Shortcuts and Shortfalls in Meta's Content Moderation Practices:

A Glimpse from its Oversight Board's First Year of Operation

Janny Leung[1]

Abstract: Social media companies regulate more speech than any government does, and yet how they moderate content on their platforms receives little public scrutiny. Two years ago, Meta (formerly Facebook) set up an oversight body, called the *Oversight Board*, that handles final appeals of content moderation decisions and issues policy recommendations. This article sets out to examine Meta's approach to content moderation and the role of the Board in steering changes, as revealed by the first 20 decisions that the Board published during its first year of operation. The study identifies interpretive shortcuts that Meta's content moderators frequently deployed, which led to pragmatic deficiency in their decisions. These interpretive shortcuts are discussed under the notions of decontextualisation, literalisation, and monomodal orientation. Further analysis reveals that these shortcuts are design features rather than bugs in the content moderation system, which is geared toward efficiency and scalability. The article concludes by discussing the challenge of adopting a universal approach to analysing speaker intentionality, warning against a technochauvinistic approach to content moderation, and urging the expansion of the Board's power to not only focus on outcomes but also processes.

Keywords: Content moderation, Social media, Meta, Oversight Board, Context.

1. Private governance of online speech

In January 2021, when the now former president of the United States Donald Trump was blocked from accessing his social media accounts, the world began to wake up to how much power private companies wield in controlling public discourse. Meta[2], which owns the social media platform Facebook, has been policing the speech of its 3.64 billion users[3]—at a larger scale than any national government or intergovernmental organisation has ever done. As noted by law professor Jeffrey Rosen, "Facebook has more power in determining who can speak and who can be heard around the globe than any Supreme Court justice, any king or any president".[4]

[1] *Janny H. C. Leung—Wilfrid Laurier University.*

[2] The company Facebook changed its name to Meta in October 2021. For the sake of consistency, I will use the term Meta to refer to the company and Facebook to refer to the product, except in citations.

[3] As of Q1 of 2022. This only takes into account Facebook. Meta also owns and moderates content on Instagram.

[4] M. HELFT, *Facebook Wrestles With Free Speech and Civility*, in *The New York Times*, 13 December 2010, https://www.nytimes.com/2010/12/13/technology/13facebook.html.

Also happened in the same month is another notable event concerning the private governance of online speech: the Oversight Board (hereafter the Board), newly established by Meta, issued its "rulings" for the first time. The Board represents an attempt to strengthen Meta's public accountability in its content moderation practices. Decisions published by the Board also offer a peek at Meta's internal operation, as content moderation in social media companies typically happens behind closed doors under the companies' supervision. Other than through journalistic investigations[5] or cases that caught media attention, the public rarely have access to how content moderation rules are applied in practice.

This paper sets out to examine Meta's approach to content moderation and the role of the Board in steering changes, as revealed by the first 20 decisions that the Board published during its first year of operation. Section 2 of this paper reviews content moderation practices at Meta and its Board, providing background information about how content moderation is done and what appeal mechanisms exist. Section 3 examines the first 20 decisions that the Board published. These decisions reveal not only the outcome of individual cases, but also offer a glimpse of how Meta moderates content. Drawing from linguistic analyses, I identify interpretive shortcuts frequently taken in Meta's content moderation practices, which result in interpretations that could deviate from intended meaning. Section 4 discusses the extent to which these shortfalls are attributable to the poverty of context in the online communication environment or to the design of the content moderation process, as well as Meta's policy of defaulting towards removal. Section 5 concludes by critiquing Meta's normalization of pragmatic deficiency and evaluating how the shortfalls could be met.

2. Content moderation and the appeal system

Many legislative bodies are concerned with the need to regulate online expressions and how social media companies moderate content and treat their users. In the European Union, the General Data Protection Regulation (GDPR) targets privacy and data collection breaches, and the Digital Services Act (DSA), which came into force in 2022, imposes a set of obligations on gatekeeping digital platforms, prohibiting unfair practices, and also obliging platforms to cooperate with "trusted flaggers" of illegal content (Art. 19) and to offer users the opportunity to challenge content moderation decisions (Art. 18). Both France and Germany have recently enacted laws that combat illegal hate speech on social media, and the United Kingdom is in the process of enacting an Online Safety Bill that requires platforms to assess risks associated with some categories of legal but harmful speech. A similar legislation on online harms is also being discussed and developed in Canada.[6] Even in the US, where Section 230 of the Communication Decency Act grants online intermediaries broad immunity from liability for user-generated content posted on their platforms, there is no shortage of advocacy for content-based regulation.[7] Apart from national governments, civil society groups, the media, and intergovernmental organizations also put pressure on platforms to hold them publicly accountable.

As external pressure mounts, online intermediaries like Meta have developed more and more elaborate content moderation structures. Although Meta originates from the United States and its company culture is deeply rooted in American free speech norms, it operates globally and needs to navigate local regulation. Through "geo-blocking", which determine whether users can post or view certain content based on their internet protocol (IP) addresses, the company puts geographical restrictions to content in order to comply with local laws.

---

[5] Such as S. T. ROBERTS, *Behind the Screen: Content Moderation in the Shadows of Social Media*, New Haven, 2019, pp. 1–266.

[6] https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content.html.

[7] See for example, M. A. FRANKS, *The Cult of the Constitution*, Stanford, 2020. The First Amendment implications on content moderation are an unsettled debate, centring on whether online intermediaries act like the state and are therefore constrained by First Amendment, whether they function like a speech conduit like radio and television and therefore attract regulation, or whether the companies enjoy First Amendment protection as speakers. See K. KLONICK, *The New Governors: The People, Rules, and Processes Governing Online Speech*, in *Harvard Law Review*, 131, 2018, pp. 1958–1670.

