DSA, conceptualizing it as a "virtuous loop"—a process which empowers civil society and will help solve the conundrum of regulating content moderation. It analyses what role different players should and should not play in that loop, including the Commission, the Board for Digital Services, platforms and auditors, and explains why civil society stakeholders should play a central in the process of assessing and mitigating risks. It describes how civil society involvement already constitutes the core of another approach to tame the power of platforms, of an approach referred to as "multistakeholderism." To test the proposed model, the Section lays out the standard criticism of conventional multi-stakeholder consultancy processes, and then argues that the "virtuous loop" of risk assessments provides responses to this criticism. Finally, focusing on social media councils, it outlines one among many possible visions of how to strengthen civil society involvement in risk assessments in more detail.<sup>60</sup> It describes how social media councils could contribute and strengthen civil society involvement in risk assessments and further mitigate conventional shortcomings of multistakeholderism.

##### I. The Process as Established in the DSA

Beyond defining sources of risk, kinds of risks, and the mitigation measures platforms must consider, the DSA outlined responsibilities and processes around systemic risk assessments. It basically creates a loop in which, once a year and before launching a new product, platforms' efforts to assess and mitigate risks are evaluated and recalibrated.<sup>61</sup> Platforms are supposed to

<sup>57</sup>See Amre Metwally, *The Governors' Advisors: Experts and Expertise as Platform Governance*, 24 YALE J.L. & TECH. 510 (2022) (showing the role of experts, including Trust & Safety experts, in content moderation more generally).

<sup>58</sup>See Douek, *supra* note 54 (Explaining limitations of human rights responsibilities as normative basis for content moderation decisions). See also Dvoskin, *supra* note 54.

<sup>59</sup>See Metwally, *supra* note 57, at 6.

<sup>60</sup>See *A Framework for Meaningful Engagement*, EUR. CTR. FOR NOT-FOR-PROFIT L. (Mar. 8, 2023), <https://ecn1.org/publications/framework-meaningful-engagement-human-rights-impact-assessments-ai> (discussing other proposals on what form engagement should take—not specifically tailored to the DSA but to AI systems).

<sup>61</sup>See Douek, *supra* note 1, at 71 (explaining relatedly, Douek to her model her model of self-regulation as a "virtuous cycle of regulatory, public and industry learning").

work with civil society organizations and independent experts to assess risks and develop mitigation measures and are obliged to submit a yearly report on their efforts to the EU Commission, the Board for Digital Services and auditors. They will then assess whether platforms comply with their obligations under the DSA, or whether they need to change and improve their risk assessment efforts and take alternative mitigation measures. After the Commission and the Board for Digital Services received the first round of risk assessments from platforms, they are expected to begin developing best practices and guidelines which platforms, in their future risk assessments, will have to account for. In addition, the Commission and the Board for Digital Services are tasked with facilitating the development of codes of conduct, including on issues pertaining to systemic risks.<sup>62</sup>

Some roles, namely the roles of the platforms, the Commission and the Board for Digital Services, are formally set. However, all three are somewhat ill-suited to ultimately decide what systemic risks are and how they should be mitigated. Platforms are ill-suited, because self-assessments by corporations whose ultimate goal is to maximize profits, lack credibility.

The Commission and the Board for Digital Services, consisting of the National Digital Services Coordinators, are ill-suited, because they are both public institutions, which raises the concerns explained above in Section B.

With platforms, the Commission and the Board for Digital Services not being well suited, the question arises what other actors can contribute to transform systemic risk assessment into something meaningful and legitimate. Three actors might be especially relevant: Auditors, civil society organizations and, not outlined in the DSA, but always relevant in the EU, national courts and the European Court Justice (ECJ).

Despite the rather ambitious aspirations of the EU Commission laid out in its recently delegated act,<sup>63</sup> it is still unclear how audits can add value to risk assessments rather than only creating revenue for profit-oriented consultancy firms.<sup>64</sup> Auditors are usually large companies, likely with no significant expertise in content moderation and certainly with no particular legitimacy with regard to defending the public interest. Until shortcomings of auditing processes are overcome, we cannot rely on auditing to create true accountability. Many of the submissions the Commission received for the public consultation of its delegated act make this very point.<sup>65</sup> Even platforms themselves<sup>66</sup> and large consulting companies<sup>67</sup> urge the Commission to provide more concrete standards and argue that it should not be the auditors who define these standards.

Courts, and the ECJ in particular, might, over time, play their role in defining standards which need to be considered in systemic risk assessments. The text of the DSA obliges platforms to assess the effects on fundamental rights of curation practices such as recommendations, the newsfeed and amplification, more generally. National courts and the ECJ have both the competence and expertise to weigh in on questions pertaining to the horizontal application of fundamental rights. How they might do so, however, remains rather unpredictable. As explained in Section C.III, one can construct the questions of assessing the systemic fundamental rights implications of curation

<sup>62</sup>See Commission Regulation 2022/2065, *supra* note 2, at art. 45(2).

