Israel Law Review (2026), **59**, 96–133
doi:10.1017/S0021223726100156

CAMBRIDGE UNIVERSITY PRESS

ARTICLES

Bending the Rules: On Large Language Models and Content Moderation

Renana Keydar[1], Noa Mor[2], Yuval Shany[2] and Omri Abend[3]

[1] Hebrew University of Jerusalem Faculty of Law and Center for Digital Humanities, Israel; [2] Hebrew University of Jerusalem Faculty of Law, Israel and [3] School of Computer Sciences and Engineering and Department of Cognitive Sciences, Hebrew University of Jerusalem Faculty of Science, Israel

Corresponding author: Yuval Shany; Email: yshany@mscc.huji.ac.il

Abstract

This article examines the transformative impact of large language models (LLMs) on online content moderation, revealing a critical gap between platforms' rule-based policies and their AI-driven enforcement mechanisms. Using Facebook's hate speech moderation policies and practices as a case study, we identify a paradox: while content policies are increasingly rule-oriented, AI-driven enforcement seems to operate in a standard-like manner. This disconnect creates transparency, consistency and accountability challenges relating to the delineation of online freedom of expression that are not addressed in the literature, and require attention and mitigation. In this specific context, we introduce the concept of 'rules by the millions' to describe how AI systems actually operate through generating vast networks of micro-rules that evade traditional regulatory oversight. This phenomenon disrupts the conventional rules-versus-standards framework used in legal theory, raising urgent questions about the adequacy of current AI governance mechanisms. Indeed, the rapid adoption of LLMs in content moderation has outpaced the human capacity to monitor them, creating a pressing need for adaptive frameworks capable of managing the evolving capacities of AI.

Keywords: content moderation; artificial intelligence; large language models; rules v. standards; online hate speech

1. Introduction

The content moderation policies and enforcement practices of the major online platforms represent a critical arena for shaping public discourse and balancing freedom of expression against competing values and interests.[1] Yet, the landscape of content moderation is undergoing a major transformation. In January 2025, Meta's CEO, Mark Zuckerberg, announced a fundamental shift in how digital platforms govern

[1] Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2017) 51 *UC Davis Law Review* 1149; Barrie Sander, 'Freedom of Expression in the Age

© The Author(s), 2026. Published by Cambridge University Press in association with the Faculty of Law, the Hebrew University of Jerusalem. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

online speech,[2] pivoting away from centralised enforcement to a largely decentralised, user-driven model of flagging violations and discontinuing its fact-checking program, replacing it with 'Community Notes'. In adopting this reform, Meta has followed the trend of crowd-sourced moderation seen on platforms like X (formerly Twitter), which relies on community-based contextualisation to address content moderation challenges.[3] This is a paradigm shift for Meta, which for years invested heavily in building a highly centralised enforcement infrastructure, deploying vast teams of human moderators and fact checkers that proactively curb harmful content, and elaborating precise policy frameworks to monitor online discourse.[4] The move comes at a time in which platforms like X, YouTube and Facebook are increasingly facing scrutiny for their content moderation policies and practices. The 2025 announcement by Zuckerberg also comes after several years in which the exponential growth of online content and the proliferation of harmful materials have prompted platforms to adopt advanced AI technologies for content moderation purposes.[5]

Albeit slightly relaxed, as a result of the recent changes Meta has adopted, AI technologies still play an important role in the company's content moderation processes

of Online Platforms: The Promise and Pitfalls of a Human Rights-based Approach to Content Moderation' (2019) 43 *Fordham International Law Journal* 939; Evelyn Douek, 'Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability' (2021) 121 *Columbia Law Review* 759; Anastasia Kozyreva and others, 'Resolving Content Moderation Dilemmas between Free Speech and Harmful Misinformation' (2023) 120 *Proceedings of the National Academy of the Sciences of the United States of America* e2210666120, https://www.pnas.org/doi/10.1073/pnas.2210666120. For a view on the development of content moderation enforcement practices, see generally Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (Yale University Press 2018); Tarleton Gillespie, 'Content Moderation, AI, and the Question of Scale' (2020) 7(2) *Big Data & Society*; Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 *Big Data & Society*.

[2] Joel Kaplan, 'More Speech and Fewer Mistakes', *Meta Newsroom*, 7 January 2025, https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes. As part of the changes announced by Meta, the company is also to relax its proactive, automated moderation of certain content categories, including, so it seems, hate speech. Instead, the company will rely on user reports to identify possible violations. While this will influence how AI is incorporated into the content moderation landscape, we believe the centrality of AI in this context will remain. First, AI, including LLMs, will still be ubiquitous in Meta's content moderation practices. Considering the scale of future users' reports, AI will probably be used to make automated decisions following such reports. Moreover, as mentioned in Meta's announcement, AI will be used in additional content moderation decisions, including reviewing users' appeals on removals of their content. Secondly, the reduced reliance on proactive AI removal will not apply to all content fields: '[I]llegal and high-severity violations, like terrorism, child sexual exploitation, drugs, fraud and scams', will still be subject to proactive AI-based practices. Thirdly, the global application of Meta's recent change is doubtful, considering the regulatory frameworks that apply to content moderation in different territories, and particularly Europe: Robert Booth, Dan Milmo and Jennifer Rankin, 'Meta's Changes to Policing Will Lead to Clash with EU and UK, Say Experts', *The Guardian*, 8 January 2025, https://www.theguardian.com/technology/2025/jan/08/metas-changes-to-social-media-policing-will-lead-to-clash-with-eu-and-uk-say-experts.

[3] X Corp, 'About Community Notes on X' (X Help Center), https://help.x.com/en/using-x/community-notes.

[4] See below Sections 2 and 3.

[5] See below Section 3.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

in general,[6] and regarding hate speech in particular.[7] This article will examine broad governance questions regarding resort to these automated technologies, and will not explore the specific details of their reduced use by Meta, the new reliance on 'Community Notes' and user reports, or the implications (including the social, political and human rights consequences) these developments entail.

The turn to AI in content moderation by Meta and its counterparts has been driven by multiple factors, including the exponential spread of hate speech towards different audiences, a public outcry over the working conditions of human content moderators,[8] the need for cost-efficiency in running a large-scale content moderation operation,[9] and growing demands from companies to curb dangerous health and political online misinformation, while preserving freedom of expression.[10]

The onset of the COVID-19 pandemic further expedited the shift towards automation. The disruption in workplace attendance caused by the pandemic, coupled with the rampant spread of misinformation,[11] strained the capacities of human content moderators.[12] Amid these growing pressures, AI-driven models, particularly large language models (LLMs), have emerged as a linchpin in automated moderation systems.[13]

[6] Kaplan (n 2) ('we're going to continue to focus these systems on tackling illegal and high-severity violations, like terrorism, child sexual exploitation, drugs, fraud and scams').

[7] Meta, 'Community Standards Enforcement Report: Hateful Conduct', Meta Transparency Center, under 'Proactive Rate', https://transparency.meta.com/reports/community-standards-enforcement/hateful-conduct (pointing out that 87.5% of the violating content that Facebook actioned for hateful conduct from April to June 2025 was proactively found and acted upon by the company rather than by users. Given the high-volume of such content, this figure indicates considerable reliance on automated tools.

[8] The use of human moderation has been strongly criticised for the exploitative labour conditions and adverse impact on moderators' mental health: Casey Newton, 'Facebook Will Pay $52 Million in Settlement with Moderators Who Developed PTSD on the Job', *The Verge*, 12 May 2020, https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health.

[9] Mark Scott and Laura Kayali, 'What Happened when Humans Stopped Managing Social Media Content', *POLITICO*, 21 October 2020, https://www.politico.eu/article/facebook-content-moderation-automation.

[10] Guy Rosen, 'Moving Past the Finger Pointing', *Meta Newsroom*, 17 July 2021, https://about.fb.com/news/2021/07/support-for-covid-19-vaccines-is-high-on-facebook-and-growing; Thierry Breton, 'DSA: Fighting Disinformation', European Parliament, 18 October 2023, https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_23_5126; Michael Landon-Murray, Edin Mujkic and Brian Nussbaum, 'Disinformation in Contemporary U.S. Foreign Policy: Impacts and Ethics in an Era of Fake News, Social Media, and Artificial Intelligence' (2019) 21 *Public Integrity* 512.

[11] In February 2020, facing the surge in global cases of the COVID-19 virus, the Director-General of the World Health Organization (WHO) said: 'We're not just fighting an epidemic; we're fighting an infodemic': World Health Organization, 'WHO Publishes Public Health Research Agenda for Managing Infodemics', Departmental update, 2 February 2021, https://www.who.int/news/item/02-02-2021-who-public-health-research-agenda-for-managing-infodemics.

[12] Scott and Kayali (n 9).

[13] Deepak Kumar, Yousef AbuHashem and Zakir Durumeric, 'Watch Your Language: Large Language Models and Content Moderation', Cornell University, 25 September 2023, https://arxiv.org/abs/2309.14517; Mirko Franco, Ombretta Gaggi and Claudio E Palazzi, 'Analyzing the Use of Large Language Models for Content Moderation with ChatGPT Examples', 4 September 2023, https://dl.acm.org/doi/10.1145/3599696.3612895; Ke-Li Chiu, Annie Collins and Rohan Alexander, 'Detecting Hate Speech with GPT-3', Cornell University, 24 March 2022, https://arxiv.org/abs/2103.12407; Lilian Weng, Vik Goel and Andrea Vallone, 'Using GPT-4 for Content Moderation', OpenAI, 15 August 2023, https://openai.com/index/using-gpt-4-for-content-moderation.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

Reliance on AI systems in content moderation, however, introduces critical questions about the ongoing tension between rule-oriented content policies and their increasingly standard-like enforcement mechanisms, which in turn raises concerns about the accuracy and transparency of automated content moderation. On the one hand, platforms continue to rely on clear, prescriptive rules to articulate what constitutes acceptable speech. On the other, the integration of AI in content moderation systems marks a shift towards flexible and unpredictable enforcement – a hallmark of standards-based approaches.[14] We maintain below that the challenges inherent in the tension between clear rules and less-than-clear enforcement provide an important context for evaluating Meta's online content moderation policies and practices, including its most recent change in policy. In addition, we stipulate that the tension is unlikely to be resolved for as long as generative AI-based LLMs (such as Facebook's Llama, OpenAI's GPT and Anthropic's Claude) continue to play a central role in the enforcement of the platforms' content moderation policies, without adequate policy, regulatory and oversight adjustments.[15]

Furthermore, we argue in the article that use of LLMs drives platforms to apply to distinct cases vast numbers of micro-rules – which we term 'rules by the millions'– which adapt to specific contexts at scale.[16] While this approach offers unprecedented efficiency and capacity in the realm of decision making – exceeding by far the data-collecting and processing capabilities of human decision makers – it also blurs the traditional boundaries between rules and standards, creating a paradoxical coexistence of the two that raises questions about transparency, fairness and regulatory oversight with regard to content moderation decisions and, more broadly, online freedom of expression.

Meta's new shift towards 'more speech, fewer mistakes', in fact, amplifies these tensions. By relaxing restrictions on controversial topics, switching back to user-flagging of violations rather than proactive moderation, and empowering users to provide contextual information in the 'Community Notes' model,[17] Meta's approach decentralises the content moderation process while introducing additional unpredictability into the policy enforcement outcomes generated through reliance on AI systems.

This article examines these developments through the lens of Facebook's (now Meta's)[18] evolving hate speech policies, tracing the company's evolution from a predominantly standard-based governance model to a rule-centric

[14] See Sections 3 and 4.

[15] See Section 4. See also 'Using LLMs to Moderate Content: Are They Ready for Commercial Use?', *Tech Policy Press*, 3 April 2024, https://www.techpolicy.press/using-llms-to-moderate-content-are-they-ready-for-commercial-use.

[16] See Section 4.

[17] Kaplan (n 2).

[18] In October 2021, Facebook, Inc. changed its name to 'Meta Platforms, Inc.'. The company's social media platform, however, is still called 'Facebook': Meta, 'Introducing Meta: A Social Technology Company', *Meta Newsroom*, 28 October 2021, https://about.fb.com/news/2021/10/facebook-company-is-now-meta/#:~:text=Today%20at%20Connect%202021,%20CEO,find%20communities%20and%20grow%20businesses. We will mostly use the name 'Facebook' in this article.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

governance model, and then to one increasingly reliant on AI-based enforcement mechanisms.[19]

We argue that the juxtaposition of rule-oriented content moderation policies directed towards users and standard-like enforcement deployed by algorithms creates significant risks and concerns. These include challenges related to algorithmic biases, transparency and accountability, as well as meeting user expectations and legitimacy concerns, which invite urgent legal and policy attention.

This article proceeds as follows: In Section 2 we explore how Facebook's content moderation policies on hate speech have evolved over the years, shifting towards more rule-based structure, form and content, and whether recent changes in policy reverse the general trend towards increased rule specification. In addition, we include in this section a theoretical discussion, exploring the distinction between rules and standards in contemporary legal theory, with a focus on the balance of opportunities and risks associated with each approach. Section 3 discusses content moderation enforcement, highlighting the development of new AI technologies, in particular LLMs, and their alignment with standard-like enforcement approaches in the context of managing hate speech. We also identify and conceptualise the tension between the rules-versus-standards narrative and the resort by LLMs to 'rules by the millions'. Section 4 explores the ramifications of the disparity we have identified between the manner in which the content policies are articulated and their AI-based enforcement. The article concludes with a few modest policy recommendations aimed at fostering a more fair, consistent and informed content moderation governance landscape.