Meta moderates much more content than is needed to comply with local laws. In other words, a significant portion of content moderation is done in accordance with the platform's internal rules. My description of Meta's content moderation below is largely based on Klonick's work[8], which offers a comprehensive account of Meta's content moderation practices and its creation of the Board. Based on its Community Standards[9], Meta restricts speech that involves violent and criminal behaviour or poses safety concerns, as well as speech that it considers objectionable or inauthentic.

Content moderation may be conducted ex ante before content is published, or it may be ex post, after it has been published. Ex ante moderation is done through automated detection by a combination of automated tools that screen for extremism and hate speech, and "hash technology" that compares the newly uploaded content with a database of known impermissible content. Apart from an army of 15000 human content moderators who manually look for and delete content that violates its Community Standards, Facebook and Instagram users also contribute to the moderation process by reporting content, which will then be reviewed by a human content moderator; outcome of the moderation feeds back into Meta's algorithms as data points.

Meta publishes a Community Standards Enforcement Report every quarter, as an effort to demonstrate transparency in content moderation. For example, in Q1 of 2022, for the violating category of Adult Nudity and Sexual Activity alone, Meta acted on 31M pieces of content. Of the violating content they acted on, 96.7% of them were identified proactively before being reported by its users.[10] In the same quarter, 287k of the actioned content was restored after being confirmed to be false positives. Since the data represent the combined result of human and automated content moderation, they do not reveal how much work was performed by each and how much correction human moderators made to automated decisions. Neither do they tell us how much time passes before action is taken, and how many impressions pieces of violating content made before they are taken down. Despite the volume and regularity of data-sharing, such data do not always reveal the efficacy of content moderation processes.[11]

In 2018, Zuckerberg acknowledged that moderation decisions were wrong in more than 10% of cases.[12] Meta's automated detection works relatively well in detecting image-based copyright violations and child pornography, which are based on similarity matching with existing databases, but struggles more with text-based content such as hate speech and bullying, which involves the open texture and contextual dependence of language. As Eric Goldman observes, content is particularly difficult to classify if understanding it requires "extrinsic information"—that is, information *outside* the image, video, audio, or text.[13] What he refers to—meaning that arises from context and negotiated *in situ*, is known as pragmatics in the study of language and communication. As we will see, pragmatic deficiency is indeed a problem in content moderation.

According to Klonick, Meta's human moderators are organised into three tiers: Tier 3 moderators are employees and contract workers around the world who do the bulk of day-to-day reviewing; at Tier 2 are experienced or specialised moderators who review escalated or prioritized content, as well as a

[8] K. KLONICK, *The New Governors: The People, Rules, and Processes Governing Online Speech*, cit., pp. 1958–1670; K. KLONICK, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, in *The Yale Law Journal*, 129, 2020, pp. 2418–99.

[9] Until April 2018 Community Standards were different from the internal rules actually used by content moderators.

[10] Community Standards Enforcement Report, Q1 2022. Available at https://transparency.fb.com/data/community-standards-enforcement/. The proactive rate is calculated based on the number of pieces of content actioned that they found and flagged before users reported them, divided by the total number of pieces of content actioned.

[11] See discussion of transparency theatre E. DOUEK, *Content Moderation as Systems Thinking*, in *Harvard Law Review*, vol. 136, number 2, 2022, pp. 526-607.

[12] P. M. BARRETT, *Who Moderates the Social Media Giants? A Call to End Outsourcing*, New York, June 2020, pp. 1–32.

[13] J. VINCENT, *AI Won't Relieve the Misery of Facebook's Human Moderators*, in *The Verge*, 27 February 2019, https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms.

randomized sample of Tier 3 decisions; Tier 1 moderation happens at the legal or policy headquarters.[14] Currently, although users could tell Meta why they disagreed with the platform's content removal decision, they are not always given the option to appeal.

To improve Meta's public accountability in their regulation of online speech, the Board began its operation in 2020 as an independent body that selectively reviews Meta's content moderation decisions. Created through a trust funded by Facebook, the Board operates at arm's length from Meta, working like a Supreme Court.[15] Its members include law professors, journalists, a former Prime Minister, and a Nobel laureate, coming from diverse geographical locations. Meta's move to set up the Board may be seen as a form of power sharing, just as it could be seen as a public relations stunt, a convenient scapegoat for controversial decisions, a way of deflecting regulatory pressures, or an attempt to build or retain user trust. As Klonick suggests, Meta has "myriad incentives" in creating an oversight body.[16]

The Board's powers are set out in its Charter and Bylaws. Since the Board only selects a very limited number of cases to review, it seeks to consider cases that have the greatest potential to guide future decisions and policies (Art. 2.1 of Charter). The Board's decisions to allow or remove content are binding on Meta; it can also make policy recommendations, which Meta is not obliged to accept but has committed to considering (Art. 4 of Charter).

The Board's decision-making is informed by Facebook's Community Standards, its values, and relevant Human Rights Standards. Facebook's values, as outlined in the introduction to the Community Standards, include the paramount value of "Voice", which may be limited in service of four other values: "Authenticity", "Dignity", "Privacy", and "Safety". In terms of Human Rights Standards, the Board draws from the UN Guiding Principles on Business and Human Rights (UNGPs), which articulate a voluntary framework for the human rights responsibilities of private businesses. The international human rights standards that the Board frequently relies on include the right to freedom of expression (Art. 19 of the International Covenant on Civil and Political Rights, or ICCPR; General Comment No. 34 of the Human Rights Committee 2011), the right to non-discrimination (ICCPR Art. 2 and 26), the right to life and security (ICCPR Art. 6 and 9). According to Douek, companies are quick to adopt the language of International Human Rights Law into their content moderation governance, but the impact of such adoption is quite limited.[17]

3. The Board's first year of adjudication

3.1. Overview

The Board started to work on cases in 2020 and issued its first decisions in January 2021. It published 20 decisions in 2021[18], averaging 1.6 cases per month. These cases were selected among over a million user appeals and a few dozens of referrals from Meta. In its first year of operation, the Board overturned Meta's decision in 14 out of 20 of cases, or 70% of the time (see Fig. 1). As an independent grievance mechanism, the Board takes pride in the frequency at which it overturns Meta's decision—it publishes a similar figure (16 overturned decisions out of the first 22 cases) on its website to illustrate "the Power of the Board" (https://oversightboard.com/appeals-process/).

