REDUCTION: AI, POWER, AND OPACITY IN CONTENT MODERATION'S BACKSTAGE

NOA MOR[†]

In recent years, social media platforms have been quietly transforming the content moderation framework by increasingly relying on the AI-driven reduction of content visibility rather than on its outright removal. Initially applied to clickbait, misinformation, and sensitive content, reduction is now used by these platforms to demote users' exposure to information across all content categories. Among other types of content, this moderation strategy is now applied to the vast realm of content that borders with the platforms' removal policy (but does not reach it) or is likely to violate it (but its violation is not confirmed). It thereby elevates the entire normative threshold for permissible content and erodes the scope of information available to users. Alongside its widespread application, reduction's impact stems from its efficacy in limiting views and its flexible and multifaceted nature. Unlike removal, reduction employs various methods, including downranking content, adjusting the recommendation system, excluding content from dominant areas in the platform, combining reduction with other sanctions, integrating designated choice architecture, and outsourcing reduction options to users.

Despite its far-reaching implications for the informational landscape, digital platforms implement reduction using patchwork, short, and opaque guidelines of doubtful legitimacy. The platforms also fail to adequately provide data concerning reduction through their Transparency Reports, inform sanctioned users, offer explanations, or allow appeals. The unaccountable and sweeping application of reduction also undermines the rule of law, procedural fairness, freedom of expression and other human rights. Its undisturbed development relies, to a great extent, on diverting our attention toward a more celebrated direction: removal and the policies governing it, which introduce a more detailed, carefully updated, and publicly scrutinized measure for guiding behavior. This Article aims to cast light on the evolution and application of reduction, its dramatic impact, and the way it is being concealed in

[†] Faculty of Law, The Hebrew University of Jerusalem. The author would like to thank Renana Keydar, Lando Kirchmair, Tamar Megiddo, Yuval Shany, the participants of the Biennial Conference of the Standing Group on Regulatory Governance at the University of Antwerp 2023, and the participants of the Workshop on Algorithmic Injustice at the University of Amsterdam 2023, for their invaluable suggestions. The author also thanks Yuval Abramovich and Noam Nadler for excellent research assistance. The research was conducted with the support of ERC Grant No. 101054745: Three Generations of Digital Human Rights, and the Alfred Landecker Lab for Digital Humanities at the Hebrew University of Jerusalem.

the backstage of content moderation. It also examines the legitimacy of the motivations behind reduction and the legal and AI-related challenges it poses. Finally, the Article offers a way forward, outlining how we can tackle reduction’s challenges while harnessing its sophisticated nature to benefit our future digital sphere.

I. INTRODUCTION

A. Setting the Stage

In 2018, Mark Zuckerberg, Facebook's (now Meta's) founder and CEO, released a post that addressed different content moderation challenges that the company was facing during that year. One of the topics that Zuckerberg explored was titled "Discouraging Borderline Content." He wrote that

> One of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content. This is not a new phenomenon. . . . At scale it can undermine the quality of public discourse and lead to polarization. In our case, it can also degrade the quality of our services.[1]

To tackle this problem, Zuckerberg noted, "[w]e train AI systems to detect borderline content so we can distribute that content less."[2] Borderline content, as it arises from the post, is content that does not violate the company's famous content removal policy, known as the Community Standards,[3] but "gets close" to the policy's limits.[4]

Zuckerberg explained that "[t]he category we're most focused on is click-bait and misinformation." However, he added "[i]nterestingly, our research has found that this natural pattern of borderline content getting more engagement applies . . . to almost every category of content."[5] He further stated that the company's efforts around reduction will provide users with control over the content they consume and that these efforts represent some of the company's "most important work."

[1] Josh Constine, *Facebook Will Change Algorithm to Demote "Borderline Content" That Almost Violates Policies*, TECHCRUNCH (Nov. 15, 2018), https://techcrunch.com/2018/11/15/facebook-borderline-content/ [https://perma.cc/56JG-TNKX]. Costine's article includes the text of Zuckerberg's original post on borderline content. The revised version of this post is available at Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement,* META (May 6, 2021), https://www.facebook.com/notes/751449002072082/ [https://perma.cc/SJ9J-WCFQ].

[2] Constine, *supra* note 1.

[3] *Facebook Community Standards*, META TRANSPARENCY CENTER, https://transparency.meta.com/policies/community-standards/ (last visited Feb. 5, 2024) [https://perma.cc/FD7L-MNZH**]**.

[4] Constine, *supra* note 1.

[5] *Id*.

He acknowledged "significant progress in the last year," while noting there is still "a lot of work ahead."[6]

Zuckerberg's post is important because it is one of Meta's earliest public communications regarding the wide application of reduction,[7] a powerful strategy through which the company limits the visibility of content and restricts users' engagement with it without removing it.[8] The post also matters because it encapsulates some of the challenges that still surround reduction today, namely partial information that is provided tardily, deep into the integration of the practice; use of undefined, abstract, almost demagogical terms such as "borderline," "sensationalism," "polarization," and "control"; the absence of publicly available data to support such content restriction and to facilitate public discussion; an inclination to apply reduction in a sweeping fashion; and utter silence on oversight, redress or due process mechanisms.[9]

Furthermore, Zuckerberg's post marks a worrying missed opportunity for policymakers and society at large. Building on the lessons learned during the last decade concerning platforms' central role in shaping the information and communication landscape, [10] and the backlash against unaccountable and nontransparent content moderation processes, one would expect Zuckerberg's post to invoke vigilance and

[6] *See id.*; Zuckerberg, *A Blueprint*, *supra* note 1 .

[7] Another early communication is a blog post that was published in May 2018, where the company shortly mentioned reduction when describing their "three-pronged" (remove, reduce, inform) approach to misleading or harmful content. Tessa Lyons, *The Three-Part Recipe for Cleaning up Your News Feed*, META (May 22, 2018), https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/ ("There are . . . types of problematic content that, although they don't violate our policies, are still misleading or harmful and that our community has told us they don't want to see on Facebook—things like clickbait or sensationalism. When we find examples of this kind of content, we *reduce* its spread in News Feed using ranking," emphasize in original) [https://perma.cc/J8ZV-RE6L]. In later communications, however, the company stated that it had begun applying reduction even earlier. *See Remove, Reduce, Inform: New Steps to Manage Problematic Content*, META (Apr. 10, 2019), ttps://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/ [https://perma.cc/8HAL-23AH].

[8] Tarleton Gillespie, *Do Not Recommend? Reduction as a Form of Content Moderation*, 8 SOC. MEDIA SOC. 1, 1 (2022) (explaining that when reduction is applied "[t]he offending content remains on the site, still available if a user can find it directly. However, the platform limits the conditions under which it circulates . . . . ").

[9] *See infra* Parts III & IV.

[10] *See, e.g.*, Jack M. Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011, 2015 (2018); NICOLAS P. SUZOR, LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES 6–9 (2019); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 1598 (2017); TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 5 (2018); Ira Steven Nathenson, *Super Intermediaries, Code, Human Rights*, 8 INTERCULTURAL HUM. L. REV. 19, 158 (2013).

prompt urgent scrutiny around reduction and its implications. Unfortunately, this did not happen. Since 2018 reduction was vigorously developed by Meta, largely free from scrutiny. As will be shortly shown, during this time, reduction has quietly transformed content moderation as we knew it.

B. Reduction Today

One cannot overstate the current dramatic effect of reduction on the content users consume. In one of the rare instances in which Meta provided data in this regard, the company explained that reduction led to about an 80% decrease in average views.[11] Moreover, in some cases, reduction might be applied to a wider scope of content compared to removal.[12] Meta's response to the U.S. administration's accusations regarding the part played by digital platforms in spreading misinformation related to the COVID-19 vaccines[13] revealed that reduction was applied to approximately nine times more pieces of content compared to removal (167 million pieces of content were reduced while only 18 million pieces of content were removed during a certain period in the pandemic).[14]

In addition, reduction is now applied across all fields of content. Originating with clickbait, misinformation, and sensitive content,[15] the current application of reduction encompasses the vast landscape of any content likely to violate the Community Standards (but that the company cannot confirm that it is indeed violating), as well as content that borders on these Standards (but does not reach them). Reduction also extends to multiple additional types of content, including "low quality" comments and events, comments likely to be reported, and privacy-violating content.[16]

[11] Tessa Lyons, *Hard Questions: How Is Facebook's Fact-Checking Program Working?*, META (June 14, 2018), https://about.fb.com/news/2018/06/hard-questions-fact-checking/ [https://perma.cc/8P9T-KZ25]. See *infra* Part III.B, for elaboration on the impact of reduction in suppressing exposure of content.

[12] Guy Rosen, *Moving Past the Finger Pointing*, META (July 17, 2021), https://about.fb.com/news/2021/07/support-for-covid-19-vaccines-is-high-on-facebook-and-growing/ [https://perma.cc/8WW6-U3WF].

[13] Zolan Kanno-Youngs & Cecilia Kang, *'They're Killing People': Biden Denounces Social Media for Virus Disinformation*, N.Y. TIMES, July 16, 2021, https://www.nytimes.com/2021/07/16/us/politics/biden-facebook-social-media-covid.html [https://perma.cc/K5X6-D89D].

[14] Rosen, *supra* note 12. *See also* discussion *infra* Section III.B.

[15] *Id.*

[16] Types of Content We Demote, META TRANSPARENCY CTR., https://transparency.meta.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/ (Oct. 16, 2023) [https://perma.cc/9PZ2-NSCR]; Our Approach to Facebook Feed Ranking, META TRANSPARENCY CTR. (Nov. 28, 2023),

The picture that emerges is that through reduction, Meta has effectively raised the entire normative threshold determining what content we are allowed to access and share. Even though the company consistently presents the Community Standards as those that "outline what is and isn't allowed on Facebook,"[17] in reality, it is the reduction's enforcement that makes that call.

Besides its wide application, reduction also evolved into a highly sophisticated and versatile practice from the platform's perspective. It could be performed in various ways, including downranking content in the News Feed and adjusting the recommendation system; excluding content from dominant areas in the platform (like Facebook's "Search," or Instagram's "Explore" and "Hashtag" pages); and being combined with other sanctions.[18] In certain cases, reduction relies on a designated choice architecture[19] and is outsourced to the users themselves through tools such as the option to automatically hide comments that include pre-selected words.[20] Lastly, unlike removal, which, in Meta's words, applies "equally to everyone, everywhere,"[21] reduction may not necessarily be applied globally. Indeed, Meta acknowledges that they may "temporarily adjust . . . [their] enforcements in a specific region or during a critical event."[22] This too highlights the elastic and multifaceted nature of reduction.

A significant challenge posed by reduction concerns the secrecy that surrounds it. While removal is publicly and willingly discussed, Meta's approach to reduction is much more low-key. Unlike removal, users whose content was reduced are often not notified and are not provided with an explanation. Even if they become aware or suspect that their content was reduced, they cannot appeal to either Meta's

https://transparency.meta.com/features/ranking-and-content/ [https://perma.cc/KER9-SP5N]. See also discussion infra Part III.B.

[17] *Facebook Community Standards*, *supra* note 3. *See* discussion *infra* Part II.

[18] *About Fact-Checking on Facebook and Instagram*, META BUS. HELP CTR., https://www.facebook.com/business/help/2593586717571940?id=673052479947730 (last visited Feb. 5, 2024) [https://perma.cc/K5VM-4WCR]. *See also* discussion *infra* Part III. B.

[19] *See* discussion *infra* Part III.B.

[20] *Types of Content We Demote*, *supra* note 16. *See also* Garnier v. O'Connor-Ratcliff, 41 F.4th 1158, 1164 (9th Cir. 2022). *How Do I Block Certain Words from Appearing in Comments on My Facebook Page?*, FACEBOOK HELP CTR., https://www.facebook.com/help/131671940241729?helpref=faq_content (last visited Feb. 5, 2024) [https://perma.cc/P24U-FX3L]. *See also* discussion *infra* Part III.B.

[21] *The Facebook Community Standards Apply the Same to Everyone Everywhere*, META TRANSPARENCY CTR., https://transparency.meta.com/policies/improving/policies-apply-to-everyone-everywhere (Jan. 19, 2022) [https://perma.cc/AZ8T-UTFS].

[22] *Types of Content We Demote*, *supra* note 16. See *infra* Part III.B, for the opportunities and challenges this creates.

internal review mechanism or the company's Oversight Board.[23] In contrast with the Community Standards, the "Content Distribution Guidelines" that govern reduction are abstract and brief, reflecting an incoherent assortment of vastly unrelated content types. Meta does not present these guidelines as carrying normative or behavior-guiding weight; they are usually not featured in reports to enforcement authorities, nor are they subject to open, nuanced, and consistent amendment processes. Additionally, conversely to the Community Standards, data concerning the enforcement of the "Content Distribution Guidelines" is largely absent from the company's voluntary Transparency Reports.[24] Unlike removal, reduction is also missing from the principal part of the regulation that governs content moderation around the world.[25]

However, the troubling strategy of reduction and its associated policy remain shielded from scrutiny, concealed behind the responsible and desirable image of the Community Standards and the sanction of removal. Indeed, the Community Standards and content removal represent much more detailed, human-rights-respecting, publicly scrutinized, and openly updated sources, as will be elaborated below. It appears that Meta encourages a focus on these governance elements, steering attention away from the chaotic characteristics of reduction.[26]

The origins of this masking strategy can be traced back to Zuckerberg's 2018 post, which began with a comprehensive account of the Community Standards. It stated "[e]very community has standards, and since our earliest days we've also had our Community Standards—the rules that determine what content stays up and what comes down on Facebook."[27] Zuckerberg then proceeded to describe various virtues of these policies, including their carefully articulated provisions, designed to encourage consistent enforcement, and the thorough, diverse, and frequent updates they undergo.[28] This approach, which highlights removal and the Community Standards as the normative framework for content moderation and takes pride in their refined attributes, has allowed the company to develop and apply reduction undisturbedly in the "backstage" of content moderation, transforming it into the powerful and troubling practice it is today.[29]

[23] *See infra* Parts II, III.B & IV.B.

[24] *Id.*

[25] *Id.* See *infra* Parts III.B & IV.B., for a discussion of the inclusion of reduction in recent regulations.

[26] *See infra* Parts II & III.B. *See also* Gillespie, *supra* note 8, at 2 ("It is not that reduction techniques are hidden entirely, but platforms benefit from letting them linger quietly in the shadow of removal policies.").

[27] Zuckerberg, *supra* note 1.

[28] *Id.*

[29] *See infra* Parts II & III.B.

Under these conditions, reduction has matured into a widespread and influential practice, characterized by opacity, lack of accountability, and user disempowerment. The secrecy that has marked it from the outset led to the coining of the term "shadow banning," which is frequently used by users and others to describe various manifestations of reduction.[30]

Such a lack of transparency also surrounds the incentives and interests that drive reduction. As will be discussed below, Meta provided a puzzling and incomplete set of considerations in this regard.[31] One significant factor that is absent from Meta's account concerns the pressure applied by governments. A judgment passed by the Court of Appeals for the 5th Circuit found that, at least since 2020, federal officials "coerced social-media platforms into censoring certain social media content, in violation of the First Amendment."[32] This censorship included not only the removal of content but also its reduction.[33] The Supreme Court recently reversed the judgment due to lack of standing, but some of its findings, which I will later address, are nonetheless worrying.[34]

The current application of reduction poses a series of additional challenges: first, it conflicts with the rule of law and undermines procedural fairness; second, it disproportionately hinders freedom of expression and other human rights; third, as an AI-driven strategy, reduction is susceptible to various vulnerabilities, including biases, limitations in contextual understanding, low performance in minority languages, and a lack of transparency and explainability.[35]

[30] Jesselyn Cook, *Instagram's CEO Says Shadow Banning 'Is Not A Thing.' That's Not True.*, HUFFPOST UK (2020), https://www.huffpost.com/entry/instagram-shadow-banning-is-real_n_5e555175c5b63b9c9ce434b0 (addressing shadow banning as "the secret censorship of a person, topic or community on social media") [https://perma.cc/DT6K-P885]. *See also* Chanté Joseph, *Instagram's Murky 'Shadow Bans' Just Serve to Censor Marginalised Communities*, THE GUARDIAN (Nov. 8, 2019), https://www.theguardian.com/commentisfree/2019/nov/08/instagram-shadow-bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive ("Shadow banning refers to when images aren't outright removed from the platform, but instead strategically hidden from users") [https://perma.cc/QSX5-SN6Q]. *See also* Eric Goldman, *Content Moderation Remedies*, 28 MICH. TECH. L. REV. 1, 31 (2021).

[31] *See infra* Part IV.A.

[32] *See generally* Missouri v. Biden, 80 F.4th 641 (5th Cir. 2023). *See also* discussion *infra* Part IV.A.

[33] *Id.* at 2. *See also* Brendan Pierson, *Unusual orders sow confusion in case over Biden social media contacts,* REUTERS (Sept. 27, 2023), https://www.reuters.com/legal/government/unusual-orders-sow-confusion-case-over-biden-social-media-contacts-2023-09-26/ [https://perma.cc/53HD-7YUD].

[34] Murthy v. Missouri, 144 S. Ct. 1972, 1985 (2024).

[35] *See infra* Part IV.B.