2. Content moderation policies and the rules–standards continuum

This section will focus on Facebook's move from reflecting standard-like content moderation policies to those of a more rule-oriented nature. It will also include a short discussion on the relevant theoretical legal framework that is applicable to the distinction between rules and standards.

2.1. *The rules-versus-standards continuum: A theoretical discussion*

The rules-versus-standards debate has received much attention in modern legal philosophy, and is still the subject of a considerable body of legal literature.[20] The debate surrounds the conduct-guiding features of the law which, in certain application areas, can be conceptualised as a series of *If-Then* conditional directives, comprising a 'trigger' ('If') and a required response ('Then').[21] The trigger identifies a given

[19] Meta, 'Hateful Conduct', https://transparency.meta.com/policies/community-standards/hateful-conduct. We will not be exploring content policies of other platforms owned by Meta, such as Instagram's Community Guidelines. It should also be noted that, according to Meta, '[c]ontent that is considered violating on Facebook is also considered violating on Instagram': Meta, 'Community Standards Enforcement Report', https://transparency.meta.com/reports/community-standards-enforcement.

[20] Pierre Schlag, 'Rules and Standards' (1985) 33 *UCLA Law Review* 379, 379–81.

[21] ibid 381. See also Kevin M Clermont, 'Rules, Standards, and Such' (2020) 68 *Buffalo Law Review* 751, 756. Hart views laws as 'orders backed by threats': HLA Hart, *The Concept of Law* (2nd edn, Clarendon Press 1994) 9, 18, 27, 32, 79. See also Duncan Kennedy, 'Form and Substance in Private Law Adjudication' (1976) 89 *Harvard Law Review* 1685, 1690–91.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

phenomenon, and the response describes the legal consequence that must follow.[22] The distinction between rules and standards underscores the existence of a choice by lawmakers regarding the level of specificity used to articulate the legal directives at hand.[23]

Whereas, in reality, the formulation of legal directives and the process of their interpretation and application tend to feature a mixture of rule-like and standard-like elements, the two categories continue to exist in legal theory as 'ideal types', representing two ends of a continuum of lawmaking choices.[24] Furthermore, it is possible to evaluate, as we do below, across a number of comparators, whether a change introduced in a legal directive has rendered it, as a practical matter, more rule-like or more standard-like.

One key difference between rules and standards (also referred to at times as 'principles')[25] lies in their manner of articulation, or, more specifically, in 'their clarity prior to an incident that invokes the legal system'.[26] A rule is commonly perceived as a directive that 'establish[es] legal boundaries based on the presence or absence of well-specified triggering facts'.[27] A standard, on the other hand:[28]

> indicates the kinds of circumstances that are relevant to a decision on legality and is thus open-ended. That is, it is not a list of all the circumstances that might be relevant but is rather the criterion by which particular circumstances presented in a case are judged to be relevant or not.

This formal difference between rules and standards is often exemplified in the legal literature by speed limit provisions: when trying to deter drivers from driving too fast, one approach may be to create a clear and specific speed limit rule (e.g., 50 mph), and to declare driving beyond that limit unlawful, with specific legal consequences (such as a fine, or suspension of one's driving licence). Another approach is to create a more flexible standard of conduct, by stipulating that driving at an unreasonable speed or reckless manner is unlawful.[29]

However, the formal difference between rules and standards is just the beginning of the story. The choice between the two options (or, more likely, the point

[22] Schlag (n 20) 381.

[23] ibid. Nonetheless, unlike Hart, who did not consider standards to be part of law as a system of rules, Dworkin believed both (rules and standards) to constitute the law: Ronald M Dworkin, 'The Model of Rules' (1967) 35 *University of Chicago Law Review* 14, 29; Michael D Bayles, *Hart's Legal Philosophy: An Examination* (Kluwer Academic 1992) 167–68.

[24] See, for instance, Kathleen M Sullivan, 'The Supreme Court, 1991 Term – Foreword: The Justices of Rules and Standards' (1992) 106 *Harvard Law Review* 22, 57–58.

[25] Melvin A Eisenberg, *Legal Reasoning* (1st edn, Cambridge University Press 2022) 60–63.

[26] Russell B Korobkin, 'Behavioral Analysis and Legal Form: Rules vs. Standards Revisited' (2000) 79 *Oregon Law Review* 23, 25.

[27] ibid.

[28] Isaac Ehrlich and Richard A Posner, 'An Economic Analysis of Legal Rulemaking' (1974) 3 *Journal of Legal Studies* 257, 258. See also Frederick Schauer, 'The Convergence of Rules and Standards' (2003) *New Zealand Law Review* 303, 304–5. cf Joseph Raz, 'Legal Principles and the Limits of Law' (1972) 81 *Yale Law Journal* 823, 838.

[29] Ehrlich and Posner (n 28) 257.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

chosen on the continuum between them)[30] is far from a mere technical matter of legal style. Indeed, it is very much a political choice, with possibly far-reaching societal and institutional ramifications, and power-distribution consequences.[31] Duncan Kennedy noted in this regard that the choice between rules and standards 'cannot be made in a neutral fashion. Each choice affects the balance of economic power, to the advantage of one side and the disadvantage of the other'.[32] As we show below, the choice between rules and standards for regulating online content in the present context implicates the rights and interests of different groups of platform users (e.g., those engaging in online speech – including those engaging in what is potentially prohibited hate speech – and those adversely affected by certain types of online speech) and of the platforms themselves.

(a) *Power and discretion*

One ramification of the choice between rules and standards involves the allocation of power between rule-makers (typically, the legislature) on the one hand, and rule-interpreters (typically, the courts) and rule-appliers (typically, the police and governmental bodies), on the other. Such power-allocation implications arise largely because the choice between rules and standards determines 'whether the law is given content ex ante or ex post'. Rules determine the 'specification of what conduct is permissible' in advance, whereas standards may leave this discretion to adjudicators and enforcers.[33] Russell Korobkin explains that '[r]ules state a determinate legal result that follows from one or more triggering facts', and that '[n]o other circumstances are relevant to the legal consequences' of the act. Standards, by contrast, 'require legal decision makers to apply a background principle or set of principles to a particularized set of facts in order to reach a legal conclusion'.[34] Given their open-ended nature, they do not *necessarily* predetermine a specific result.[35]

In their purest form, rules reflect the substantive choices made by the drafter of a given directive at the time of drafting, with the interpreters and those who apply the law making largely mechanical decisions by applying easily ascertainable facts to clearly formulated instructions.[36] Conversely, standards 'leave most of the important choices to be made by ... the enforcer, or the interpreter, and leave them to be made at the moment of application'.[37]

As the distinction between rules and standards pertains to questions of allocation of power and discretion between various legal and political institutions – legislatures, courts, police, and the like – the choice of the form used is often linked to

[30] ibid. See also Sullivan (n 24) 57–58.

[31] Sullivan (n 24). See also Kennedy (n 21) 1710 ('the choice of form is seldom purely instrumental or tactical').

[32] Kennedy (n 21) 1751.

[33] Louis Kaplow, 'Rules versus Standards: An Economic Analysis' (1992) 42 *Duke Law Journal* 557, 559–60. See also Clermont (n 21) 758, and Schauer (n 28) 305. See also Cass R Sunstein, 'Problems with Rules' (1995) 83 *California Law Review* 953, 961. See also Sullivan (n 24) 57.

[34] Korobkin (n 26) 23.

[35] See Bayles (n 23) 166.

[36] Frederick Schauer, 'The Tyranny of Choice and the Rulification of Standards' (2005) 14 *Journal of Contemporary Legal Issues* 803, 803.

[37] ibid 804. See also Schauer (n 28) 305–9, and Sullivan (n 24) 58–59. See also Kennedy (n 21) 1752.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

weighty questions of separation of powers and checks and balances between the different branches of government.[38]

(b) *Manner of application*

Another aspect of the distinction between rules and standards concerns the legal technique that these two legal models invite into their manner of application. While applying rule-like directives often entails the use of 'categorisation' – the classification of data or cases into predetermined categories – choosing standards normally entails the use of 'balancing' considerations or principles by those who interpret and those who apply the law. Kathleen Sullivan stated that '[c]ategorization is taxonomic. Balancing weighs competing rights or interests'. A rule, she explains, 'defines bright-line boundaries and then classifies fact situations as falling on one side or the other. When categorical formulas operate, the key move in litigation is to characterize the facts to fit them into the preferred category'.[39] Standard-based balancing, on the other hand, 'explicitly considers all relevant factors with an eye to the underlying purposes or background principles or policies at stake'.[40] Among other considered factors, one may find the importance of the value or right that had been infringed, the seriousness of that infringement, and the factors supporting the infringing conduct.[41] This balancing/categorising distinction echoes Dworkin's view that principles have 'weight', and that they can be added up and balanced against one another. This is not the case with rules, which are not assigned a weight and are not balanced by other rules: they either apply or do not.[42] The difference between categorisation and balancing also carries transparency and accountability implications. While the application of rules through categorisation is typically easier to understand and monitor, it also tends to be accompanied by a shorter justification than that associated with standard-based balancing.[43] This is because balancing requires the law-interpreter or applier to discuss the rationales, values and interests they have taken into consideration, while categorising mostly does not necessitate an engagement with the broader notions that underlie the relevant rules.[44]

(c) *Predictability in the application of the law*

The distinction between rules and standards is also significant from a legal certainty and conduct-guidance standpoint. Justice Scalia, in this regard, voiced support for articulating specific rules: 'We can less and less afford protracted uncertainty

[38] Sullivan (n 24) 64–65. See also Justice Scalia's view in Antonin Scalia, 'The Rule of Law as a Law of Rules' (1989) 56 *University of Chicago Law Review* 1175.

[39] Sullivan (n 24) 59.

[40] ibid 60.

[41] ibid 60.

[42] Bayles (n 23) 166–67.

[43] However, the drafting of a rule may include publicly available reasoning regarding, say, the scope and purposes of the rule. Moreover, the drafting stage sometimes involves certain levels of civic participation: Cary Coglianese, Heather Kilmartin and Evan Mendelson, 'Transparency and Public Participation in the Federal Rulemaking Process: Recommendations for the New Administration' (2008) 77 *George Washington Law Review* 924.

[44] Sullivan (n 24) 67.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

regarding what the law may mean. Predictability … is a needful characteristic of any law worthy of the name. There are times when even a bad rule is better than no rule at all'.[45] At the same time, some argue that the uncertainty surrounding standards generates a positive 'chilling effect' that deters people from engaging in undesirable activity which might conceivably be regarded by a law-interpreter or law-applier as covered by it, while the clear boundaries of rules allow people to 'walk the line'.[46]

The different degrees of predictability of rules and standards also imply a difference in cost externalities. Rules may be more costly than standards to articulate, but standards can be more 'costly for individuals to interpret when deciding how to act and for an adjudicator to apply to past conduct'.[47] In addition, some argue that standards may be 'more precise in reaction to the particular case, avoiding the costs of over- and under-inclusion',[48] while others believe that the more 'precise and detailed' the provision – the higher is the probability that the activity will be deemed illegal if it is in fact undesirable (the kind of activity the legislature wanted to prevent) and the lower is the probability that the activity will be deemed illegal if it is desirable. Thus, the expected punishment cost for undesirable activity is increased, and that of desirable activity is reduced.[49]

(d) *Effectiveness, fairness and legality*

Another consequence of the distinction between rules and standards concerns effectiveness in handling specific policy problems. Whereas rules are often helpful in offering clear-cut solutions to specific problems like exceeding the speed limit, standards can offer answers to complicated legal questions that rules cannot effectively address (such as other manifestations of reckless driving), either because the case classification is unclear, or because the rules do not offer a suitable solution for the case at hand.[50] In line with this latter approach, some commend the flexibility that standards afford in providing appropriate legal answers, and their ability to accommodate uncertainty and adapt to new developments that were not foreseen by the legislature when articulating rules.[51]

The choice between rules and standards also involves a choice between competing dimensions of fairness. Those who support rules might argue that they encourage equality and consistency in that they apply the same requirements across all cases; standards, on the other hand, may lead to biased decisions by those who apply the law, who may be influenced by irrelevant considerations such as personal sympathy or antipathy towards the parties in litigation or specific political agendas.[52] Yet, given that rules might be over- or under-inclusive, their application could also raise fairness concerns that can be avoided by invoking standards. Questions of fairness may also pertain to each of the implications of the choice between rules and

[45] Scalia (n 38) 1179.
[46] Kennedy (n 21) 1695–96.
[47] Kaplow (n 33) 557.
[48] Clermont (n 21) 759. See also Kennedy (n 21) 1695.
[49] Ehrlich and Posner (n 28) 262.
[50] Cf. Bayles (n 23) 174.
[51] Sullivan (n 24) 66.
[52] ibid 62.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

standards that we have listed in this section. From the predictability angle, for example, the distinction between rules and standards may involve concerns about human liberty, procedural regularity, people's ability to exert control over their lives, and arbitrariness in the decision-making process.[53]

Finally, the choice between rules or standards also implicates the principles of the rule of law and substantive fairness. Being a central tenet of democracy, the rule of law requires that legal provisions that obligate individuals are known and published in advance,[54] and that they are applied and enforced in a consistent manner.[55] Reliance on rules appears to conform more closely to this rule of law ideal. Still, the crude nature of rules and their tendency to become outdated over time may render the use of standards more compatible with principles of substantive fairness. For example, the same treatment of different cases under the same rule may lead to unjust results.