---

[14] K. KLONICK, *The New Governors: The People, Rules, and Processes Governing Online Speech*, cit., pp. 1639–1641.
[15] Zuckerberg stated in an interview that he envisioned the Oversight Board as "a Supreme Court, that is made up of independent folks who don't work for Facebook, who ultimately make the final judgment call on what should be acceptable speech in a community that reflects the social norms and values of people all around the world." E. KLEIN, *Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next*, in *Vox*, 2 April 2018, https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge.
[16] K. KLONICK, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, cit., pp. 2426–27.
[17] E. DOUEK, *The Limits of International Law in Content Moderation*, in *UC Irvine Journal of International, Transnational, and Comparative Law*, 6(1), 2021, pp. 37–76.
[18] This excludes a case (2020-001-FB-UA) on hate speech in Malaysia that the Board selected but did not adjudicate on, as it became unavailable for review after the user deleted the post.

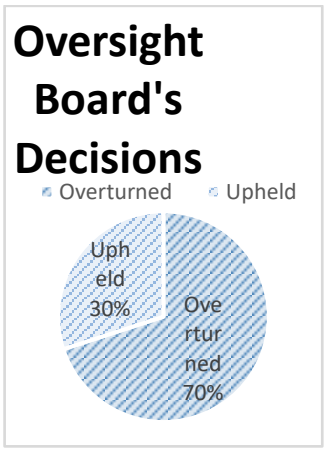

Fig. 1. Percentage of Meta's decisions that were overturned or upheld by the Oversight Board between Q4 2020 and Q4 2021

Of the 20 cases that the Board selected, hate speech is the most frequent type of violation, occurring in 50% of the cases, as shown in Fig. 2. The two other most frequent violations are Dangerous Individuals and Organisations and Violence and Incitement, occurring in 25% and 20% of the cases respectively. Hate speech is also the type of violation that generated most user appeals (36%), according to the Oversight Board's Transparency Reports.[19] However, the same reports indicate that Violence and Incitement and Dangerous Individuals and Organisations account for 13% and 6% of user appeals only. The Board has only handled one case involving Bullying and Harassment (5%), even though this type of violation generated 31% of user appeals. The Board's case selection reflects its priorities and its perception of policy areas that require more urgent guidance.

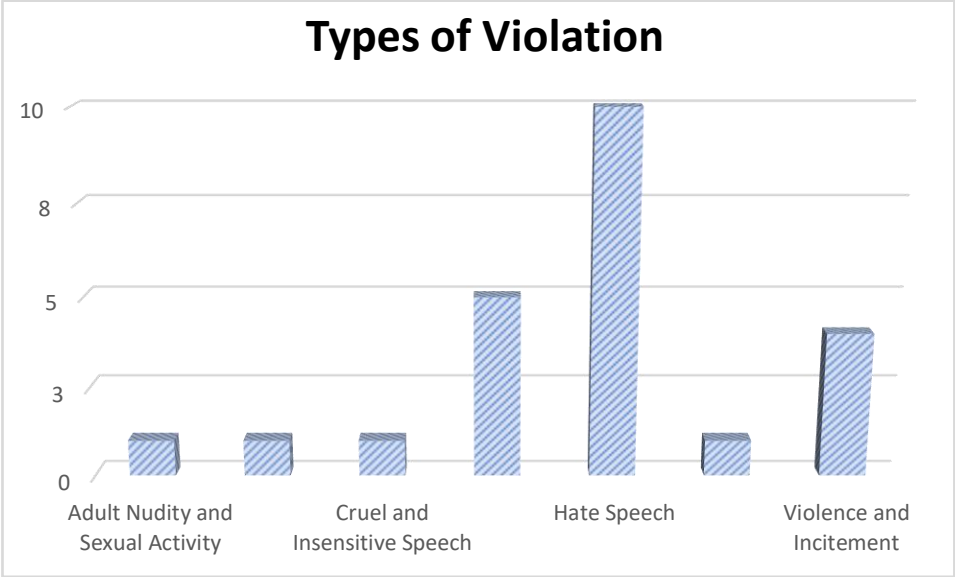

Fig. 2. Type of violation in the first 20 cases decided by the Oversight Board, by number of cases. Note that there may be multiple types of violation involved in a single case

Closely related to types of violation, it is observed that the allegedly violating content in almost all of the 20 cases was overtly political speech. The only exceptions are 2020-004-IG-UA and 2021-013-IG-UA, which involve a breast cancer awareness campaign[20] and discussion of non-medical drugs. The

[19] The Oversight Board, *Oversight Board Transparency Reports Q4 2020, Q1 & Q2 2021*, October 2021.
[20] But of course, the Community Standard that censors of female nipples lies at the core of gender politics, as it reflects the oversexualisation of the female body.

Board has noted in various cases that the ICCPR gives heightened protection to political expressions. It is not clear whether such prominence of political speech in the 20 cases reflects the general vulnerability of political speech to Meta's content moderation practices, the cases' potential impact on public discourse, or other priorities of the Board.

3.2. Enforcement errors, misplaced policy, and errors perpetuated through automation

While it is not unexpected that Meta and the Board disagree about the outcome of cases, what is striking about these cases is the frequency at which errors in Meta's content moderation processes are discovered as a result of the Board's review. Of the 20 cases the Board adjudicated on during its first year of operation, 4 of them were referred to the Board by Meta (or Facebook Referral, FBR, cases), and 16 of them were User Appeal (UA) cases. Unsurprisingly, Meta's moderators have given comparatively thorough attention to the FBR cases before referring them to the Board for further guidance. Most of the errors revealed were found in the UA cases. Among the 16 UA cases that the Board handled, Meta reversed its decision in 6 of them after the cases were selected by the Board for review. This represents 38% of the UA cases selected. Moreover, as shown in Table 1 below, most of these cases had been reviewed more than once internally in Meta before the "enforcement errors" were discovered.