Does all this suggest that reduction is inherently detrimental? Not necessarily. In fact, reduction may offer a versatile and flexible approach for tackling the challenges posed by the immense volume of content being produced and shared, as well as the variety of content formats, such as video-streaming and AI-generated imagery.[36] Reduction, which does not involve the physical removal of the content, also aligns with the growing recognition of user-generated content as valuable data, and with the advantageous potential applications that depend on its availability.[37] Furthermore, the adaptable nature of reduction could help manage content moderation errors and shifts in societal perceptions regarding what constitutes unacceptable content.[38]

However, to properly leverage these potential benefits, the considerable difficulties associated with reduction must be addressed. In an era where the influence of digital platforms is compared to "states,"[39] "governors,"[40] "custodians,"[41] and "information fiduciaries,"[42] with the U.S. Supreme Court depicting social media as the primary means for "exploring the vast realms of human thought and knowledge,"[43] and the UN Secretary-General describing social media as increasingly dominant in "how individuals access and share information and ideas,"[44] overlooking the drawbacks of reduction could inflict significant harm.

The focus of this Article is on the reduction strategy implemented by Meta, analyzed as a case study, and encompassing both Facebook and Instagram. However, many of the outlined challenges

[36] For a discussion of these challenges, see *infra* Part III.A.

[37] *See, e.g.*, Paul Sawers, *Meta Reignites Plans to Train AI Using UK Users' Public Facebook and Instagram Posts*, TECHCRUNCH (Sep. 13, 2024), https://techcrunch.com/2024/09/13/meta-reignites-plans-to-train-ai-using-uk-users-public-facebook-and-instagram-posts/ [https://perma.cc/MP3K-Y4EC].

[38] *See infra* Part IV.C.

[39] Noa Mor, *No Longer Private: On Human Rights and the Public Facet of Social Network Sites*, 47 HOFSTRA L. REV. 651, 691 (2018).

[40] Klonick, *supra* note 10, at 1603.

[41] Gillespie, *supra* note 10.

[42] Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49(4) U.C. DAVIS L. REV. 1183, 1186 (2016); *see also* Jonathan Zittrain, *How to Exercise the Power You Didn't Ask For*, HARV. BUS. REV. (Sept. 19, 2018), https://hbr.org/2018/09/how-to-exercise-the-power-you-didnt-ask-for [https://perma.cc/WT3K-S84D].

[43] Packingham v. North Carolina, 582 U.S. 98, 99 (2017).

[44] U.N. Secretary-General, *Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 10, U.N. Doc. A/73/348 (Aug. 29, 2018); *see also* Frank La Rue, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 44, U.N. Doc. A/HRC/17/27 (May 16, 2011).

apply more broadly, resonating with other platforms, like X (previously Twitter) and YouTube.[45]

The Article unfolds as follows: Part II explores how Meta has been cultivating removal and the Community Standards as the "window dressing" of content moderation. Part III opens with a glimpse into the evolving landscape of content moderation against which reduction has been embraced. It then examines the ascent of reduction the challenges its policy poses, and how it is intentionally relegated to the "backstage" of content moderation. This Part concludes with a comparison of reduction to other attention control mechanisms used by Meta. Part IV looks at the motivations powering the reduction strategy and the pressing legal challenges it raises concerning the rule of law and procedural fairness, along with freedom of expression and additional human rights. It also discusses the implication of conducting reduction in light of current AI vulnerabilities and limitations. Finally, it charts a path forward and proposes several recommendations to address the concerns surrounding reduction while harnessing its benefits.

II. REMOVAL: CONTENT MODERATION'S WINDOW DRESSING

> *"Facebook's normal penalties include removing the violating content, imposing a time-bound period of suspension, or permanently disabling the page and account."*[46]

Removing content and disabling accounts have been the first and primary moderation tools used by Meta and other digital platforms to control the content displayed and shared on their services.[47] Over time, additional sanctions were introduced, namely territorial blockage of content that violated local laws, warning screens, and labels.[48] However, removal, or the "binary approach" as Eric Goldman terms it, is still perceived by many as "the conventional wisdom" of what content moderation is.[49]

[45] Gillespie, *supra* note 8; *see also* Missouri v. Biden, 80 F.4th 641, 660 (5th Cir. 2023).

[46] *Case Decision 2021-001-FB-FBR*, OVERSIGHT BD., https://www.oversightboard.com/decision/FB-691QAMHJ/ (last visited Feb. 5, 2024) [https://perma.cc/FC4H-CNC9].

[47] Klonick, *supra* note 10, at 1638–1648.

[48] Gillespie, *supra* note 8. Territorial blockage applies to a much smaller scope of content than removal, *see Content Restrictions Based on Local Law*, META TRANSPARENCY CTR., https://transparency.meta.com/reports/content-restrictions/ (last visited Feb. 5, 2024) [https://perma.cc/N6HE-59DN].

[49] Goldman, *supra* note 30, at 5.

Indeed, in recent years, the sanction of removal has garnered considerable attention and scrutiny from various actors, including civil society organizations and policymakers.[50] This focus arises from several reasons, many of which—though not all—are promoted by the platforms themselves through their efforts to position the Community Standards and the removal sanction as the emblematic representation of content moderation.[51]

First, as mentioned earlier, removal is a binary, intuitive, and easy-to-understand sanction.[52] We all share a common understanding of this "taking down" action and its implication—that the content has been deleted and is no longer accessible on the platform. This seemingly trivial characteristic of removal is significant, as it facilitates the engagement and discussion of the public, activists, and policymakers regarding this sanction.

Second, users whose content has been removed are notified and provided with an explanation (although this explanation typically addresses the broad Community Standard violated, rather than the specific rule within that standard).[53] Meta emphasized the significance of this approach, stating "[w]e'll let you know when something you posted goes against our Community Standards. . . . [W]e'll reference which part of the Community Standards you didn't follow, as well as a brief description of why the content isn't allowed, so you can avoid having content removed in the future."[54]

The notification and explanation provided to users can ignite public criticism and deliberation concerning the sanction's legitimacy and regarding possible errors or biases.[55] These discussions may unfold across various avenues, including users' social media, and through news and NGO reporting.

[50] *See infra* Part II.
[51] *Id.*
[52] Goldman, *supra* note 30, at 5.
[53] *Taking down violating content*, META TRANSPARENCY CTR., https://transparency.meta.com/enforcement/taking-action/taking-down-violating-content/ (last visited Feb. 5, 2024) [hereinafter *Taking down violating content*] [https://perma.cc/JBM3-Q2VN]; *What happens when Facebook removes my content?*, FACEBOOK HELP CTR., https://www.facebook.com/help/260743102021762 (last visited Feb. 5, 2024) [https://perma.cc/48N9-VR9V].
[54] S*upra* note 53. Meta even provides a "Takedown experience" that simulates such notification (and the channel to appeal it, as will be explored below). *Taking down violating content*, *supra* note 53.
[55] See *Why Was My Simple Post Removed for Cybersecurity Reasons by Meta When There Was Nothing Related to a Cybersecurity Threat?*, R/FACEBOOK (2023), www.reddit.com/r/facebook/comments/18atov4/why_was_my_simple_post_removed_for_cybersecurity/ (last visited Oct 13, 2024) [https://perma.cc/MT6Y-SWJU], for a discussion of this.

Third, users whose content has been removed have the opportunity to appeal the decision.[56] Meta has emphasized the importance of this review process, stating that it "allows people to let us know if they think we've made a mistake, which is essential to help us build a fair system."[57] For Meta users whose appeals are declined, an additional appealing avenue exists: the Oversight Board.[58] Though the Board hears a limited number of cases each year,[59] its decisions attract considerable public engagement and receive widespread exposure, as will be explored further in this Part.

Fourth, removal practices are deeply embedded within regulatory frameworks, thereby gaining international attention and legitimacy. One example is the 2016 Code of Conduct for Countering Illegal Hate Speech ("The Code of Conduct" or "The Code"),[60] an impactful agreement between the European Commission and digital companies, including Meta, X, and YouTube. The Code governs the moderation of hate speech and terror-related content, stipulating[61] that "[t]he IT companies [commit] to have in place Rules or Community Guidelines clarifying that they prohibit the promotion of incitement to violence and hateful conduct. Upon receipt of a valid removal

[56] *I don't think Facebook should have taken down my post*, FACEBOOK HELP CTR., https://www.facebook.com/help/2090856331203011?helpref=faq_ (last visited Feb. 5, 2024) [https://perma.cc/X4US-Z848]; *see also Appealed content*, META TRANSPARENCY CTR., https://transparency.meta.com/policies/improving/appealed-content-metric/ 18, 2022) [https://perma.cc/BS32-984B]. The option to ask for a review may not be given in certain cases. *Id*.

[57] *Appealed content*, *supra* note 56.

[58] *Appeal a Facebook content decision to the Oversight Board*, FACEBOOK HELP CTR., https://www.facebook.com/help/346366453115924 (last visited Feb. 5, 2024) [https://perma.cc/KUQ2-69H8]; Not all cases reviewed by Meta qualify for appeal to the Board. *Id.*

[59] 2*022 Annual Report: Oversight Board Reviews Meta's Changes to Bring Fairness and Transparency to its Platforms*, OVERSIGHT BD. (June 6, 2023), https://www.oversightboard.com/wp-content/uploads/2023/11/795921088637952.pdf.

[60] *EU Code of conduct on countering illegal hate speech online*, EUR. COMM'N, https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (last visited Feb. 5, 2024) [https://perma.cc/ET8Z-TLHS].

[61] The initial signatories of the Code of Conduct included these three companies, along with Microsoft. Subsequently, other companies such as LinkedIn, Snapchat, and TikTok also joined the Code. *EU Code of Conduct Against Illegal Hate Speech Online: Results Remain Positive but Progress Slows Down*, PUBAFFAIRS BRUXELLES, https://www.pubaffairsbruxelles.eu/eu-institution-news/eu-code-of-conduct-against-illegal-hate-speech-online-results-remain-positive-but-progress-slows-down/ (last visited Feb. 5, 2024) [https://perma.cc/2LPQ-7V6M].

notification, the IT Companies [commit] to review such requests against their rules and community guidelines."[62]

Subsequent regulations, such as the German Act to Improve Enforcement of The Law in Social Networks ("NetzDG")[63] and the Austrian "Communications Platform Law" ("KoPl-G"),[64] further underscore the centrality of removal by treating it as the primary moderation tool. It is also important to note that Meta issues periodic and detailed public reports on its compliance with some of these regulations, thereby enhancing the scrutiny and accountability directed at removal practices.[65]

Fifth, removal is closely linked to the Community Standards. Meta has consistently emphasized "[i]f your content goes against our Community Standards, we'll remove it."[66] In this context, the Community Standards and removal are the same. Since the Community Standards are subject to extensive transparency, deliberation, and even endorsement, as will be further discussed below, some of these advantages extend to the practice of removal. The prominence and visibility of the Community Standards are supported by a robust array of channels:

[62] *The EU Code of Conduct on Countering Illegal Hate Speech Online*, *supra* note 60. Another aspect that emerges from the Code is that the platforms' policies, which position removal as the primary sanction, serve as the normative standard for assessing the legality of content. *See id. See also* Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power Over Online Speech*, HOOVER INST. 6, https://s3.documentcloud.org/documents/5699593/Who-Do-You-Sue-State-and-Platform-Hybrid-Power.pdf; Mor, *supra* note 39, at 679–680.

[63] Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [NetzDG] [Act to Improve Enforcement of the Law in Social Networks], Oct. 1, 2017, Netzwerkdurchsetzungsgesets vom 1 (Ger.), https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html. For periodic reports submitted by Meta in accordance with this law, see R*egulatory and Other Transparency Reports*, META TRANSPARENCY CTR., https://transparency.meta.com/reports/regulatory-transparency-reports/ (last visited Feb. 5, 2024) [https://perma.cc/A3BP-P939].

[64] RIS—Kommunikationsplattformen-Gesetz, [Commuications Platform Act] Feb. 16, 2024 No. 151/2020, https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20011415&FassungVom=2024-02-16&ShowPrintPreview=True (Austria). *See also* Gabriela Staber & Angelika Stütz, *Communication platforms face new obligations and high fines in Austria*, LEXOLOGY, https://www.lexology.com/library/detail.aspx?g=fcf46df4-4694-4f10-b11b-67564a824470 (last visited: Nov. 18, 2024) [https://perma.cc/8C9F-7NAF]. See *Regulatory and Other Transparency Reports*, *supra* note 63, for periodic reports submitted by Meta in accordance with this law.

[65] *See infra* Part II.

[66] *What happens when Facebook removes my content?*, *supra* note 53; *See also How Meta's third-party fact-checking program works*, META (June 1, 2021), https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works [https://perma.cc/4J6H-ASFC].

(1) The Community Standards are presented and administered by Meta as the normative policy, delineating allowed and disallowed content on Facebook and Instagram.[67] As such, they are formulated as detailed, user-centric provisions that incorporate "do not post" examples and graphics to facilitate understanding.[68] Meta ensures that the Community Standards are not only comprehensible and easy to follow but also readily accessible to users and other stakeholders. For instance, these standards are prominently displayed on the landing page of Meta's Transparency Center and are listed first under the "Policy" section there.[69]

(2) Data regarding removal is provided in Meta's voluntary Transparency Reports, and more specifically, in its quarterly "Community Standards Enforcement Report."[70] First published in 2018,[71] this report offers a range of metrics related to the enforcement of the Community Standards, including the prevalence

[67] Facebook Community Standards, *supra* note 3; *See also Corporate Human Rights Policy*, META, https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf ("Our Community Standards . . . outline what user-generated content is and is not allowed on Facebook") (last visited Feb. 5, 2024) [https://perma.cc/KX7V-S3ZB]; The Community Standards are also applicable to Instagram users, despite Instagram having its distinct content policy, called the "Community Guidelines." *See* Oversight Board Overturns Original Facebook Decision in Breast Cancer Symptoms and Nudity Case, OVERSIGHT BD., (Jan. 28, 2021), https://oversightboard.com/news/682162975787757-oversight-board-overturns-original-facebook-decision-case-2020-004-ig-ua/ [https://perma.cc/925E-K8CH].

[68] *See, e.g.*, *Hate Speech*, META TRANSPARENCY CTR., https://transparency.meta.com/policies/community-standards/hate-speech/ (last visited Feb. 5, 2024) [https://perma.cc/7P3U-Z7JH].

[69] *Policies*, META TRANSPARENCY CTR., https://transparency.meta.com/policies/ (last visited Feb. 5, 2024) [https://perma.cc/ABE2-AEZR].

[70] *Community Standards Enforcement*, META TRANSPARENCY CTR., https://transparency.fb.com/reports/community-standards-enforcement/ (last visited Feb. 5, 2024) [https://perma.cc/5WTV-358X]; Meta also submits other types of Transparency Reports. They include, *inter alia*, "Government Requests for User Data," "Intellectual Property," "Internet Disruptions," "Content Restrictions" (territorial blockage of content that violates local law), "Regulatory and Other transparency Reports" (including reports mandated by state regulations), and the "Widely Viewed Content Report (WVCR)." *See Transparency Reports*, META TRANSPARENCY CTR., https://transparency.meta.com/reports/ (last visited Feb. 5, 2024) [https://perma.cc/A98A-2WVF]. For the WVCR, see *Widely Viewed Content Report: What People See on Facebook*, META TRANSPARENCY CTR. (Feb. 6, 2024), https://transparency.fb.com/data/widely-viewed-content-report/ [https://perma.cc/UD8Q-4JWV].

[71] *Facebook Publishes Enforcement Numbers for the First Time*, META, (May 15, 2018), https://about.fb.com/news/2018/05/enforcement-numbers/ [https://perma.cc/3P4M-QMGH].

of such content, the volume of sanctioned content,[72] proactive enforcement rate,[73] appealed content, and restored content.[74]

(3) In recent years, particularly since 2018,[75] Meta has been diligently updating and refining its Community Standards, making them more comprehensive, nuanced, and consistent.[76] This process engages an array of stakeholders, including representatives from academia and civil society.[77] Meta publicly discusses this enhancement process, emphasizing the company's commitment to a collaborative, informed, and transparent approach to content moderation.[78]

A pivotal role in refining the Community Standards is played by Meta's Oversight Board. Conceived and established by Meta itself in 2018, the Board operates with a considerable degree of independence, yet its mandate, along with other aspects of its

[72] The "Content actioned" metric within the report provides information on content that was removed, disabled accounts, and content marked with warnings. *See Content Actioned*, META TRANSPARENCY CTR., https://transparency.meta.com/policies/improving/content-actioned-metric/ (Nov. 7, 2023) [https://perma.cc/E8DU-4LWS].

[73] The proportion of content on which Meta took action before it was reported by users. See *Proactive Rate*, META TRANSPARENCY CTR., https://transparency.meta.com/policies/improving/proactive-rate-metric/ (Feb. 22, 2023) [https://perma.cc/7PE3-2XN3].

[74] *Community Standards Enforcement*, *supra* note 70.

[75] In 2018, the Community Standards underwent a significant transformation when the company revised them to include the platform's internal guidelines for content moderation. This revision followed the leakage of these guidelines and the subsequent public outcry over the lack of transparency in content moderation. *See Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process*, META (Apr. 24, 2018), https://about.fb.com/news/2018/04/comprehensive-community-standards/ [https://perma.cc/C2FJ-98JK]; Casey Newton, *Facebook Makes Its Community Guidelines Public and Introduces an Appeals Process*, THE VERGE (Apr. 24, 2018), https://about.fb.com/news/2018/04/comprehensive-community-standards/ [https://perma.cc/6389-CWA3].

[76] Tarleton Gillespie, *Facebook's Improved "Community Standards" Still Can't Resolve the Central Paradox*, SOC. MEDIA COLLECTIVE (Mar. 18, 2015), https://socialmediacollective.org/2015/03/18/facebooks-improved-community-standards-still-cant-resolve-the-central-paradox/ [https://perma.cc/P89T-YBPL]; Klonick, *supra* note 10, at 1631–1635.