As we demonstrate below, the formulation of a legal policy governing content moderation invites a choice between rules and standards. This is, in effect, a complex choice between competing values, and different forms of allocation of power, conduct guidance and law enforcement, with significant effectiveness, fairness and legality implications. The specific case study we explore – Facebook's hate speech (now referred to as hateful conduct) policy – shows a gradual shift from standards to rules in the manner in which the policy is articulated (at least, until 2025). Still, we also claim that the increased reliance on LLMs to enforce the applicable content moderation policy introduces into the regulatory mix a new decision-making model – 'rules by the millions'. This new model stands in tension with the traditional rules-versus-standards dichotomy and invites further evaluation of its normative implications.

2.2. *Facebook content moderation policies in the lens of the rules-versus-standards continuum*

When founded in 2004, Facebook had no clear content moderation policy regarding permissible content on its platform. Facing growing challenges as it rapidly evolved, the company assembled a content moderation team in 2009 and tasked it with formulating its content policies.[56] In 2011, the team introduced the 'Community Standards' – the name that the company's content policies bear to this day.[57]

Early versions of the Community Standards were short, rather informal, and reflected big ideas, guidelines and perceptions regarding unacceptable and acceptable content.[58] The current Community Standards are, however, very different.[59] In recent years Facebook has constantly tweaked, updated and revised

[53] Scalia (n 38) 1179. See also Sullivan (n 24) 63–64.

[54] Jeremy Waldron, 'The Rule of Law in Contemporary Liberal Theory' (1989) 2 *Ratio Juris* 79, 84.

[55] ibid 81.

[56] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 *Harvard Law* Review 1598, 1620.

[57] Jillian C York, 'Community Standards: A Comparison of Facebook vs. Google+', 30 June 2011, https://jilliancyork.com/2011/06/30/google-vs-facebook.

[58] Tarleton Gillespie, 'Facebook's Improved 'Community Standards' Still Can't Resolve the Central Paradox', Social Media Collective, 18 March 2015, https://socialmediacollective.org/2015/03/18/facebooks-improved-community-standards-still-cant-resolve-the-central-paradox.

[59] Meta (n 19) 'Hateful Conduct'.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

these policies for various reasons[60] – including public pressure,[61] legal requirements introduced by the countries in which it operates,[62] and recommendations of Meta's Oversight Board.[63] When updating its policies, Facebook often consults with a wide range of stakeholders, which include academics and civil society organisations.[64]

As a result of these changes, the current content moderation policies no longer resemble a general standard-like pronouncement of the types of content that are allowed or not allowed on the platform. Instead, they have become much more detailed, specific and clear – much more rule-oriented.[65]

The recent changes to the company's hate speech/hateful conduct policy announced by Zuckerberg have seemingly sought to reverse this trend. However, as is further discussed in this section, the current policy still largely reflects a rules-based approach, and the general direction of travel through the policy's evolution over the years continues to be rule-oriented.[66]

2.2.1. *Inside the policies*

A close inspection of the current policies, compared to their earlier versions, reveals that, at least up until 2025, the policies have generally become more and more rule-oriented.

(a) *Length*

The first indicator of the movement of the policies towards being more rule-like lies in their changing length. Growth in length often reflects a more detailed form of regulation.[67]

[60] Meta, 'How We Update the Facebook Community Standards', 20 January 2020, https://transparency.meta.com/policies/improving/deciding-to-change-standards. Since 2018, the Community Standards changes are documented in a 'Change Log', which is shown next to the current version of the policies. For the Hate Speech change log, see Meta (n 19) 'Hateful Conduct'.

[61] See, for instance, the adjustments of the Community Standards pertaining to photos depicting breastfeeding and women's organised fight in this regard (Gillespie (2018) (n 1)) and Sarah Myers West, 'Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Platforms' (2017) 5(3) *Media and Communication* 28, 33. For changes in the policies following the backlash over the removal of a post that included the iconic 'Napalm Girl' photo see Lucy Baily, 'Facebook Slammed for Censoring Iconic "Napalm Girl" Photo, Deleting Posts', *NBC NEWS*, 9 September 2016, https://www.nbcnews.com/tech/social-media/facebook-under-fire-censoring-iconic-napalm-girl-photo-n645526, and Joel Kaplan, 'Input from Community and Partners on Our Community Standards', Facebook, 21 October 2016, https://about.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-standards.

[62] See discussion in Section 4.2.

[63] See discussion in Section 4.2.

[64] Monika Bickert, 'Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process', Meta, 24 April 2018, https://about.fb.com/news/2018/04/comprehensive-community-standards; see also Meta (n 60).

[65] For the current Community Standards see Meta, 'Policies', https://transparency.meta.com/policies. It should be noted that the Community Standards do not include all measures of content moderation and, inter alia, reduction of content visibility; in this regard see Gillespie (2018) (n 1).

[66] See below in this section.

[67] Conversely, excessively verbose policies might lead to the opposite result, and be inaccessible for users: David Lazarus, 'Column: Want to Read a Tech Company's User Agreements? Got 90 Minutes

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

As mentioned above, Meta's first version of the Community Standards was introduced in 2011. At that time, the entire section on hate speech was about 50 words long:[68]

> Facebook does not tolerate hate speech. Please grant each other mutual respect when you communicate here. While we encourage the discussion of ideas, institutions, events, and practices, it is a serious violation of our terms to single out individuals based on race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability, or disease.

By May 2018, this section of the policies had grown to about 300 words[69] and, by the end of 2024, had expanded to 1,600 words – over thirty times the scope of the 2011 policies. As part of Meta's recent content moderation change, this figure was reduced to roughly 1,300 words.[70] Nonetheless, the general trend still indicates a significant increase in the length of this policy over the years.

An important milestone in the ever-growing word count of Facebook's policies (and their rule-orientation, as explored below) was an update that occurred in April 2018, when the platform incorporated its detailed internal protocols on content moderation into the Community Standards.[71]

(b) *Breaking down 'big terms' into smaller ones and providing detailed definitions*

Looking into the changes undergone by the content moderation policies, it is evident that 'big terms' were broken into smaller ones, and that detailed definitions were offered for newly created terms. Consider, for instance, the 2011 version of the policies cited above. Although the term 'hate speech' is not formally defined there, in the closing sentence of the paragraph it is said to involve the singling out of individuals 'based on race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability, or disease'.

Conversely, in the current version of the policy, this term (adjusted to 'hateful conduct') is carefully defined as:[72]

> [D]irect attacks against people – rather than concepts or institutions – on the basis of what we call protected characteristics (PCs): race, ethnicity, national

to Spare?', *Los Angeles Times*, 24 August 2024, https://www.latimes.com/business/story/2021-08-24/column-consumer-contracts.

[68] Future Free Speech, 'Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies', 2023, 28, https://futurefreespeech.org/wp-content/uploads/2023/07/Facebook.pdf. See also 'Facebook Community Standards', Facebook Archive, https://web.archive.org/web/20110518113654/https:/www.facebook.com/communitystandards.

[69] Meta (n 19) 'Hateful Conduct', Change Log.

[70] Kaplan (n 2).

[71] Bickert (n 64). See also David Ingram, 'Facebook Releases Long-Secret Rules on How It Polices the Service', *Reuters*, 24 April 2018, https://www.reuters.com/article/us-facebook-abuse-idUSKBN1HV0VR (citing Monika Bickert: 'Now, the company is providing the longer document on its website to clear up confusion and be more open about its operations'). We thank Tarleton Gillespie for this eye-opening insight regarding the 2018 update to the Community Standards.

[72] Meta (n 19) 'Hateful Conduct'.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

> origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease … Additionally, we consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants, and asylum seekers from the most severe attacks … though we do allow commentary on and criticism of immigration policies. Similarly, we provide some protections for non-protected characteristics.

The policy includes additional content that the company will remove, including 'harmful stereotypes' and 'slurs', and both are also given definitions. The term 'harmful stereotypes' is defined by Meta as 'dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence'.[73] Slurs are defined as 'words that inherently create an atmosphere of exclusion and intimidation against people on the basis of a protected characteristic, often because these words are tied to historical discrimination, oppression, and violence'. Other examples are the definitions given to 'targeted cursing' (discussed below) and to '[c]alls or support for exclusion or segregation or statements of intent to exclude or segregate'.[74]

It should be noted that Meta's recent policy change included the deletion of several terms.[75] However, the upshot of the changes we have described in this subsection still indicates a general shift of the policies, over time, from broad terminology – which invites discretion and a standard-like case-by-case application – to more detailed and specific directives that resemble rules.

(c) *Expanding the scope of definitions and providing examples*

Over the years, Facebook has also expanded the scope of some of the definitions in its hate speech policies. In the 2011 version, 'protected characteristics', which were cited in the Community Standards (though not explicitly defined), included the following: 'race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability, or disease'.

The current version defines 'protected characteristics' as also covering caste, along with a clarification that the protection includes age ('when referenced along with another protected characteristic'). The company states that it also protects 'refugees, migrants, immigrants and asylum seekers' from severe attacks, as mentioned above.[76] These newly enumerated 'protected characteristics', or at least some of them, may have been implicit within the scope of the terms of previous versions (such as 'national origin'), but listing them explicitly adds clarity and certainty to the scope of prohibited content.

A further development that reflects a change towards more rule-like policies concerns the inclusion of very detailed examples that are now provided in the policies,

[73] ibid.

[74] ibid.

[75] For instance, the definitions given to 'mental characteristics' and 'moral characteristics' used in generalisations that state inferiority was deleted; see Kaplan (n 2).

[76] This provision is then limited; see discussion regarding the definition of 'hateful speech' above in this section.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

including specific prohibited language. Facebook's current policies relating to 'content targeting a person or group of people' demonstrate this point. The company explains that it is forbidden to post content using the following:[77]

- Dehumanizing speech in the form of comparisons to or generalizations about animals, pathogens, or other sub-human life forms, including:
  - Insects (including but not limited to cockroaches, locusts)
  - Animals in general or specific types of animals that are culturally perceived as inferior (including but not limited to: Black people and apes or ape-like creatures; Jewish people and rats; Muslim people and pigs; Mexican people and worms)
  - Bacteria, viruses, or microbes
  - Subhumanity (including but not limited to savages, devils, monsters).

An additional illustration of using detailed examples of 'hateful speech' can be found in the policy's definition of 'targeted cursing', which the company describes as: 'Targeted use of "fuck" or variations of "fuck" with intent to insult, such as "Fuck the [Protected Characteristic]!"' or 'Terms or phrases calling for engagement in sexual activity, or contact with genitalia, anus, feces or urine, including but not limited to: "suck my dick", "kiss my ass", "eat shit"'.[78]

(d) *Exceptions*

A further indication that content policies are becoming more rule-oriented is their inclusion of specific exceptions – a law-making technique that characterises rules, not standards. Compared with the 2011 version, the current version includes more exclusions of certain types of well-defined content carved out from the scope of the prohibition of hate speech.

One example relates to slurs. After stating that slurs are removed, the policy further reads:[79]

> We recognize that people sometimes share content that includes slurs or someone else's speech in order to condemn the speech or report on it. In other cases, speech, including slurs, that might otherwise violate our standards is used self-referentially or in an empowering way. *We allow this type of speech where the speaker's intention is clear*. Where intention is unclear, we may remove content.

This part then continues as follows, much of which was added in Meta's recent change:[80]

[77] Meta (n 19) 'Hateful Conduct'.
[78] ibid.
[79] Meta (n 19) 'Hateful Conduct' (emphasis added).
[80] ibid (emphasis added).

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

> People sometimes use sex- or gender-exclusive language when discussing access to spaces often limited by sex or gender, such as access to bathrooms, specific schools, specific military, law enforcement, or teaching roles, and health or support groups. Other times, they call for exclusion or use insulting language in the context of discussing political or religious topics, such as when discussing transgender rights, immigration, or homosexuality. Finally, sometimes people curse at a gender in the context of a romantic break-up. *Our policies are designed to allow room for these types of speech*.

Another exception removes certain people from the scope of protected groups under the policy. For instance, the policy states that '[c]ontent targeting a person or group of people ... on the basis of their ... protected characteristic(s)' is prohibited, but exempts 'groups described as having carried out violent or sexual crimes or representing less than half of a group)'.[81]

(e) *Sub-sectioning*

As previously noted, the policies have become much longer over the years. At the same time, they are now organised into subsections that make them easier to navigate. Present policies now include a 'policy rationale', which provides a general formulation of the general directive and relevant definitions. The rationale is followed by two additional sections, the first being a 'Do not post' section, which adds more specific explanations and examples to the policy rationale. The 'Do not post' section is divided into two tiers of severity of content which targets people on the basis of their protected characteristic(s): 'Tier 1' covers 'dehumanizing speech'; 'Tier 2' focuses on 'calls or support for exclusion or segregation', 'insults', and 'targeted cursing'.[82]

The second section that follows the rationale concerns content where additional information or context is required. This section includes, for instance, '[c]ontent attacking concepts, institutions, ideas, practices, or beliefs associated with protected characteristics, which are likely to contribute to imminent physical harm, intimidation or discrimination against the people associated with that protected characteristic'.[83] This last section and the 'Do not post' section are marked and distinguished by two signals: a red sign that may resemble a stop sign and a yellow sign containing an exclamation mark. The use of such graphic design further highlights the accessibility and clarity of these rule-oriented policies.[84]

[81] ibid.