<sup>63</sup>See generally *Delegated Regulation on Independent Audits Under the Digital Services Act*, *supra* note 13 (explaining the delegated act).

<sup>64</sup>See, e.g., Anna Morandini, *DSA Audits: Procedural Rules Leave Some Uncertainties*, DSA OBSERVATORY (Nov. 28, 2023) <https://dsa-observatory.eu/2023/11/28/dsa-audits-procedural-rules-leave-some-uncertainties/> (explaining the auditing process).

<sup>65</sup>See *Digital Services Act: Delegated Regulation on Independent Audits Now Available for Public Feedback*, EUR. UNION COMM'N (May 05, 2023), <https://digital-strategy.ec.europa.eu/en/news/digital-services-act-delegated-regulation-independent-audits-now-available-public-feedback> (discussing the draft and call for public feedback). See also *Digital Services Act – Conducting Independent Audits*, EUR. UNION COMM'N (June 2, 2023), [https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13626-Digital-Services-Act-conducting-independent-audits\\_en](https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13626-Digital-Services-Act-conducting-independent-audits_en) (referring to the submissions from The European Tech Alliance 2023, the Annenberg Public Policy Center 2023, Global Network Initiative 2023, AlgorithmWatch and AI Forensics 2023, Centre for Democracy and Technology Europe 2023, and the Mozilla Foundation).

<sup>66</sup>See *id.* (referring to the submissions by Google, Booking.com, and the industry association Digital Online Tech Europe).

<sup>67</sup>See *id.* (referring to the submissions by PriceWaterhouseCoopers and Deloitte).



practices as a matter of concretizing *positive fundamental rights obligations on a horizontal level*. This is not entirely implausible, as it is only a step further from the analogy which led to applying negative rights on a horizontal level. However, such a construction entails uncertainties, faces objections from the horizontal as well as federal separation of powers within the EU, and might ignore the fundamental rights protections which platforms themselves enjoy.<sup>68</sup> In the near future, the role of ECJ in addressing systemic fundamental rights implications may remain constraint to its assessments of competing interests which justify individual rights encroachments, such as when it considered in its ruling on “the right to be forgotten” whether the general public had a legitimate interest in accessing information provided by search engine providers.<sup>69</sup>

This leaves one more role open, which could contribute to solving some of the challenges of systemic risk assessments, which is the role of civil society organizations. This role is only established in the recitals of the DSA, indicating minor relevance, per recital Nb. 90. However, taking into account the practical and theoretical concerns raised here, stakeholders might be best suited to concretize the standards and processes based on which systemic risks should be measured and mitigated.<sup>70</sup> Many submissions to the Commission’s delegated act on independent auditing make that point with view to both, systemic risk assessments and independent auditing.<sup>71</sup>

## II. “Multistakeholderism” and its Flaws

The idea that civil society organizations should fill the gap which regulators cannot close is not new, and has previously been described as the “‘least-worst’ option.”<sup>72</sup> It was dubbed with the ugly yet somehow fitting term of “multistakeholderism.”<sup>73</sup> Before systemic risk assessments were introduced, multistakeholderism was conceived as the second large approach, besides rule of law and individual remedy mechanisms, to hold platforms accountable. It aims at increasing civil society’s influence in platform governance, through increasing transparency and assuring consultation and participation of stakeholders.<sup>74</sup> It “pursues democratic accountability . . . but reflects concerns about direct state regulation of online communications.”<sup>75</sup> Building usually only on voluntary commitments, it does not raise the same concerns as direct state regulation.

However, as indicated above, not only is the idea of civil society organizations helping regulate content moderation not new, there is also persistent skepticism that this indeed provides useful solutions. Let us look at four major objections against multistakeholderism. One, the voluntary nature of civil society involvement leaves platforms too much discretion on when, how and for what questions they should consult stakeholders. Two, even where platforms involve civil society, let us assume for a moment that civil society involvement creates concrete solutions, they are not required to implement these solutions. Three, let us challenge that assumption, and acknowledge that civil society participation often does not produce compromises and implementable solutions. Different civil society organizations will bring different views to the table—which is their very task—and often favor different results. Stakeholder consultation helps to inform decision making, which is immensely important, but does not preempt the choice between different possible

<sup>68</sup>See Wendel, *supra* note 21.

<sup>69</sup>See, e.g., Jean-Marie Chenou & Roxana Radu, *The “Right to Be Forgotten”: Negotiating Public and Private Ordering in the European Union*, 58 BUS. & SOC’Y 74, para. 75, (June 28, 2019) (explaining the discussion).

<sup>70</sup>See Brenda Dvoskin, *Representation Without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance*, 67 VILL. L. REV. 447, 457 (2021) (talking about the link between risk assessments in the DSA and civil society involvement). See also Griffin, *supra* note 6, at 53–54.

<sup>71</sup>See Article text accompanying *supra* note 63. See also sources collected *supra* note 65 (viewing the submission by Dr. Christian Djefal, Technical University of Munich, to the Commission’s call for feedback on the delegated auditing act).

<sup>72</sup>See Douek, *supra* note 43.