Case Number	**Type of Violation**	**Number of "Enforcement Errors" Prior to Reversal**	**Content Moderation Performed by**
2020-004-IG-UA	Adult Nudity and Sexual Activity	1	Automated System (1)
2021-003-FB-UA	Dangerous Individuals and Organisations	1	Human (1)
2021-006-IG-UA	Dangerous Individuals and Organisations	2	Human (2)
2021-009-FB-UA	Dangerous Individuals and Organisations	2	Human (2)
2021-012-FB-UA	Hate Speech	4	Automated System (2) + Human (2)
2021-014-FB-UA	Hate Speech; Violence and Incitement	3	Automated System (1) + Human (2)

Table 1. Decisions that Meta reversed after the Oversight Board selected the cases

The Board's transparency reports revealed that during its first year of operation, Meta actually reversed its original decision in 38 cases after they were shortlisted by the Board.[21] The majority of these

[21] The Oversight Board, *op. cit*.

reversals concern Hate Speech (47.4%) and Dangerous Individuals and Organisations (31.6%). The Board only proceeded to adjudicate on 6 of them, as tabled above. In all but one instance, the Board agreed with the reversal rather than the original decision. All 6 cases involved restoring content after removal, rather than removing content after Meta decided to leave it up.

These enforcement errors are significant because the erroneously removed content would not be discovered and restored if the Board had not selected these cases. Many of these errors appeared unambiguously to the Board as mistakes that should not have been made. If Meta's moderators had reviewed the content carefully, they would not have made such errors, raising questions about the adequacy of the moderation process. Moreover, in multiple cases, the impact of the erroneous decisions was amplified as they became training data in automated moderation processes (see 2021-006-IG-UA, 2021-007-FB-UA and 2021-012-FB-UA).

The rate of reversal is much higher than the rate at which Meta restores content after removal as indicated in its transparency reports.[22] The rate of reversal is however not the same as error rate, as Board might have selected these UA cases for review precisely because the action taken blatantly contradict Meta's Community Standards. That said, it is still alarming that Meta was not aware of and could not explain how these errors occurred.

The first time such a reversal happened (in 2020-004-IG-UA), Meta claimed that the Board should decline to hear the case, as there was no longer disagreement between the user and the company. The Board refused, arguing that it was empowered to hear the case provided that the disagreement existed when the user exhausted Facebook's internal appeal process (Art. 2, Section 1 of the Charter). This is reasonable because hearing the case could bring impact beyond the content of the case. Once it was decided that the Board could still hear cases after moderation decisions are reversed, it is clear that reversals do not stop enforcement errors from being publicised. Why would Meta want to reverse decisions prior to the Board's review then? One possible motivation is that the reversals allow Meta to focus its rationale on the revised decision, rather than on how the error happened. As the Board notes in 2021-012-FB-UA, "(i)t is unhelpful that in these cases, Meta focuses its rationale entirely on its revised decision, explaining what should have happened to the user's content, while inviting the Board to uphold this as the company's 'ultimate' decision".

Apart from enforcement errors, the Board's queries also led to the discovery of communication errors within Meta's content moderation teams and with platform users. For example, in 2021-012-FB-UA and 2021-014-FB-UA, the users were not informed that their appealed content had been restored, and Meta did not send the notifications until the Board asked for the content of the messages. In 2021-013-IG-UA, the user received a wrong message about their appeal. In 2021-006-IG-UA, for three years until discovered by the Board, an internal guidance on policy exception for Dangerous Individuals and Organisations was misplaced, not shared within the policy team and therefore not applied. This means that content that should have fallen within the exception had been removed for three years with no accountability whatsoever.

3.3. Interpretive shortcuts and pragmatic deficiency

The decisions published by the Board provide a rare opportunity to examine not only content moderation decisions made by Meta but what went into the decision-making process: what factors were considered, what were not, and how competing considerations were weighed. A recurrent criticism the Board makes about Meta's content moderation practices concerns the deficiency of its contextual analysis. Drawing from relevant linguistic concepts, the analysis presented here breaks down the nature of such pragmatic deficiency by outlining the interpretive shortcuts that its content moderation took. I discuss these shortcuts under the headings of decontextualisation, literalisation, and monomodal orientation. While these shortcuts are conceptually distinct, they are interrelated in practice.

[22] See https://transparency.fb.com/data/community-standards-enforcement/.

1. Decontextualisation

Decontextualisation refers to the interpretation of a sign or a text in isolation from the context that it is embedded in. Here I will focus on two types of context: discourse context and situational context. Discourse context is the larger text that an utterance[23] is part of. Situational context refers to the time, place, and other aspects of the environment in which an utterance takes place, such as relationships among discourse participants and socio-political climate.

First let us consider discourse context, which Meta seems unwilling to engage with in some of the cases examined. In October 2020, a user posted a quote which was incorrectly attributed to Joseph Goebbels, claiming that arguments should appeal to emotions and instincts rather than to intellectuals[24]. The quote further stated that truth does not matter and is subordinate to tactics and psychology. The post was a plain text, written in English, without any accompanying visual representation of Goebbels or Nazism. In a statement submitted to the Board, the user explains that their post was meant to be a political commentary, which draws a comparison between fascism and the presidency of Donald Trump. Comments to the post indicate that the user's friends understood his intention. Meta tells the Board that it keeps an internal list of individuals and organizations that "proclaim a violent mission or are engaged in violence" and removes content that expresses support or praise for these individuals and organisations in order to prevent and disrupt "real-world harm". It has designated the Nazi party as a hate organisation and Joseph Goebbels, the Reich Minister of Propaganda in Nazi Germany, as a dangerous individual. Although the post was flagged for mentioning Goebbels, there was no explicit indication or contextual cue which suggested that the author supported or praised him. According to its submission, Meta treats all content that quotes (regardless of accuracy) a designated dangerous individual as an expression of praise or support for that individual unless the user provides additional context to make their intent explicit[25]. Meta states that they only review the post itself when making a moderation decision, without considering reactions or comments to the post—even though they could provide important clues to intentionality as speakers orient their speech towards their target audience.[26] Since our ability to draw inferences about utterances relies on contextual enrichment, ignoring discourse context will severely limit our ability to understand an utterance.[27] Interestingly, as a response to the Board's decision in this case, Meta updated its policy to more explicitly require "people to clearly indicate their intent" when discussing dangerous individuals and organisations and warns that "if the intent is unclear, we may remove content"[28], which actually increases Meta's discretion in cases where intent is not stated clearly.