[82] ibid.

[83] ibid.

[84] Interestingly, this corresponds with the Digital Services Act stipulation that '[p]roviders of intermediary services may use graphical elements in their terms of service': Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC [2022] OJ L 277/1, art 45 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2065 (Digital Services Act or DSA).

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

(f) *Beyond the polices*

Another indication that the policies are becoming more rule-like in their orientation is the growing incidence of explicit statements of company representatives regarding its content moderation policies. Addressing the 2015 update to Facebook's Community Standards, for instance, Mark Zuckerberg explained that '[p]eople rightfully want to know what content we will take down, what controversial content we'll leave up, and why'.[85] Monika Bickert, then head of Facebook's Global Policy Management, further explained that '[w]e're not changing anything about the policies ... We're just trying to explain what we do more clearly'.[86]

With regard to one of the 2018 Community Standard updates, the company stated:[87]

> One of the questions we're asked most often is how we decide what's allowed on Facebook ... For years, we've had Community Standards that explain what stays up and what comes down. Today we're going one step further and publishing the internal guidelines we use to enforce those standards ... [T]he guidelines will help people understand where we draw the line on nuanced issues.

Bickert also stated in this regard: 'You should, when you come to Facebook, understand where we draw these lines, and what's OK and what's not OK'.[88]

The recent change in Meta's content moderation policy has been accompanied by a long statement by Joel Kaplan, Meta's new Chief Global Affairs Officer, in which he expressed concern about the over-complexity of the rules and the ensuing propensity of the system to generate 'false positives'. The new policy was designed to render the rules less restrictive and less prone to misapplication:[89]

> Over time, we have developed complex systems to manage content on our platforms, which are increasingly complicated for us to enforce. As a result, we have been over-enforcing our rules, limiting legitimate political debate and censoring too much trivial content and subjecting too many people to frustrating enforcement actions ... We want to undo the mission creep that has made our rules too restrictive and too prone to over-enforcement. We're getting rid of a number of restrictions on topics like immigration, gender identity

[85] Alexei Oreskovic, 'Facebook Clarifies Rules on Banned Content', *Reuters*, 16 March 2015, https://www.reuters.com/article/idUSL3N0WI4U1.

[86] Hayley Tsukayama, 'Facebook Rolls Out More Readable Rules for Its Network', *Washington Post*, 16 March 2015, https://www.washingtonpost.com/news/the-switch/wp/2015/03/16/facebook-rolls-out-more-readable-rules-for-its-network. See also Meta (n 60).

[87] Bickert (n 64). See Bickert's further explanation in this regard in David Ingram, 'Facebook Releases Long-Secret Rules on How It Polices the Service', *Reuters*, 24 April 2018, https://www.reuters.com/article/technology/facebook-releases-long-secret-rules-on-how-it-polices-the-service-idUSKBN1HV0VR. See also Meta (n 60).

[88] Ingram (n 71). For more statements by the company regarding its commitment to create more clarity around the Community Standards see Meta, 'Case on Former President Trump's Suspension from Facebook', 12 June 2023, https://transparency.meta.com/oversight/oversight-board-cases/former-president-trump-suspension-from-facebook.

[89] Kaplan (n 2).

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

> and gender that are the subject of frequent political discourse and debate. It's not right that things can be said on TV or the floor of Congress, but not on our platforms.

Still, even in its new, less expansive configuration, the content moderation policy aims to be rule-oriented. In fact, Kaplan expressed a wish that enforcement action would mirror as closely as possible the applicable rules.

3. AI-based enforcement of content moderation policies

While much has been written on the pros and cons of algorithmic and automated content moderation from a normative-legal perspective,[90] relatively little attention has been given in the legal scholarship to changes in the underlying technology that enables automated content moderation enforcement. However, practices of algorithmic content moderation cannot be understood without accounting for their inextricable connection with developments in the field of AI and the revolutionary changes brought about by LLMs.[91]

The introduction of deep learning models – and, in particular, of LLMs – pushes moderation practices towards standard-like enforcement, changing the division of labour between human and machine as a result of the advanced capacities for learning and understanding content and context, and for reacting to it, featured by these computational models. Moreover, the application of LLMs in moderating content profoundly challenges the balance of opportunities and risks typically associated with the choice between rules and standards discussed in Section 2.[92] We explain below how these developments result in a normative gap between the rule-like nature of the policies and their standard-like enforcement by the new language models. In fact, we maintain that the new enforcement practices possibly introduce a new decision-making paradigm, which we term 'rules by the millions'.

3.1. NLP and the LLM revolution

Natural language processing (NLP) represents a critical subfield of computer science focused on enabling machines to understand, interpret and generate human language.[93] The field emerged at the intersection of linguistics, artificial intelligence and machine learning,[94] aiming to

[90] Gillespie (2018) (n 1)

[91] Zhouxing Shi and others, 'Red Teaming Language Model Detectors with Language Models', Cornell University, 19 October 2023, https://arxiv.org/abs/2305.19713. See also Bertie Vidgen and others, 'Challenges and Frontiers in Abusive Content Detection', Proceedings of the Third Workshop on Abusive Language Online 80, 2019, https://aclanthology.org/W19-3509.

[92] Weng, Goel and Vallone (n 13).

[93] Julia Hirschberg and Christopher D Manning, 'Advances in Natural Language Processing' (2015) 349 *Science* 261; Erik Cambria and Bebo White, 'Jumping NLP Curves: A Review of Natural Language Processing Research' (2014) 9 *IEEE Computational Intelligence Magazine* 48; Eric Brill and Raymond J Mooney, 'An Overview of Empirical Natural Language Processing' (1997) 18(4) *AI Magazine* 13.

[94] Brill and Mooney (n 93). Critics have called attention to the conflation of the buzzword 'AI' with the more concrete applications of machine learning, arguing that frequent news of 'major AI breakthroughs' overly inflates expectations and distracts from the precise way in which machine learning will improve business operations: Eric Siegel, 'The AI Hype Cycle Is Distracting Companies' (2023) *Harvard Business Review* 2.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

bridge the gap between human communication and computer processing capabilities.[95]

In its early stages, NLP relied primarily on rule-based systems and handcrafted linguistic rules.[96] These approaches aimed to explicitly define language patterns and grammatical rules for machines to follow, relying more on logic than statistics. However, these early AI systems faced significant challenges in capturing the intricacies and ambiguities of natural language. The inability of the systems to cope with variations in sentence structure, idiomatic expressions and context-dependent meanings highlighted the limitations of purely rule-based approaches.[97]

A significant transformation occurred with the advent of statistical methods and machine learning in the 1990s and 2000s.[98] Rather than relying on explicit rules, researchers began leveraging large corpora of annotated data and statistical models to uncover patterns in language.[99] This shift marked a fundamental change in approach – instead of trying to analyse languages on the basis of predetermined rules, systems could now learn patterns from data directly.

The field underwent another revolutionary advancement with the emergence of deep learning techniques, powered by unprecedented amounts of data and computing resources.[100] 'Deep learning' refers to training artificial neural networks with multiple layers (hence the term 'deep') to automatically learn hierarchical representations of data. While these models often require large amounts of annotated training data and substantial computational resources, their ability also to learn directly from raw text data, without relying on handcrafted features or linguistic rules, has profoundly transformed the field.[101] In parallel with this development, newer context-sensitive models of word embedding emerged (word embedding is a technique used for representing words in the form of real-valued dense numerical vectors, which effectively capture semantic relationships between words). Contextualised word embedding was the first instance of another transformative idea: pre-training. Pre-training involves exposing the model to large corpora of text (such as books, articles and websites), allowing it to learn grammar, syntax and semantics, along with contextual nuances. Once pre-trained, these models can be fine-tuned in new tasks, such as language translation, sentiment analysis and

[95] Mirjana Kocaleva and others, 'Pattern Recognition and Natural Language Processing: State of the Art' (2016) 5 *TEM Journal* 236.

[96] Daniel Martin Katz and others, 'Natural Language Processing in the Legal Domain', Cornell University, 23 February 2023, 4, https://arxiv.org/abs/2302.12039.

[97] Anton Chernyavskiy, Dmitry Ilvovsky and Preslav Nakov, 'Transformers: "The End of History" for Natural Language Processing?' in Nuria Oliver and others (eds), *Machine Learning and Knowledge Discovery in Databases: Research Track*, European Conference, Bilbao (Spain), 13–17 September 2021, Part 3 ECML PKDD 2021 (Springer 2021) 677.

[98] Katz and others (n 96) 4.

[99] Joseph W Picone, 'Continuous Speech Recognition Using Hidden Markov Models' (1990) 7 *IEEE ASSP Magazine* 26.

[100] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, 'Deep Learning' (2015) 521 *Nature* 436.

[101] Ashish Vaswani and others, 'Attention Is All You Need', Cornell University, 12 June 2017, https://arxiv.org/abs/1706.03762. Tatwadarshi P Nagarhalli, Vinod Vaze and NK Rana, 'Impact of Machine Learning in Natural Language Processing: A Review' in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (IEEE 2021) 1529.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

question answering. While fine-tuning requires human involvement and labelled data, pre-training is largely unsupervised.[102]

The November 2022 release of OpenAI's ChatGPT – a chatbot powered by a generative pre-training transformer (hence the acronym 'GPT')[103] – has drawn scholarly and popular attention to the revolutionary potential of pre-trained language models, and particularly LLMs.[104] Driven by the need for even more powerful and capable models, LLMs (such as GPT by OpenAI,[105] Claude by Anthropic,[106] Gemini by Google[107] or Jamba by AI21[108]) have pushed the boundaries of what was previously possible by significantly increasing the size and capacity of the model and its understanding of language.[109]

3.2. The transformation of content moderation systems: From rules to language models

The seemingly technical choice of model used for enforcing content moderation policies bears significant normative implications for power allocation and the scope of discretion exercised by humans and machines. At first glance, the evolution from simple rule-based systems to sophisticated LLMs appears to represent a progression from clear rules towards vague standard-like 'black boxes' on the traditional legal theory continuum. However, a deeper analysis reveals that LLMs may actually represent something entirely new: a system of 'rules by the millions' that transcends the traditional rules–standards dichotomy.

The evolution of automated systems used for content moderation can be understood in three distinct phases, each representing a different approach to the balance between rules and standards.

[102] Xu Han and others, 'Pre-Trained Models: Past, Present and Future', Cornell University, 14 June 2021, https://arxiv.org/abs/2106.07139.

[103] Alec Radford and others, 'Improving Language Understanding by Generative Pre-Training', OpenAI, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[104] Luciano Floridi, 'AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models' (2023) 36(15) *Philosophy & Technology* 14. The 'largeness' of a model is usually measured by the number of parameters it contains; Google's PaLM is the largest known model, with 540 billion parameters; GPT-4 by OpenAI may be bigger – its specifics are unknown: Stanford Institute for Human-Centered Artificial Intelligence, 'AI Index Report 2023 – Measuring Trends in Artificial Intelligence', Stanford University, April 2023, https://hai.stanford.edu/ai-index/2023-ai-index-report. Generative AI technologies are closely related to LLMs. The latter are a type of generative AI system that has been designed specifically to generate natural language texts.

[105] OpenAI, 'GPT-4 is OpenAI's Most Advanced System, Trained to Provide Safer and More Accurate Responses', https://openai.com/index/gpt-4; Tom B Brown and others, 'Language Models are Few-Shot Learners', Cornell University, 28 May 2020, https://arxiv.org/abs/2005.14165; OpenAI and others, 'GPT-4 Technical Report', Cornell University, 15 March 2023, https://arxiv.org/abs/2303.08774.

[106] Yuntao Bai and others, 'Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback', Cornell University, 12 April 2022, https://arxiv.org/abs/2204.05862.

[107] Sundar Pichai and Demis Hassabis, 'Introducing Gemini: Our Largest and Most Capable AI Model', Google, 6 December 2023, https://blog.google/innovation-and-ai/technology/ai/google-gemini-ai.

[108] 'Introducing Jamba: AI21's Groundbreaking SSM-Transformer Model', A121Labs, 28 March 2024, https://www.ai21.com/blog/announcing-jamba.

[109] Josh A Goldstein and others, 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations', Cornell University, 10 January 2023, https://arxiv.org/abs/2301.04246.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

3.2.1. Phase 1 – Rule-based systems: The era of binary rules

Rule-based automated moderation systems represent the clearest analogue to legal rules on the rules–standards continuum. These systems operate on explicit, predefined rules, such as lists of offensive words or phrases, created by experts. The appeal of these systems lies in their transparency and predictability. Whenever a post was removed, moderators could point to specific rules that triggered the action, making decisions easily explainable to users and platform administrators alike. This transparency helped to guide user behaviour by clearly communicating what would and would not be allowed on the platform.