<sup>73</sup>See, e.g., Griffin, *supra* note 6.

<sup>74</sup>See Griffin, *supra* note 6, at 50–54.

<sup>75</sup>See Griffin, *supra* note 6, at 52.

solutions. Platforms remain the “single point of authority” and ultimately decide what to make of consultations.<sup>76</sup> Four, although multistakeholderism aims at introducing public interest and purports to contribute to democratic accountability, it can lead to unequal participation, the underrepresentation of poorly funded or marginalized groups and unfair outcomes which only represent the interest of those who organized more effectively.<sup>77</sup>

The conclusion which is often drawn is thus that multistakeholderism cannot provide democratic accountability and, therefore, has limited legitimizing capacity.<sup>78</sup> Perhaps, one could distill the criticism, the ugliness of the term captures well the thing itself, which is complicated, unpractical, unfair and ultimately useless. Why, then, should multistakeholderism help respond to challenges posed by systemic risk assessments under the DSA? Because, maybe, systemic risk assessments can fix multistakeholderism, and in turn, multistakeholderism can fix what would otherwise be performative and ineffective systemic risk assessments. This leads us to the decisive question: How would that work? The final Sections of this Article propose an answer to that question, first focusing on the mandatory nature of risk assessments and then outlining a potential role for social media councils.

### *III. Voluntary and Selective Involvement Versus Mandatory and Pre-Defined Involvement*

The Commission could empower civil society organizations—emphasis on *could*, as it is by no means certain that it will. It could decide to give civil society involvement significant weight in its evaluations of the submissions of platforms, and it could, together with the Board for Digital Services, require substantial civil society involvement in its best practices and guidelines. Perhaps this is the true ingenuity of the DSA’s systemic risk assessment provisions: They create a loop in which the Commission can use its enforcement powers to empower civil society and independent expertise. This loop provides solutions to some of the most important objections against multistakeholderism—and might untimely help solve the conundrum regulating content moderation.

Best practices and guidelines, in which the Commission and the Board for Digital Services specify what civil society involvement as outlined in the recitals of the DSA concretely entails, would not create legal obligations, but would have quasi-legal effects: Whoever risks ignoring these guidelines and best practices risks breaching their due diligence obligations with view to assessing and mitigating risks. Best practices and guidelines would thus ultimately remove the cardinal flaw of multistakeholderism, its voluntary nature. Platforms could no longer simply ignore outcomes of civil society engagement, as this might indicate to the Commission that they fail to adequately assess and mitigate risks. They can also no longer arbitrarily decide in which areas to involve civil society, as Article 34 DSA defines the sources and kinds of risks that risk assessment must address—and, if refined by best practices and guidelines, would be expected to involve civil society in all major areas.

The loop conveys the task of concretizing and applying substantive standards based on which systemic risks should be assessed and mitigated to civil society actors, while the task of enforcing obligations remains with the Commission. It assures that platforms will have to account for public interest as articulated in the stakeholder process, as imperfect as it might be, without public actors having to define what that concretely means themselves. While multistakeholderism has the potential to respond to a difficult issue raised by systemic risk assessments, namely how to define and apply substantive standards and reconcile competing interests, it will by no means provide a panacea, understood as a comprehensive solution to all the challenges posed by risk assessments. Instead, we should see it as one piece of the puzzle, one component of what makes up an effective

<sup>76</sup>See Griffin, *supra* note 6, at 50.

<sup>77</sup>See generally Dvoskin, *supra* note 70.

<sup>78</sup>See Griffin, *supra* note 6, at 58–76.

risk assessment ecosystem. With good reason, academics and regulators are working to develop data-driven analyses, thorough empirical methodologies and clear audit standards to measure harm and the effectiveness of mitigation measures.<sup>79</sup> This work is crucial to the success of risk assessments, too. However, we should be careful not to confuse the hard substantive questions underlying risk assessments—such as how to draw red lines between what is acceptable and unacceptable risk, how to reconcile competing fundamental rights, what forms of content moderation harm democracy—with questions about how to measure harm. We should be cautious not to drown difficult normative questions in methodological and technocratic standard-setting.

The DSA already foresees that civil society should contribute to the development of codes of conduct.<sup>80</sup> While codes of conduct may play an increasingly important role over time in concretizing obligations in certain risk areas, developing these codes is a tedious process which takes a lot of time, and they, too, remain relatively abstract and thus have limited impact on the concrete handling of risks.<sup>81</sup> Providing civil society with a role in the process of assessing risks would give them a voice where it really matters—in the evaluation of concrete risks and the decision on mitigation measures—and would ensure the practical impact of their work. Best practices and guidelines on risk assessments issued by the Commission and the Board for Digital Services can do what codes of conduct likely cannot do, which is to shape a procedural framework for risk assessments and help create the virtuous loop outlined here.

