COMPARATIVE LAW — FREEDOM OF EXPRESSION — OVERSIGHT BOARD FINDS A FACEBOOK RULE'S APPLICATION VIOLATES INTERNATIONAL HUMAN RIGHTS LAW. — *Case Decision 2021-004-FB-UA*, OVERSIGHT BD. (MAY 26, 2021), https://oversightboard.com/decision/FB-6YHRXHZR [https://perma.cc/4JKV-NLPX].

Social media platforms have a great deal of power to regulate speech and face challenges doing so on a global scale. Facebook recently committed to respecting International Human Rights Law (IHRL) through the voluntary framework set forth in the United Nations Guiding Principles on Business and Human Rights (UNGPs).[1] Facebook also created the Oversight Board, a quasi-independent adjudicatory body, to review Facebook's content moderation decisions, putting IHRL's application to social media to the test. In *Case Decision 2021-004-FB-UA*[2] ("*Cowardly Bot Case*"), Facebook removed political content that contained an insult under the platform's rules on bullying and harassment, which are deferential to self-reporting targets. The Oversight Board found that Facebook's content moderation decision complied with the company's rules but violated IHRL. During and after the case, Facebook staunchly defended its practices.

The *Cowardly Bot Case* showcases a key disconnect in Facebook's commitment to IHRL: Facebook's rules and IHRL use different methodologies for adjudication. Proportionality and categorization are the two most notable methodologies in fundamental rights adjudication.[3] For speech rights, IHRL uses proportionality, balancing interests within *individual* cases. By contrast, Facebook uses categorization, balancing interests to create rules that govern *all* cases.

While neither methodology is intrinsically superior, this comment argues that categorization better suits major platforms for several reasons. First, on major platforms, categorical rules can produce more accurate decisions across groups of cases for reasons including moderators' lack of judicial experience. Second, categorization allows for rules to be tailored per class of content, which helps platforms manage their high caseloads and allows platforms to design rules with risk preferences to govern issues for which moderators cannot access relevant information, like bullying. Third, repeated adjudication in an area tends to produce

---

[1] *Corporate Human Rights Policy*, FACEBOOK, https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf [https://perma.cc/GE7L-8G2T]; *see also* evelyn douek, *The Limits of International Law in Content Moderation*, 6 U.C. IRVINE J. INT'L TRANSNAT'L & COMPAR. L. 37, 38–39 (2021).

[2] OVERSIGHT BD. (May 26, 2021) [hereinafter *Cowardly Bot Case*], https://oversightboard.com/decision/FB-6YHRXHZR [https://perma.cc/4JKV-NLPX].

[3] *See* AHARON BARAK, PROPORTIONALITY 502 (Doron Kalir trans., 2012).

rules, which makes major platforms' use of categorical rules natural given their unprecedented volumes of cases.

The Oversight Board is a quasi-independent adjudicatory body with the power to review Facebook's content moderation decisions and to issue policy recommendations.[4] The Board assesses whether Facebook's decisions comply with three sources of authority: Facebook's private rules (known as Community Standards), Facebook's company "values," and IHRL.[5] Although nations' and companies' IHRL obligations are not identical under the UNGPs, the Board has followed a U.N. Special Rapporteur's position that platforms committed to IHRL generally assess "the same kind of questions about protecting their users' right to freedom of expression" that "[g]overnments" consider.[6]

In the *Cowardly Bot Case*, a user posted about protests in Russia after the jailing of an opposition leader against the state government.[7] Another user ("the Critic") added a comment, claiming that the protesters in Moscow were all "shamelessly used" schoolchildren, not the voice of the people.[8] After others challenged the assertions, the Critic stated that those who brought elderly people to the protests were "morons."[9] Yet another user ("the Protester"), who self-identified as elderly, added commentary in support of the opposition that ended by calling the Critic a "cowardly bot."[10] The Critic reported the Protester's comment under the Community Standard on Bullying and Harassment.[11]

Facebook removed the comment. Under the Bullying and Harassment policy, "Facebook removes negative character claims aimed at a private individual when the target reports the content."[12] A moderator found that "cowardly" was a "negative character claim" and that a private "target" filed the report.[13] On appeal, Facebook swiftly affirmed.[14]

The Oversight Board overturned the decision. First, the Board affirmed that the content moderation decision complied with the Bullying

---

[4] Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L.J. 2418, 2481–83 (2020).