However, these systems encountered the same limitations that legal scholars have long identified with rigid rules: they proved to be both over- and under-inclusive, unable to adapt to context or handle edge cases, struggling with the inherent complexity of human communication. For example, they might automatically flag or remove a post containing the n-word, regardless of context (such as reclamation of the terms by African Americans), or target the term 'tr*nny', which, while clearly offensive in most contexts, represents a legitimate technical term in automotive discussions.[110] Similarly, these systems often failed to catch deliberate attempts to evade detection, such as when users employed creative misspellings (changing 'f*ck' to 'ph*ck') or used coded language that carried harmful meaning without triggering explicit word filters. Consequently, rule-based enforcement may inadvertently allow malicious users to exploit these limitations and result in many false negatives and positives.

3.2.2. Phase 2 – Classic supervised machine learning models: The hybrid approach

The introduction of supervised machine learning models like Naive Bayes or SVMs represented a first step towards more flexible enforcement, marking a crucial transition in how discretion was allocated in content moderation. In this type of model, humans retained significant discretion through their role in creating training data and defining classification features.

These models process text classification through a synergy of human and machine effort, exercising limited machine discretion. By learning from human-labelled datasets to differentiate between categories like Hate Speech and Non-Hate Speech, these systems began to bridge the gap between strict rules and context-aware standards. Their ability to generalise from labelled data allows them to predict outcomes for unseen text, introducing an element of standard-like flexibility while maintaining rule-like predictability. The machine's 'discretion' is thus limited to pattern matching within these human-defined parameters. For example, while a rule-based system might miss a post containing antisemitic messaging because it did not use any explicitly banned terms, a supervised machine learning model trained on appropriate examples could learn to recognise how certain combinations of words or phrases (like 'global bankers' combined with specific stereotypes) often signal hate speech.

[110] Some of the examples in this subsection are adapted from Sara Parker and Derek Ruths, 'Is Hate Speech Detection the Solution the World Wants?' (2023) 120 *Proceedings of the National Academy of Sciences* e2209384120, https://www.pnas.org/doi/10.1073/pnas.2209384120.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

The system's decisions became less deterministic than rule-based approaches in this phase, but they still remained anchored in traceable human judgement introduced through the training process. However, this dependence on human-labelled data also created scalability challenges and potential inconsistencies in how standards were applied.

3.2.3. Phase 3 – LLMs: Machine discretion and standard-like enforcement

The emergence of LLMs represents a fundamental shift in the balance of power between humans and machines. Unlike earlier automated systems, LLMs develop their own representations of language and context through pre-training on vast datasets, enabling them to infer patterns, intentions and nuances with minimal human intervention.[111] As a result, LLMs can be presented with content policies and asked directly whether a given piece of content constitutes hate speech or another policy violation. They can also provide what is often described as 'reasoning' for their decisions, giving (only) the impression of explainability.[112] This capability allows them to exercise what appears to be independent judgement, *bringing content-moderation practices closer to the logic of standard-like enforcement rather than rule-following.*

Now, it could be argued that this discretion can and should be constrained through fine-tuning – that is, by training the pre-trained LLM on carefully curated examples that embody a platform's content rules (for instance, hate speech standards).[113] Such fine-tuning could enable the model to internalise these rules and apply them more consistently during moderation tasks. However, if carried out in an overly literal or highly restrictive manner (for example, as is commonly done by generating rule-obeying data in fairly large amounts and training the models on it), this risks severely narrowing the contextual dependence of the LLM's prediction, thereby undermining the very rationale for integrating LLMs into the content-moderation pipeline in the first place.

In the cases described above, for example, LLMs may recognise a reclaimed use of a prohibited slur (such as the n-word) or a context-dependent usage of terms like 'tr*nny', even if they have never seen those precise expressions in that specific context. Similarly, when a post refers to 'Jews' through implied or coded language (such as 'Global Bankers', 'Soros', or 'Benjamins'), the ability of LLMs to infer linguistic context may enable them to apply a form of 'discretion-like' judgement and flag

[111] Meta, 'Self-supervised Learning: The Dark Matter of Intelligence', 4 March 2021, https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence.

[112] Huan Ma and others, 'Adapting Large Language Models for Content Moderation: Pitfalls in Data Engineering and Supervised Fine-Tuning', Cornell University, 5 October 2023, https://arxiv.org/abs/2310.03400; Kumar, AbuHashem and Durumeric (n 13); Franco, Gaggi and Palazzi (n 13). However, see Peter Hase and others, 'Reasoning Models Don't Always Say What They Think, for Faithfulness Challenges in the Faithfulness of LLMs' Chain-of-Thought Reasoning', *Anthropic*, 3 April 2025, https://www.anthropic.com/research/reasoning-models-dont-say-think.

[113] But see also Mahi Kolla and others, 'LLM-Mod: Can Large Language Models Assist Content Moderation?', CHI EA (2024), https://doi.org/10.1145/3613905.3650828 (in which the authors show that off-the-shelf LLMs without fine-tuning struggle in accurately moderating content).

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

such posts, even where the platform's guidelines do not explicitly list these terms as violations.[114]

This demonstrates an important distinction: although platform policies are articulated as rules, LLMs do not actually operate by straightforward rule-following. Strict rule enforcement could, in principle, be achieved through highly literal inference-time constraints or through extremely restrictive fine-tuning, but LLMs, when used in their intended capacity, tend instead to function more like *standard-based decision-makers*.

Consider the following simplified illustration from online tutorials for LLM-based content-moderation systems:[115]

```
# Example guidelines
example_guidelines = "'BLOCK CATEGORY:
- Promoting violence, illegal activities, or hate speech
- Explicit sexual content
- Harmful misinformation or conspiracy theories
ALLOW CATEGORY:
- Most other content is allowed, as long as it is not explicitly disallowed
"'
user_comments = [
"This movie was great, I really enjoyed it. The main actor really killed it!",
"Delete this post now or you better hide. I am coming after you and your family.",
"Stay away from the 5G cellphones!! They are using 5G to control you.",
"Thanks for the helpful information!",
]
for comment in user_comments:
classification = moderate_text(comment, example_guidelines)
print(f"Comment: {comment}\nClassification: {classification}\n")
```

The prompt does not specify what hate speech is. Unlike rule-based systems in which the instructions are detailed and categorical, here the language of the prompt leaves discretion and interpretation for the model to determine what is '[p]romoting violence, illegal activities, or hate speech'. The model generalises from this instruction, weighing contextual cues and nuances. In this sense it is much like legal standards, such as the 'reasonable person' standard, which require decision-makers to balance multiple factors in real time. This shift in the decision-making process manifests itself in several ways that closely resemble the operation of legal standards. Firstly, LLMs can evaluate multiple contextual dimensions simultaneously and implicitly,

[114] Ma and others (n 112); Franco, Gaggi and Palazzi (n 13).

[115] Devesh Surve, 'Beginner's Guide to LLMs: Build a Content Moderation Filter and Learn Advanced Prompting with Free Groq API', *Medium*, 5 June 2024, https://deveshsurve.medium.com/beginners-guide-to-llms-build-a-content-moderation-filter-and-learn-advanced-prompting-with-free-87f3bad7c0af. For more examples of prompts used in mainstream LLM-based content moderation see Cong Chen and others, 'A Comprehensive Review of LLM-based Content Moderation: Advancements, Challenges, and Future Directions' (2025) 330(C) *Knowledge Based Systems* 114689, Figure 4.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

producing judgements that approximate the flexible balancing characteristic of legal standards. Secondly, their decisions adapt to specific situational nuances, rather than applying rules in a mechanistic or literal fashion.

Yet this expansion of machine discretion raises significant concerns about accountability and oversight. When an LLM flags content as harmful, it may be drawing on implicit standards learned during training that may diverge from the platform's explicit rules. This creates a potential mismatch between the *policy as written* and the *policy as enforced*, echoing the classic legal distinction between 'law on the books' and 'law in action'. Kaplan's concern about systematic over-enforcement of platform rules can be understood in this light.

As with other forms of standard-based enforcement, the normative question remains: is broader-than-necessary removal desirable because it reduces the total volume of harmful content, including false negatives (i.e., content that violates policy but would otherwise slip through)? Or is it problematic because it increases false positives (i.e., the removal of benign content), thereby over-curtailing users' freedom of expression? The answer turns not only on empirical performance, but also on normative commitments regarding speech, safety, and the appropriate degree of machine discretion.

3.3. The promises and perils in LLM-based content moderation

The advancement of LLMs has expanded the scale and scope of algorithms capable of detecting and filtering hate speech.[116] LLMs' sophisticated understanding of language and context and their ability to capture nuanced contextual information enable them to identify hate speech in complex scenarios involving sarcasm, satire, coded language and metaphors – areas where traditional machine learning models consistently fall short.[117]

Transfer learning provides these models with a comprehensive foundation in language understanding, facilitating zero-shot and few-shot learning that reduces the need for extensive domain-specific training data, enabling them to apply their knowledge to new and varied contexts.[118] The multilingual capabilities of LLMs and their adaptability to emerging patterns of harmful speech also allow them to effectively filter hate speech across different languages and leverage their understanding of linguistic structures and cross-lingual transfer learning, to identify hate speech in languages where labelled data may be scarce.[119] LLMs can easily adjust to

[116] Chiu and others (n 13); Franco, Gaggi and Palazzi (n 13); Ma and others (n 112).

[117] Fan Huang, Haewoon Kwak and Jisun An, 'Is ChatGPT Better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech', Cornell University, 15 March 2023, https://arxiv.org/abs/2302.07736. Chen and others (n 115) (arguing that LLM's 'leap in capability fundamentally redefines the paradigm of content moderation – shifting the focus from mere detection to nuanced interpretation, transparent justification, and even proactive engagement'). However, see Kolla and others (n 113) (in which the authors argue that off-the-shelf large language models can assist human moderators by reliably identifying compliant ('rule-passing') posts and providing reasoned explanations, but currently perform poorly at flagging complex rule-violating content).

[118] Lanqin Yuan and others, 'Transfer Learning for Hate Speech Detection in Social Media', Cornell University, 29 October 2023, https://arxiv.org/abs/1906.03829.

[119] Janis Goldzycher and others, 'Evaluating the Effectiveness of Natural Language Inference for Hate Speech Detection in Languages with Limited Labeled Data', Cornell University, 10 June 2023, https://arxiv.org/abs/2306.03722.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

changing patterns and new forms of hate speech over time. As they continually learn from new data, they can update their understanding and detection capabilities.[120] Additionally, their exposure to vast amounts of constantly updated data allows them to study rare or emerging patterns of hate speech – such as novel phrases, slang or evolving terminology used by hate groups.[121]

Still, the integration of LLMs into content moderation systems raises profound concerns that mirror and amplify known challenges in LLM deployment. A fundamental challenge of using LLMs is the issue of explainability. The 'black box' nature of LLM decision-making becomes particularly problematic in content moderation, where transparency and accountability relating to limitations of freedom of expression or curbing hate speech are crucial. When an LLM flags content as harmful (or non-harmful), the complexity of its internal processes makes it difficult, if not impossible, to provide users with a clear explanation for the decision. This opacity challenges fundamental principles of due process and fairness in content moderation.[122]

The phenomenon of LLM 'hallucination'[123] – in which, among other manifestations, the models may address knowledge gaps by suggesting fabricated or unreliable content – takes on new significance in content moderation contexts. While hallucination in general applications might lead to incorrect information, in content moderation it can result in false positives that directly and illegitimately restrict user freedom of expression. Unlike rule-based systems where errors follow predictable patterns, LLM hallucinations can produce seemingly arbitrary decisions that are difficult to systematically identify or correct.

In addition, training data biases manifest themselves in particularly concerning ways in content moderation. LLMs trained on internet-scale data inevitably absorb societal biases, potentially leading to discriminatory moderation practices. For example, these models might disproportionately flag content from minority communities while missing subtle forms of harassment that use majority-culture coded language. This risk of algorithmic discrimination becomes especially acute when moderation decisions affect the ability of individuals to participate in online discourse.[124]

The 'alignment problem' – ensuring that AI systems behave in accordance with human values and intentions – also presents unique challenges in content moderation. LLMs may develop their own implicit standards for identifying harmful content

[120] Paul Röttger and Janet B Pierrehumbert, 'Temporal Adaptation of BERT and Performance on Downstream Document Classification: Insights from Social Media', Cornell University, 8 September 2021, https://arxiv.org/abs/2104.08116.

[121] Chiu and others (n 13); Yau-Shian Wang and Yingshan Chang, 'Toxicity Detection with Generative Prompt-based Inference', Cornell University, 24 May 2022, https://arxiv.org/abs/2205.12390.

[122] Chen and others (n 115). There are efforts in the technical community that explore ways of improving LLM explainability. For a comprehensive survey of such efforts see Ahsan Bilal, David Ebert and Beiyu Lin, 'LLMs for Explainable AI: A Comprehensive Survey', Cornell University, 31 March 2025, https://arxiv.org/abs/2504.00125; Haiyan Zhao and others, 'Explainability for Large Language Models: A Survey' (2024) 15(2) *ACM Transactions on Intelligent Systems and Technology* 1.