In another case, Meta ignored discourse context that would have helped to resolve a critical ambiguity in the post. In 2021-007-FB-UA, Meta removed a Burmese post based on its Hate Speech

[23] An utterance is a unit of speech in context; it is used in contrast with a sentence in formal linguistics. A sentence can be repeated but an utterance cannot, because the context necessarily changes.

[24] Case Decision 2020-005-FB-UA, Oversight Board, available at https://oversightboard.com/decision/FB-2RDRCAVQ/.

[25] This is the language used in its Community Standards, which is different from the rules used internally in the company: "We do not allow symbols that represent any of the above organizations or individuals to be shared on our platform without context that condemns or neutrally discusses the content". Available at https://www.facebook.com/communitystandards/dangerous_individuals_organizations. A similar presumption is adopted for Hate Speech as well: "We recognize that people sometimes share content that includes someone else's hate speech to condemn it or raise awareness. In other cases, speech that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If intention is unclear, we may remove content". Available at https://www.facebook.com/communitystandards/hate_speech.

[26] A. HALEVY et al., *Preserving Integrity in Online Social Networks*, in *Proceedings of Facebook AI*, n. ACM, New York, 2020, http://arxiv.org/abs/2009.10311 gives an example where user reaction to a suicide post can be much more telling than the language of the post, for the user's immediate social network often has knowledge of the urgency of the situation.

[27] J. H. C. LEUNG, *The Audience Problem in Online Speech Crimes*, in *Journal of International Media & Entertainment Law*, 9, n. 2, 2021, pp. 189–234.

[28] See discussion in Case Decision 2021-009-FB-UA.

Community Standard. The violating part translates into English as "Hong Kong people, because the fucking Chinese tortured them, changed their banking to UK and now (the Chinese), they cannot touch them." The question is whether "fucking Chinese" constitutes hate speech, which under Meta's Community Standard refers to content targeting a person or group of people based on their race, ethnicity, or national origin with "profane terms or phrases with the intent to insult". At the crux of the case is the lexical ambiguity of the Burmese word "ta-yote" ("Chinese"), which could be used to refer to China as a country and/or Chinese as a people. Four Burmese-speaking content reviewers at Meta found the content to be hate speech. Meta stated that because of difficulties in "determining intent at scale", it considers the phrase "fucking Chinese" as referring to both Chinese people and the Chinese government unless the user provides additional context that suggests otherwise. The Board's analysis suggests that the additional context is right there: the immediate discourse context refers to China's policies in Hong Kong, and the wider post discusses ways of limiting financing to the Myanmar military, following the coup that happened on 1 February 2021. The Board's translators also identified terms commonly used by the Myanmar government and the Chinese embassy to address each other, which are lexical cues in the post that provide further evidence that the Chinese state is the target referent. The Board concludes that the phrase clearly targets the Chinese state rather than Chinese people, and therefore does not constitute hate speech. The intention of the post is to discuss the Chinese government's role in Myanmar, not to attack Chinese people based on their race, ethnicity or national origin. Given that Meta's four content reviewers all found the post to be violating and missed all the discourse contexts that could have resolved the lexical ambiguity, the Board "questions the adequacy of Facebook's internal guidance, resources and training provided to content moderators".

By contrast, the divergence between Meta and the Board in the following two cases can be largely attributed to how they approached situational context. Both were cases that Meta referred to the Board. Case Decision 2020-006-FB-FBR[29] concerns a post that Meta removed for violating its misinformation and imminent harm rule (part of its Community Standard on Violence and Incitement). The post, shared in a public Facebook group related to Covid-19 with 500,000 members, contained a video and an accompanying text in French, which criticized the Agence Nationale de Sécurité du Médicament (the French agency responsible for regulating health products) for not authorizing the combined use of hydroxychloroquine and azithromycin as a cure for Covid-19. The user questioned what the society had to lose by allowing the emergency use of a "harmless drug". Meta argued that the claim that there is a cure for Covid-19 could lead people to ignore health guidance or attempt to self-medicate. The Board overturned Meta's decision and ordered that the content be reinstated, arguing that Meta has failed to demonstrate that the post rises to the level of imminent harm, and that the platform could have chosen a less intrusive intervention (such as labelling the content) than content removal. The misinformation and imminent harm rule also "require[s] additional information and/or context to enforce". Not all misinformation leads to imminent physical harm; context is crucial in assessing risk. According to the experts that the Board consulted, combining the drugs that the user mentioned in their post may be harmful, but these drugs are not available without a prescription in France.[30] Ultimately the Board disagreed with Meta about what context is needed in assessing imminent harm. Both engaged external assistance, though they sought different types of expertise—Meta consulted global health experts and the Board sought expertise in local context.

The case 2020-007-FB-FBR concerns a post in an Indian Muslim group, which contains a meme featuring an image depicting a Turkish television show character holding a sheathed sword. The text overlay in Hindi translates into English as "if the tongue of the kafir starts against the Prophet, then the sword should be taken out of the sheath". The post included hashtags that refer to President Emmanuel Macron of France as the devil and calls for the boycott of French products. Meta initially did not remove the post after two users reported it for hate speech and for violence and incitement. However, a third-

[29] Oversight Board, case decision 2020-006-FB-FBR, https://oversightboard.com/decision/FB-XWJQBU9A/.