Given the above-described feedback the Commission received on its delegated act for auditing, there is hope to believe that the Commission might strengthen the role of civil society stakeholders in risk assessments, potentially through best practices and guidelines. However, we also need to consider the consequences of if it does not: What would happen if the Commission did not encourage a role for civil society stakeholders? The DSA and its risk assessment obligations would likely, contrary to their objective, marginalize civil society's influence on platforms' content moderation practices. Before the DSA, assessing and mitigating societal harms was a matter of corporate social responsibility for platforms.<sup>82</sup> They could work with civil society stakeholders to develop solutions but were also under no legal pressure to implement solutions proposed by stakeholders if they believed they would harm their business practices.<sup>83</sup> As assessing and mitigating societal harms becomes a legal requirement, the dynamic between platforms and civil

<sup>79</sup>*Digital Services Act: Application of the Risk Management Framework to Russian Disinformation Campaigns*, EUR. UNION (Aug. 2023), <https://op.europa.eu/en/publication-detail/-/publication/c1d645d0-42f5-11ee-a8b8-01aa75ed71a1/language-en> (showing that the Commission published an independent study on risk assessments which focuses heavily on the empirical analysis of online data); SALLY BROUGHTON MICOVA & ANDREA CALEF, *ELEMENTS FOR EFFECTIVE SYSTEMIC RISK ASSESSMENT UNDER THE DSA* (2023). See also Naomi Shiffman, Carly Miller, Manuel Parra Yagnam, Claudia Flores-Saviaga, *Burden of Proof: Lessons Learned for Regulators from the Oversight Board's Implementation Work*, 2 J. ONLINE TR. & SAFETY (2024); Morandini, *supra* note 64.

<sup>80</sup>See Rachel Griffin & Carl Vander Maelen, *Codes of Conduct in the Digital Services Act: Exploring the Opportunities and Challenges* (May 30, 2023) <https://ssrn.com/abstract=4463874>. See also Commission Regulation 2022/2065, *supra* note 2, at art. 45(2).

<sup>81</sup>See Rachel Griffin & Carl Vander Maelen, *Twitter's Retreat from the Code of Practice on Disinformation Raises a Crucial Question: Are DSA Codes of Conduct Really Voluntary?*, DSA OBSERVATORY (June 12, 2023), <https://dsa-observatory.eu/2023/06/12/twitters-retreat-from-the-code-of-practice-on-disinformation-raises-a-crucial-question-are-dsa-codes-of-conduct-really-voluntary/> (discussing the binding nature of codes of conduct).

<sup>82</sup>See Michael Cusumano et al., *Can Self-Regulation Save Digital Platforms?*, 30 INDUS. & CORP. CHANGE 1259 (2021) (explaining the potential and limits of self-regulatory efforts). See also Evelyn Douek, *Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation*, HOOVER INST. AEGIS PAPERS SERIES 1–28 (2019), [https://www.hoover.org/sites/default/files/research/docs/douek\\_verified\\_accountability\\_aegisnstl1903\\_webready.pdf](https://www.hoover.org/sites/default/files/research/docs/douek_verified_accountability_aegisnstl1903_webready.pdf). See also Amélie Heldt, *Let's meet halfway: Sharing new responsibilities in a digital age*, 9 J. OF INF. POLICY 336 (Oct. 8, 2019) (discussing the relation between regulatory and self-regulatory efforts).

<sup>83</sup>See, e.g. Evelyn Douek, *The oversight board moment you should've been waiting for: Facebook responds to the first set of decisions*, LAWFARE (Feb. 26, 2021), <https://www.lawfareblog.com/oversight-board-moment-you-shouldve-been-waiting-face-book-responds-first-set-decisions> (analyzing how Meta reacted to the first decisions of the Oversight Board); See also Dipayan

society actors changes. Working with civil society suddenly creates compliance risks. Platforms run the risk of having to justify before regulators why they did not implement solutions proposed by stakeholders, and could even be fined for not doing so. The consequence of transforming corporate social responsibility matters into hard law is that the legal rather than policy departments of platforms take the lead. The task of legal teams is to minimize compliance risks. To minimize compliance risks, legal teams will likely advise against any civil society engagement that is not required by regulators. Because of the DSA, civil society stakeholders risk losing the impact they had gained during the era where assessing the societal risks of platforms was a matter of corporate social responsibility. The demise of policy and governance teams, evidenced by the recent waves of layoffs of platforms, would be followed by the demise of civil society stakeholders' impact.<sup>84</sup> The proposed loop is necessary not only to strengthen civil society involvement but also to prevent its marginalization.

To sum up, the proposed loop would respond to two of the most important objections against multistakeholderism: that platforms can arbitrarily decide on where and how to involve civil society stakeholders and that platforms are not bound by the results. Two major objections remain: the problem that civil society involvement does not actually produce any implementable solutions and leaves the discretion to decide between different proposals to platforms. And finally, the problem that stakeholder involvement often fails to guarantee equal participation. The following and final Section develops a proposal for how to respond to these objections. It proposes that platforms should work with social media councils to translate civil society involvement into concrete solutions.