[123] Sebastian Farquhar and others, 'Detecting Hallucinations in Large Language Models Using Semantic Entropy' (2024) 630 *Nature* 625.

[124] Emily Sheng and others, 'Societal Biases in Language Generation: Progress and Challenges', Cornell University, 22 June 2021, https://arxiv.org/abs/2105.04054.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

that diverge from platform policies or community norms. This divergence can lead to overreach in content removal, particularly when dealing with complex topics where the line between legitimate discourse and harmful content is nuanced. Other concerns include the financial and environmental costs of training large models, as well as infringement of intellectual property rights[125] and user privacy concerns that such training processes may entail.[126] Lastly, it should be recalled that generative language models are also employed, unfortunately, to *generate*, rather than moderate, toxic content and hate speech.[127]

3.4. The use of AI and LLMs in enforcement of hate speech policies

Over the last several years, Facebook has significantly invested in AI technologies to combat hate speech on its platforms. In 2016, Facebook incorporated AI algorithms in its code to help to identify and remove content that violated its policies.[128] These algorithms could detect nudity, hate speech and other forms of explicit or harmful content, thus augmenting the work of human moderators. At the same time, Facebook trained a deep learning-based text-analysing engine, noting that 'tricky scaling and language challenges' make traditional NLP techniques 'not effective'. Using deep learning, the company explained, 'we are able to understand text better across multiple languages and use labelled data much more efficiently than traditional NLP techniques'.[129] In its 2018 review, Facebook stated that it was 'advancing AI learning through semi-supervised and unsupervised training'. Facebook then addressed the limitations of existing supervised learning in studying a particular task – specifically, the problem that the dependency on large numbers of labelled samples restricted the number of tasks that AI systems could learn – which, in turn, limited the technology's long-term potential.[130] As a result, Facebook moved

[125] Gil Appel, Juliana Neelbauer and David A Schweidel, 'Generative AI has an Intellectual Property Problem' (2023) *Harvard Business Review*. For the *New York Times* recent lawsuit against Open AI and Microsoft in this regard see Jonathan Stempel, 'NY Times Sues OpenAI, Microsoft for Infringing Copyrighted Works', *Reuters*, 28 December 2023, https://www.reuters.com/legal/transactional/ny-times-sues-openai-microsoft-infringing-copyrighted-work-2023-12-27.

[126] Matt Burgess, 'ChatGPT Has a Big Privacy Problem', *Wired*, 4 April 2023, https://www.wired.com/story/italy-ban-chatgpt-privacy-gdpr.

[127] James Vincent, 'Twitter Taught Microsoft's AI Chatbot to be a Racist Asshole in Less than a Day', *The Verge*, 24 March 2016, https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist; Christianna Silva, 'It Took Just One Weekend for Meta's New AI Chatbot to Become Racist', *Mashable*, 8 August 2022, https://mashable.com/article/meta-facebook-ai-chatbot-racism-donald-trump; Hugo Touvron and others, 'Open Foundation and Fine-Tuned Chat Models', Cornell University, 19 July 2023, https://arxiv.org/abs/2307.09288 (reporting toxicity testing for Llama).

[128] Guy Rosen, 'F8 2018: Using Technology to Remove the Bad Stuff Before It's Even Reported', Meta, 2 May 2018, https://about.fb.com/news/2018/05/removing-content-using-ai; Mike Schroepfer, 'Update on Our Progress on AI and Hate Speech Detection', Meta, 21 February 2021, https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection.

[129] Ahmad Abdulkader, Aparna Lakshmiratan and Joy Zhang, 'Introducing DeepText: Facebook's Text Understanding Engine', Engineering at Meta, 1 June 2016, https://engineering.fb.com/2016/06/01/core-infra/introducing-deeptext-facebook-s-text-understanding-engine.

[130] Jerome Pesenti, 'AI Year in Review', Engineering at Meta, 8 January 2019, https://engineering.fb.com/2019/01/08/ai-research/ai-2018.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

to reduce the amount of supervision necessary for training, including by embarking on projects that demonstrate the benefits of learning from semi-supervised, or even unsupervised, data. It claimed, in this regard, that training automatic translation models on unsupervised data resulted in performance comparable to that of systems trained on supervised data.[131] In November 2020, Facebook stated it was '[t]raining AI to detect hate speech in the real world'.[132] Citing its billions of users, the company said that it relies on AI 'to scale our content review work and automate decisions when possible'. It added that 'AI now proactively detects 94.7 percent of hate speech we remove from Facebook, up from 80.5 percent a year ago and up from just 24 percent in 2017'.[133]

In January 2022, Facebook announced that it was developing supercomputer infrastructure for AI research, to support the training of 'increasingly large, complex, and adaptable models', needed for 'critical use cases like identifying harmful content'.[134] In February 2023, Facebook announced the release of Llama – a 65 billion-parameter LLM – and, in July 2023, released Llama 2, which had been trained on 40 per cent more data than Llama 1, and has double the context length. In the report accompanying the release of the Llama model, the research team explained that the pre-training was conducted in a manner that would 'allow Llama 2 to be more widely usable across tasks (e.g., it can be better used for hate speech classification)'.[135]

Zuckerberg's January 2025 policy announcement has been seen by critics as 'an abandonment of his technological project'. The radical policy shift marks a move away from years-long investment in automation as the future of content moderation, accompanied by quarterly reports on the automated systems' improvements in proactive detection and removal of harmful content, to crowd-sourcing content moderation techniques of the kind used in platforms such as X.[136] Still, it is hard to imagine Meta's massive scale of content moderation returning to rely primarily on human oversight. With billions of daily users generating an overwhelming volume of posts, videos and interactions, automation remains a core necessity for the company.

[131] Myle Ott, Marc Aurelio Ranzato and Guillaume Lampe, 'Unsupervised Machine Translation: A Novel Approach to Provide Fast, Accurate Translations for More Languages', Engineering at Meta, 31 August 2018, https://engineering.fb.com/2018/08/31/ai-research/unsupervised-machine-translation-a-novel-approach-to-provide-fast-accurate-translations-for-more-languages.

[132] Meta, 'Training AI to Detect Hate Speech in the Real World', 19 November 2020, https://ai.meta.com/blog/training-ai-to-detect-hate-speech-in-the-real-world.

[133] Meta, 'How AI is Getting Better at Detecting Hate Speech', 19 November 2020, https://ai.meta.com/blog/how-ai-is-getting-better-at-detecting-hate-speech. In 2023, Facebook stated that for most violation categories AI technology finds over 90% of the content that the company removes before anyone reports it: Meta, 'How Technology Detects Violations', 18 October 2022, https://transparency.meta.com/en-gb/enforcement/detecting-violations/technology-detects-violations.

[134] Meta, 'Introducing the AI Research SuperCluster – Meta's Cutting-Edge AI Supercomputer for AI Research', 24 January 2022, https://ai.meta.com/blog/ai-rsc.

[135] Touvron and others (n 127) 20. See also Meta, 'Introducing Llama: A Foundational, 65-Billion-Parameter Language Model', 24 February 2023, https://ai.meta.com/blog/large-language-model-llama-meta-ai.

[136] 'Meta Goes MAGA Mode + A Big Month in A.I + HatGPT', *The New York Times*, Hard Fork podcast transcript, 10 January 2025, https://www.nytimes.com/2025/01/10/podcasts/hardfork-meta-zuck-o3.html?showTranscript=1.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

This is especially the case as the company continues to be legally exposed to the requirements of significant anti-hate speech regulations applicable in the European Union (EU) and a number of other jurisdictions.[137]

Indeed, in May 2025, Meta announced that it was testing LLMs for content-enforcement tasks by training them on the company's Community Standards. Meta emphasised that LLMs offer 'significant opportunities to counter high-severity and illegal content, at scale'. According to the company's Integrity Report, these models not only outperformed existing machine-learning systems but, in certain policy areas, operated 'beyond that of human performance'. Meta further reported that LLMs were already being used to remove content from review queues and to support other moderation-related functions, such as reviewing user-submitted bug reports.[138]

4. Governance implications of AI-enabled enforcement of hate speech policies

4.1. The governance gap between rule-based content policies and standard-based AI enforcement

The disparity between Facebook's rule-oriented Community Standards and their standard-like enforcement through AI tools, like advanced NLP models, presents several notable challenges. These challenges pertain to the balance of advantages and disadvantages inherent in rule-based policies and the shift towards a more discretionary, standard-like approach in enforcement.

Firstly, the reliance on advanced NLP models in content moderation alters the anticipated benefits of rule-oriented policies. As explained above, rules are seen typically as directives that factor in advance the relevant considerations. Conversely, standards leave most of the important choices to be made later, 'at the moment of application'.[139] The policies of digital platforms, while often articulated as rules, are enforced in a manner more akin to standards by NLP models, factoring in various post-legislative considerations in ways that are constantly evolving. This shift in enforcement style raises institutional questions about power allocation. Traditionally, rule-makers (policy drafters) hold substantive decision-making powers, with enforcers (e.g., AI models) playing a more mechanical role. NLP-based enforcement disrupts this, assigning more discretion to AI models. This shift carries significant implications, given, inter alia, the differences in the transparency and scrutiny mechanisms and incentives associated with policy drafting and technology development.

[137] See, eg, Digital Services Act (n 84) art 35(1)(c); Online Safety Act 2023 (UK), s 10; and Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (FRG), s 4a (Netzwerkdurchsetzungsgesetz – NetzDG).

[138] Meta Transparency Center, 'Integrity Reports, First Quarter 2025', 29 May 2025, https://transparency.meta.com/reports/integrity-reports-q1-2025; Akos Lada and others, 'How Facebook Leverages Large Language Models to Understand User Bug Reports and Guide Fundamental Improvements', *Medium*, 12 March 2025, https://medium.com/@AnalyticsAtMeta/how-facebook-leverages-large-language-models-to-understand-user-bug-reports-and-guide-fundamental-70ab26475850.

[139] See Schauer (n 36) 804.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

Furthermore, the technique used in content moderation affects the disparity between policy articulation and enforcement. Legal techniques following rule articulation involve technical categorisation, while standards require complex balancing and the weighing of various considerations.[140] NLP-based moderation, contrary to the impression of straightforward categorisation, relies on a multitude of constantly updated parameters and weights. This complexity also challenges predictability, a key advantage of rule-based approaches, as NLP-driven enforcement in digital platforms introduces unpredictability, undermining legal certainty and behavioural guidance.

Cost considerations also play a role in the analysis. As previously noted, rules may be more costly than standards to articulate, but standards can be more 'costly for individuals to interpret when deciding how to act and for an adjudicator to apply to past conduct'.[141] In Facebook's case, significant costs could arise at the enforcement stage as a result of the uncertain nature of NLP-driven enforcement. This extends to users who face monetary and time costs in challenging moderation decisions, through internal appeals processes, Meta's Oversight Board decisions, and litigation, exacerbated by the opacity and lack of transparency in NLP technologies.

The tension between rule-oriented policies and standard-like enforcement also has an impact on fairness considerations. Rule-based approaches promote equality and consistency in enforcement, but NLP enforcement raises concerns of bias, performance disparities and low transparency, making it challenging to ensure similar treatment across multiple cases and to hold accountable platforms that disparately treat their users. Additionally, the opaque and potentially biased nature of NLP enforcement may weaken rule-of-law principles, which rely on clear directives that are applied evenly. Nonetheless, it should be noted that standards may offer flexible and adaptable approaches to addressing emerging hate speech challenges (including new vocabulary, as well as new social, cultural and political developments) that are not covered in Meta's content policies.[142]

While dilutive of the advantages of rule-based approaches, the use of AI in content moderation does not fully capture the benefits of standard-based methods either. As demonstrated, advanced NLP technologies like LLMs significantly enhance the effectiveness of content moderation in problem-solving by adeptly handling a wide range of cases, particularly in detecting and filtering hate speech. They excel in interpreting context and subtleties, adapt to new languages, and continually evolve to identify emerging patterns of hate speech. However, these technologies do not inherently guarantee desirable outcomes or fully reasoned policy applications, on a par with those generated through the exercise of human discretion when applying legal standards. Their sophisticated and nuanced operation relies on computational approaches rather than a meaningful normative engagement with competing values and interests, introducing an intricate disruption to the application of the rules-versus-standards paradigm in digital contexts. Indeed, the operation of LLMs is based on algorithms and data-driven processes; they analyse and process language by recognising patterns in the vast amounts of data on which they have been trained.

[140] See Bayles (n 23).

[141] See Kaplow (n 33) 557.

[142] See discussion in Section 2.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

This computational approach enables them to make decisions – like identifying hate speech – based on statistical likelihood and correlations found in their training data. Their decision-making capabilities are rooted in patterns and examples from their training data rather than a deep, principled understanding of norms and ethics. In fact, the deep learning process underlying LLMs could result in unpredictable and unexplainable over-assignment of weight to certain rules extrapolated from the applicable policies and training data and under-weighing of other factors, leading to nonsensical or hallucinatory outcomes. This underscores the importance of a balanced and thoughtful application of AI in the dynamic field of online content moderation, where the capabilities of LLMs can be leveraged effectively while being mindful of their limitations and the complexities of moderating online discourse.