[30] The most elaborate reason that the Board gives for deciding that the misinformation does not meet the standard of "imminent" harm is that the alleged cure (unlike other alleged cures such as cold water or bleach) is not readily available to the audience vulnerable to the message. However, the Board immediately and rightly notes that there may well be French speakers outside of France in the public group concerned.

party partner flagged it, and Meta's local policy team agreed that the post was potentially threatening. Meta interpreted the post as a veiled threat against "kafirs" (a pejorative term referring to non-believers) and removed it under its Community Standard on Violence and Incitement, but also referred the case to the Oversight Board for guidance. The contexts that Meta was concerned with include religious tensions in India related to the Charlie Hebdo trials in France and to elections that were happening in the Indian state of Bihar. It also noted anti-Muslim sentiment following the Christchurch attack in New Zealand. The Board was not satisfied with how Meta arrived at the implicit meaning of the post, however. Despite the visual reference to a sword in the post, the Board considered the call for a boycott of French products a call for non-violent action. Similarly, the Board found that protests in reaction to the French trials were not reported to be violent, and that the Bihar elections were not marked by religiously motivated violence. In other words, Meta's contextual analysis focused on major global events and broad climate, while the Board devoted more attention to scrutinising the immediate discourse context (including the identity of the user and the audience) as well as the relevance of situated contexts.

If we think about some types of contexts as being closer and more immediate to the speech event of interest and others being wider and broader, then discourse context belongs to the former and situational context belongs to the latter. Shuy recommends an approach to contextual interpretation that begins from wider and ends with closer context, like an inverted pyramid.[31] He observes that in police investigations or legal interpretation, words are sometimes taken as "smoking gun" evidence against criminal suspects. Moving systematically from macro to micro contexts helps with disambiguation and improves the accuracy of interpretation.[32] In other words, for our purpose here, Meta's content moderators are understandably confused about wider situational contexts if they do not then narrow the interpretations down by analysing more immediate and more local contexts. It is laudable that Meta consults external experts, but sociocultural and geopolitical expertise needs to be followed up with proper construction of the speech event and its immediate contexts.

1. Literalization

Literalization may be understood as the tendency to focus on the denotation of a word or phrase, at the cost of neglecting non-literal meaning such as indirect and implied meanings, which is often the intended meaning conveyed. The interpretation of non-literal meaning is dependent on context.

An illustrative example is 2021-005-FB-UA, where Meta removed a post containing an adaption of the "two buttons" meme, firstly for violating its Cruel and Insensitive Community Standard and upon appeal for violating its Hate Speech Community Standard. The meme features a cartoon character whose face has been substituted for a Turkish flag, sweating in front of a split screen, with a red button on each side accompanied respectively by the following statements in English: "The Armenian Genocide is a lie" and "The Armenians were terrorists that deserved it". For Meta, the meme could be viewed as either condemning or embracing the two statements featured. While the company did consider whether the content shares hate speech to condemn it or raise awareness of it, which is an exception to hate speech, it concluded that the user did not make their intention clear. The company found the statement "The Armenians were terrorists that deserved it" to be hate speech because it claims that all members of a protected characteristic are criminals. This view ignores the contradictory nature of the statements, which is precisely the basis of the meme's mockery of contemporary Turkey. The exclusive focus on the literal meaning of the statements ignores the effect of their juxtaposition and their visual context. The expectation for users to explicitly state their intent also defies the genre of satire, which sometimes

[31] R. W. SHUY, *Linguistics and Terrorism Cases*, in M. COULTHARD, A. JOHNSON (eds.), *Routledge Handbook of Forensic Linguistics*, London, 2010, pp. 558–75.

[32] From the perspective of Critical Discourse Analysis, Van Dijk (2008) also points out that not all situational contexts have the same value. He suggests understanding context not as any social situation that influences discourse but as how discourse participants subjectively construe such situation. This is to say that the speech event and its immediate contexts should limit the scope of situational contexts that are relevant. T. A. VAN DIJK, *Discourse and Context: A Sociocognitive Approach*, Cambridge, 2008, https://doi.org/10.1017/CBO9780511481499.

uses words to convey the opposite of their meaning. Meta seems to have decided that humour is not something their content moderation practices cope well with, claiming that "creating a definition for what is perceived to be funny was not operational for Facebook's at-scale enforcement". Although it may be difficult to analyse humour at scale, it is an important form of political expression.

Another striking case is an even clearer example of counter speech, where hate speech is referenced to resist oppression and discrimination. In 2021-012-FB-UA, a user posted a picture of Indigenous artwork with accompanying text in English. The artwork is a wampum belt, which is a traditional means of documenting history, with shells or beads that depict "the Kamloops story", based on the discovery of unmarked graves at a former residential school for First Nations children in Canada. The title of the artwork is "Kill the Indian/Save the Man". The text also contains the following phrases which correspond to depictions on the belt: "Theft of the Innocent", "Evil posting as Saviours", "Residential School/Concentration Camp", "Waiting for Discovery" and "Bring Our Children Home". The post also explicitly states that "its sole purpose is to bring awareness to this horrific story". Meta's automated systems identified the content as violating its Hate Speech Community Standard and a human reviewer confirmed the violation and removed the post. After the user appealed, a second human reviewer also assessed the content as violating. The phrase that triggered the content removal was "Kill the Indians", which when considered out of context constitutes violent speech targeting people based on a protected characteristic. However, Meta reversed its decision after the Board selected the case for review, acknowledging that its policy permits sharing someone else's hate speech to condemn it or raise awareness. The title of the artwork is an intertextual reference to "kill the Indian in him and save the man", a phrase with a long history in the colonial project of "civilizing" indigenous peoples in North America. It did not help that the two human reviewers Meta assigned to the case are based in the Asia-Pacific region and may not be familiar with the relevant history. "Kill the Indians" would only be read as hate speech if it is read literally and in isolation from context in this case. Moreover, the title of the work was used with quotation marks, which should have given further cues to the reviewer that it is not to be read literally.