[77] *See, e.g.*, *Meta's Annual Human Rights Report*, META (2023), https://humanrights.fb.com/annual-human-rights-report/ [https://perma.cc/NP3J-YXAM].

[78] *See, e.g.*, *Input From Community and Partners On Our Community Standards*, META (Oct. 21, 2016), https://about.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-standards/ [https://perma.cc/4KRR-G5VN]. *See also* RISKommunikationsplattformen-Gesetz, *supra* note 64, *and Meta Human Rights Report: Insights and Actions 2021–2022*, FACEBOOK 20 (2022), https://humanrights.fb.com/wp-content/uploads/2024/09/2023-Meta-Human-Rights-Report.pdf

operation, is defined by the company.[79] The Oversight Board's authority is primarily focused on reviewing content removal cases that have undergone Meta's internal appeal process.[80] Recently, the Board's jurisdiction was expanded to include decisions regarding the application of warning screens to content Meta chooses to retain.[81] Facebook's Community Standards are instrumental in the Board's deliberations.[82] The Board conducts thorough reviews of the policies' phrasing and interpretation, evaluates their application against international human rights norms and Meta's values,[83] and recommends amendments to their provisions.[84] Elucidating the Oversight Board's role in refining the Community Standards, David Wong and Luciano Floridi noted:

> By reviewing Meta's policies, the [Oversight Board] can highlight blind spots in [the] Community Standards . . . . [T]he [Oversight Board]'s decisions, investigations, and findings can provide Meta with the insights to address blind spots. Such roles can also enable the public to understand and discuss platform content moderation decisions, and how platforms balance concerns of

[79] Brent Harris, *Establishing Structure and Governance for an Independent Oversight Board*, META (Sep. 17, 2019), https://about.fb.com/news/2019/09/oversight-board-structure/ [https://perma.cc/QM43-S2Q7].

[80] *Oversight Board Charter*, OVERSIGHT BD. 5 (Feb. 2023), https://www.oversightboard.com/wp-content/uploads/2023/11/3427086457563794.pdf. According to the Board's Charter, it is authorized to "interpret Facebook's Community Standards and other relevant policies." While the phrasing allows for the inclusion of more content policies, the Board's decisions to date primarily addressed the Community Standards.

[81] 2022 Annual Report, *supra* note 59, at 17; *Oversight Board publishes transparency report for second quarter of 2022 and gains ability to apply warning screens*, OVERSIGHT BD. (Oct. 20, 2022 https://www.oversightboard.com/news/784035775991380-oversight-board-publishes-transparency-report-for-second-quarter-of-2022-and-gains-ability-to-apply-warning-screens/ [https://perma.cc/N79W-F6UE]. *See also Oversight Board Bylaws*, OVERSIGHT BD. 25 (Feb. 2023) https://www.oversightboard.com/wp-content/uploads/2024/03/Oversight-Board-Bylaws.pdf.

[82] For these decisions, see *Decision*, OVERSIGHT BD., https://www.oversightboard.com/decision/ (last visited Feb. 5, 2024) [https://perma.cc/9TTZ-VMAU]. For tables summarizing the decisions, *inter alia*, according to the Community Standard that was discussed, see *2022 Annual Report*, *supra* note 59 at 39; *2021 Annual Report*, OVERSIGHT BD., 24–25 (Jun. 2022), https://www.oversightboard.com/wp-content/uploads/2023/11/425761232707664.pdf .

[83] *See generally Oversight Board Charter*, *supra* note 80.

[84] These suggestions could be introduced under the Board's authority to issue nonbinding "policy advisory opinions." *See Oversight Board Charter*, *supra* note 80, at 5.

> freedom of expression with other values such as safety and diversity.[85]

As clarified in this text, the Board's decisions not only improve the formulation of the Community Standards but also encourage public involvement in this regard. This influence of the Board is magnified by extensive media coverage[86] and academic research focusing on its decisions.[87] Moreover, Meta publishes quarterly reports on its adherence to the Board's recommendations[88] and addresses them via additional channels, like the company's Newsroom.[89] An additional factor contributing to the Board's impact on the Community Standards' evolution is the comments that are submitted by individuals and organizations worldwide on the cases being reviewed, many of which delve into the Standards' clauses.[90] In 2022 alone, for example, the Board received over 600 such comments.[91]

(4) Other avenues for enhancing transparency around the Community Standards include various reports submitted by Meta to regulators

[85] David Wong & Luciano Floridi, *Meta's Oversight Board: A Review and Critical Assessment*, 33 MINDS MACH. 261, 266 (2023).

[86] Shannon Bond, *In 1st Big Test, Oversight Board Says Facebook, Not Trump, Is the Problem*, NPR (May 7, 2021), https://www.npr.org/2021/05/07/994436847/what-we-learned-about-facebook-from-trump-decision [https://perma.cc/DJ29-5TPY]; Wong & Floridi, *supra* note 85.

[87] *See, e.g.*, *id*. *See generally* Andreas Kulick*, Meta's Oversight Board and Beyond—Corporations as Interpreters and Adjudicators of International Human Rights Norms*, 22 L. & PRAC. INT'L CTS. & TRIBS. 161 (2022).

[88] *Oversight Board*, META TRANSPARENCY CTR., https://transparency.meta.com/oversight/overview (last visited Feb. 5, 2024) [https://perma.cc/4X2F-VJKY].

[89] *See, e.g.*, *Facebook Newsroom: Oversight Board Upholds Facebook's Decision to Suspend Donald Trump's Accounts*, FACEBOOK FOR GOV'T, POL. & ADVOC. (May 5, 2021), https://www.facebook.com/government-nonprofits/blog/facebook-newsroom-oversight-board-decision [https://perma.cc/L3RN-6QCQ]. Observe the company's pride in adhering to the Oversight Board's feedback: "Our Transparency Center provides a hub for Facebook's and Instagram's integrity and transparency work, acting as a central destination for all updates on how we enforce Facebook's Community Standards and how we respond to decisions, recommendations, and case updates from the Oversight Board." *2022 Annual Report*, *supra* note 59, at 20.

[90] See, *e.g.*, Justin Hendrix, *Civil Rights and Watchdog Groups React to Facebook Oversight Board Decision on Donald Trump*, TECH. POL'Y PRESS (May 6, 2021), https://www.techpolicy.press/civil-rights-and-watchdog-groups-react-to-facebook-oversight-board-decision-on-donald-trump/ [https://perma.cc/H4FF-TW44]; Jillian C. York, *EFF's Comment to the Meta Oversight Board on United States Posts Discussing Abortion*, ELEC. FRONTIER FOUND. (July 10, 2023), https://www.eff.org/deeplinks/2023/07/effs-comment-meta-oversight-board-united-states-posts-discussing-abortion [https://perma.cc/T83H-W9QC]

[91] *2022 Annual Report*, *supra* note 59.

and other stakeholders. In these documents, the company aims to demonstrate its commitment to legal obligations, consistently showcasing the Community Standards as a testament to its responsible and carefully crafted policy framework. Meta's first Annual Human Rights Report highlights this approach.

> We look to international human rights experts when developing our standards for what content is and is not allowed on our social media platforms, and when deciding how to implement these standards in practice. These rules are known as the Community Standards for Facebook and Community Guidelines for Instagram. They were—and are—developed based on feedback from our community and the advice of experts in fields such as technology, public safety and human rights. In seeking to ensure that everyone's voice is valued equally, we take care to create standards that include different views and beliefs, especially from people and communities that might otherwise be overlooked or marginalized.[92]

The same modus operandi is mirrored in other reports, such as the NetzDG Transparency Report, where the discourse on the Community Standards occupies a central role in the introductory section.[93]

III. REDUCTION AND THE EVOLUTION OF CONTENT MODERATION

A. The Changing Landscape of Content Moderation

Over the last two decades since the emergence of Meta (then Facebook) and other major online platforms,[94] dramatic changes have fundamentally reshaped the content moderation landscape. These transformations have paved the way for the recent adoption of the reduction strategy. As will be explored below, some developments address emerging content moderation challenges that reduction is

[92] *Meta Human Rights*, *supra* note 78, at 30.

[93] *NetzDG Transparency Report*, META 1 (Jan.–June 2022), https://about.fb.com/de/wp-content/uploads/sites/10/2022/07/Facebook-NetzDG-Transparency-Report-July-2022.pdf.

[94] Facebook was established in 2004, YouTube in 2005, X (originally known as Twitter) in 2006, and Instagram, which was later acquired by Facebook, in 2010. *See* Alexandra Samur & Colleen Christison, *The History of Social Media in 33 Key Moments*, MEDIA MKTG. & MGMT. DASHBOARD (Apr. 6, 2023), https://blog.hootsuite.com/history-social-media/ [https://perma.cc/BTR2-N4ZF].

designed to tackle, while others provide the technological foundation necessary for reduction's implementation.

A key development in the digital landscape concerns the dramatic expansion of these platforms' user base. The number of social media users has risen meteorically,[95] with Meta's daily users now nearing 3.3 billion.[96] This surge weaves into a broader tapestry of changes, including expansion into non-U.S. markets and a growing presence of users of different cultures, representing diverse perceptions and beliefs.[97] These trends are driven by various factors, with the widespread adoption of smartphones for social media access being a crucial force.[98]

With the growing number of users, many of whom remain constantly logged in,[99] there has also been a significant increase in the scope of user-generated content (UGC).[100] Instagram users, for instance, upload approximately 95 million photos and videos every day.[101] As of 2021, Meta was selecting daily content for each Facebook user from approximately a thousand candidate posts; considering its billions of users, the company handled, in total, trillions of posts every

[95] Belle Wong, *Top Social Media Statistics And Trends Of 2024*, FORBES (May 18, 2023), https://www.forbes.com/advisor/business/social-media-statistics/ [https://perma.cc/6FG7-8A9C]; *Social Media Fact Sheet*, PEW RSCH. CTR. (Jan. 31, 2024), https://www.pewresearch.org/internet/fact-sheet/social-media/ [https://perma.cc/B627-J2RB].

[96] *Meta Reports Third Quarter 2024 Results*, META (Oct. 30, 2024), https://investor.fb.com/investor-news/press-release-details/2024/Meta-Reports-Third-Quarter-2024-Results/default.aspx#:~:text=Revenue%20%E2%80%93%20Total%20revenue%20was%20%2440.59,%25%20year%2Dover%2Dyear [https://perma.cc/V3H3-HGXR].

[97] Jacob Poushter, *Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies*, PEW RSCH. CTR. 21–22 (2016), https://www.pewresearch.org/wp-content/uploads/sites/2/2016/02/pew_research_center_global_technology_report_final_february_22__2016.pdf; DENNIS BROEDERS, THE PUBLIC CORE OF THE INTERNET 21–23 (AUP 2016).

[98] Monica Anderson & Jingjing Jiang, *Teens, Social Media and Technology 2018*, PEW RSCH. CTR. (May 31, 2018), https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ [https://perma.cc/UQV5-K2LK]; Poushter, *supra* note 97, at 3–6.

[99] Andrew Perrin & Sara Atske, *About three-in-ten U.S. adults say they are 'almost constantly' online*, PEW RSCH. CTR. (Mar. 26, 2021), https://www.pewresearch.org/short-reads/2021/03/26/about-three-in-ten-u-s-adults-say-they-are-almost-constantly-online/ [https://perma.cc/6C7E-5QVV].

[100] *Data Never Sleeps 10.0*, DOMO, https://www.domo.com/data-never-sleeps (last visited Feb. 5, 2024) [https://perma.cc/K54C-TPXB].

[101] Jack Flynn, 3*0+ Instagram Statistics [2023]: Facts About This Important Marketing Platform*, ZIPPIA (Mar. 23, 2023), ://www.zippia.com/advice/instagram-statistics/ [https://perma.cc/8JAJ-LAAL].

day.[102] Concurrently with the growth in UGC, the scale of harmful content—including, misinformation, copyright violations, bullying, incitement to violence and racism, hate speech, and terror-related content—has also dramatically grown.[103] UGC not only increased in volume but also diversified significantly across languages, dialects, designs, and formats. This includes text and imagery combinations; 3D images;[104] videos;[105] video-streaming;[106] short videos like "Stories" or "Reels";[107] and AI-generated content,[108] including deepfakes and other synthetic media.[109] These shifts pose substantial and multifaceted challenges for content moderation processes.[110]

An additional change in the content moderation landscape is the evolving regulatory framework. Since the mid-2010s, new regulations have introduced various obligations for digital platforms to curb harmful content, thereby intensifying the challenges they face in the content moderation context. These new regulations include the

[102] Akos Lada, Meihong Wang & Tak Yan, *How Does News Feed Predict What You Want to See?*, META (Jan. 26, 2021), https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/ [https://perma.cc/M6SW-Y7AA].

[103] BIG DATA SOC'Y, 1 (2020).

[104] *How do I create a 3D photo on Facebook?*, FACEBOOK HELP CTR., https://www.facebook.com/help/iphone-app/414295416095269 (last visited Feb. 5, 2024) [https://perma.cc/K3YY-B95S].

[105] Amanda Silberling, *Meta says its metaverse biz lost another $3b ... but the 2030s will be 'exciting,'* TECHCRUNCH (Apr. 27, 2022), https://techcrunch.com/2022/04/27/meta-facebook-q1-2022-earnings/?guccounter=1 (reporting that video "accounts for 50% of the time that users spend on Facebook") [https://perma.cc/VW88-E5LU].

[106] *Facebook Live*, META, https://www.facebook.com/formedia/tools/facebook-live (last visited Feb. 5, 2024) [https://perma.cc/DX82-5B2A].

[107] Romain Dillet, *Facebook launches Stories in the main Facebook app*, TECHCRUNCH (Mar. 28, 2017), https://techcrunch.com/2017/03/28/facebook-launches-stories-in-the-main-facebook-app/ [https://perma.cc/T8AL-KSUJ]; *Differences between Stories and Reels on Facebook*, FACEBOOK HELP CTR., https://www.facebook.com/help/1026380301307372 [https://perma.cc/7DLD-LUKC].

[108] *Introducing Make-A-Video: An AI System That Generates Videos*, META (Sep. 29, 2022), https://ai.meta.com/blog/generative-ai-text-to-video/ [https://perma.cc/TP2G-A7LV]; Moomal Shaikh, *What's AI Got To Do With It?*, THE MOD. SCIENTIST (Dec. 7, 2022), https://medium.com/the-modern-scientist/whats-ai-got-to-do-with-it-1240367e900e [https://perma.cc/HS4N-Z5H3].

[109] Xi Yin & Tal Hassner, *Detecting the Models Behind Deepfakes*, META (June 16, 2021), https://about.fb.com/news/2021/06/detecting-the-models-behind-deepfakes/ [https://perma.cc/CR8F-XPXW].

[110] *See, e.g.* Guy Rosen, *A Further Update on New Zealand Terrorist Attack*, META (Mar. 20, 2019), https://about.fb.com/news/2019/03/technical-update-on-new-zealand/ [https://perma.cc/4CQJ-W6M7] (discussing Facebook's difficulties in detecting the live streaming of the New Zealand terror attack in 2019).

previously mentioned EU Code of Conduct,[111] the German NetzDg,[112] and the Austrian KoPl-G,[113] among others.[114]

Finally, another dramatic development affecting content moderation pertains to AI and the automation of the enforcement process. In the early years of social media platforms, content moderation, mainly in the form of removal, relied on user reports and human moderators. Tarleton Gillespie noted that in 2009, 150 of Facebook's employees were tasked with moderation "one click at a time."[115] Over time, the company increasingly relied on algorithmic and AI-driven tools for enforcement.[116] This shift led to a reduced reliance on human labor and the incorporation of proactive, automatic flagging and removal of content.[117]

Digital platforms have considerably invested in R&D and the acquisition of advanced AI technologies, including tools based on computer vision and NLP (natural language processing).[118] One notable area of focus has been, and continues to be, semi-supervised and self-supervised deep learning, which, among other benefits, reduces dependency on human labeling for the training processes of the

[111] *See* discussion *supra* Part II. *See also Countering illegal hate speech online #NoPlace4Hate*, EUR. COMM'N (Mar. 18, 2019), https://ec.europa.eu/newsroom/just/items/54300 [https://perma.cc/L745-S8AA]; Mor, *supra* note 39, at 679–680. *See also* Damien Cave, *Australia Passes Law to Punish Social Media Companies for Violent Posts*, N.Y. TIMES (Apr. 4, 2019), https://www.nytimes.com/2019/04/03/world/australia/social-media-law.html [https://perma.cc/C2JP-AL4W].

[112] *See* discussion *supra* Part II. *See also* Natasha Lomas, *Germany tightens online hate speech rules to make platforms send reports straight to the feds*, TECHCRUNCH (Jun. 19, 2020), https://techcrunch.com/2020/06/19/germany-tightens-online-hate-speech-rules-to-make-platforms-send-reports-straight-to-the-feds/ (discussing how the Act has been modified since enacted in 2017) [https://perma.cc/CKE6-942A].

[113] *See supra* Part II.

[114] *See*, *e.g.*, Cave, *supra* note 111.

[115] Tarleton Gillespie, *The Scale Is Just Unfathomable*, LOGICS(S) (Apr. 1, 2018), https://logicmag.io/scale/the-scale-is-just-unfathomable/ [https://perma.cc/HHU9-QCRS].

[116] *Id.*; *How technology detects violations*, META TRANSPARENCY CTR., https://transparency.meta.com/enforcement/detecting-violations/technology-detects-violations/ (Oct. 18, 2023) [https://perma.cc/X87B-PPZ2].