[5] *See, e.g.*, *Case Decision 2020-003-FB-UA*, OVERSIGHT BD. § 4 (Jan. 28, 2021), https://oversightboard.com/decision/FB-QBJDASCV [https://perma.cc/EF55-KV77]. The *Cowardly Bot Case* presented the first conflict between the authorities. In prior cases, the Board's analyses under the three sources had aligned. *See, e.g.*, *id.* §§ 8.1–.3.

[6] *Id.* § 4 (quoting David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 41, U.N. Doc. A/74/486 (Oct. 9, 2019)).

[7] *Cowardly Bot Case*, *supra* note 2, § 2. The jailing was of Alexei Navalny. *Id.*

[8] *Id.*

[9] *Id.*

[10] *Id.*

[11] *Id.*

[12] *Id.* § 8.1.

[13] *Id.* § 2.

[14] *Id.*

and Harassment policy, as "cowardly" could be "construed as a negative character claim."[15] Nonetheless, the Board criticized the rule itself, lamenting that the "case illustrates that Facebook's blunt and decontextualized approach can disproportionately restrict freedom of expression."[16] The Board highlighted that Facebook's rule failed to "*balance*" the speech interests in the debate "against the reported bullying."[17] By contrast, Facebook provided statements for the case explaining that the "*balancing* [of competing interests] is undertaken when the Community Standards are drafted."[18] Facebook asserted that, as a general matter, negative character claims "prevent people from feeling safe and respected on the platform."[19]

In a detailed analysis, the Oversight Board found that the content moderation decision "was not consistent with Facebook's human rights responsibilities."[20] Under Article 19 of the International Covenant on Civil and Political Rights (ICCPR),[21] speech restrictions must meet principles of "legitimate aim," "proportionality," and "legality."[22] The policy's goal "to protect" others was a "legitimate aim."[23] However, the Board found that the moderation failed the proportionality test, which requires that speech restrictions are "proportionate to the interest to be protected."[24] The Board cited authority on "hate speech," which indicated that political speech warrants heightened protections, and the Board declared that "[t]his approach may be extended to assessments of bullying and harassment."[25] The Board discussed sociopolitical context and reasoned the term "cowardly bot" was "unlikely to cause harm."[26]

The Board also found that the content moderation decision violated Facebook's "values" for "fail[ing] to balance" "'Dignity' and 'Safety' against 'Voice'"[27] — rejecting Facebook's argument that the moderation was "in line with its values of 'Dignity' and 'Safety'" and that requiring self-reporting by targets "ensures everyone's 'Voice' is heard."[28]

---

[15] *Id.* § 8.1.

[16] *Id.*

[17] *Id.* (emphasis added).

[18] *Id.* (emphasis added).

[19] *Id.* § 6.

[20] *Id.* § 8.3. Facebook provided a short counterargument about IHRL that did not address the Protester's speech interests. *Id.* § 6.

[21] *Adopted* Dec. 16, 1966, 999 U.N.T.S. 171 [hereinafter ICCPR].

[22] *Cowardly Bot Case*, *supra* note 2, § 8.3 (citing ICCPR, *supra* note 21, art. 19(3)).

[23] *Id.* (citing ICCPR, *supra* note 21, art. 19(3)).

[24] *Id.* (quoting Hum. Rts. Comm., General Comment No. 34, ¶ 34, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011)).

[25] *Id.* (citing Kaye, *supra* note 6, ¶ 47(d)).

[26] *Id.* The Board also found the policy failed the "legality" principle, requiring "clear and accessible" rules, due to its complexity, organization, and failure to define certain terms. *Id.*

[27] *Id.* § 8.2. The Board suggested that "political speech" was central to "Voice." *Id.*

[28] *Id.* § 6.