We regularly communicate more than what we literally say. An explicit statement of intention could help clarify what might otherwise be an ambiguous message, but it could easily also be used to convey an exact opposite message. Just as one can ridicule an idea by explicitly endorsing it, one can explicitly condemn an idea while actually supporting it. The most common trope of overt untruthfulness is irony[33]. Explicit statements of intention can conflict with context, such as tone of voice, facial expressions and gestures, speaker identity, audience characteristics, and shared knowledge, leading the audience to look for an alternative meaning that is not stated but implicated.

For a marginalized group trying to raise awareness about atrocities committed against them to then be censored for hate speech adds insult to the injury. Even though it is not an explicit policy at Meta to prefer literal meaning over intended meaning, both their automated systems and human reviewers seem to be geared towards literal meaning. Meta's policies also default towards content removal, which we will discuss further in Section 4. As our examples show, such content moderation practices could end up restricting the speech of those they set out to protect.

1. Monomodal orientation

[33] Irony is overt untruthfulness used not to deceive others but to implicate meaning reversal. In the framework of Gricean conversation analysis, the speaker flouts the maxim of Quality by expressing something that s/he believes to be false, and prompts the audience to look for an alternative, implicated meaning. M. DYNEL, *Irony, Deception and Humour: Seeking the Truth about Overt and Covert Untruthfulness*, 1st ed., Boston/Berlin, 2018. Other than irony, there are other situations where an explicit statement of intention can misalign with the actual intention and where the maxim of Quality is flouted. While irony contradicts reality, hyperbole or meiosis distorts reality by overstating or understating it. Another example is metaphor, such as "you are the cream in my coffee", where the audience translate metaphorical expressions into literal expressions through world knowledge and pragmatic reasoning. All these rhetorical devices could add poetic and humorous quality to language.

Content moderation decisions may be based on a cue from a singular modality in the content, which becomes the smoking gun evidence for violation, while other modalities are ignored. This monomodal orientation may be related to the limited time and resources that human content moderators were given, and to the limitation on multimodal processing[34] by Meta's automated content moderation.

An illustrative case[35] concerns Meta's removal of a post on Instagram for violating the company's Community Standard on Adult Nudity and Sexual Activity. "Nudity" in the Community Standard is defined to include "…uncovered female nipples[36] except in the context of […] health-related situations (for example, post-mastectomy, breast cancer awareness […])". The post, with a title in Brazilian Portuguese which clearly started that its purpose is to raise awareness about breast cancer, contains eight photographs of breast cancer symptoms with corresponding descriptions (such as "ripples", "clusters", and "wounds"). Five of these photographs included visible and uncovered female nipples. Meta has "a machine learning classifier trained to identify nudity" that promptly detected the nudity in the image. The post was removed despite a policy exception that expressly allows the display of nudity used to "raise awareness about a cause or educational or medical reasons". In a statement submitted to the Board, the user explains that they posted the content as part of the national "Pink October" campaign for breast cancer prevention. Promoting awareness of main signs of breast cancer is useful for early detection and can save lives. Since this purpose squarely falls within Meta's policy exception, how did its content moderation process fail to identify the content as such? According to the Board, Meta's automated systems failed to recognise the words "Breast Cancer" in Brazilian Portuguese ("Câncer de Mama"). News reports abound about how Meta's algorithms may overfit to the English language and struggle to locate contextualized meaning in other languages[37]. Although Meta urges the Board to focus on the outcome of enforcement, not the method, the case clearly raises questions about the use of automation. Meta's engineers have noted that classifiers that work with multiple modalities are prone to overfitting to one of the modalities[38], and in the present case the system might have overfitted to the nude images at the expense of the text. Over-reliance on semantic cues can also generate false positives, as evident in the mass removal of posts (including those from years ago) containing the sarcastic expression "kill me" and related suspension of accounts on Twitter deemed as glorifying self-harm.[39] One visual cue to the purpose of the post in question is the colour pink, in line with "Pink October", an international campaign that raises awareness of breast cancer. Failure in word recognition aside, if Meta's system had the intelligence to connect the colour of the image with real world knowledge, the timing of the post (October 2020), or the likely cooccurrence of similarly themed images at the time, it would have had an additional contextual cue that helps with its interpretation. Facebook's Quarterly Update (2021 Q1) on the Oversight Board[40] denies that its systems failed to identify the keywords; instead, the company explains, the systems are not trained to ignore all content that contains the

[34] Human communication has always been largely multimodal—people combine the use of language with a diverse range of semiotic resources (including gesture, gaze, and posture) in everyday communication. Online communication is no different. A popular form of digital expression—memes—uses a combination of text and static or moving image.

[35] Oversight Board, case decision 2020-004-IG-UA, https://oversightboard.com/decision/IG-7THR3SI1/.

[36] In its analysis, the Board points out that Meta's differential treatment of male and female nipples raises discrimination concerns, but does not follow up on this issue in its Policy Advisory Statement.

[37] VILLE DE BITCHE*: Facebook Mistakenly Removes French Town's Page*, in *BBC News*, 13 April 2021, https://www.bbc.com/news/world-europe-56731027 (reporting that Facebook's algorithm removed the page of the French town Ville de Bitche, confusing it with the English insult); J. COBIAN, C. SCURATO, and B. V. CASTILLO (eds.)*, Facebook and the Disinformation Targeting Latinx Communities*, in *Colorlines*, 19 March 2021, https://www.colorlines.com/articles/op-ed-facebook-and-disinformation-targeting-latinx-communities (making the case that the Spanish word "parense" was mistranslated by Facebook as "stop" rather than "stand up", which dilutes the violence-inciting potential of a call-to-arms message).

[38] A. HALEVY et al., *Preserving Integrity in Online Social Networks*, *op. cit.*, pp. 1–32.