[117] *How technology detects violations*, *supra* note 116 (stating that for most violation categories "our technology finds more than 90% of the content we remove before anyone reports it"). *Proactive Rate*, *supra* note 73. *See also Meta and Microsoft Introduce the Next Generation of Llama*, META (Jul. 18, 2023), https://about.fb.com/news/2023/07/llama-2/ [https://perma.cc/U2Y6-UQ5R]; Zuckerberg, *supra* note 1.

[118] *See Bringing the World Closer Together with a Foundational Multimodal Model for Speech Translation*, META (Aug. 22, 2023), https://ai.meta.com/blog/seamless-m4t/ [https://perma.cc/XBU8-UJ8B]; *Our New AI System to Help Tackle Harmful Content*, META (Dec. 8, 2021), https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/ [https://perma.cc/8NPS-L7PV].

models.[119] Recently, LLMs (large language models) and generative AI have entered the scene, introducing revolutionizing qualities.[120] AI, thus, became an indispensable factor in detecting, translating, and assessing content, as well as routing and escalating it to different review nodes[121] and taking action on it.[122] This groundbreaking development is linked to a broader shift in the "data economy" where data is considered "the new oil" by platforms and various other stakeholders, impacting its usage and application.[123]

Nonetheless, the incorporation of AI in content moderation processes has also raised troubling concerns relating to the lack of transparency and accountability, biased decision-making, and challenges in identifying context and understanding low-resource languages. An additional set of concerns addresses the novel and subtle ways of controlling user attention and nudging their behavior through, for instance, personalized recommending and content downranking systems.[124] These dual facets of AI—as both a technological enabler and a potential source of concerns—are also valid in the context of reduction, as will be discussed in further detail later.[125]

B. The Rise of Reduction

1. Reduction: Early Phases and Evolution

i. How Reduction Grew to Encompass All Content Fields

The strategy of content reduction has been gradually developed and implemented over the last several years, initially targeting limited and specific content categories. One of the earliest instances of Meta discussing and deploying reduction concerned spammy content and

[119] *The Self-Supervised Learning Cookbook*, META (Apr. 25, 2023), https://ai.meta.com/blog/self-supervised-learning-practical-guide/ [https://perma.cc/AL3B-X474].
[120] *Meta and Microsoft, supra* note 117.
[121] *How Meta Prioritizes Content for Review*, META TRANSPARENCY CTR., https://transparency.meta.com/policies/improving/prioritizing-content-review/ (Jan. 26, 2022) [https://perma.cc/CQ6C-8LK6].
[122] Robert Gorwa, Reuben Binns & Christian Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, 7 BIG DATA & SOC'Y 1, 6 (2020).
[123] Joris Toonders, *Data Is the New Oil of the Digital Economy*, WIRED, https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/ (last visited Feb. 5, 2024) [https://perma.cc/XEJ7-S5U2]; Susie Alegre, *Regulating around Freedom in the "Forum Internum,"* 21 ERA FORUM 591, 598 (2021).
[124] *See supra* Parts III.B & IV.B.3.
[125] *Id.*

clickbait.[126] Since then, the company has continuously and publicly addressed its efforts to mitigate clickbait through reduction.[127] Even today, clickbait is often the first example cited by the company when relating to reduction.[128]

The sustained focus on clickbait can be linked to the lack of controversy surrounding the efforts to combat it. Nonetheless, like with other types of content, the identification of clickbait, which relies on a complex set of AI-driven criteria, is susceptible to errors, biases, and additional challenges.[129] Therefore, it is important to ensure transparency in how these processes are executed and to actively address their vulnerabilities, even though the need to curb clickbait might appear straightforward or obvious in principle.[130]

Another early-adopted area of reduction is misinformation. In a 2018 article,[131] set against the backdrop of public backlash regarding the foreign interference in the 2016 U.S. election,[132] Meta explained that as part of their third-party fact-checking program, they display content that was identified as false lower in News Feed, and that this leads to "reducing future views by over 80% on average."[133]

[126] Geoffrey A. Fowler, *Shadowbanning is Real: Here's how you End up Muted by Social Media*, WASH. POST (Dec. 27, 2022), https://www.washingtonpost.com/technology/2022/12/27/shadowban/ [https://perma.cc/5XPY-4MXR]. *See also News Feed FYI: Click-baiting*, META (Aug. 25, 2014), https://about.fb.com/news/2014/08/news-feed-fyi-click-baiting/ ("Click-baiting," the company explained, "is when a publisher posts a link with a headline that encourages people to click to see more, without telling them much information about what they will see." Meta listed two factors to identify clickbait: (1) the time users spend reading it, with a short duration indicating clickbait; (2) the way users engaged with the content, with lack of action such as "liking" or commenting suggesting clickbait.) [https://perma.cc/QF9P-VWUV].

[127] *See, e.g., Further Reducing Clickbait in Feed*, META (Aug. 4, 2016), https://about.fb.com/news/2016/08/news-feed-fyi-further-reducing-clickbait-in-feed/ [https://perma.cc/4S3X-67EH].

[128] *See, e.g.*, *Reducing the Distribution of Problematic Content*, META TRANSPARENCY CTR., https://transparency.meta.com/enforcement/taking-action/lowering-distribution-of-problematic-content/ (May 18, 2023) [https://perma.cc/3HVN-MK77]. *See also* Zuckerberg, *supra* note 1 (discussing concerns on Zuckerberg's post on borderline content).

[129] *See* Gorwa et al., *supra* note 122; *see also Further Reducing Clickbait in Feed, supra* note 127.

[130] *See infra* Parts IV.B & IV.C, regarding the need for transparency around the application of reduction.

[131] Lyons, *supra* note 11.

[132] *Id.*

[133] *Id. See also About Fact-Checking on Facebook and Instagram*, *supra* note 18 (discussing the impact of reduction on the volume of views).

Meta also addressed the role of reduction in tackling COVID-19-related misinformation.[134] The company emphasized the part played by fact-checkers in identifying misleading content subjected to reduction, but it is important to note that most content deemed misleading was not directly reviewed by human fact-checkers. Instead, Meta reduced the visibility of a large volume of content using AI tools, based on a much smaller set of fact-checked items.[135] While this approach is understandable given the vast amount of content to be reviewed, it nonetheless highlights potential mistakes and vulnerabilities in moderating misinformation, including its reduction.

As previously mentioned, following accusations from the American administration against digital platforms for their role in the spread of misinformation concerning COVID-19 vaccines,[136] Meta issued a rebuttal titled "Moving Past the Finger Pointing."[137] The response unveiled the extensive scope of reduction compared to removal, with reduction (combined with labeling), being applied on roughly nine times more content.

> [W]hen we see misinformation about COVID-19 vaccines, we take action against it. Since the beginning of the pandemic we have **removed over 18 million** instances of COVID-19 misinformation. We have also **labeled and reduced the visibility of more than 167 million** pieces of COVID-19 content debunked by our network of fact-checking partners so fewer people see it and—when they do—they have the full context. (emphasis added).[138]

Such figures underscore the pivotal impact of reduction in the content moderation framework and the urgent need to promote clarity, accountability, and public discourse around the practice.

It should also be noted that reduction of misinformation, as reflected in the above quote, is often implemented alongside other

[134] Guy Rosen, *An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19*, META (Apr. 16, 2020), https://about.fb.com/news/2020/04/covid-19-misinfo-update/ [https://perma.cc/3W7C-9D7J].

[135] *See, e.g.*, *id.* ("Once a piece of content is rated false by fact-checkers, we reduce its distribution and show warning labels with more context. Based on one fact-check, we're able to kick off similarity detection methods that identify duplicates of debunked stories. For example, during the month of March, we displayed warnings on about 40 million posts related to COVID-19 on Facebook, based on around 4,000 articles by our independent fact-checking partners").

[136] *See* Kanno-Youngs & Kang, *supra* note 13.

[137] Rosen, *supra* note 12.

[138] *Id.*

sanctions, namely—warning screens and labels.[139] These additional measures may further diminish the visibility of reduced content.[140] Meta may also extend the combination of warning screens and reduction to other areas of content beyond misinformation, such as content depicting violence.[141] In two recent decisions, Meta's Oversight Board, which rarely addresses reduction, stated that the two sanctions "serve separate functions" and that in some instances, they should be decoupled.[142]

A third category of content subjected to Meta's reduction, starting around 2019, was "sensitive content,"[143] where Instagram played a significant role. Meta explained "[y]ou can think of sensitive content as posts that don't necessarily break our rules, but could potentially be upsetting to some people—such as posts that may be sexually suggestive or violent."[144]

Unlike the reduction of clickbait and even misinformation, the reduction of sensitive content sparked considerable backlash.[145] To a great extent, this was a button-up resistance, spurred by user complaints about their content being unseen,[146] the secretive nature of the practice, and its vaguely articulated guidelines.[147] Moreover, with no notification provided to users whose content was reduced, many users were left

[139] *See supra* Part I; *see also About Fact-Checking on Facebook and Instagram*, *supra* note 18.

[140] For the impact of warning screens, *see* Rosen, *supra* note 134 (arguing, in the context of COVID-19-related warning screens: "When people saw those warning labels, 95% of the time they did not go on to view the original content").

[141] *See, e.g.*, *Hostages Kidnapped from Israel*, OVERSIGHT BD. (2023), https://www.oversightboard.com/decision/FB-M8D2SOGS/ (last visited Nov. 22, 2024) [https://perma.cc/V96A-WM4T]; *Al Shifa Hospital*, OVERSIGHT BD., https://www.oversightboard.com/decision/IG-WUC3649N/ (last visited Nov. 22, 2024) [https://perma.cc/7BK3-ZLX8].

[142] *Al Shifa Hospital*, *supra* note 141. While the Board's review of reduction in these decisions is desirable, it is nonetheless limited. First, the cases qualified for the Board's review because the original sanction applied by Meta—and subsequently appealed by the user—was the removal of content. Second, the lion's share of the decision focused on the Community Standard. The discussion regarding reduction was brief and did not include a review of its policy.

[143] *See* Carolina Are, *The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram*, 22 FEMINIST MEDIA STUD. 2002, 2 (2022).

[144] *Introducing Sensitive Content Control*, META (Jul. 20, 2021), https://about.fb.com/news/2021/07/introducing-sensitive-content-control/ [https://perma.cc/VP8G-7QSP]. *See also Why Certain Posts on Instagram are not Appearing in Explore and Hashtag Pages*, INSTAGRAM HELP CTR., https://help.instagram.com/613868662393739 (last visited Feb. 5, 2024) [https://perma.cc/V9FW-JHCM].

[145] *See, e.g.*, Are, *supra* note 143; Cook, *supra* note 30.

[146] *See* Joseph, *supra* note 30.

[147] *See*, *e.g.*, Cook *supra* note 30 ("[T]he company has yet to explain how it defines 'inappropriate,' or to give users even a vague idea, instead threatening to secretly curtail their visibility should they fail to follow an undefined rule").

second-guessing and suspecting the platform's interference.[148] Even when Instagram admitted to reducing content, the company sometimes attributed this to an error.[149]

Addressing the reduction of sensitive content as "shadow-banning," Chanté Joseph expressed frustration over the ambiguity of the guidelines governing this practice, writing "the vagueness of Instagram's shadow-banning policy is perhaps the most frustrating part. It leaves users confused as to what is and isn't appropriate . . . .The reluctance to properly define what it means to be 'sexually suggestive' and a refusal to acknowledge the nuances around it, are unfair."[150]

Moreover, there were alarming concerns about the practice excluding vulnerable or marginalized groups and individuals, including people of color, members of the LGBTQ+ community, and activists. It was also claimed to silence unpopular ideas and perspectives, such as those affirming plus-sized bodies.[151]

In her autoethnography on Instagram's reduction of her pole dancing-related content, Carolina Are argued that reduction leads to the "othering of a variety of user groups," and that it replicates "puritan, conservative values that conflate trafficking with sex, sex with a lack of safety and a lack of safety with women's bodies." [152] Are also highlighted the asymmetry in how Instagram and Facebook intensely regulate female bodies compared to the more allowing approach towards male bodies. She pointed out the internalizing of the "male gaze" in the governance framework chosen by these platforms.[153]

In July 2021, Instagram launched the "Sensitive Content Control" feature, which, according to the company, aims to provide users aged 18 and older with control over the sensitive content shown on the "Explore" page.[154] Instagram explains that this feature is now

[148] *See, e.g.*, Fowler, *supra* note 126 ("First we have to agree that shadowbanning exists. Even victims are filled with self-doubt bordering on paranoia: How can you know if a post isn't getting shared because it's been shadowbanned or because it isn't very good?" and "There are signs, but rarely proof—that's what makes it shadowy").

[149] *See* Are, *supra* note 143, at 2011.

[150] *See* Joseph, *supra* note 30.

[151] *See* Are, *supra* note 143, at 2; Fowler, *supra* note 126; *see generally* Ej Dickson, *Why Did Instagram Confuse These Ads Featuring LGBTQ People for Escort Ads?*, ROLLING STONE (Jul. 11, 2019), https://www.rollingstone.com/culture/culture-features/instagram-transgender-sex-workers-857667/ (discussing AI-driven censorship of marginalized groups) [https://perma.cc/HHS8-A2WA].

[152] Are, *supra* note 143, at 14.

[153] *Id.* at 5. See also Sarah Myers West, *Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Platforms*, 5 MEDIA AND COMMC'N 28, 32 (2017).

[154] *Introducing Sensitive Content Control*, *supra* note 144.

applied more broadly, encompassing "Search, Reels, Accounts You Might Follow and Recommendations."[155]

The Sensitive Content Control mechanism raises several concerns. First, it builds on an architecture that subtly nudges users towards limiting their exposure to non-violating content,[156] while ironically employing the terminology of "choice" and "control."[157] The original version of "Sensitive Content Control" included three options: "Allow," "Limit" (the default), and "Limit even more."[158] The "Limit" option was presented as a moderate middle ground but actually led to users seeing less content permitted under Instagram's Community Guidelines (and Facebook's Community Standards, which also apply to Instagram).[159] The subsequent option, "Limit even more," further confined the content displayed to users.[160] Instagram has since renamed these three categories, but the underlying principles and the secrecy surrounding the scope of each category remain unchanged.[161] Second, there appears to be a one-size-fits-all threshold applied to vastly different types of content, including potentially violent, sexually suggestive, or those promoting regulated goods.[162] Whether the

[155] *About Sensitive Content Control on Instagram*, INSTAGRAM HELP CTR., https://help.instagram.com/1055538028699165/?helpref=related_articles (last visited Feb. 5, 2024) [https://perma.cc/M9QE-DFVQ].

[156] *Limit Sensitive Content That You See on Instagram*, INSTAGRAM HELP CTR, https://help.instagram.com/251027992727268 (last visited Feb. 5, 2024) [https://perma.cc/N8RA-J3PN].

[157] *Id* "[Sensitive content] may be considered upsetting, offensive, or sensitive and we may make it harder to find, rather than removing it from Instagram. We've previously limited content like this, but you can also choose to see more or less content that could be upsetting or offensive using the **Sensitive Content Control**." Emphasis in original. *See About Sensitive Content Control on Instagram*, *supra* note 155; Gillespie, *supra* note 8, at 4.

[158] *Introducing Sensitive Content Control*, *supra* note 144.

[159] *See Oversight Board Overturns*, *supra* note 67.

[160] *Introducing Sensitive Content Control*, *supra* note 144 The nature of the caption placed next to each option also reduced the likelihood of users actively changing these defaults. The caption next to the "Allow" option reads: "You may see more photos or videos that could be upsetting or offensive," the "Limit" option reads: "You may see some photos or videos that could be upsetting or offensive," and the "Limit even more" option reads: "You may see fewer photos or videos that could be upsetting or offensive." *Id.*

[161] *Updates to the Sensitive Content Control*, INSTAGRAM, https://about.instagram.com/blog/announcements/updates-to-the-sensitive-content-control (Aug. 25, 2022) (The categories' names are now "More," "Standard," and "Less." Instagram explains: "Standard' is the default state and will prevent people from seeing some sensitive content and accounts. 'More' enables people to see more sensitive content and accounts, whereas 'Less' means they see less of this content than the default state"). [https://perma.cc/XZT7-WCJ7]. Teens under 16 years of age are not provided with the "More" option. *Id.*

[162] *Limit Sensitive Content That You See on Instagram*, *supra* note 156.

company enforces specific internal thresholds for each type of content behind the scenes remains undisclosed. These concerns are amplified considering that the reduction of "sensitive content" is not exclusive to Instagram, but is also a default practice on Facebook, as will be elaborated on later.[163]

After being applied to clickbait, misinformation, and "sensitive content," the implementation of reduction has expanded to encompass nearly all content categories. As will be explored later, reduction now targets, among other types, a vast array of content that approaches any point along the spectrum of the Community Standards without crossing them. Additionally, it applies to content "likely" to violate these standards, even when such violations cannot be confirmed by Meta.[164]

ii. The Multifaceted Nature of Reduction

Besides its broad application, another key factor that has shaped the impact and evolution of reduction lies in the variety of methods it employs. Reduction can now be facilitated through downranking content in users' feeds, tweaking the recommendation system, or excluding content from dominant areas on the platform, like the "Search," "Explore," or "Hashtag" pages. Some of these methods might be paired with other sanctions, such as warning screens or labels.[165]

In addition, there exists what I term "reduction by proxy," which is the outsourcing of certain reduction powers from the platforms to the users. This includes capabilities such as blocking other users or hiding their comments. Hiding comments can be automated if the users include pre-selected banned words. For instance, Facebook allows Page operators to blacklist up to 1000 such words. The company, in addition, automatically hides comments that include variations of these words.[166] Hiding comments and blocking other users carry reduction characteristics, in my view, *inter alia*, since they are carried out covertly, without the awareness of the individuals whose content has been restricted. Moreover, akin to the reduction implemented by Meta, these actions do not involve the removal of the content. When used *en masse*, these user-enabled options may have a significant impact on the informational and communicative landscape. Moreover, these outsourcing measures are often adopted by public figures, including world leaders, underscoring the substantial implications of this content

[163] *See infra* Part III.B.2.
[164] *Id.*
[165] *See supra* Part III.B.1.
[166] *How Do I Block Certain Words from Appearing in Comments on My Facebook Page?*, *supra* note 20.

moderation form.[167] Nonetheless, Meta does not provide any data on this influential moderation channel.