Lastly, the Board provided recommendations for "compl[iance] with international human rights standards."[29] In line with the proportionality analysis, the Board recommended: "Facebook should . . . require an assessment of [content's] social and political context" upon which to "reconsider the enforcement of [the] rule."[30] In a response revealing a sharp disconnect, Facebook declined to commit to the proposal:

> This recommendation proposes that we scale the ability to moderate potentially violating content differently depending on the social or political context within which a user posts. By its nature, though, content moderation at scale requires *principled criteria* for our content moderators designed to ensure *speed*, *accuracy*, *consistency*, and *non-arbitrary* content moderation.[31]

In the *Cowardly Bot Case*, the Board's and Facebook's positions sharply conflicted. The Board's stance on platforms' IHRL obligations was in step with authority from a U.N. Special Rapporteur and numerous scholars,[32] and the IHRL analysis was fair. A platform governance scholar even called it "an easy case."[33] Yet Facebook staunchly defended its practices. This comment proceeds by: (1) explaining a disconnect between Facebook's rules and IHRL — they use different methodologies, respectively, categorization and proportionality — and (2) arguing that categorization is superior for major platforms.

The conflict between Facebook's rules and IHRL can be understood by reference to methodologies in fundamental rights adjudication. While private platforms prohibit a great deal of speech that the First Amendment protects, they have used categorization, a methodology familiar to the First Amendment[34] that contrasts proportionality.

Methodologically, modern First Amendment jurisprudence generally uses categorization: classifications of speech determine degrees of protection with corresponding rules.[35] For speech rights, non-American regimes by and large employ proportionality, which balances interests in a structured inquiry.[36] Categorization and proportionality both balance

---

[29] *Id.* § 10.

[30] *Id.*

[31] *Case on a Comment Related to the January 2021 Protests in Russia*, META (Jan. 19, 2022) (emphases added), https://transparency.fb.com/oversight/oversight-board-cases/comment-related-to-january-2021-protests-in-russia [https://perma.cc/3LJN-9QAW].

[32] *See, e.g.*, Kaye, *supra* note 6, ¶ 47; Emma J. Llansó, *No Amount of "AI" in Content Moderation Will Solve Filtering's Prior-Restraint Problem*, BIG DATA & SOC'Y, Jan.–June 2020, at 1, 4.

[33] evelyn douek (@evelyndouek), TWITTER (May 26, 2021, 1:00 PM), https://twitter.com/evelyndouek/status/1397598502776152064 [https://perma.cc/KAE4-RMVS].

[34] evelyn douek, *Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability*, 121 COLUM. L. REV. 759, 770–76 (2021).

[35] Frederick Schauer, *The Exceptional First Amendment*, *in* AMERICAN EXCEPTIONALISM AND HUMAN RIGHTS 29, 53–54 (Michael Ignatieff ed., 2009).

[36] *Id.*; Adrienne Stone, *The Comparative Constitutional Law of Freedom of Expression*, *in* COMPARATIVE CONSTITUTIONAL LAW 406, 410 (Tom Ginsburg & Rosalind Dixon eds., 2011).

interests, but the key distinction is that categorization transforms underlying interests into rules for *all* cases, while proportionality weighs interests within *individual* cases.[37] In legal theory, this distinction aligns with the distinction between "rules" and "standards."[38]

The *Cowardly Bot Case* showcased this methodological divide. The dispute implicated both speech and the harm of bullying and harassment. Facebook used categorization, stating that the "*balancing* [of competing interests] is undertaken when the Community Standards are drafted."[39] By contrast, the Board's proportionality analysis balanced speech interests and the harm of language *within* the case.

The Bullying and Harassment policy helps to illustrate categorical rulemaking based on interests. Because adjudicators are imperfect at case-by-case balancing, First Amendment rules contain risk preferences that favor speech over other interests.[40] Facebook's Bullying and Harassment rules also contain risk preferences, but not always in favor of speech. Years ago, Facebook concluded that moderators could not detect bullying on content alone due to the behavior's inherently personal nature; thus, rules would result in either over- or underregulation.[41] Influenced by popular demands, Facebook crafted risk-averse rules that generally require self-reporting from targets but are deferential.[42]