#### IV. Situating Social Media Councils in the Systemic Risk Framework

The term “social media council” is broadly used to describe bodies which typically include a committee of experts or civil society representatives and which work with platforms to improve their content moderation practices.<sup>85</sup> The idea that such councils—whatever form or shape they might take—can contribute to content moderation accountability has been supported not only by academics but also by the former Special Rapporteur on Freedom of Expression of the UN, David Kaye,<sup>86</sup> and they feature in government programs to regulate platforms.<sup>87</sup> Social media councils have been extensively discussed in academia in recent years, especially the most prominent one, the Oversight Board.<sup>88</sup> Discussions centered on questions including whether they

---

Ghosh, *Facebook's Oversight Board Is Not Enough*, HARV. BUS. REV. 16 (Oct. 16, 2019) <https://hbr.org/2019/10/facebooks-oversight-board-is-not-enough> (analyzing the limitations of the impact of self-regulatory effort such as the Oversight Board).

<sup>84</sup>See Hayden Fields & Jonathan Vanian, *Tech layoffs ravage the teams that fight online misinformation and hate speech*, CNBC (May 26, 2023), <https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html#>; See also Jenny Domino, *Why Facebook's Oversight Board is Not Diverse Enough*, JUST SECURITY (May 21, 2020), <https://www.justsecurity.org/70301/why-facebooks-oversight-board-is-not-diverse-enough/>.

<sup>85</sup>See, e.g., *Social Media Councils One piece in the puzzle of content moderation*, ARTICLE 19 (Oct. 12, 2021), <https://www.article19.org/wp-content/uploads/2021/10/A19-SMC.pdf> (showing early studies on social media councils); See also Matthias C. Kettemann & Martin Fertmann, *Platform Proofing Democracy Social Media Councils as Tools to Increase the Public Accountability of Online Platforms*, FRIEDRICH NAUMANN STIFTUNG (May 2021); See also Matthias C. Kettemann, *PLATFORM://DEMOCRACY – PERSPECTIVES ON PLATFORM POWER, PUBLIC VALUES AND THE POTENTIAL OF SOCIAL MEDIA COUNCILS* (Wolfgang Schulz ed. 2023).

<sup>86</sup>See David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/38/35, para. 41–43 (Apr. 6, 2018), <https://digitallibrary.un.org/record/1631686?ln=en>.

<sup>87</sup>See *Koalitionsvertrag: Mehr Fortschritt wagen*, WOFÜR WIR KÄMPFEN (Nov. 24, 2021), <https://www.gruene.de/artikel/koalitionsvertrag-mehr-fortschritt-wagen>.

<sup>88</sup>See OVERSIGHT BOARD, <https://oversightboard.com/>; See also Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L. J. 2418 (July 20, 2020); See also Douek, *supra* note 43; See also David Wong & Luciano Floridi, *Meta's Oversight Board: A Review and Critical Assessment*, MINDS AND MACHINES (Oct 24, 2022); See also Lorenzo Gradoni, *Twitter Complaint Hotline Operator: Will Twitter join Meta's Oversight Board?*, VERFASSUNGSBLOG (Nov. 10, 2022), <https://verfassungsblog.de/musk-ob/>; See also Helfer & Land, *supra* note 48; See also

actually are credible mechanisms to hold platforms accountable, and with view to the Oversight Board, whether it is truly independent from Meta, whether its human rights approach is compelling, whether it is composition and processes indeed provide an additional and legitimate way to take important decisions and account for the public interest. Originally described as Meta's "Supreme Court," social media councils tended to be contextualized in the world of individual remedy. This Section revisits that characterization, and situates social media councils in the world of systemic risk assessments and the regulatory landscape the DSA creates.

It argues that, with the DSA, we have a clear language and framework to talk about systemic issues and societal harms of these platforms, and that this allows to strengthen the added value of social media councils. It paints a vision for a contribution social media councils could make, and outlines how they could help respond to the two outstanding shortcomings of multistakeholderism: Its unfairness and incapability to reconcile competing positions and produce implementable solutions.

Let us begin by briefly reassessing what social media councils do based on the systemic risks assessment framework, taking the Oversight Board as an example. The Oversight Board has been described as "a body of experts, drawn from geographically, culturally, and professionally diverse backgrounds, funded by an independent trust created under Delaware law, which Meta has empowered to make decisions on content that will affect billions of people worldwide."<sup>89</sup> The goal of the Oversight Board is to decide emblematic cases, cases which are difficult and significant. When deciding a case, the Board digs deep, analyzing the processes, automated systems and policies which decided the particular case, but which equally affect other cases. The Board ultimately aims at discovering systemic issues and flaws in Meta's content moderation and takes individual cases as a vehicle to advocate for systemic changes. Based on the analyses of an individual case, it then recommends changes to matters such as policies, enforcement processes and transparency. And it tracks whether Meta implements the recommended changes. While the overall number of cases the Board decides is low, it is not really the number that matters but the systemic changes the Board achieves through cases. Engaging with individual cases, the Board is still part of the world of individual remedy, yes, but it is also part of the world of systemic risk assessments, as it aims at structurally improving the way Meta treats users. The confusing part here is that we can also talk about individual rights and remedy on a systemic level, and that is what the Board is doing in its cases.