Lastly, another aspect to consider when evaluating the nuanced nature of reduction is its flexible territorial application. Unlike removal which the company implements globally,[168] the application of reduction can vary based on geographical location. This localized approach to reduction is also time-bound. In Meta's words, "[w]hile the majority of our reduced distribution enforcements are applied around the world equally, we also recognize that in certain situations we cannot always take a one-size-fits-all approach to enforcement. For example, we may temporarily adjust our enforcements in a specific region or during a critical event." [169]

Such territorial flexibility may allow Meta to more precisely address specific and localized needs. However, it also introduces significant challenges. Particularly when coupled with the opaque manner in which reduction is applied, this flexibility could, for instance, render the reduction methods more vulnerable to pressure from national authorities, or facilitate their exploitation by hegemonic and conservative factions seeking to preserve their dominance.[170]

2. Reduction Policy Morass

In parallel with its development and the broadening of its application, reduction was institutionalized within Meta's content moderation framework as part of an approach that the company describes as "remove, reduce, and inform."[171] It was not until September 2021, however, that the company published the initial version of the

[167] *See also*, Knight First Amend. Inst. at Columbia Univ. v. Trump, 928 F.3d 226, 240 (2d Cir. 2019); Joanne Chianello, *Watson Changes Tune on Twitter Clash,* CBC NEWS (Nov. 2, 2018), https://www.cbc.ca/news/canada/ottawa/mayor-watson-unblocks-critics-twitter-lawsuit-1.4887540 [https://perma.cc/8VZP-KN9U]; Garnier v. O'Connor-Ratcliff, 41 F.4th 1158, 1164 (9th Cir. 2022).

[168] *Facebook Community Standards*, *supra* note 3 ("Our Community Standards apply to everyone, all around the world, and to all types of content"). The company can apply territorial blockage if the content violates local law. *See Transparency Reports*, *supra* note 70.

[169] *Types of Content We Demote, supra* note 16.

[170] *Content Restrictions, supra* note 48.

[171] *Remove, Reduce, Inform, supra* note 7 (stating that reduction includes false news, groups that repeatedly shared misinformation, and content that enjoys much attention on Facebook but not outside the platform); *People, Publishers, the Community*, META (Apr. 10, 2019), https://about.fb.com/news/2019/04/people-publishers-the-community/ [https://perma.cc/TS5Q-HDEX].

reduction policy, [172] occasionally referred to as the "Content Distribution Guidelines."[173]

Nonetheless, despite Meta's assurance that these guidelines would "go into detail about the types of content that receive reduced distribution, and explain why we've decided to reduce distribution for each particular type of content,"[174] this policy still emerges as extremely vague, incoherent, and perplexing. It includes opaque terms, clauses of doubtful legitimacy, subcategories that do not align with their respective categories, and overlapping provisions.[175]

The Content Distribution Guidelines comprise three very broad categories: **The first category** is titled "Responding to People's Direct Feedback." The company explained "[w]e're always eager to receive people's feedback about what they do and don't like seeing on Facebook and make changes to Feed in response."[176] This category includes content types such as "Clickbait links" and "Pages predicted to be spam," as well as "Sensationalist Health Content and Commercial Health Posts."[177] This category also includes "Comments that Are Likely to Be Reported or Hidden" (*inter alia*, because similar content is frequently reported), [178] and "Links to Websites Requesting Unnecessary User Data," used to "harvest people's personal

[172] The Change log (for guidelines that were updated) indicates that the guidelines were first published in September 2021. *Policies*, *supra* note 69. The log does not cover guidelines that were entirely removed by the company. *See infra* text accompanying note 197.

[173] Meta sometimes uses this name when referring to reduction's policy, see *Widely Viewed Content Report: Companion Guide*, META TRANSPARENCY CTR., https://transparency.fb.com/data/widely-viewed-content-report/companion-guide (Aug. 23, 2022) (where a link to the guidelines is also provided) [https://perma.cc/LQ2V-MBB6]. However, the guidelines themselves are titled differently. *See Types of Content We Demote*, *supra* note 16.

[174] *Widely Viewed Content Report*, *supra* note 173.

[175] *See infra* Part III.B.2.

[176] *Types of Content We Demote*, *supra* note 16.

[177] *Sensationalist Health Content and Commercial Health Posts*, META TRANSPARENCY CTR., https://transparency.fb.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/sensationalist-health-content-commercial-health-posts (last visited Feb. 5, 2024) [https://perma.cc/K62Z-SCTR].

[178] *Comments That Are Likely to Be Reported or Hidden*, META TRANSPARENCY CTR., https://transparency.fb.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/comments-likely-reported-hidden (last visited Feb. 5, 2024) [https://perma.cc/LUM3-UC58].

information."[179] Finally, this category also encompasses "low-quality" comments and events.[180]

As observed, the first category of the Content Distribution Guidelines amalgamates various types of content, including clickbait, privacy-violating content, and repeatedly reported content. It also covers low-quality content and health-related misinformation, despite these issues being specifically tackled within the second category of the guidelines, which will be elaborated on further below.[181] The first category provokes additional concerns, *inter alia*, regarding the "Comments that Are Likely to Be Reported or Hidden" subcategory. This provision targets content for reduction without a substantive review, merely based on its propensity to be reported (or hidden). However, users' reports may be influenced by a range of reasons, including foreign intervention and the implemented choice architecture, which are not necessarily indicative of harmful content.[182] This moderation approach is also problematic, since it tends to favor mainstream ideas, potentially further sidelining marginalized and unpopular perspectives and groups.[183]

The second category is titled "Incentivizing Creators to Invest in High-Quality and Accurate Content."[184] This category encompasses, for instance, "Domains with Limited Original Content," such as those that contain "high volumes of low-quality content for the purposes of inflating virality and driving traffic."[185] It also covers "Inauthentic Sharing."[186] Additionally, this category addresses "Unoriginal News

[179] *Links to Websites Requesting Unnecessary User Data*, META TRANSPARENCY CTR., https://transparency.meta.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/links-websites-requesting-unnecessar-user-data (last visited Feb. 5, 2024) [https://perma.cc/LGP9-H567].

[180] *See also Low Quality Comments*, META TRANSPARENCY CTR., https://transparency.meta.com/features/approach-to-ranking/content-distribution-guidelines/low-quality-comments (last visited Feb. 5, 2024) [https://perma.cc/7HSR-7N55].

[181] *See infra* Part III.B.2.

[182] For the impact of the choice architecture on users' reports, see William Echikson & Olivia Knodt, *Germany's NetzDG: A Key Test for Combatting Online Hate*, THE CTR. FOR EUR. POL'Y STUD. 7–9, 11 (Research Report No. 2018/09, 2018), http://aei.pitt.edu/95110/1/RR_No2018-09_Germany's_NetzDG.pdf.

[183] See *supra* Part III.B.1.

[184] *Types of Content We Demote*, *supra* note16.

[185] *Domains with Limited Original Content*, META TRANSPARENCY CTR., https://transparency.meta.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/domains-with-limited-original-content (last visited Feb. 5, 2024) [https://perma.cc/Z9TP-QCUW].

[186] *Inauthentic Sharing*, META TRANSPARENCY CTR., https://transparency.fb.com/features/approach-to-ranking/content-distribution-guidelines/inauthentic-sharing (last visited Feb. 5, 2024) [https://perma.cc/6D6G-XXJ5].

Articles," described as articles that "do not contain new, original reporting or analysis."[187] The company explains regarding the latter that "[t]he more extensive original reporting an article contains, the more distribution it will receive in Feed. Original reporting includes things like exclusive source materials, significant analysis, new interviews, or the creation of original visuals."[188] Lastly, this category includes "Fact-Checked Misinformation," referring to "Content that has been debunked as 'False, Altered, or Partly False.'"[189] In this case, as mentioned above,[190] reduction may be carried out in tandem with the addition of labeling to the content.[191]

Similar to the first category, the second category of the guidelines also groups together different content types, including misinformation and unoriginal content. It raises additional challenges concerning Meta's role in determining what constitutes "low-quality" content and journalism, along with questions on the justification for suppressing such content.[192]

Notwithstanding these concerns, it is **the third category**, titled "Fostering a Safer Community," that, in my view, most distinctly illustrates the disruptive impact of reduction.

This category—which, according to Meta, includes content that "may be problematic for our community, whether or not it's intended that way"[193]—exemplifies the all-encompassing nature of reduction. It demonstrates how reduction subtly elevates the entire threshold of permissible content and erodes the breadth of information accessible to users.

This category includes several types of content, one of which is "Content Likely Violating Our Community Standards."[194] When

[187] *Unoriginal News Articles*, META TRANSPARENCY CTR., https://transparency.fb.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/unoriginal-news-articles (last visited Feb. 5, 2024) [https://perma.cc/P9GJ-MV8V].

[188] *Id.*

[189] *Fact-Checked Misinformation*, META TRANSPARENCY CTR., https://transparency.fb.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/misinformation (last visited Feb. 5, 2024) [https://perma.cc/XN79-ULXK].

[190] *See supra* Parts I & III.B.1.

[191] *Fact-Checked Misinformation*, *supra* note 189.

[192] *See infra* Part IV.A.

[193] *Types of Content We Demote*, *supra* note 16.

[194] *Content Likely Violating Our Community Standards*, META TRANSPARENCY CTR., https://transparency.fb.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-likely-violating-our-community-standards (last visited Feb. 5, 2024) [https://perma.cc/AX3K-LQ86]. The remaining types of content listed under the third category are "Posts that Indicate Suspicious Virality," "Unsafe Reporting About Suicide," and "Content Posted by Repeat Violators of Our Policies." *Types of*

enforcing this subcategory, the company aims to reduce the visibility of content that their AI-driven systems predict to likely violate the Community Standards. However, such content "has not been confirmed" to constitute a violation, and thus cannot be removed outright.[195] In other words, the company targets content for which it has a lower degree of confidence that it constitutes an actual violation. This subcategory is extensive, covering content "likely to violate" all areas forbidden by the Community Standards, including hate speech, incitement to violence, bullying and harassment, graphic violence, adult nudity, sexual activity, content distributed by fake accounts, and spam.[196]

Until October 2023, the third category of the Content Distribution Guidelines also included "Content Borderline to the Community Standards." Such content, Meta stated, is not prohibited by the Community Standards, but "come[s] close to the lines drawn by those policies." [197] Akin to "Content Likely Violating" the Community Standards, this subcategory extends to various fields of content covered in the Community Standards, while seamlessly raising the bar for permissible content. Take borderline hate speech, for instance. While the Community Standards include strict requirements for removing such content, including that which targets people on the basis of their protected characteristics, borderline content extends to "content that dehumanizes individuals or groups who are *not* defined by their protected characteristics."[198]

Despite Meta's formal removal of borderline content from its Content Distribution Guidelines, such content is still subject to reduction. On the guidelines landing page, Meta states that it employs "personalized ranking" to reduce the distribution of borderline

Content We Demote, *supra* note 16. The latter type of content may stem from governmental requirements, see Missouri v. Biden, 80 F.4th 641, 660 (5th Cir. 2023).

[195] *Content Likely Violating Our Community Standards*, *supra* note 194.

[196] *Id.*

[197] *Content Distribution Guidelines: Changes, Correction, and Adjustments*, META TRANSPARENCY CTR., https://transparency.fb.com/features/approach-to-ranking/cdgs-changes-corrections (Dec. 19, 2023) (indicating the removal of "borderline" content from the guidelines) [https://perma.cc/3LF7-FDNU]. While the content that appeared under this removed subcategory is no longer available on the guidelines page, its key components can be found here: *Publisher Content and Facebook Community Standards*, META BUS. HELP CTR., https://www.facebook.com/business/help/201148151829614?id=208060977200861 (last visited Feb. 28, 2024) [https://perma.cc/LU8H-3L3R]. This subcategory included "Borderline Adult Nudity and Sexual Activity," "Borderline Violent & Graphic Content" and "Borderline Bullying & Harassment, Hate Speech, and Violence & Incitement." *Id.*

[198] *Publisher Content and Facebook Community Standards, supra* note 197.

content.[199] It then directs readers to a different set of guidelines, those governing its "personalized ranking approach," activated by adjustments users make in their settings.[200]

While acknowledging the challenges that the delegation of adjusted reduction powers to users introduces, a closer examination of the platform's settings architecture reveals that the reduction of borderline content largely remains a predetermined content moderation practice, rather than the personalized option Meta claims it to be.[201] First, the extent to which users can tailor content through the settings is restricted to a few areas (e.g., "Low-quality content," "Unoriginal content and problematic sharing," and "Sensitive content").[202] Second, within the limited content fields where users presumedly have control, reduction is implemented *by default*. To deactivate this feature, users must actively change their settings, [203] a step that many will probably not take.[204]

While the extent and characteristics of borderline content are marked by a troubling lack of clarity, the implications for content moderation could be far-reaching. This practice establishes a completely new policy boundary that precedes the Community Standards, thereby reshaping the informational landscape and influencing which communities, ideas, and narratives are deemed legitimate. These concerns also apply to content "likely violating" the Community Standards, since the enforcement of such content is performed with a lower degree of assurance regarding its infringing nature, compared to content removed for violating the Community Standards.

[199] *Types of Content We Demote*, *supra* note 16.

[200] *Our Approach to Facebook Feed Ranking*, *supra* note 16.

[201] Such examination unveils additional challenges such as opacity and overlap with the Content Distribution Guidelines, see *Content Distribution Guidelines*, *supra* note 197.

[202] *Manage How Content Ranks in Your Feed Using Reduce*, FACEBOOK HELP CTR., https://www.facebook.com/help/543114717778091 (last visited Feb. 5, 2024) [https://perma.cc/5SZX-PRV9]. The three options appear under the "Reduce" settings, mentioned in the Personalized Ranking Guidelines. It is unclear whether other types of content that may border on violating the Community Standards, such as hate speech and inciting content, are still being reduced by Meta, just as they were before the removal of the "borderline content" subcategory from the reduction policy. It should also be noted that "low-quality" content and "unoriginal content" are also covered in the Content Distribution Guidelines, thereby creating a puzzling overlap.

[203] *Id*.

[204] First, finding these particular settings is not an easy task (changing the default settings cannot be reached through the control options attached to the posts themselves, for instance). Second, the lack of transparency and public attention regarding content reduction may also adversely influence the likelihood of users actively seeking and adjusting these settings.

3. The Subtle Art of Keeping the Lion Quiet (or: How Reduction is Held in Content Moderation's "Backstage")

Reduction, as previously explored, has grown to become an opaque, yet potent strategy of content moderation, emerging, in certain cases, as more influential than removal in terms of scope, versatility, and the overall power it provides. Moreover, unlike removal, the company invests significant effort in keeping reduction in the "backstage" of content moderation, ensuring it remains a less conspicuous aspect of this task. In the subsequent paragraphs, I will discuss the factors that facilitate this "under the radar" approach, many of which—though not all—are orchestrated by Meta. This veil of secrecy obstructs public discourse and oversight concerning reduction.

First, in contrast to removal, fully comprehending reduction can be challenging for both users and other stakeholders, such as decision-makers. As previously outlined, reduction utilizes an array of measures, including downranking and excluding content from dominant areas of the platform, and is sometimes accompanied by additional sanctions like labels and warning screens. Furthermore, reduction capabilities can be outsourced to users, supported by designated choice architecture and adapted on a territorial basis.[205]

Second, users whose content is subjected to reduction are often not notified by Meta, nor are they provided with an explanation.[206] Consequently, they may remain unaware that their content has been penalized or merely suspect as much.[207] This subtlety stands as "a key characteristic" of reduction.[208] As mentioned earlier, the "vagueness" surrounding reduction is perceived as its "most frustrating part," and as an approach that "leaves users confused as to what is and isn't appropriate."[209]

Third, even if Meta's users become aware or suspect that their content was reduced, they cannot challenge this decision. Meta does not offer a mechanism for users to appeal against content reduction. Furthermore, since utilizing Meta's internal appeal mechanism is a prerequisite for escalating a case to the Oversight Board, users affected

[205] *See supra* Part III.B.1.

[206] *See, e.g.*, *Instagram's CEO*, *supra* note 30 (arguing that Instagram is conducting reduction "without alerting affected users, who are often left to wonder why their content's engagement is lower than usual"). *See also* Paddy Leerssen, *An End to Shadow Banning? Transparency Rights in the Digital Services Act Between Content Moderation and Curation*, 48 COMPUT. L. SECUR. REV. 1, 3 (2023); Goldman, *supra* note 30, at 30.

[207] *See supra* Part III.B.1.

[208] Instagram's CEO, *supra* note 30.