Like First Amendment doctrine, the Bullying and Harassment policy now contains subcategorical distinctions to account for additional interests. For instance, the rules give public figures less protection than private individuals "to allow discussion, which often includes critical commentary of [public figures]."[43] The distinction and rationale track *New York Times Co. v. Sullivan*[44] and its progeny, which provide public figures less protection from libel to allow "debate on public issues" that can include "sharp attacks."[45]

---

[37] *See* BARAK, *supra* note 3, at 509; Iryna Ponomarenko, *The Unbearable Lightness of Balancing: Towards a Theoretical Framework for the Doctrinal Complexity in Proportionality Analysis in Constitutional Adjudication*, 49 U.B.C. L. REV. 1103, 1129–31 (2016); *see also* John Hart Ely, Comment, *Flag Desecration: A Case Study in the Roles of Categorization and Balancing in First Amendment Analysis*, 88 HARV. L. REV. 1482, 1493 n.44 (1975).

[38] Ponomarenko, *supra* note 37, at 1122; Stefan Sottiaux & Gerhard van der Schyff, *Methods of International Human Rights Adjudication: Towards a More Structured Decision-Making Process for the European Court of Human Rights*, 31 HASTINGS INT'L & COMPAR. L. REV. 115, 118 (2008).

[39] *Cowardly Bot Case*, *supra* note 2, § 8.1 (emphasis added).

[40] ADRIAN VERMEULE, THE CONSTITUTION OF RISK 41–42 (2014).

[41] Thomas E. Kadri & Kate Klonick, *Facebook v.* Sullivan*: Public Figures and Newsworthiness in Online Speech*, 93 S. CAL. L. REV. 37, 60 (2019).

[42] *Id.*

[43] *Bullying and Harassment*, META, https://transparency.fb.com/policies/community-standards/bullying-harassment [https://perma.cc/EL32-D5NV].

[44] 376 U.S. 254 (1964).

[45] *Id.* at 270; *see* Kadri & Klonick, *supra* note 41, at 60–61.

Next, this comment argues that — while neither proportionality nor categorization is intrinsically superior[46] — categorization better suits major platforms for several reasons. First, on major platforms, categorical rules can produce more accurate decisions. Second, platforms benefit from tailoring rules for different categories of content. Third, regimes tend to produce rules from repeated adjudications, and platforms adjudicate at unprecedented volumes.

First, categorical rules can produce more accurate decisions on major platforms. Both categorization and proportionality have strengths and weaknesses for decisionmaking, but a regime's features affect the suitability of a methodology. In this respect, the circumstances of a major platform, like Facebook, bear hallmarks of where well-designed categorical rules can produce superior outcomes.

Proportionality's benefits include structured and transparent judicial inquiry as well as flexibility to respond to the interests in individual cases and unforeseen circumstances.[47] Proportionality also preserves notions of substantive justice, allowing all features of a case to be considered.[48] However, proportionality is criticized for giving adjudicators too much discretion.[49] Studies find that political ideology influences judges' case-by-case balancing, and human rights courts adjudicate speech cases with great inconsistency.[50]

Categorization cannot perfectly mediate interests but reduces disadvantages of case-by-case balancing. Decisionmakers are error prone, and categorical rules constrain how decisions can be made.[51] Thus, a rule can be designed such that it functions suboptimally within *individual* cases yet produces more accurate results across *all* cases.[52] In other words, rules "accept the benefits of comparative closeness of getting it right in exchange for the aspirations of getting it right all the time."[53]

A regime's features affect the suitability of the methodologies. Rules are particularly appropriate for nonjudicial officials who lack judges'

---

[46] *See* BARAK, *supra* note 3, at 526; Vicki C. Jackson, *Constitutional Law in an Age of Proportionality*, 124 YALE L.J. 3094, 3193–94 (2015).

[47] Stone, *supra* note 36, at 410.

[48] Sottiaux & van der Schyff, *supra* note 38, at 121.

[49] *See* BARAK, *supra* note 3, at 487.