However, we wanted to look beyond individual remedy, and focus on more systemic issues such as amplification or recommendations. How could social media councils, as one player among many, contribute to addressing the systemic risks of content moderation? Perhaps it is not that big of a leap. Perhaps it is already happening, in a modest way. Let us look at the Oversight Board again as an example. When deciding individual cases, the Board can enquire about the algorithmic systems Meta uses to moderate content, and make recommendations pertaining to those systems.<sup>90</sup> Additionally, the Board does not only decide individual cases but also publishes "Policy

---

Thomas E. Kadri, *Juridical Discourse for Platforms*, 136 HARV. L. REV. 163 (Dec. 29, 2022); See also Gulati, *supra* note 55; See also Brenda Dvoskin, *Expertise and Participation in the Facebook Oversight Board: From Reason to Will*, 47 TELECOMM. POLICY (June 2023); See also *Introducing the TikTok Content Advisory Council*, NEWSROOM, (Mar. 18, 2020), <https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council>; See also *Meet TikTok's European Safety Advisory Council*, NEWSROOM (Mar. 1, 2021), <https://newsroom.tiktok.com/en-gb/tiktok-european-safety-advisory-council>, *TikTok's Diversity & Inclusion Council: How We #ChooseToChallenge*, TIKTOK (Mar 3, 2021); See also Dawn Chmielewski, *Spotify forms council to deal with harmful content*, REUTERS (June 13, 2022), <https://www.reuters.com/technology/exclusive-spotify-forms-safety-advisory-council-2022-06-13/>; See also *Safety Advisory Council*, TWITCH, [https://safety.twitch.tv/s/article/Safety-Advisory-Council?language=en\\_US](https://safety.twitch.tv/s/article/Safety-Advisory-Council?language=en_US).

<sup>89</sup>See Helfer & Land, *supra* note 48, at 1.

<sup>90</sup>See Edward Pickup, *The Oversight Board's Dormant Power to Review Facebook's Algorithms*, 39 YALE J. ON REG., 1 (2021).



Advisory Opinions.”<sup>91</sup> Meta can decide to send difficult content moderation questions to the Board, which then, after a few months of acquiring relevant information, analyzing public comments, engaging with civil society stakeholders, deliberating and drafting, answers the questions and recommends a concrete way forward. The recommendations are, again, not binding, but the Board tracks how Meta reacts to them. In terms of substance, the scope of Policy Advisory Opinions is not limited. They can focus on issues which the DSA describes as sources of systemic risks, such as how Meta amplifies or demotes content. Building on the DSA, we can conceive this process as a potential way to assess the systemic risks of the most important components of platforms’ content moderation, including the amplification of content and recommender systems. Mechanisms in which platforms work with social media councils, such as Policy Advisory Opinions or what we can perhaps, more generally, call “risk advisory opinions,” may offer a way forward for social media councils to address all kinds of hard questions pertaining to systemic risks.

Moving away from the concrete example and existing practices, we can more generally assert that one credible way for platforms to conduct risk assessments could be to rely on social media councils to analyze risks, answer hard questions and decide between different risk mitigation measures. Let us leave the example of the Oversight Board behind. The potential skepticism of the credibility and independence of the Oversight Board, the endless discussion about whether the Oversight Board is a legitimate actor or a PR stunt, should not distract us from the task ahead: Envisioning a future in which social media councils, however you want them to be structured and funded, could help fix shortcomings of multistakeholderism and make systemic risks assessments a mechanism of effective accountability.

The cooperation between platforms and social media councils can take different shapes and different degrees of intensity. Enabled by the Commission’s implementation of the DSA, different models could emerge. On one end of the spectrum could be cooperation which leaves control largely in the hands of the platform, allowing them to decide what questions the social media council should answer, what access they get to information and leaving them discretion if and how to implement solutions suggested by the council. On the other end of the spectrum, the cooperation could consist of platforms conveying certain competencies to social media councils, providing them decay on priorities for risk assessments, giving them access to relevant data, allowing them to choose the questions they wish to engage with, to develop their own frameworks and to make recommendations on some of the most important questions. The platforms would not necessarily have to follow all recommendations that social media councils make, but at least would commit to respond to them—and justify before the Commission and the Board for Digital Services if they did not follow the recommendations the social media councils made. Social media councils could track whether platforms implement the proposed measures and assess the effectiveness of those measures.

We do focus on the potential of social media councils to contribute to the virtuous loop of risk assessments, but we also need to account for their limitations. Social media councils would fulfill what can be conceived as a public function, and they, too, require legitimacy.<sup>92</sup> The topics that social media councils will have to deal with are so diverse and complex, that we cannot seriously assume that the members of these councils will be experts on all of these issues or that the staff of the social media council could develop the necessary expertise to assess systemic risks in a variety

<sup>91</sup>See, e.g., *Oversight Board Publishes Policy Advisory Opinion on the Removal of COVID-19 Misinformation*, OVERSIGHT BOARD (Apr. 20, 2023), <https://oversightboard.com/news/739141534555182-oversight-board-publishes-policy-advisory-opinion-on-the-removal-of-covid-19-misinformation/> (Showing the Policy Advisory Opinion on Meta’s Cross Check program and misinformation).