[209] Joseph, *supra* note 30.

by content reduction find themselves also unable to refer their case to that higher tribunal.[210]

Fourth, unlike removal, reduction does not play a significant role in content moderation regulation and has not garnered much attention or legitimization through it.[211] When the Code of Conduct (2016) and the NetzDG (2017) were formulated, reduction was a relatively insignificant tool in moderation. These two regulations proved to significantly influence content moderation governance, primarily framing removal as the central content moderation sanction. Subsequent regulations, such as the KoPl-G (2020), followed suit, further strengthening this trend.[212] However, the EU Digital Service Act (DSA),[213] which became fully applicable in February 2024,[214] marks a positive shift by including reduction in its definition of content moderation.[215] The DSA imposes varying obligations on platforms, including the requirement for "clear and unambiguous language" in their Terms and Conditions.[216] It also mandates that content moderation sanctions be accompanied by a "Statement of Reasons,"[217] and that content moderation decisions be reviewable.[218] While this new regulation may not tackle all the

[210] *See supra* Part II. *See also* Noa Mor, *On Facebook's New "Oversight Board", Accountability, and Control*, DLI-CORNELL-TECH, https://www.dli.tech.cornell.edu/post/on-facebook-s-new-oversight-board-accountability-and-control (Apr. 11, 2020) [https://perma.cc/HV3L-79LN]. In addition, the Board's mandate does not encompass reduction policies, *see id.* and *supra* Part II. Note that according to the second report submitted by Facebook in compliance with the Very Large Platforms DSA requirements, users whose content was demoted following other users' reports can request a review of that decision. *See Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook*, FACEBOOK 14–15 (June 13, 2024) https://transparency.meta.com/sr/dsa-transparency-report-apr2024-facebook [https://perma.cc/VZ3W-PQZP].

[211] Leerssen, *supra* note 206, at 5 ("Earlier content moderation laws have concerned themselves almost exclusively with content removal and account suspension").

[212] *See supra* Part II. *See also* Gillespie, *supra* note 8.

[213] *See* The *Digital Services Act Package Shaping Europe's Digital Future*, EUR. COMM'N (2023), https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package (last visited Feb. 6, 2024) [https://perma.cc/6PZG-6GTT].

[214] *Questions and Answers on the Digital Services Act*, EUR. COMM'N, https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348 (last visited Feb. 6, 2024) [https://perma.cc/Q82H-SSD6]; *The Digital Services Act Package*, *supra* note 213.

[215] *Regulation 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/21/EC (Digital Services Act)*, OFF. J. EUR. UNION, 2022 O.J. (L 277) 1, 43 [hereinafter *Digital Services Act*].

[216] *Digital Services Act*, *supra* note 215, at 49.

[217] *Id.* at 51–52.

[218] *Id.* at 53–54.

accountability challenges surrounding reduction,[219] it is nonetheless anticipated to have a positive effect.

Fifth, unlike the case of removal and the Community Standards, reduction receives very limited transparency through Meta's Content Distribution Guidelines.[220] These guidelines cannot function as a normative, behavior-guiding tool, nor seem intended to. Initially, they are not listed under the "policy" section in Meta's Transparency Center (which starts with the "Community Standards"),[221] making them less accessible to users.[222] Moreover, as detailed above, in contrast to the Community Standards, the Content Distribution Guidelines are articulated as a confusing, chaotic, and vague report.[223] These guidelines are also considerably shorter in comparison to the Community Standards.[224] Additionally, while the Community Standards include graphics and are user-friendly, the Content Distribution Guidelines are presented in a plain, text-based document. Furthermore, it appears that the Content Distribution Guidelines themselves refer to the Community Standards as the compelling normative benchmark, as reflected in provisions regarding the reduction of content that is "likely to violate the Community Standards."[225]

In addition, the company provides significantly less information about the process for updating the Content Distribution Guidelines as well as the participants involved.[226] A major factor contributing to the limited visibility of these guidelines and the challenges they present is that their review falls outside the Oversight Board's mandate, unlike the Community Standards. Consequently, they are almost entirely absent from the Board's decisions and from the impactful public debate that follows.[227] Furthermore, the company often avoids mentioning the

[219] *See infra* Part IV. B.1.

[220] *Types of Content We Demote*, *supra* note 16. *See also supra* Part II (regarding removal). The second report submitted by Facebook in compliance with the Very Large Platforms DSA requirements does include some data about reduced content. *See Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook*, *supra* note 210 at 12–13.

[221] *Policies*, *supra* note 69.

[222] *Id.*

[223] *See supra* Part III.B.2.

[224] *See supra* notes 3, 16.

[225] *See supra* Part III.B.2.

[226] *Compare* the instances where the company discussed the modification processes of the Community Standards, *with supra* Part II.

[227] *Id*. *See supra* Part III.B, *supra* note 142, for two recent decisions in which the Board swiftly addressed reduction (without relating to its policy). See also *Content Moderation in a New Era for AI and Automation*, OVERSIGHT BD. (Sep. 2024), https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation/ (last visited Nov. 22, 2024) [https://perma.cc/LS5F-MR44], for the Oversight Board's recent report on AI in content moderation where it addressed

Content Distribution Guidelines in contexts where content moderation is discussed, such as reports it submits to regulators. In these instances, by contrast, the Community Standards are heavily emphasized.[228]

Lastly, and notably concerning, is the fact that, unlike data on removals and violations of the Community Standards, information pertaining to the enforcement of the Content Distribution Guidelines is conspicuously absent from Meta's voluntary Transparency Reports.[229]

C. How Digital Platforms Control User Attention

Practices through which platforms determine the preference given to certain content are by no means a new phenomenon. As I will explore shortly, organizing content and deciding which will receive more views and engagement is the "bread and butter" of social media,[230] and for Meta in particular, these are processes that were adopted early in the company's history.[231]

On the individual user level, Meta has long been personalizing both commercial and noncommercial information based on constant surveillance of users' behavior, in an attempt to encourage users' engagement with content, spend time on the platform, and consume ads.[232] As Adam Musuri, then an executive at Facebook, has put it,

reduction; and Noa Mor, *Meta's Oversight Board's Report on AI: What's Left Unpacked*, THREE GENERATIONS OF DIGIT. HUM. RTS., https://3gdr.huji.ac.il/sites/default/files/threegenerationsofdigitalhumanrights/files/metas_oversight_boards_report_on_ai_-_whats_left_unpacked_fi.pdf (last visited Nov. 17, 2024), for my analysis of this report.

[228] See *supra* Part II.

[229] See *id.*, for a discussion of the "Community Standard Enforcement Transparency Report." The company has also been publishing the "Widely Viewed Content Report" ("WVCR"), which "aims to provide more transparency and context about what people are seeing on Facebook by sharing the most-viewed domains, links, Pages and posts for a given quarter on Feed in the United States." *See Widely Viewed Content Report*, *supra* note 70. Although this report offers insights into content preference issues, it does not extend to cases of content reduction.

[230] ONLINE PROFILING: A REPORT TO CONGRESS, F.T.C. 9–10 (2000).

I. [231] SEE, *E.G.*, *FACEBOOK UNVEILS FACEBOOK ADS*, META (NOV. 6, 2007), HTTPS://ABOUT.FB.COM/NEWS/2007/11/FACEBOOK-UNVEILS-FACEBOOK-ADS/ [HTTPS://PERMA.CC/CM3J-G2RK].

[232] *Our Approach to Facebook Feed Ranking*, Meta TRANSPARENCY CTR., *supra* note 16; *How Does News Feed Predict What You Want to See?*, *supra* note 102 ("Put simply, the system determines which posts show up in your News Feed, and in what order, by predicting what you're most likely to be interested in or engage with. These predictions are based on a variety of factors, including what and whom you've followed, liked, or engaged with recently."). *See also* Mor, *supra* note 39 at 667; *What Most Relevant means on a Facebook Page post*, FACEBOOK HELP CTR. (Feb. 4, 2024),

> Our aim is to deliver the types of stories we've gotten feedback that an individual person most wants to see. We do this not only because we believe it's the right thing but also because it's good for our business. When people see content they are interested in, they are more likely to spend time on News Feed and enjoy their experience.[233]

Personalization of content on Facebook is implemented across all areas of the platform, including the main feed, stories, search results, recommendations, pages, and individuals' accounts. [234] Such personalization has been critically addressed in academic literature as generating "Filter Bubbles" [235] and "Echo Chambers"; [236] and as contributing to segmentation and polarization.[237]

Content Preferences also apply across the userbase. This is reflected, for instance, in the favoring of certain types of media, such as live video streams, over other content formats like plain text.[238] Here too, the reason for the platforms' intervention lies in its predictions of increased engagement with the preferred type of content.[239]

https://www.facebook.com/help/539680519386145 [https://perma.cc/XPF5-Y5FK]; Nicole B. Ellison et al., *Cultivating Social Resources on Social Network Sites: Facebook Relationship Maintenance Behaviors and Their Role in Social Capital Processes*, 19 J. COMPUT.-MEDITATED COMMC'N 855, 866 (2014); Dan Levy, *Building the Next Era of Personalized Experiences*, META FOR BUS. (Jul. 7, 2021), https://www.facebook.com/business/news/building-the-next-era-of-personalized-experiences [https://perma.cc/3PE5-8ES3].

[233] *Building a Better News Feed for You*, META (Jun. 29, 2016), https://about.fb.com/news/2016/06/building-a-better-news-feed-for-you/ [https://perma.cc/7GH9-82B].

[234] Raghav Bharadwaj, *AI for Social Media Censorship—How It Works at Facebook, YouTube, and Twitter*, EMERJ A.I. RSCH., https://emerj.com/ai-social-media-censorship-works-facebook-youtube-twitter/ (Feb. 10, 2019) [https://perma.cc/C8QD-3VUP].

[235] *See generally* ELI PARISER, THE FILTER BUBBLE: HOW THE NEW PERSONALIZED WEB IS CHANGING WHAT WE READ AND HOW WE THINK (Reprint ed. 2012).

[236] *See generally* CASS R. SUNSTEIN, ECHO CHAMBERS: BUSH V.GORE, IMPEACHMENT, AND BEYOND (Princeton Univ. Press, 2001), *and* Matteo Cinelli et al., *The echo chamber effect on social media*, 118 PNAS 1 (2021).

[237] Christina Pazzanese, *To Combat Endless Feeds of One-Sided Data, Sunstein Suggests an 'Architecture of Serendipidy,'* THE HARVARD GAZETTE (Mar. 21, 2017), https://news.harvard.edu/gazette/story/2017/03/cass-sunsteins-republic-explores-dangers-of-social-media-curation/ [https://perma.cc/FY26-YRZQ]; CASS R. SUNSTEIN, #REPUBLIC: DIVIDED DEMOCRACY IN THE AGE OF SOCIAL MEDIA 8 (2017).

[238] Vibhi Kant & Jie Xu, *Taking into Account Live Video When Ranking Feed*, META (Mar. 1, 2016), https://about.fb.com/news/2016/03/news-feed-fyi-taking-into-account-live-video-when-ranking-feed/ ("As a first step, we are making a small update to News Feed so that Facebook Live videos are more likely to appear higher in News Feed when those videos are actually live") [https://perma.cc/5FFY-Q2D3].

[239] *Id.*

Sometimes platforms make more holistic and far-reaching decisions regarding the priorities given to content across their userbase. See, for example, Mark Zuckerberg's statement.

> [W]e're making a major change to how we build Facebook. I'm changing the goal I give our product teams from focusing on helping you find relevant content to helping you have more meaningful social interactions. . . . The first changes you'll see will be in News Feed, where you can expect to see more from your friends, family and groups. As we roll this out, you'll see less public content like posts from businesses, brands, and media.[240]

Lastly, content preference methods are also significantly facilitated by various choice architecture measures, including default settings and the notification system.[241]

Is reduction any different from these existing forms of attention preference control? Well, yes and no. Reduction does share common ground with these methods, as it significantly impacts the informational landscape and the nature of the communicative processes. However, reduction is distinct because, unlike other ranking methods that aim to highlight specific content, it focuses on the *penalizing* of content. Therefore, while reduction and other attention control methods may result in a somewhat similar distribution of visibility, they diverge in the normative implications they carry. Reduction, distinct from these methods, encapsulates an inherent judgment concerning the content, marking it as "upsetting," "problematic," or otherwise inappropriate.[242] Thus, even though preferring certain content inevitably leads to less

[240] Mark Zuckerberg, *One of areas for 2018 is* [...], FACEBOOK (Jan. 12, 2018), https://www.facebook.com/zuck/posts/one-of-our-big-focus-areas-for-2018-is-making-sure-the-time-we-all-spend-on-face/10104413015393571/ [https://perma.cc/276Z-TUL6]. With relation to X, see Kari Paul, *Elon Musk Reportedly Forced Twitter Algorithm to Boost His Tweets after Super Bowl Flop*, THE GUARDIAN, (Feb. 15, 2023), https://www.theguardian.com/technology/2023/feb/15/elon-musk-changes-twitter-algorithm-super-bowl-slump-report [https://perma.cc/E8TE-AWLC].

[241] *See generally*, WOODROW HARTZOG, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES (2018); *FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook*, F.T.C. (Jul. 24, 2019) https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions-facebook ("Facebook repeatedly used deceptive disclosures and settings to undermine users' privacy preferences…") [https://perma.cc/ZEK4-VKSY].

[242] *See supra* Parts I & III.B.

visibility for other types of content due to limited user attention,[243] reduction stands apart and necessitates a separate and careful study. Moreover, while the attention control methods mentioned earlier are often confined to specific users or types of content, reduction is proving to be a comprehensive content moderation strategy that encompasses all areas of content.

IV. LEGITIMACY, LEGAL, AND TECHNOLOGICAL CONCERNS

A. The Doubts Surrounding the Motivations Behind Reduction

The doubts and uncertainty concerning reduction further extend to the motivations that drive this strategy. Over the years, Meta has provided a range of reasons for performing reduction. The company has claimed, for instance, that reduction curbs content that degrades the quality of their services and public discourse. As explored earlier, Meta has stated that reduction targets content that is "sensationalist," "problematic," "offensive," "upsetting," "sensitive," or "low-quality," and that such content could lead to polarization.[244] Additionally, a narrative that Meta has emphasized since the beginning of applying reduction refers to the company's intention to align with people's content preferences, suggested by statements along the lines of "people told us they do not want to watch this and that content,"[245] emphasizing the provision of "choice" and "control."[246]

However, as these goals are expressed in extremely abstract terms and lack data support, which will be discussed later,[247] I would like to explore several alternative options regarding the actual dynamics driving this strategy.

First, it can be safely assumed that one of the motivations behind reduction lies in its innovative and powerful channels for curating content presented to users.[248] While the flexibility and effectiveness of reduction can indeed offer authentic opportunities for content moderation, as long as its execution remains unaccountable, the power this strategy provides

[243] Lada et al., *supra* note 102 (explaining that each day, every user has about a thousand candidate posts that the company's ranking system narrows down to a few hundred).

[244] *See supra* Part III.B. *See also Gillespie*, *supra* note 8, at 4.

[245] *See The Three-Part Recipe for Cleaning up Your News Feed*, *supra* note 7. *See also Types of Content We Demote*, *supra* note 16 (regarding the "Responding to People's Direct Feedback" category in the "Content Distribution Guidelines").

[246] *See supra* Part III.B and discussion infra Section IV.B. *See also* Mor, *supra* note 210.

[247] *See infra* Part IV.B.2.

[248] *See supra* Part III.

may be prone to arbitrariness, errors, and abuse, as well as susceptibility to pressure from various stakeholders.[249]

Second, in accordance with the previous point, there are concerns that enforcement bodies and other governmental actors may be pressuring digital platforms to implement reduction. This is illustrated in the judgment by the aforementioned U.S. Court of Appeals for the 5th Circuit, which determined that some federal officials' intervention in social media content moderation constituted coerced censorship, in violation of the First Amendment.[250] Such intervention concerned the flagging of "problematic" content, demanding data on the impact of reduction, mandating changes to moderation policies,[251] and alluding to potential future legal liability for the platforms.[252] The pressure executed by these officials had significant consequences in the integration of reduction.[253] The court noted, for example, that "one platform sent out a post-meeting list of 'commitments' including a policy 'change[] [sic]' 'focused on reducing the virality' of anti-vaccine content even when it 'does not contain actionable misinformation.'"[254] In a separate case, "one email from Facebook stated that although a group of posts did not 'violate our community standards,' it 'should have demoted them before they went viral.'"[255] In another instance, Facebook recognized that a popular video did not qualify for removal under its policies, yet promised that it was "labeled" and "demoted" following its flagging by officials.[256] The Supreme Court, as mentioned above, has recently reversed the judgment due to lack of standing.[257] In a 6-3 ruling, the Court asserted, *inter alia*, that "the Government

[249] Gabriel Nicholas, *Shedding Light on Shadowbanning*, CDT 16 (Apr. 16, 2022), https://cdt.org/insights/shedding-light-on-shadowbanning/ ("Social media services face difficult trade-offs in their content moderation design choices because they face multiple competing incentives and have many stakeholders to manage, including posters, viewers, advertisers, shareholders, and governments…") [https://perma.cc/4QVQ-65JB].

[250] Missouri v. Biden, 80 F.4th 641, 650, 654 (5th Cir. 2023) (addressing, *inter alia*, a case where one official told a platform it would be "good to have from you all . . . a deeper dive on [misinformation] reduction." The court also stated that "one White House official demanded more details and data on Facebook's internal policies at least twelve times, including to ask what was being done to curtail 'dubious' or 'sensational' content, what 'interventions' were being taken, what 'measurable impact' the platforms' moderation policies had, 'how much content [was] being demoted,' and what 'misinformation' was not being downgraded")

[251] *Id.* at 652.

[252] *Id*.

[253] *Id*. at 650.

[254] *Id*.

[255] *Id*.

[256] *Id*.