[50] *See* Jacob Mchangama & Natalie Alkiviadou, *Hate Speech and the European Court of Human Rights: Whatever Happened to the Right to Offend, Shock or Disturb?*, 21 HUM. RTS. L. REV. 1008, 1010 (2021); Raanan Sulitzeanu-Kenan et al., *Facts, Preferences, and Doctrine: An Empirical Analysis of Proportionality Judgment*, 50 LAW & SOC'Y REV. 348, 362, 376 (2016).

[51] LAURENCE H. TRIBE, AMERICAN CONSTITUTIONAL LAW 794 (2d ed. 1988).

[52] Mark V. Tushnet, *The Hardest Question in Constitutional Law*, 81 MINN. L. REV. 1, 14–18 (1996); *see also* Jackson, *supra* note 46, at 3167; Frederick Schauer, *The Second-Best First Amendment*, 31 WM. & MARY L. REV. 1, 16–17 (1989).

[53] Schauer, *supra* note 52, at 17.

training in decisionmaking and deliberative environments.[54] Legal decisionmaking by actors without judicial experience is also particularly susceptible to bias,[55] and speech cases are prone to biases affecting case-by-case balancing, as regulated messages are often divisive.[56] In addition, inconsistencies caused by not using rules are exacerbated in the United States, "a large country[] with highly decentralized opportunities for judicial review" of constitutional claims, in contrast to nations like Germany with specialized constitutional courts.[57]

Given these considerations, well-designed categorical rules stand to produce more accurate results on major platforms, given their scale and decentralization, moderators' training, and focus on speech. In 2021, Facebook took action on over 585 million pieces of content.[58] For the task, Facebook has enlisted over 15,000 moderators globally, with great reliance on outsourcing to adjust staffing quickly if tumultuous regional events occur.[59] Content moderators lack judicial training and deliberative environments,[60] and moderation principally concerns speech, which is especially challenging for case-by-case balancing. While a given rule might benefit from revision, major platforms have hallmarks of where categorization can generate superior results.

A second reason that categorization is preferable for major platforms is that it permits tailoring rules per class of content, unlike how a proportionality test governs "all speech restrictions" under IHRL.[61] This affordance helps platforms manage their high caseloads and build rules with desirable risk preferences for the context of social media.

Categorization helps adjudicators with high caseloads manage resources by allowing methods to correspond with issues' complexities.[62] Hate speech is a complex issue in free speech theory,[63] and many argue that content moderation of hate speech requires contextual evaluations by humans.[64] This aim can be achieved in the design of categorical

---

[54] FREDERICK SCHAUER, PLAYING BY THE RULES 150–51 (Tony Honoré & Joseph Raz eds., 1991); Jackson, *supra* note 46, at 3155.

[55] *See* Dan M. Kahan et al., *"Ideology" or "Situation Sense"? An Experimental Investigation of Motivated Reasoning and Professional Judgment*, 164 U. PA. L. REV. 349, 410–11 (2016).

[56] *See* Ely, *supra* note 37, at 1501.

[57] Jackson, *supra* note 46, at 3167; *see id.* at 3110 n.75.

[58] *See Community Standards Enforcement Report*, META, https://transparency.fb.com/data/community-standards-enforcement [https://perma.cc/NT23-TDRM] (click "Download (CSV)"). This figure excludes spam and fake accounts.

[59] PAUL M. BARRETT, WHO MODERATES THE SOCIAL MEDIA GIANTS? 3–4 (2020).

[60] *Id.*

[61] Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26, 58 (2018); *see also* Llansó, *supra* note 32, at 2.

[62] Sottiaux & van der Schyff, *supra* note 38, at 124–25.

[63] *See* Frederick Schauer, *Freedom of Expression Adjudication in Europe and the United States: A Case Study in Comparative Constitutional Architecture*, *in* EUROPEAN AND US CONSTITUTIONALISM 49, 60 (Georg Nolte ed., 2005).

[64] *See* Kaye, *supra* note 6, ¶ 50; douek, *supra* note 34, at 793–94.

methods. After all, First Amendment tests can be highly contextual, considering factors like whether imminent harm is likely.[65] Meanwhile, platforms might conclude that the justifications for rights in free speech theory, such as "self-government" and "truth,"[66] do not support protecting spam, and thus that spam can be deleted upon identification — perhaps comparable to how the First Amendment categorically does not protect false commercial speech.[67] As a result, categorization offers platforms a principled justification for differing analyses based on issues' complexities to manage their unprecedentedly high caseloads.