4.2. The regulatory asymmetry between policies and their enforcement

Current regulatory and oversight mechanisms provide inadequate tools for mitigating the gap between digital platforms' policies and their AI-driven enforcement. Existing national and international regulation, coupled with self-instituted oversight mechanisms (such as Meta's Oversight Board), focus and rely on the policies and their wording, rather than on their manner of enforcement and their consistency with the articulated policies.[143]

Indeed, national and international regulatory frameworks that currently apply to online platforms prompt them to adopt policies that are detailed and clear. Moreover, transparency requirements deal primarily with the policies rather than with their enforcement. While policies are available for everyone to review – and the platforms take pride in them being public, detailed and clear – the algorithmic models used for enforcement are usually inaccessible and incomprehensible to the public and decision-makers.[144]

The tension between the rule-oriented policies and their standard-like enforcement is illustrated by the different treatment that Meta's Oversight Board affords each of these two components. The Board – which was conceived and created by Facebook itself – was given a confined mandate to review users' appeals regarding decisions of Facebook and Instagram to remove content on the grounds that it violated their content policies (Facebook's Community Standards and Instagram's Community Guidelines).[145] The Board is authorised to determine whether the

[143] See below in this section.

[144] See Section 4.2.

[145] In 2024, the Oversight Board's mandate was expanded to include 'Threads': Oversight Board, 'The Oversight Board Expands to Threads', 22 February 2024, https://www.oversightboard.com/news/1376739096292499-the-oversight-board-expands-to-threads. The mandate did not include a review of other content moderation methods – such as account suspensions, visibility reduction, or removal of content for violations of local state law. In addition, the Board can only consider cases that have undergone Meta's internal appeal processes: Noa Mor, 'On Facebook's New "Oversight Board," Accountability, and Control', *Cornell Tech*, 11 April 2020, https://dli.tech.cornell.edu/post/on-facebook-s-new-oversight-board-accountability-and-control. Recently, the Board's jurisdiction was expanded to include decisions concerning the application of warning screens: Oversight Board, 'Oversight Board Publishes Transparency Report for Second Quarter of 2022 and Gains Ability to Apply Warning Screens', 20 October 2022, https://www.oversightboard.com/news/784035775991380-oversight-board-publishes-transparency-report-for-second-quarter-of-2022-and-gains-ability-to-apply-warning-screens. See also

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

content moderation decisions made by Facebook or Instagram were in line with their policies and values, and with international human rights law. Such decisions are binding on Meta.[146] The Board can also issue non-binding recommendations concerning other issues, such as the articulation of policies or the training and review process conducted by the platform's moderation staff.[147]

For example, in a case relating to the removal of a post that was shared by a news outlet page in Colombia and concerned the term 'slur,' as covered in the policy,[148] the Board recommended, inter alia, that:

> [Facebook will] publish illustrative examples from the list of slurs it has designated as violating under its Hate Speech Community Standard. These examples should be included in the Community Standard and include edge cases involving words which may be harmful in some contexts but not others, describing when their use would be violating. Facebook should clarify to users that these examples do not constitute a complete list.

Meta committed to adopting that recommendation and to updating its hate speech Community Standard,[149] but what about enforcement of the policy? Here, too, as in the previous example, the Board did make some recommendations regarding

Oversight Board, 'Oversight Board Bylaws', https://www.oversightboard.com/wp-content/uploads/2025/07/Oversight-Board-Bylaws-June-2025.pdf.

[146] Facebook's stated values are voice, authenticity, privacy, safety and dignity: Meta, 'Updating the Values that Inform Our Community Standards', 12 September 2019, https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards. According to Facebook, these values inform their Community Standards and serve as the 'guidepost for what is and isn't allowed on Facebook'.

[147] Mor (n 145). In May 2022, for example, the Board published a decision regarding Facebook's removal of a meme in which Turkey was described as having to choose between two options, 'The Armenian Genocide is a Lie' and 'The Armenians Were Terrorists Who Deserved It': Meta, 'Case on a Comment related to the Armenian People and the Armenian Genocide', 12 June 2023, https://transparency.meta.com/oversight/oversight-board-cases/comment-related-to-armenian-people-and-the-armenian-genocide (*Armenian* case). See also Oversight Board, '"Two Buttons" Meme', 20 May 2021, https://www.oversightboard.com/decision/FB-RZL57QHJ (*Two Buttons* decision). One of the recommendations that the Board issued was to allow for satire exceptions in Facebook's hate speech policy (which, at the time, were applied in practice by the company, but not included in its Community Standards). Meta committed to carrying out the recommendation. The company explained: 'We'll add information to the Community Standards that makes it clear where we consider satire as part of our assessment of context-specific decisions': *Armenian* case, ibid, Recommendation 2. The Board also recommended that Facebook '[m]ake sure that it has adequate procedures in place to assess satirical content and relevant context properly'. This additional recommendation, however, focused on the human moderation teams, not on the automated enforcement measures used by the company. For further decisions of the Board requiring enhanced clarity of the policies, with less (or no) attention to automated enforcement means to pair with such policies, see Meta, 'Oversight Board Selects a Case Regarding a Post with a Photo of a Topless Person and Shirts for Sale', 12 June 2023, https://transparency.meta.com/oversight/oversight-board-cases/post-containing-photo-of-topless-individual.

[148] Oversight Board, 'Colombia Protests', 27 September 2021, https://www.oversightboard.com/decision/FB-E5M6QZGA.

[149] ibid. However, Meta noted that '[a]lthough we recognize the need for transparency and clarity in our Community Standards, we are still considering the extent to which we will publish specific slur words, given that the mere mention of these terms can be demeaning and can lead to an environment of intimidation and exclusion: Oversight Board, 'Colombian Protest Video', 16 April 2024, https://transparency.meta.com/oversight/oversight-board-cases/protests-colombia-while-using-slur.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

enforcement, but focused on processes involving content reviewers rather than algorithmic mechanisms,[150] or addressed automated enforcement concerns very partially.[151] Facebook's reserved agenda in this respect is well reflected in the response it offered to one of the earliest cases that the Board heard, in which it very clearly signalled that the Board's meddling in the automated process was not appreciated. The Board noted in this regard that:[152]

> Facebook ... claims that it is not relevant to the Board's consideration of the case whether the content was removed through an automated process, or whether there was an internal review to a human moderator. *Facebook would like the Board to focus on the outcome of enforcement, and not the method*.

This regulatory asymmetry between the policies and their automated enforcement is addressed, however, in some spaces within the online content moderation ecosystem. The Digital Services Act, for instance, imposes various obligations on online actors and embraces both the platforms' policies and their manner of enforcement, thereby requiring platforms to address the friction between these two components (but offers limited guidance in this regard).[153]

The EU Code of Conduct on Countering Hate Speech Online, which steers the platform's policies and their enforcement with regard to fighting hate speech, is another example; this underscores the need to draw attention to the tension between articulation and application of content moderation policies, and to consider ways of tackling its implications.[154]

4.3. 'Rules by the millions': Implications of LLMs on the rules-versus-standards debate

Although originally established within the world of legal theory, the distinction between rules and standards has proven relevant to laying the theoretical groundwork for recognising and tackling key challenges presented by online content moderation. However, with the emerging role of LLMs in enforcing content moderation, we may need to revisit our traditional understanding of this distinction.

[150] ibid. Nonetheless, it seems that the Oversight Board is increasingly addressing some automated enforcement aspects of Meta's content moderation; see, for instance, Oversight Board, 'Sudan's Rapid Support Forces Video Captive', 11 April 2024, https://www.oversightboard.com/decision/fb-14uy7pvn.

[151] Noa Mor, 'Meta's Oversight Board's Report on AI: What's Left Unpacked', Three Generations of Digital Human Rights ERC Project, Hebrew University of Jerusalem, n.d., 2, https://3gdr.huji.ac.il/sites/default/files/threegenerationsofdigitalhumanrights/files/metas_oversight_boards_report_on_ai_-_whats_left_unpacked_fi_01.pdf.

[152] Oversight Board, 'Breast Cancer Symptoms and Nudity', 28 January 2021, https://www.oversightboard.com/decision/IG-7THR3SI1 (emphasis added).

[153] See Digital Services Act (n 84) Recital, para 49.

[154] European Commission and Facebook, Microsoft, Twitter and YouTube, 'Code of Conduct on Countering Illegal Hate Speech Online', https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985. In January 2025, the European Commission announced the adoption of a revised version of the Code of Conduct, now called 'The Code of Conduct on Countering Hate Speech Online +': European Commission, 'The Code of Conduct on Countering Illegal Hate Speech Online +', Shaping Europe's Digital Future, 20 January 2025, https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

Unlike rule-based methods, which apply overarching generalisations to classify examples, LLMs, with their advanced natural language processing capabilities, are context-aware and adaptive.[155] In that sense, it may be argued that they feature attributes of standard-based decisions, as they offer less certainty and more room for factoring in context-sensitive considerations. However, they can also be likened to rules in that they are based on the mechanical activation of masses of calculations, micro-prescriptions and specific instructions on how to weigh them together.[156] Furthermore, it is currently an open question among AI scholars whether enormous volumes of data, coupled with unprecedented computing power running LLMs, amount to a qualitative difference of language models, which is sometimes referred to as the 'emergent abilities' of LLMs.[157]

These normative and computational considerations, we submit, indicate that the traditional understanding of the rules–standards distinction in digital settings must be revisited, at least with regard to LLM-based decision-making.

Indeed, LLMs not only tip the scale from rule-like to standard-like enforcement but may pave the way for a nascent, third category, implementing what we might call 'rules by the millions' – a novel paradigm that transcends the traditional rules–standards dichotomy. In this new paradigm, LLMs operate through millions of micro-rules learned during training, each contributing to a complex web of pattern recognition.

The term 'rules by the millions' underscores the massive scale and granular level at which these AI models operate. Rather than traditional rules that prescribe clear-cut, universal directives, or standards that provide overarching principles that are open to interpretation and to all-things-considered analysis, this new paradigm operates by generating an immense number of highly specific, context-sensitive rules. These micro-rules evolve dynamically and adapt based on the vast amounts of data on which the models are trained.[158]

When an LLM evaluates, for example, whether a post referring to 'Zionists' in an extremely negative manner constitutes hate speech, it is not simply applying a

[155] Chen and others (n 115).

[156] Another example of such a view of LLMs may be found in recent attempts by NLP scholars to decompose transformers into large collections of patterns: Mor Geva and others, 'Transformer Feed-Forward Layers Are Key-Value Memories', Cornell University, 5 September 2021, https://arxiv.org/abs/2012.14913.

[157] Jason Wei and others, 'Emergent Abilities of Large Language Models', Cornell University, 26 October 2022, https://arxiv.org/abs/2206.07682 (defining the emergent abilities of LLMs as 'abilities that are not present in smaller-scale models but are present in large-scale models; thus they cannot be predicted by simply extrapolating the performance improvements on smaller-scale models'). For a criticism of this approach see Rylan Schaeffer, Brando Miranda and Sanmi Koyejo, 'Are Emergent Abilities of Large Language Models a Mirage?', Cornell University, 22 May 2023, https://arxiv.org/abs/2304.15004 (arguing that for a given task and model family, when analysing fixed model outputs, emergent abilities appear as a result of the user's choice of metric rather than to qualitative changes in model behaviour with scale). For a survey of the current state of the field see Leonardo Berti, Flavio Giorgi and Gjergji Kasneci, 'Emergent Abilities in Large Language Models: A Survey', Cornell University, 15 March 2025, https://arxiv.org/abs/2503.05788.

[158] Compare Anthony J Casey and Anthony Niblett, 'Self-Driving Laws' (2016) 66 *University of Toronto Law Journal* 429 (positing that AI 'micro-directives' resolve the rules–standards dilemma via precise, intended commands, with the proposed concept of 'rules by the millions', where enforcement is emergent and unintended). We argue that rather than optimising law, these hidden rules function as a 'shadow legislator', which evades oversight and severs the link between policy and enforcement.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

broad standard of 'hateful content'. Instead, it is simultaneously drawing on millions of learned patterns related to language use, context and social dynamics.[159] Each of these patterns functions like a micro-rule, and their vast number and concurrent application create behaviour that *resembles* the flexibility of a standard while retaining the fine-grained precision of rule-like mechanisms.

Crucially, however, these millions of learned patterns do not contribute equally or transparently. The model's output emerges from a highly complex computation in which different constraints, signals and criteria are weighted and combined in varying degrees. As a result, the decision process is neither uniform nor capable of easy interpretation, even though it produces judgements that appear cohesive and principled from the outside.

To demonstrate how complex, quasi-judicial behaviour emerges from the aggregation of micro-rules, we conducted a controlled experiment using Google's Gemini (see the online Supplementary Materials for documentation). We presented the model with a core statement containing a common antisemitic trope regarding Jewish ownership of media, but varied the attribution of the statement to test the model's context sensitivity.

To ensure that our methodology aligned with existing practices in LLM-based content moderation research, we adopted a standardised prompt structure.[160] The prompt provided a static definition of hate speech ('Attacks or incites hatred against a group based on race, religion, gender, sex, sexual orientation') and instructed the LLM to classify the input as either 'Hate Speech' or 'Not Hate Speech' with a brief explanation. The results revealed distinct variations based on attribution:

Scenario A
Input: 'Nearly every major TV channel and advertising medium in the US is Jewish owned'.
Output: [Hate Speech]
Model Explanation: 'The statement promotes a conspiracy theory about Jewish control over media, which can incite hatred and discrimination against Jewish people'.