[257] Murthy v. Missouri, 144 S.Ct. 1972, 1977 (2024).

defendants played a role in at least some of the platforms' moderation choices. But the Fifth Circuit, by attributing *every* platform decision at least in part to the defendants, glossed over complexities in the evidence" (emphasis added). [258] Nonetheless, and as emphasized in the Supreme Courts' dissenting opinion, some of the lower court's findings remain concerning.[259]

Third, reduction may offer Meta a questionable yet potent means to meet its regulatory obligations pertaining to the removal of harmful content. The Code of Conduct, for example, mandates that Meta and other platforms promptly remove content violating their (removal) policies upon user reports.[260] Reduction could effectively decrease the volume of potentially reportable content, particularly when applied to content deemed "likely to violate" the policies or content that is "borderline." [261] As a result, this approach may assist Meta to unaccountably comply with the Code's requirements.

Fourth, unlike removal, reduction leaves the content up, resulting in more data, an invaluable asset for Meta. This data serves as essential raw material for (1) ongoing surveillance supporting the personalization of content and other objectives, and (2) fueling Meta's AI enterprise.[262] Meta is a dominant force in the AI realm, with influence expanding far beyond its content moderation efforts.[263] Its advanced AI-driven products and processes rely on an ever-growing corpus of data.[264] Moreover, the data encapsulated in "problematic" and "borderline" content could be particularly valuable in the training of AI technologies, making it a resource Meta may be reluctant to forego.[265]

[258] *Id.* at 1988.

[259] *See id.* at 1997. The dissenting opinion concluded that "For months, high-ranking Government officials placed unrelenting pressure on Facebook to suppress Americans' free speech." *Id.* at 2015.

[260] *See supra* Part II. *See also* Didier Reynders (Commissioner for Justice), *Countering Illegal Hate Speech Online, 7th Evaluation of the Code of Conduct*, EUR. COMM'N (Nov. 2022), https://commission.europa.eu/document/download/5dcc2a40-785d-43f0-b806-f065386395de_en?filename=Factsheet%20-%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf (emphasizing a platform's removal rate of reported content in the European Commission's evaluation of the Code of Conduct).

[261] *See supra* Part III.B.

[262] *See supra* Part III.A.

[263] *See, e.g.*, *Inside the lab: Building for the metaverse with AI*, TECH AT META (Feb. 23, 2022), https://tech.facebook.com/artificial-intelligence/2022/2/building-for-the-metaverse-with-ai/ [https://perma.cc/23NC-LW4X].

[264] *See Self-Supervised Learning Cookbook, supra* note 119.

[265] For example, see Rosen, *supra* note 110, for Meta's explanation of their AI-based technologies' failure to detect and stop the live streaming of the New Zealand Terror attack in 2019 ("AI systems are based on 'training data' . . . . [T]his particular video did not trigger our automatic detection systems. To achieve that we will need to

B. Legal and Technological Challenges

The development and implementation of reduction behind the scenes of content moderation, distanced from transparency and accountability, pose significant challenges to facilitating a robust public debate about this strategy and its legitimacy. However, the fashion in which reduction is currently designed and executed by Meta brings further challenges. In the following section, I will explore how reduction: (1) conflicts with the principles of the rule of law and procedural fairness; (2) disproportionately impedes freedom of expression and other human rights; and (3) should be evaluated in the context of AI-vulnerabilities.

1. Reduction Conflicts with the Rule of Law and Procedural Fairness

Reduction, as currently applied by Meta, pushes content moderation further from the fundamental tenets of the rule of law and procedural fairness. Although distinct, both principles are instrumental in governing and restraining the exercise of power, preventing arbitrariness, and empowering individuals.[266]

The rule of law aims to establish regimes where rulers are bound and guided by the law.[267] This principle, originally conceived with the nation-state in mind, has attracted attention for its relevance to digital platforms, particularly due to their profound impact on users' human rights and the informational landscape.[268] Key requirements of the rule of law are that rules be known and publicly available, clear and feasible

provide our systems with large volumes of data of this specific kind of content, something which is difficult as these events are thankfully rare").

[266] Rebecca Hollander-Blumoff & Tom R. Tyler, *Procedural Justice and the Rule of Law: Fostering Legitimacy in Alternative Dispute Resolution*, 2011 J. DISP. RESOL. 1, 8–9 (2011).

[267] *Id*. *See also* Jeremy Waldron, *The Rule of Law and the Importance of Procedure*, 50 NOMOS 3, 81 (2011).

[268] Nicolas Suzor, *Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms*, 4(3) SOC. MEDIA SOC'Y. 1, 4 (2018). *See also* Leerssen, *supra* note 206, at 6. There is a large body of literature advocating for the application of public law norms and state duties also to digital platforms. *See generally* Mor, *supra* note 39, at 651–52.

to follow, consistently applied to all,[269] and provide a foundation upon which people can "plan their lives."[270]

However, reduction, as currently implemented by Meta, starkly contradicts these criteria. Consider, for instance, Meta's Content Distribution Guidelines on "Content Likely Violating the Community Standards."[271] It is not clear where this blurred threshold is located, what separates "likely" violating content from confirmed violations, and how this threshold varies across different Community Standards (e.g., the "Hate Speech" standard versus the "Bullying and Harassment" standard). Even more ambiguity surrounds the application of borderline content, which was omitted from the "Content Distribution Guidelines" but is still applied, as explored above.[272]

While the Community Standards are detailed and explicit, the reduction of content labeled as "likely violating" and "borderline" in relation to these standards creates an opaque boundary that runs somewhere before the threshold established by these Community Standards. This results in a policy that remains largely unknown to users, hampering their ability to adjust their behavior accordingly. Such ambiguity is evident in the experience of users whose content has been reduced, with some reporting that "the hardest part is simply not understanding what they've done wrong," or how to adjust their content to the platform's "liking."[273] This confusion is unsurprising, given that, as previously described, Meta does not seem to ascribe the same normative value to its reduction policy as it does to the Community Standards.[274]

Procedural Fairness focuses on the means used in decision-making.[275] Its application is tied, *inter alia*, to securing people's

[269] Joseph Raz, *The Rule of Law and its Virtue, in* THE AUTHORITY OF LAW: ESSAYS ON LAW AND MORALITY 210, 214 (1979). *See also* Hollander-Blumoff & Tyler, *supra* note 266, at 8.

[270] Waldron, *supra* note 267. *See also* Raz, *supra* note 269, at 213 ("[T]he law should be such that people will be able to be guided by it."). Some scholars also attribute a substantive meaning to the rule of law, encompassing the protection of human rights. *See* Hollander-Blumoff & Tyler, *supra* note 266, at 8. *See generally* Evan Fox-Decent, *Is the Rule of Law Really Indifferent to Human Rights* 27 L. AND PHIL. 533 (2008).

[271] *See supra* Part III.B.2. *See also Content Likely Violating our Community Standards*, *supra* note 194.

[272] *See supra* Part III.B.2.

[273] Cook, *supra* note 30.

[274] *See supra* Part III.B.3.

[275] Jerald Greenberg & Tom R. Tyler, *Why Procedural Justice in Organizations?*, 1 SOC. JUST. RSCH. 127, 129 (1987); Liangtie Dai & Haixin Xie, *Review and Prospect on Interactional Justice*, 4(1) OPEN J. SOC. SCI. 55–61 (2016).

voices,[276] fostering trust, legitimizing a governing regime,[277] and enhancing people's willingness to cooperate with policies and decisions.[278] Procedural fairness requires, among other components, that individuals be aware of sanctions imposed on them, be provided with reasons for such sanctions, and have the opportunity to challenge these decisions.[279] As noted, these opportunities are currently largely unavailable to Meta's users in the context of implementing reduction.[280]

The recent application of the DSA is expected to introduce some positive developments in this domain. However, beyond its limited territorial scope, there are concerns about whether procedural fairness requirements will be thoroughly met, even after this regulation is adopted by digital platforms. Consider, for example, the requirement to provide an explanation. Even in cases of removal, the explanation currently offered to affected users tends to be quite general, often referring to the supposedly violated standard rather than specifying the particular rule that was breached.[281] The opacity surrounding the Content Distribution Guidelines and the application of reduction only compounds this issue. A mere general reference to these guidelines, although a step in the right direction, will not suffice to provide adequate reasoning.

2. Reduction Disproportionately Impedes Freedom of Expression and Other Human Rights

i. Reduction's Adverse Effect on Freedom of Expression and Additional Human Rights

As enshrined, *inter alia*, in Article 19 of the International Covenant on Civil and Political Rights (ICCPR), freedom of expression encompasses the right to speak and be heard, as well as the right to access information.[282] Reduction, it appears, obstructs these critical aspects of freedom of expression.

[276] Hollander-Blumoff & Tyler, *supra* note 266, at 5–6.

[277] Marcia Grimes, *Organizing Consent: The Role of Procedural Fairness in Political Trust and Compliance*, 45 EUR. J. POLIT. RSCH. 285, 285 (2006).

[278] Robert J. MacCoun, *Voice, Control, And Belonging: The Double-Edged Sword of Procedural Fairness*, 1 ANN. REV. L. SOC. SCI. 171, 180 (2005).

[279] Leerssen, *supra* note 206. For the application of procedural fairness to digital platforms, see Nathenson, *supra* note 10, for example. See Greenberg & Tyler, *supra* note 275, for a more general discussion on its application to private companies

[280] *See supra* Part III.B.3.

[281] *Id.*

[282] G.A. Res. 2200A (XXI), International Covenant on Civil and Political Rights, at 19, (Dec. 16, 1966). *See also* G.A. Res. 217 (III) A, Universal Declaration of Human

Reduction acts as a sweeping mechanism that "turns down" voices, ideas, and perceptions, affecting both creators of content and their potential audience. Those whose expression is reduced are not truly heard, and society may receive a watered-down, distorted, and partial selection of information and communication opportunities.[283] This compounded effect of reduction also contradicts the two rationales underpinning freedom of expression. The first rationale centers on the individual, focusing on autonomy; self-fulfillment; dignity; and the ability to express oneself, consume information, and freely shape personal views.[284] It revolves around people's *forum internum*: their "inner realm of thoughts, beliefs, and convictions."[285] The second rationale is epistemic, safeguarding "public debate, open dialogue, and the foundations of democracy itself."[286] Given the extensive and unregulated limitation of content imposed by Meta in its reduction strategy, neither rationale is fully realized.

Another troubling aspect of reduction is that it often targets content that is unpopular and disliked.[287] However, the insistence on preserving space for such content is the very essence of freedom of expression. Exposure to disagreements and different viewpoints, including such that are not common, allows people to revisit their thoughts and beliefs on various matters, [288] and contributes to the

Rights (Dec. 10, 1948). Private companies are expected to respect human rights, as outlined in Off. of the High Comm'r, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, U.N. Doc. HR/PUB/11/04 (2011). Meta has explicitly expressed their commitment to comply with these legal sources. *See Corporate Human Rights Policy*, *supra* note 67.

[283] *See supra* Part III.B. *See also* Complaint for Declaratory and Injunctive Relief at 23–24, Knight First Amend. Inst. at Columbia Univ. v. Trump, 302 F. Supp. 3d 541 (S.D.N.Y. 2018) (plaintiffs contended that the blocking of users from the then-U.S. President, Donald Trump's Twitter account, violates the First Amendment "by distorting the expressive forum" in which unblocked users participate. This argument is also applicable in cases of content reduction, in my view).

[284]Ahron Barak, *Freedom of Speech and its Limitations*, M(A) THE ATT'Y 5, 6–10 (1990) (Heb.). *See also* G. Michael Parsons, *Fighting for Attention: Democracy, Free Speech, and the Marketplace of Ideas*, 104 MINN. L. REV. 2157, 2158 (2020).

[285] Gehan Gunatilleke, *Justifying Limitations on the Freedom of Expression*, 22 HUM. REV. 91, 93 (2021).

[286] Parsons, *supra* note 284, at 2158. *See also* Gunatilleke, *supra* note 285, at 93. This rationale is emphasized in the following words of John Stuart Mill: "[T]he peculiar evil of silencing the expression of an opinion is, that it is robbing the human race; posterity as well as the existing generation; those who dissent from the opinion, still more than those who hold it." JOHN STUART MILL, ON LIBERTY 33 (1st ed. 2018).

[287] *See supra* Parts III.B.1–2.

[288] Joshua Cohen & Archon Fung, *Democracy and the Digital Public Sphere*, *in* DIGITAL TECHNOLOGY AND DEMOCRATIC THEORY 29, 30 (Lucy Bernholz, Héléne Landemore, & Rob Reich eds., 2021). *See also id.* at 51 ("Platform architects should seek to expose users to ideas that lie outside their familiar territory and to content that

prevention of "tyranny of the majority within a democracy, especially considering how strongly people tend to believe that their own views are correct."[289] The U.S. Supreme Court has long stated, "[i]f there is a bedrock principle underlying the First Amendment, it is that the government may not prohibit the expression of an idea simply because society finds the idea itself offensive or disagreeable."[290] Restricting the visibility of content merely because it is considered "problematic," "inappropriate," or "sensational," inherently contradicts this approach.

It should also be noted that due to the intrinsic social virtues of social media platforms, restricting content not only limits the informational resources that are available to users but also undermines the social-communal resources that could develop from this content. In other words, when content concerning vulnerable groups or controversial issues is silenced, the potential for interpersonal connections that might build upon this content may also be hampered.[291]

Moreover, the exclusion of content, people, and communities that do not conform to the mainstream norms, as is currently done through reduction, may label them as unacceptable "others,"[292] fostering their stigmatization[293] and isolation.[294] Indeed, according to a survey by the Center for Democracy and Technology (CDT), 54% of users reported that reduction "made them feel isolated and removed from their social group, community, or society at large."[295]

The secretive and sweeping character of reduction might also lead to a chilling effect, encouraging users to overly self-regulate the content

is visibly common. Broad adherence to such common-good-oriented behaviors would foster greater access, expression, and perhaps diversity in the digital public sphere."). For the obligation to maintain diversity in Media, see *General comment No.34 on Article 19: Freedoms of opinion and expression*, CCPR/C/GC/34 OHCHR (Jul. 29, 2011), https://documents.un.org/doc/undoc/gen/g11/453/31/pdf/g1145331.pdf (last visited Feb. 6, 2024).

[289] Melina Constantine Bell, *John Stuart Mill's Harm Principle and Free Speech: Expanding the Notion of Harm*, 33 UTILITAS 162, 164 (2021) (addressing the background against which Mill wrote *On Liberty*. *See supra* note 286).

[290] Texas v. Johnson, 491 U.S. 397, 414 (1989). In *New York Times Co. v. Sullivan*, 376 U.S. 254 (1964), the court emphasized the importance of including unpopular perspectives in discussions on public matters ("debate on public issues should be uninhibited, robust, and wide-open, and that it may well include vehement, caustic, and sometimes unpleasantly sharp attacks on government and public officials"). This stance is part of a broader notion governing content restrictions in the U.S., wherein government limitations on content "because of its message, its ideas, its subject matter, or its content," is generally prohibited." For a related discussion on the application of the proportionality requirement, see infra Part IV.B.2.

[291] Mor, *supra* note 39, at 656–62

[292] Are, *supra* note 143, at 2. *See also* Cohen & Fung, *supra* note 288, at 45.

[293] *See supra* Part III.B.

[294] Nicholas, *supra* note 249, at 30.

[295] *Id.*

they post and share. This hampers their freedom of expression, autonomy, and ability to participate in the digital sphere.[296] The U.S. Court of Appeals for the 5th Circuit noted that plaintiffs previously censored by social media stated that these sanctions "caused them to self-censor and carefully word social-media posts moving forward in hopes of avoiding suspensions, bans, and censorship in the future."[297] The court emphasized that the plaintiffs' fears were "far from hypothetical" and that their self-censorship "is a cognizable, ongoing harm resulting from their past censorship injuries, and therefore constitutes injury-in-fact."[298]

In addition, the opacity characterizing reduction could make it more susceptible to pressure from governments and enforcement bodies. The Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression warned that "[b]roadly worded restrictive laws on 'extremism', blasphemy, defamation, 'offensive' speech, 'false news' and 'propaganda' often serve as pretexts for demanding that companies suppress legitimate discourse."[299]

Lastly, before moving on to the discussion of justified limitation on freedom of expression, it's important to note that freedom of expression is a gateway to the realization of many other liberties and "a basis for the full enjoyment of a wide range of other human rights."[300] The UN Special Rapporteur stated when addressing online domains that

> The right to freedom of opinion and expression is as much a fundamental right on its own accord as it is an "enabler" of other rights, including economic, social and cultural rights, such as the right to education and the right to take part in cultural life and to enjoy the benefits of scientific progress and its applications, as well as civil

[296] *See generally* Jonathon W. Penney, *Chilling Effects: Online Surveillance and Wikipedia Use*, 31 BERKELEY TECH. L.J. 117,] (2016); Jonathon W. Penney, *Internet Surveillance, Regulation, and Chilling Effects Online: A Comparative Case Study*, 6 INTERNET POL'Y REV. 8 (2017); *Chilling Effects: NSA Surveillance Drives US Writers to Self-censor*, PEN AMERICA CENTER (2013). For a broader discussion regarding the chilling effect and its impact on users' choices and behavior, *see* MICHEL FOUCAULT, DISCIPLINE AND PUNISH: THE BIRTH OF THE PRISON 187 (Alan Sheridan trans., Vintage Books 2d ed. 1995) (1977); *see generally* Frederick Schauer, *Fear, Risk and the First Amendment: Unraveling the Chilling Effect*, 58 B.U. L. REV. 685 (1978).

[297] Missouri v. Biden, Missouri v. Biden, 80 F.4th 641, 655 (5th Cir. 2023).

[298] *Id. But see* Murthy v. Missouri, 144 S. Ct. 1972, 1987–88 (2024) (doubting the ability to refer such self-censoring to the defendants.