<sup>92</sup>See, e.g., Burchardt, *supra* note 55, 443–45 (discussing their public function). See also, Dvoskin, *supra* note 70 (discussing the legitimacy of social media councils from various angles).

of fields. Expertise and fundamental rights, we found earlier, provide an imperfect basis to resolve hard questions in a legitimate way.

This is why social media councils will have to work with civil society organizations who do the hard work of empirically and qualitatively assessing the practical effects of curation. The added value of social media councils as envisioned in the “virtuous loop” proposed here not so much consists in expertise but in providing a process which involves a variety of civil society organizations, thus enabling participation and consultation of affected groups.<sup>93</sup> Multistakeholderism can help fix the shortcomings of an otherwise expert-led, and ultimately illegitimate, process concretizing how risks should be assessed and mitigated. Social media councils, in turn, can help make civil society involvement fairer and translate it into concrete solutions.

Scholars criticized multistakeholderism as the impact that civil society organizations have in such processes depends on their funding and ability to effectively organize.<sup>94</sup> Civil society involvement, so the concern, thus does not actually represent the public interest and does not actually provide democratic accountability. The deliberative processes of social media councils could help mitigate this concern. We can imagine social media councils to work like courts, where parties are expected to all facts and concerns, but where the complex task of ultimately assessing their weight and relevancy for the legal evaluation falls to the judge. The idea that “the court knows the law” implies that parties presenting their concerns need not to and the court should give equal considerations to all issues raised, even if they are not presented in a comprehensive and formal manner. Social media councils should, in turn, know the basic legal parameters governing systemic risk assessments and situate relevant facts and concerns wherever they are the most relevant. Where need be, they could commission additional research to dive deeper into underrepresented positions. Social media councils could serve as a corrective for otherwise potentially skewed civil society involvement and could thus help to make it fairer, providing a solution to another shortcoming of conventional multistakeholderism.

Social media councils would not only oversee the civil society involvement but would also serve as the new “point of authority”<sup>95</sup> which decides what conclusions to draw from the various positions presented in stakeholder engagement. The remaining objections against “multistakeholderism” consist in the problem that civil society involvement ultimately only surfaces various divergent positions, failing to reconcile those and producing concrete implementable measures. The solution to this problem lies in the deliberative processes of social media councils and the reflection of these deliberations in its reasoned decisions.

We acknowledged that human rights, legal reasoning, and expertise do not actually alleviate anyone from making hard calls, deciding between different tradeoffs and resolving conflicting interests. The next best thing to legal reasoning and expertise are deliberative processes: Processes in which a group of people discusses, and weighs affected interests, to finally come to a conclusion. Yes, the conclusions reached in these deliberations and solutions proposed by social media councils will be imperfect. But at least a reasonable conversation that was detached from economic interests and well-informed took place. And we can read about the considerations which led to these solutions and criticize the reasoning behind them. This allows for academic and public conversations which can feed into the next round of risk assessments the following year. This solution is not perfect, but, if we get the set-up of social media councils and their cooperation with civil society right, it could be less worse than the existing “least-worst” options.

For social media councils to become credible actors and add value in the loop of risk assessments, we need to think more about how they should be structured and funded, how the decision-makers should be chosen and how the body should operate. We can build on the experiences of and the debate on the Oversight Board to develop proposals. We also need to think

<sup>93</sup>See Griffin, *supra* note 6, at 50–54 (discussing the legitimacy of multistakeholderism).

<sup>94</sup>See Dvoskin, *supra* note 70.

<sup>95</sup>See Griffin, *supra* note 6, at 51.

about how the research and the work of contributing civil society organizations should be funded. Much speaks in favor of the EU providing resources for this work, as the quality of risk assessments will largely depend on the quality of this research. This would be money well spent, and the EU has sufficient resources based on the fee it charges to platforms for the implementation of the DSA, which consist of 0.05% of their worldwide annual net income.<sup>96</sup> The delegated act by the Commission on methodologies and procedures regarding the supervisory fees mentions “studies and external consultants referring to a given designated service, including . . . or analyzing a given category of risk resulting from the risk assessment” as one potential expenditure.<sup>97</sup> This may provide starting points for funding required under the proposed model.

Many open questions remain, and not much time before the Commission and the Board for Digital Services will begin issuing best practices and guidelines which will likely shape risk assessments for years to come. How should social media councils be set up for them to be independent from both platforms and European regulators? Who should sit in these councils and how exactly will their processes integrate civil society organizations? How exactly should they be plugged into the risk assessment process and what should their decision-making process look like? How can they base decisions on fundamental rights, and legitimately exercise discretion which unavoidably remains?