Categorization enabling rules to be tailored per class of content also allows for rules to contain risk preferences, unlike IHRL's proportionality-based approach.[68] Risk preferences may be desirable for issues where moderators cannot access relevant information. Moderators often cannot detect bullying and harassment on the basis of content alone due to the behavior's inherently personal nature; thus, Facebook heavily relies on self-reporting by users while addressing dozens of millions of reports per year.[69] As Facebook noted after the *Cowardly Bot Case*, the Oversight Board's recommendation would have *weakened* the rules' firm and prompt enforcement. However, people generally want bullying and harassment moderated *even more* strictly than it is currently.[70] While the Board "extended" an "approach" from the hate speech context to bullying and harassment in the IHRL analysis, platforms have sound reasons to take categorically different approaches to the issues.

A third reason why categorization is preferable for major platforms is that repeated adjudication in an area tends to produce rules. This process of legal development helps to explain why categorization is a natural fit for major platforms, since they adjudicate at unprecedented volumes. In fact, the trajectory of content moderation has aligned remarkably well with well-known patterns of common law development.

In common law systems, rules emerge from applying balancing tests over time, as the cumulative results of a test demonstrate what the test requires.[71] When a fact pattern consistently yields the same outcome, regimes are incentivized to establish a rule, as a rule governs effectively,

---

[65] *See* Brandenburg v. Ohio, 395 U.S. 444, 447 (1969).

[66] *See* Stone, *supra* note 36, at 413–14.

[67] *See* Cent. Hudson Gas & Elec. Corp. v. Pub. Serv. Comm'n, 447 U.S. 557, 566 (1980).

[68] *See* douek, *supra* note 1, at 70; *see also* Llansó, *supra* note 32, at 4.

[69] Kardi & Klonick, *supra* note 41, at 60; *Bullying and Harassment*, *in* COMMUNITY STANDARDS ENFORCEMENT REPORT Q2 2021, META (2021), https://transparency.fb.com/data/community-standards-enforcement/bullying-and-harassment/facebook [https://perma.cc/3XZZ-M5AZ].

[70] *See* Emily A. Vogels, *The State of Online Harassment*, PEW RSCH. CTR. (Jan. 13, 2021), https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment [https://perma.cc/5X6T-MZUD].

[71] Matthew Tokson, *Blank Slates*, 59 B.C. L. REV. 591, 608, 652 (2018); *see* Michael Coenen, *Rules Against Rulification*, 124 YALE L.J. 644, 655 (2014); Mark D. Rosen, *Modeling Constitutional Doctrine*, 49 ST. LOUIS U. L.J. 691, 696 (2005).

provides benefits like consistency, and reduces costs of case-by-case balancing.[72] Scholars note that "rules" stand to "emerge even from case-by-case" proportionality due to the "draw of consistency."[73]

While First Amendment jurisprudence is over a century old and has shifted from balancing to categorization over time, speech adjudication in regimes employing proportionality is, at most, around four decades old.[74] Professor Frederick Schauer suggests that patterns of common law development can explain the methodological division, arguing that non-American regimes are likely to gravitate toward categorization over time, as encountering more *varieties* of speech at higher *volumes* may lead regimes to formalize patterns in decisionmaking.[75] Today, Schauer's hypothesis finds some support in the European Court of Human Rights's use of some categorical methods.[76] Still, a broad fruition should not be taken for granted, as many nations greatly value proportionality itself.[77] In any event, the process of repeated adjudications producing rules applies more straightforwardly to private platforms — specialty adjudicators of speech that encounter all *varieties* of online transmissions at *volumes* exponentially surpassing all nations combined.