Scenario B
Input: 'According to NYT, nearly every major TV channel and advertising medium in the US is Jewish owned'.
Output: [Not Hate Speech]
Model Explanation: 'The statement presents a claim about the ownership of media outlets but does not express hatred or incite violence against Jewish people'.

Scenario C
Input: 'According to Steve Bannon, nearly every major TV channel and advertising medium in the US is Jewish owned'.
Output: [Hate Speech]

[159] Empirical support to this effect can be found in Geva and others (n 156).

[160] This prompt is based on the prompts featured in Chen and others (n 115) Figure 4 (focusing only on the category of hate speech for purposes of demonstration).

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

Model Explanation: 'The statement promotes an antisemitic trope about Jewish control of media'.

These outputs demonstrate a phenomenon we term 'systematic context-sensitivity'. The model is sensitive in that granular adjustments to the input, such as changing the source of attribution, can invert the legal classification entirely. Crucially, however, this variance is systematic: the model reliably alters its verdict in response to materially relevant context, while remaining stable in the face of irrelevant noise, mirroring a form of judicial discretion, which is often perceived as a standard-based application of a policy. However, critical opacity remains. While we can observe that a change in attribution triggers a reclassification, the internal logic dictating why this specific shift occurs – namely, which rule among the millions 'learned' by the model in its training drove this decision – remains inaccessible.

To be clear, the discretion exercised by LLMs in this context is not necessarily what we perceive as the broad, human-guided discretion typically associated with standards, which integrates articulated rules, competing interests, values, moral intuitions and social expectations.[161] Rather, it is an AI-powered, data-driven discretion that sifts through an immense volume of nuanced, context-specific micro-rules at a speed and scale beyond human capabilities, integrating them through a highly complex function induced from training data. The (third) category we call 'rules by the millions' thus emphasises not only the sheer volume of content that can be moderated by the machine, but also the complexity of the underlying computation and the amount of contextual information that the model can take into account in enforcing content policies. This unparalleled level of granularity and context-sensitivity in rule application strains the existing rules–standards dichotomy. From the perspective of platform users, the opacity of contemporary LLMs may entail levels of unpredictability comparable to those associated with the application of standards, but the differences between human and machine decision-making render the latter even less explainable to those subject to it.

In other words, the transition from rule-based or classic machine learning moderation to current AI-driven moderation has shown that the traditional rule-versus-standard dichotomy may not fit well into this new context, because while AI-driven moderation may present considerable similarity in its prediction to that of human (standard-based) 'judges' (indeed, it is exactly what it is trained to do), it is ultimately a simulation that differs from human judgement in important ways, as explained above. This development highlights a need to think differently about how rules and standards function in an environment that is increasingly dominated by machine learning and AI technologies. Reconceptualising the normative frame is not merely a theoretical challenge but is also a significant practical challenge to ensuring that the regulatory environment keeps pace with technological advancements, and to advancing effective, scrutinised and accountable content moderation.

[161] We do not argue that human decision-makers – whether judges, moderators or users – neatly embody the classic rules–standards dichotomy; nor do we deny that human cognition may also operate through large constellations of tacit, context-sensitive heuristics. Our argument instead is that LLM-based moderation systems make salient a specific *engineering* realisation of such 'rules by the millions', the scale, opacity and technical form of which raise regulatory and accountability challenges that are not captured by the existing rules-versus-standards vocabulary.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

4.4. Aligning content moderation policies and AI-enabled enforcement: The way forward

Recent advancements in LLMs raise new challenges in how to align human and machine judgements and how to effectively regulate the latter. Still, there are no simple text-book solutions for bridging the governance gap between the formulation of content moderation policies and their AI-driven enforcement. In what follows, we make several recommendations relating to content moderation norms and practices that could pave the way for a more technologically informed approach to digital regulation:

(1) *Digital transparency and awareness*: It is crucial to illuminate the asymmetry between content moderation policies and their actual enforcement. This requires a concerted effort to educate and engage not only users and developers, but also civil society organisations and regulators. By enhancing our understanding of this disparity, we can pave the way for more equitable and consistent digital governance. This may involve actively sharing insights about the operational nuances of AI in content moderation and fostering a broad-based dialogue on its implications.
(2) *Collaborative and informed policymaking*: Policymaking in the realm of content moderation should be a collaborative venture, integrating insights from AI experts, technologists and other stakeholders. Such an interdisciplinary approach ensures that policies are not only practically enforceable but are also in sync with the evolving landscape of AI capabilities and constraints. This alignment is not merely about enforcing existing policies; it is about cultivating innovation and creativity in addressing gaps between content moderation policies and their manner of enforcement.
(3) *Strengthened regulatory action and oversight*: We advocate the establishment of independent regulatory bodies with expertise in AI and digital communication technologies. These entities could play a crucial role in supervising the development and deployment of AI-based enforcement tools for content moderation, with a view to ensuring they adhere to applicable legal and ethical standards and are congruent with content moderation policies. Regular mandatory audits of AI moderation tools should be conducted by oversight agencies, with the results made publicly available to enhance accountability and stimulate informed public discourse. Furthermore, we suggest reinforcing user appeal mechanisms, including empowered and informed oversight mechanisms capable of meaningfully assessing the intricacies of automated enforcement.
(4) *Revisiting legal theories in the digital age*: Our research suggests a need to re-evaluate and potentially reformulate traditional legal distinctions between rules and standards, especially in the context of AI-driven content moderation and its 'rules by the millions' features. This revision should encompass recognition of the unique technological attributes, opportunities and limitations inherent in AI-driven decision-making. Legal frameworks need to evolve to accommodate these technological realities, ensuring that they remain relevant and effective in the ever-changing digital landscape.

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

Adopting our recommendations, it is fair to assume, will not avoid all challenges. Such challenges may include, among others, fragmentation in the applicable global legal frameworks, intellectual property claims that might create hurdles for AI audits, lobbying efforts, and political incentives and agendas.[162] Still, we believe that our recommendations are feasible. The last few years have seen general progress in most categories relevant to our recommendations, spanning transparency and awareness, collaborative approaches to policy-making, and more informed regulation and oversight in digital contexts. New regulations, international law initiatives, and soft-law adoption assign responsibilities and duties to digital platforms and contribute to this progress.[163]

The regulatory efforts of digital stakeholders are particularly evident in the EU (though it should be noted that the EU approach already influences other jurisdictions[164] and may further influence them in the future).[165] One EU regulation that is particularly relevant to our recommendations is the Digital Services Act (DSA), which entered into full force in February 2024.[166] The DSA regulates online intermediaries and strives to enhance transparency and human rights, and to curb harmful content.[167] Somewhat similar to the risk-based approach of the EU AI Regulation,[168] it applies different levels of obligation to online gatekeepers, with very large online platforms (VLOPs) such as Facebook and Instagram placed at the top of this pyramid.[169]

The DSA requires digital platforms to provide users with reasons when they restrict access to users' content and to submit these reasons to the publicly available and machine-readable DSA Transparency Database.[170] This regulation also enhances transparency by allowing vetted researchers to access certain non-public VLOP data,[171] and by providing a larger group of researchers with access to publicly

[162] See, for instance, Esmat Zaidan and Imad Antoine Ibrahim, 'AI Governance in a Complex and Rapidly Changing Regulatory Landscape: A Global Perspective' (2024) 11 *Humanities & Social Sciences Communications* 1121.

[163] Stanford Institute for Human-Centered Artificial Intelligence, 'The 2025 AI Index Report', Stanford University, 325–52, https://hai.stanford.edu/ai-index/2025-ai-index-report.

[164] Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (Oxford University Press 2020).

[165] For the opinion that the principles of the DSA should be adopted in more territories, including the US, see Martin Husovec, 'Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation of Global Internet Rules' (2023) 38 *Berkeley Technology Law Journal* 883.

[166] Digital Services Act (n 84).

[167] Digital Services Act (n 84), and European Parliament, Directorate-General for Communication, 'EU Digital Markets Act and Digital Services Act Explained', 26 June 2025, https://www.europarl.europa.eu/pdfs/news/expert/2021/12/story/20211209STO19124/20211209STO19124_en.pdf.

[168] Regulation (EU) 2024/1689 of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence [2024] OJ L 2024/1689 (AI Regulation).

[169] See Digital Services Act (n 84), and European Commission, 'Supervision of the Designated Very Large Online Platforms and Search Engines under DSA', Shaping Europe's Digital Future, updated 17 February 2026, https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses. Very Large Search Engines (VLSEs) are also addressed in the same manner in the DSA.

[170] Digital Services Act (n 84) arts 17, 24(5).

[171] ibid art 40(4). The Delegated Act that was enacted under the DSA, and entered into force on 31 October 2025, sets and clarifies the procedure in this context: European Commission, Joint Research Centre, 'FAQs: DSA Data Access for Researchers', European Centre for Algorithmic Transparency,

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

available data.[172] Moreover, in alignment with the directions we offer to pursue, it obliges digital platforms to set up internal complaint mechanisms,[173] as well as out-of-court dispute-settlement avenues.[174] The DSA also contributes to more meaningful oversight, including through designated enforcement bodies, around the use of AI technologies by digital platforms.[175] In addition, its drafting processes included feedback from various stakeholders, thus generally resonating with our suggested participatory and informed policy-making processes.[176]

Current legal frameworks, including the DSA, still do not provide an appropriate regulatory solution for the tension that we point out between the content policies of digital platforms and their automated enforcement. However, they mark an encouraging prospect in this regard and strengthen the feasibility of our recommendations.

5. Conclusion

This article explores the increasing tension between rule-oriented content policies and standard-like AI enforcement in online content moderation. It focuses on Facebook's development of more precise hate speech rules (at least until 2025) and the concurrent rise of AI, particularly LLMs, revealing a paradoxical shift towards standard-like application in practice. This creates a significant governance gap in online spaces.

To address these challenges, we present policy recommendations in two main areas: enhancing awareness and integrating technology into policy frameworks on the one hand, and enforcing robust regulatory action and oversight on the other hand. On the regulatory side, we advocate establishing independent bodies specialising in AI and digital communication technologies. We also recommend robust mechanisms for user appeals and feedback on moderation decisions, to enhance fairness and accountability in digital content moderation.

Our findings indicate the need to reassess and possibly redefine the conventional legal distinctions between rules and standards, especially in the context of digital and AI-driven content moderation, and their 'rules by the millions' characteristics. This re-evaluation should consider the unique dynamics of AI-driven decision-making and adapt legal frameworks accordingly.

Although our study focuses primarily on Facebook's hate speech policies, the insights and conclusions are likely to be relevant to other major online platforms

3 July 2025, https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2025-07-03.

[172] Digital Services Act (n 84) art 40(12).

[173] ibid art 20.

[174] ibid art 21.

[175] The Commission, along with national Digital Service Coordinators (DSCs) in each Member State, are responsible for supervising and monitoring the DSA: European Commission, 'Digital Services Coordinators', Shaping Europe's Digital Future, https://digital-strategy.ec.europa.eu/en/policies/dsa-dscs. See also in this context the EU preliminary findings regarding DSA violations, including with regard to the data submitted by the platforms and to users' ability to challenge content decisions they make: European Commission, 'Commission Preliminarily Finds TikTok and Meta in Breach of Their Transparency Obligations under the Digital Services Act', Press Release IP/25/2503, 24 October 2025, https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503.

[176] European Commission, DSA Researchers Access (n 171).

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press

and various types of content regulation. Recognising this governance gap between content moderation policies and their manner of enforcement is vital for developing more effective, consistent and fair digital self-regulation schemes.

Future research should explore strategies for building more resilient digital platforms and the implications of technological advancements for traditional legal theories. This will foster innovative approaches that align with our rapidly changing technological landscape. The rise of LLMs in content moderation enforcement may signal a paradigm shift, potentially redefining the traditional distinction between rules and standards. This calls for a re-evaluation of these concepts in the digital, AI-driven context. As the digital era evolves, our understanding and application of foundational legal theories must also adapt to remain relevant and effective in emerging technological realities.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0021223726100156.

Acknowledgements. The authors would like to thank Tarleton Gillespie, Anne Meuwese, Colin Provost, Itay Ravid, Paul Röttger, Bertie Vidgen, Hannah Bloch-Wehba, the participants of the Platform Governance Research Network Conference 2023, the participants of the Biennial Conference of the Standing Group on Regulatory Governance Participants 2023, and the participants of an Oxford Ethics in AI 2024 Workshop for their helpful comments, discussions and suggestions relating to previous drafts. All views expressed here, as well as any errors, are, of course, our own.

Funding statement. Yuval Shany's involvement in the research was facilitated by ERC Grant No. 101054745: *The Three Generations of Digital Human Rights* (DigitalHRGeneration3), https://bit.ly/3rDGTT6.

Competing interests. The authors declare none.

Cite this article: Renana Keydar, Noa Mor, Yuval Shany and Omri Abend, 'Bending the Rules: On Large Language Models and Content Moderation' (2026) 59 *Israel Law Review* 96–133, https://doi.org/10.1017/S0021223726100156

https://doi.org/10.1017/S0021223726100156 Published online by Cambridge University Press