The modest goal of this Article is not to answer these questions but to demonstrate that it is worth pursuing answers, for platforms, civil society and the regulators implementing the DSA. Relying on social media councils has significant advantages for all parties involved, including platforms, civil society, and the Commission: Platforms would benefit from involving social media councils in their process of assessing and mitigating risks, as social media councils could ultimately help decide the hardest questions platforms face, weighing tradeoffs presented by other stakeholders. If platforms choose to implement the recommendations by the social media council and incorporate them into their strategy to assess and mitigate risks, this conveys credibility to the systemic risk assessments they submit to the EU. In response, platforms could expect positive evaluations and non-interference from the side of the EU institutions and would minimize the risks of fines.

The involvement of social media councils could make civil society stakeholder engagement fairer and more effective. They could help strengthen underrepresented voices, reconcile diverse positions and translate them into concrete solutions. Their deliberative processes and reasoned decisions could lead to reasonable and transparent outcomes.

The involvement of social media councils allows the EU Commission to restrain from defining the concrete standards against which to measure content moderation and curation practices, as any public institution should, while catalyzing a productive, virtuous loop which empowers civil society. Through the platforms’ work with external experts and social media councils, the Commission and the Board for Digital Services receive high-quality risks assessments, based on which they can further define substantial standards governing systems risk assessments, without unduly interfering with the workings of platforms.

## E. Conclusion

Systemic risk assessments indeed pose great, but not entirely unprecedented, challenges. Under the first pillar of content moderation, self-regulation, individual remedy, we learned a lot about institutional set-ups and normative frameworks which can help solve the conundrum of taming

<sup>96</sup>See Commission Regulation 2022/2065, *supra* note 2, at art.43(5)(c).

<sup>97</sup>See Commission Delegated Regulation (EU) 2023/1127 of 2 March 2023 supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council with the detailed methodologies and procedures regarding the supervisory fees charged by the Commission on providers of very large online platforms and very large online search engines, available at: [https://eur-lex.europa.eu/eli/reg\\_del/2023/1127/oj](https://eur-lex.europa.eu/eli/reg_del/2023/1127/oj).

content moderation. We can now leverage this experience to transform systemic risk assessments into a powerful tool of platform accountability which can help mitigate the social harms of social media. This year, in which obligations around risk assessments will be implemented for the first time, will shape the future of content moderation. If done right, systems risk assessments will defend the public interest in the space of at-scale content moderation and help protect democracy.

This Article addressed one of the issues raised by the new systemic risk assessment regime, namely how to define substantive standards and reconcile competing interests. It described one piece of the puzzle, aiming to contribute to a fuller picture of how systemic risk assessments could become a success. It argued that we should conceive the process set up by the DSA as a virtuous loop whose aim is to empower civil society stakeholders. The Commission and the Board for Digital Services could fix flaws of “multistakeholderism” by making civil society involvement mandatory through its best practices and guidelines, and social media councils could reconcile the perspectives of a variety of civil society stakeholders and contribute to fairer outcomes. Social media councils could ultimately be better suited than platforms or the Commission to answer hard questions and make difficult choices. They could be involved early in the process of risk assessments, defining priorities, and could then develop risk advisory opinions for the most important areas, assessing risks and selecting mitigation measures, and tracking whether platforms implement the proposed measures. To make the proposed model a reality, the Commission and the Board for Digital Services must empower civil society organizations, and more concretely models such as social media councils, through their best practices and guidelines, and needs to ensure their funding.

Many open questions remain. The window in which answers to these questions can meaningfully influence the implementation of the DSA might close soon. It is high time for academia and civil society to provide answers and come up with more concrete proposals on how to meaningfully involve civil society in risk assessments. The virtuous loop, as outlined in this Article, can provide a framework for the debate. And the reflections on social media councils can serve as a starting point for a detailed model—or at least as provocations for alternative proposals.

**Acknowledgements.** This paper is based on the participation in the research clinic PLATFORM://DEMOCRACY of the Humboldt Institute for Internet and Society and the short, related paper “Assessing the systemic risks of curation”, published as a result of that clinic in: Kettemann, Matthias C.; Schulz, Wolfgang (eds.) (2023): Platform://Democracy—Perspectives on Platform Power, Public Values and the Potential of Social Media Councils. Hamburg: Verlag Hans-Bredow-Institut. <https://doi.org/10.21241/ssoar.86524>, pp. 170-172.

I am grateful to Prof. Dr. Matthias Kettemann, Prof. Dr. Matthias Wendel, Amre Metwally and the entire Community of the Information Society Project at Yale for their support of this project. I also want to thank the organisers and participants of presentations at King’s College London, Solvay Brussels School of Economics and Management, the Hans-Bredow-Institut and the Humboldt Institute for Internet and Society.

**Competing Interests.** The authors declare none. Please note that I am, as explained above, working at the Oversight Board. This research has been conducted independently from my affiliation with the Oversight Board, in my private capacity as a researcher and Affiliated Fellow of the ISP. It is not funded or supported by the Oversight Board and does not express positions taken by the Oversight Board. All views in the article are strictly the author’s own.

**Funding Statement.** No specific funding has been declared for this article. See above.