Remarkably, the trajectory and landscape of content moderation have tracked the patterns of common law systems. Like early common law systems, the now-major platforms, including Facebook, began with case-by-case flexible approaches, and such approaches are still employed by smaller platforms.[78] The flexible method has tradeoffs between personalization and consistency,[79] as is true for adjudication with case-by-case balancing. Platforms have analogized the flexible approaches to a "common-law system" and to "grounded theory," a social-scientific methodology in which "*individual* cases" are "inductively buil[t] up [into] *categories*."[80] Following patterns of legal development, for the now-major platforms, the era of using case-by-case flexibility was a "period of experimentation" from which rules "develop[ed]."[81] Today, major platforms employ rules, believing that "[e]nsuring fair and consistent decisions often means breaking complex philosophical *ideals* . . . into small

---

[72] *See* Tokson, *supra* note 71, at 652.

[73] Jackson, *supra* note 46, at 3167 & n.343.

[74] *See* Schauer, *supra* note 63, at 58–59; Stone, *supra* note 36, at 410.

[75] Schauer, *supra* note 63, at 57–61; *see also* Schauer, *supra* note 35, at 53–56.

[76] *See* Alessio Sardo, *Categories, Balancing, and Fake News: The Jurisprudence of the European Court of Human Rights*, 33 CANADIAN J.L. & JURIS. 435, 443–44 (2020).

[77] Stone, *supra* note 36, at 411.

[78] ROBYN CAPLAN, CONTENT OR CONTEXT MODERATION? 17–19 (2018); *see also* Klonick, *supra* note 4, at 2435–36.

[79] CAPLAN, *supra* note 78, at 18–19.

[80] *Id.* at 18 (first emphasis omitted) (second and third emphases added).

[81] *Id.* at 23; *see id.* at 19.

components that are more likely to be interpretable"[82] — tracking how categorization transforms underlying *interests* into *rules*.

Major platforms' unprecedented caseloads may have propelled the development beyond that of any nation. As a Facebook content policy manager explained, the scale "robbed anyone . . . of the illusion that there was any such thing as a unique case. . . . On any sufficiently large social network everything you could possibly imagine happens every week."[83] Accordingly, the subcategorical distinctions of content policies have grown far more intricate than First Amendment doctrine.[84] In the *Cowardly Bot Case*, the Oversight Board's rhetoric belittled the nature of Facebook's rules.[85] However, theory on common law development supports a plausible view that, methodologically, Facebook's rules comprise the world's most mature regime for reasons that the Board's IHRL-oriented approach has yet to experience. Even without jumping to that conclusion, the trajectory and landscape of content moderation's alignment with well-known patterns of common law development supports that major platforms' use of categorical rules is natural.

In conclusion, Facebook's categorical rules clashed with IHRL's proportionality-based approach in the *Cowardly Bot Case*, but under the circumstances, a categorical approach is superior. Still, Facebook's policies can use reform on various issues, including bullying and harassment. If Facebook continues to make a sound decision not to follow established IHRL standards on speech regulation, the company should consider clarifying its intentions in its corporate policy statement on the UNGPs. Doing so could enable more constructive dialogue with the Oversight Board, which cost Facebook $130 million in initial funding.[86] For institutional design, it is profoundly unproductive to reform categorical rules that are designed in light of *millions* of cases by using proportionality to scrutinize *individual* cases.[87] Even the scholar who called the *Cowardly Bot Case* "an easy case" followed up by eerily noting that giving less deference to users self-reporting bullying and harassment was "the opposite of what a lot of people have been [requesting]."[88]

---

[82] *Id.* at 23–24 (emphases added).

[83] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET 77 (2018).

[84] *See* douek, *supra* note 34, at 782–83.

[85] *See, e.g.*, *Cowardly Bot Case*, *supra* note 2, § 8.1 ("[T]he case illustrates that Facebook's blunt and decontextualized approach can disproportionately restrict freedom of expression.").

[86] Klonick, *supra* note 4, at 2467.

[87] *Cf.* Coenen, *supra* note 71, at 644 (explaining that, unlike a natural process of common law development, making a rule more like a standard incidentally allows for outcomes that the rule intentionally prevented).

[88] evelyn douek (@evelyndouek), TWITTER (May 26, 2021, 1:04 PM), https://twitter.com/evelyndouek/status/1397599649393938432 [https://perma.cc/Y44Z-QC3K].