[299] David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, A/HRC/38/35HUMAN RIGHTS COUNCIL 6 (Apr. 6, 2018), [https://documents.un.org/doc/undoc/gen/g18/096/72/pdf/g1809672.pdf]. For a discussion about U.S. federal officials' intervention in reduction application, see *supra* Parts I & IV.A.

[300] *General comment no. 34, Article 19, supra* note 7, at ¶4.

> and political rights, such as the rights to freedom of association and assembly.[301]

Examples of affected rights, as reflected in the preceding discussion, include freedom of thought, the right to dignity and autonomy, and the right to equality. Another right that is adversely impacted by reduction is freedom of occupation, as indicated, for instance, by users who rely on their social media accounts to sustain small businesses.[302] Thus, by severely limiting freedom of expression, reduction hinders an entire array of additional human rights.

ii. Reduction and Proportionality

Notwithstanding its importance, freedom of expression is not, nor should it be, an absolute right. It is, rather, a relative right that can be restricted in appropriate cases.[303]

A widely endorsed legal measure for identifying such appropriate cases is the proportionality test.[304] This measure offers a coherent and structured formula for balancing the competing rights and interests, while still allowing for the specific consideration of each case.[305] The proportionality test has grown to "dominate the dockets of constitutional and supreme courts" in many territories[306] and is the mechanism used in the aforementioned Article 19 of the ICCPR to restrict freedom of expression.[307] Furthermore, even though the U.S.

[301] Frank La Rue, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, A/HRC/17/27 HUMAN RIGHTS COUNCIL 7 (May 16, 2011), [https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf].

[302] *See* Nicholas, *supra* note 249, at 31 (according to CDT's survey, "20% of shadowbanned users indicated that being shadowbanned affected their ability to make a living").

[303]

[304] Mor, *supra* note 39, at 692–95.

[305] *See generally* Moshe Cohen-Eliya & Iddo Porat, *The Hidden Foreign Law Debate in* Heller*: The Proportionality Approach in American Constitutional Law*, 46 SAN DIEGO L. REV. 367, 369 (2009).

[306] Alec Stone Sweet & Jud Mathews, *Proportionality Balancing and Global Constitutionalism*, 47 COLUM. J. TRANSNAT'L L. 72, 73–74, (2008).

[307] *International Covenant on Civil and Political Rights*, *supra* note 282 ("3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (ordre public), or of public health or morals."). This provision includes key components of the proportionality test. *See infra* Part IV.B.2.

adjudication does not formally apply proportionality, it manifests in American law in various ways.[308] *Inter alia*, aspects of the proportionality requirements are reflected in the Strict Scrutiny test, which assesses the constitutionality of laws that jeopardize certain fundamental rights.[309]

Proportionality does not only apply to states. Its role in online content governance is growing, and as part of this development, platforms themselves refer to proportionality constraints as applicable.[310] Approving this trend, Evelyn Douek stated that "proportionality is a mature approach to resolving the many conflicts created by the collision of varying interests online."[311] The application of proportionality to digital platforms also aligns with the growing body of literature that points out these actors' state-like characteristics.[312]

On the surface, reduction appears as a more moderate measure compared to removal, as it does not involve the outright deletion of the content. Indeed, Meta, along with other platforms, highlights this approach in public discussions about reduction. Meta explained, in the context of reducing misinformation, that "[w]e want to strike a balance between enabling people to have a voice and promoting an authentic environment. When misinformation is identified by our fact-checking partners, we reduce its distribution within Feed and other surfaces."[313]

A closer examination of the reduction currently applied, through the lens of the proportionality test, reveals a different state of affairs. The proportionality test comprises four subtests, the first of which concerns the legitimacy of the objective behind the restriction applied. It then proceeds with the following three subtests: the chosen means for the limitation must be suitable for achieving its objective ("suitability"); the means must be the least intrusive option available ("necessity"); and

[308] THOMAS SULLIVAN & RICHARD S. FRASE, PROPORTIONALITY PRINCIPLES IN AMERICAN LAW: CONTROLLING EXCESSIVE GOVERNMENT ACTIONS 6 (2009) (arguing that "proportionality review is emerging in U.S. law but is not yet a unified theory").

[309] *See* Vicki C. Jackson, *Constitutional Law in an Age of Proportionality*, 124 YALE L.J. 3094, 3094 (2015) ("[S]ome areas of U.S. constitutional law embrace proportionality as a principle, as in Eighth Amendment case law, or contain other elements of the structured 'proportionality review' widely used in foreign constitutional jurisprudence, including the inquiry into 'narrow tailoring' or 'less restrictive alternatives' found in U.S. strict scrutiny.").

[310] *See* Evelyn Douek, *Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability*, 121 COLUM. L. REV. 759, 776–85 (2021).

[311] *Id.* at 785.

[312] *See supra* notes 38–41.

[313] *How Meta's third-party fact-checking program works*, *supra* note 66. YouTube has been using the same terminology. *Accord Continuing our work to improve recommendations on YouTube*, BLOG.YOUTUBE (Jan. 25, 2019), https://blog.youtube/news-and-events/continuing-our-work-to-improve/ [https://perma.cc/63W8-QSCB].

a cost-benefit analysis must demonstrate that a proper balance between the harm to individuals' rights and the gain achieved by the restriction has been struck ("proportionality in the strict sense").[314]

I believe these requirements are not met in the context of reduction. Let us start with the legitimacy of the reduction's objective. As earlier observed, Meta presents an opaque mosaic of different goals for applying reduction, many of which raise doubts concerning their legitimacy.[315] Particularly troubling are questions regarding the legitimacy of purposes that do not aim to prevent harm, protect human rights, or secure compelling interests such as public security, public peace, or public health (as explicitly mentioned in the ICCPR).[316] Examples of such purposes include Meta's efforts to prevent people from being exposed to content they "don't like,"[317] and the company's desire to encourage "high-quality" content.[318] Other purposes mentioned by the company, like preventing polarization or fostering safety,[319] could have been appropriate, had they not been so abstractly introduced.[320]

However, even if we assume that Meta's purposes for conducting reduction are legitimate, it remains doubtful whether this practice, as currently implemented, is indeed "suitable" to further them.[321] Consider the objective of catering to peoples' wants regarding the content they consume. Is reduction, as currently implemented, an effective means to achieve this goal? Meta argues that it utilizes both direct (e.g., conducting surveys) and indirect feedback (e.g., monitoring and processing users' data) to guide the company's application of reduction.[322] Nonetheless, in the absence of detailed and publicly available data about this feedback, and the manner in which content's visibility is limited, it is difficult to determine whether this purpose can

[314] Mor, *supra* note 39, at 693; *see also* Moshe Cohen-Eliya & Iddo Porat, *American Balancing and German Proportionality: The Historical Origins*, 8 INT'L J. CONST. L. 263, 267 (2010); *see also* Dieter Grimm, *Proportionality in Canadian and German Constitutional Jurisprudence*, 57 U. TORONTO L.J. 383, 387–88 (2007).

[315] *See supra* Part IV.A.

[316] *See* La Rue, *supra* note 301.

[317] *See generally supra* Parts III.B & IV.A.

[318] *See id.*

[319] *See id.*

[320] *See id.*

[321] For the significance of the suitability subtest, see PHILIP SELZNICK, PHILIPPE NONET & HOWARD M. VOLLMER, LAW, SOCIETY, AND INDUSTRIAL JUSTICE 13 (1969) ("Rules are made arbitrarily when appropriate interests are not consulted and when there is no clear relation between the rule enunciated and the official end to be achieved.").

[322] *See Reducing the distribution of problematic content*, *supra* note 128. In the reduction policy, the "direct feedback" option is the only one included. *See supra* Part III.B.2.

indeed be realized.[323] Consider a more challenging example: Meta's goal to prevent polarization. While there is a basis to believe that reduction of certain types of content could serve as a mechanism to achieve this end,[324] the question remains: does silencing content, beyond the removal of content forbidden under the Community Standard, have a real-life, research-backed effect on polarization? Here too, without supporting data, and in light of the concerns that reduction, as currently applied, might actually *foster* segmentation, stigmatization, and isolation,[325] the answer is not straightforward.

Reduction also falls short of meeting the "necessity" requirement. Currently, it appears that reduction is conducted in a sweeping manner, encompassing all areas of content, and employing different censuring tools.[326] There is no publicly available data indicating vigilance in the application of reduction, nor is there evidence showcasing how specific types of reduction are tailored to achieve particular objectives. Furthermore, there is an absence of data indicating that reduction is a less harmful measure compared to other sanctions, such as removal or warning screens.[327] Under these conditions, reduction cannot be regarded as "the least intrusive" measure upon freedom of expression and other human rights.

The same conclusion applies to the "proportionality in the strict sense" requirement. Reduction severely undermines freedom of expression and additional human rights, dilutes the informational landscape, and excludes vulnerable groups. This is underscored by the clash between reduction and the rule of law and procedural fairness. Moreover, the absence of data on reduction's effectiveness in promoting human rights or interests further illuminates its current detrimental implementation.[328]

[323] For instance, in the context of user surveys, Meta did not reveal the sectors, nationalities, and age groups of the participants. Equally obscured were the manner in which the questions were framed, the languages employed, and the timing of these inquiries. Meta also left unclear whether there were any subsequent follow-ups, if the surveys are ongoing, and how feedback has evolved over time.

II. [324] *SEE GENERALLY* ARORA SWAPAN ET AL., *POLARIZATION AND SOCIAL MEDIA: A SYSTEMATIC REVIEW AND RESEARCH AGENDA*, 183 TECH. FORECAST. SOC. CHANGE 1, 6 (2022) (ADDRESSING THE NEXUS BETWEEN VIRALITY OF CERTAIN CONTENT AND POLARIZATION).

[325] *See supra* Part IV.B.2; *see also infra* Part IV.B.3.

[326] *See generally supra* Part III.B.

[327] *See generally supra* Part III.B; *see generally supra* Part IV.A.

[328] *Id.*

3. Reduction and AI-driven Vulnerabilities

Reduction, as observed above, is inherently dependent on AI.[329] However, AI serves not only as the technological foundation for the sophisticated and diverse manifestations of reduction, but also as the source of an additional set of vulnerabilities and limitations. These must be considered, as they further intensify the challenges previously explored.

One such difficulty is the inherent lack of transparency surrounding AI.[330] This challenge exists in traditional machine learning methods but becomes more acute with advanced, deep learning models and LLMs.[331] These AI approaches infer predictions, insights, and patterns from vast datasets, with significantly reduced human involvement.[332] When combined with the existing secrecy surrounding reduction, this lack of transparency poses a genuine obstacle to fairness, the rule of law, and human rights.

Moreover, AI's performance falls short in certain tasks, including such that involve context identification. Despite considerable advancements, AI technologies might incorrectly interpret criticism, parody, and self-referential terms as hateful content, or fail to recognize slang and abbreviations.[333] An additional weak point of AI is its limited

[329] *See Our New AI System to Help Tackle Harmful Content*, *supra* note 118; *see also* Zuckerberg, *supra* note 1; *see generally* discussion *supra* Part I.

[330] *See generally* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (Harvard University Press, 2015); *see generally* Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69(1) FLA. L. REV. 181 (2018); *see generally* Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014).

[331] See *supra* Part III.A.

[332] *See generally* Gabriel Nicholas & Aliya Bhatia, *Lost in Translation: Large Language Models in Non-English Content Analysis* (May 23, 2023) CDT, https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/ [https://perma.cc/G2WA-A6HS].

[333] *See* Bryce Hoffman, *Leaders Looking To Leverage AI Need To Think About Context*, FORBES (Mar. 31, 2023), https://www.forbes.com/sites/brycehoffman/2023/03/31/leaders-looking-to-leverage-ai-need-to-think-about-context/ [https://perma.cc/6SKS-PT78]; *see also* James Vincent, *AI won't relieve the misery of Facebook's human moderators*, THE VERGE (Feb. 27, 2019), https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms [https://perma.cc/EXY7-HL6E]; *see also* Troy Wolverton, *AI is great at recognizing nipples, Mark Zuckerberg says*, BUS. INSIDER INDIA (Apr. 25, 2018), https://www.businessinsider.com/ai-can-identify-nipples-but-not-hate-speech-mark-zuckerberg-says-2018-4 [https://perma.cc/YN6V-VUJW]. *See, e.g.* For the improvements, see, for instance, Siladitya Ghosh, *LLAMA 3: A New Frontier in Large Language Models*, MEDIUM (Aug. 19, 2024),

capabilities in languages other than English, especially those of smaller communities and sectors.[334] The opaqueness of reduction and the lack of scrutiny around this strategy may hamper efforts to expose these limitations and spark public debate about them.

Another significant concern regarding AI performance is bias and discrimination. Numerous resources indicate that AI-technologies discriminate against vulnerable groups, including women and people of color.[335] The roots of such bias extend through various stages of the AI development pipeline, from the developers' design choices to the scarcity of high-quality data on marginalized sectors, as well as the pre-existing biases in the datasets used for training AI models.[336] Here too, carrying out reduction with little transparency lowers the chances of identifying and addressing such issues. This is particularly alarming given that the current implementation of reduction already excludes and silences vulnerable voices.[337]

C. The Way Forward

Reduction, it appears, is here to stay. Furthermore, it is poised to continue evolving and, in my opinion, to become the predominant content moderation strategy employed by Meta and other online platforms. However, this does not necessarily have to be a bleak scenario.

Digital platforms are grappling with harmful content on an unprecedented scale, where this content emerges in new formats and spans various languages and dialects. AI, the driving force behind reduction, is flourishing and equips these platforms with advanced, sophisticated, and powerful tools to tackle many of these challenges. Moreover, reduction, which does not entail the physical removal of the

https://medium.com/@siladityaghosh/llama-3-a-new-frontier-in-large-language-models-d60d0edcc383 (describing improvements demonstrated by LLMs in this regard) [https://perma.cc/KQ8M-SVRK].

[334] *See* Nicholas & Bhatia, *supra* note 323, at 26–27.

[335] *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104(3) CAL. L. REV. 671, 678–80 (2016); *see also* Jeff Larson, Surya Mattu, Lauren Kirchner & Julia Angwin, *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [https://perma.cc/YW8W-3JV6]; *see also* Stella Lowry and Gordon Macpherson, *A Blot on the Profession*, 296 BRIT. MED. J. 657, 657 (1988); *see also* Michael Zhang, *Flickr Fixing 'Racist' Auto-Tagging Feature After Black Man Mislabeled 'Ape,'* PETAPIXEL (MAY 20, 2015), https://petapixel.com/2015/05/20/flickr-fixing-racist-auto-tagging-feature-after-black-man-mislabeled-ape/ [https://perma.cc/7CDK-GEUU].

[336] *See* Barocas & Selbst, *supra* note 335; *see also* Nicholas & Bhatia, *supra* note 323, at 24.

[337] *See supra* Part III.B.1.

content, aligns with our increasingly strong perception of UGC as valuable data, essential for current and future applications and technological advancements. Lastly, reduction offers a method that adapts to our evolving understanding of what constitutes "unacceptable," "acceptable," "true," or "false" content—a flexibility that removal lacks.[338]

Nonetheless, reduction is currently implemented in an extensive, opaque, and unsupervised manner that deeply endangers users' freedom of expression and other rights, including the right to dignity and the right to equality.[339] It also conflicts with the rule of law and procedural fairness.[340] For reduction to offer an appropriate way forward in the content moderation realm, a few central steps need to be undertaken.

First, reduction cannot remain concealed in the "backstage" of content moderation. Meta should substantially enhance transparency regarding this strategy and foster public discourse among various stakeholders on its scope, nature, and implications. As part of these efforts, the company should incorporate meaningful and comprehensive data about the practice in its voluntary Transparency Reports, where currently no significant information regarding reduction is provided. Additionally, since reduction may be subtly employed by Meta to fulfill its regulatory obligation concerning the removal of harmful content, as previously discussed, it is imperative that the reports Meta submits to regulators do not overlook reduction's impact.[341]

Second, to align more effectively with the rule of law and procedural fairness, Meta must develop a detailed policy that governs this strategy, ensuring it is easily understandable and actionable for users. The company should also notify users whose content has been subjected to reduction, providing them with reasons and a quantifiable measure of the sanction's implications, such as the rate of reduced views. Furthermore, Meta should establish appeal mechanisms, both internally within the company and through the Oversight Board, for content it has reduced.[342]

Third, reduction must be confined to prevent overly broad and disproportionate harm to freedom of expression and other human rights. To achieve this, the purposes of this strategy must be publicly and clearly described, with reduction being limited and calibrated to achieve

[338] *Id., supra* Part I.
[339] *See supra* Part IV.B.2.
[340] *See supra* Part IV.B.1.
[341] *See supra* Part III.B.3.
[342] *See supra* Part IV.B.1.

these stated ends. Data regarding such processes should be made publicly available.[343]

Fourth, channels for assessing the challenges posed by AI in the specific context of reduction should be established. This could involve requiring Meta to submit designated periodic reports and granting access to researchers and regulators to the relevant models in use.[344]

V. CONCLUSION

Reduction is a transformative online content moderation strategy designed to limit the visibility of content. Initially confined to specific areas, this strategy has expanded to encompass all content categories, while raising the threshold for what constitutes permissible content and diminishing the wealth of information accessible to users. Furthermore, unlike the more accountable process of content removal, the implementation of reduction is notably nontransparent and governed by incoherent and vague guidelines of dubious legitimacy. Consequently, its application stands at odds with the principles of the rule of law and procedural fairness, and disproportionally infringes upon users' human rights. To realize the significant potential of reduction's elastic and sophisticated approach in tackling harmful content and benefitting our future digital sphere, these challenges must be addressed.

[343] *See supra* Part IV.B.2.

[344] *See supra* Part IV.B.3.