[39] The Copia Institute, *Detecting Sarcasm Is Not Easy (2018), Case Study Series, Trust & Safety Foundation Project* (blog), 29 July 2020, https://www.tsf.foundation/blog/detecting-sarcasm-is-not-easy-2018.

[40] Available at https://about.fb.com/wp-content/uploads/2021/07/Facebook-Q1-2021-Quarterly-Update-on-the-Oversight-Board.pdf.

keywords. While it is true that people could convey the opposite of what they state explicitly, the public has no way of knowing what other contextual cues it takes for their automated systems to recognise the policy exception. The fact remains that its automated systems erred, and made a kind of error that human content moderators do not make.

It is more resource-intensive to analyse multimodal content than plain text posted on platforms. Another case[41] where multimodality presents interpretive challenges to Meta involves a 17-minute interview with a professor, published by a Punjabi-language online media company. The caption and text accompanying the video described the Hindu nationalist organisation Rashtriya Swayamsevak Sangh (RSS) and India's ruling party Bharatiya Janata Party (BJP) as a threat to Sikhs, a minority religious group in India. The post was removed for violating the Dangerous Individuals and Organisations Community Standard, even though none of the individuals or groups mentioned in the post are designated as "dangerous". The company conceded that the removal was made in error. Meta explains that moderation error was due to the length of the video (17 minutes), the number of speakers (2), the complexity of the content and its claims about various political groups (p. 8). It acknowledges that content reviewers do not always have time to watch videos in full.

Going back to the two-buttons meme case (2022-005-FB-UA) discussed above, Meta focused on the textual statements while ignoring the visual context of the meme. In addition to the visual elements of the meme itself, the user also posted a "thinking face" emoji preceding the meme, which is often used to express sarcasm. All these cues were ignored when Meta identified one of the statements as violating.

In sum, whether it is Meta's automated systems or human moderators, there is a tendency to focus on a single modality when analysing multimodal content. Given the prevalence of multimodality in online communication, the risk here is that the intended meaning will often be missed.

4. Discussion: design and default

Even though they are not design goals, decontextualization, literalization, and monomodal orientation are features rather than bugs in Meta's content moderation practices. These systematic failures appear to be compromises that its content moderation system makes, presumably because the identification of pragmatic features is hard to scale. As Douek suggests, content moderation is all about trade-offs. Platforms have to balance accuracy in decision-making against other competing demands such as efficiency and responsiveness.[42] That said, moderation decisions that are insensitive to context and that fail to identify speaker intention will inevitably appear arbitrary to users.

There are no doubt genuinely difficult cases that Meta deals with on a day-to-day basis, such as those involving poverty of contextual information (such as a history of contact between users offline or on another online platform), serious conflicts of values (such as the challenge in balancing between allowing for a diversity of voices and protecting the safety of users), or complicated situational contexts. However, most of the cases discussed in this paper do not fall into these categories. The cases adjudicated by the Board show that even when available and relevant, context is often excluded from Meta's content analysis. [43] Decontextualization, literalization, and monomodal orientation are interpretive shortcuts adopted to facilitate efficiency and scalability, while information that can help decipher intended meaning is ignored or suppressed. By focusing attention only on part of the context, indeterminacies that could have been resolved with relatively ease are left open.

[41] Oversight Board, case decision 2021-003-FB-IA, https://oversightboard.com/decision/FB-H6OZKDS3/.
[42] E. DOUEK, *Content Moderation as Systems Thinking*, cit.
[43] According to a Washington Post article, Facebook "moderators tasked with reviewing hate speech are not allowed to see key context around a post, such as comments, accompanying photos or a profile picture—information that would help a reviewer understand the intention of the comment". Context is excluded "to protect user privacy". E. DWOSKIN, N. TIKU, H. KELLY, *Facebook to Start Policing Anti-Black Hate Speech More Aggressively than Anti-White Comments, Documents Show*, in *Washington Post*, 3 December 2020, https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/.

Sticking to explicitly stated meaning and ignoring contextual factors conveniently accommodate the limits of so-called artificial intelligence (AI)[44], which does not know how to read between the lines. Natural language processing in artificial intelligence relies primarily on semantics (literal and pre-contextual meaning) and syntax (grammatical structure).[45] Without pragmatic competence, common sense, and real-world knowledge, it cannot reliably detect irony and sarcasm[46] through pattern matching against a database of violating content. Without sufficient context such as the source of an article (e.g., *The Onion* versus *New York Times*) or the identity of the author, even human beings may be confused about whether posts they see on social media are meant to be satirical or not.

The reality in Meta is that neither human nor machine content moderators could engage in the level of contextualisation work that the Board does. Automated systems have no pragmatic competence. The human moderators at Meta work under pressured conditions[47] and hardly have the time to study the relevant context and deliberate extensively. According to one estimate, content moderators on average have under 150 seconds to make each decision.[48] While timely decision-making is crucial to preventing harmful content from going viral, it is important to understand the nature of the trade-offs.

When an acontextual reading of potentially violating content generates an indeterminate meaning, Meta's policies default towards content removal. This risk-adverse approach could lead to over-censorship, which disproportionately harm minority groups, whose political perspectives receive limited attention in mainstream media. Default towards removal seems to be especially prevalent when hate speech and dangerous individuals and organisations are involved. In the Canadian indigenous artist case (2021-012-FB-UA) discussed above, the Board points out that an internal guidance, called "Known Questions", that Meta issues to its moderators tells them that a clear statement of intent will not always be sufficient to change the meaning of a post that constitutes hate speech. When the user's intent is not clear, moderators are instructed to err on the side of removing content. The interpretation of "Chinese" in the Burmese case (2021-007-FB-UA) is a similar example, where instead of resolving the indeterminacy through discourse context, Meta opted for removal, as reflected in four of its human reviewers' action. The same applies to the mentioning of dangerous individuals and organisations, as demonstrated in the Goebbels case (2020-005-FB-UA), a case involving the mentioning of Kurdistan Workers' Party leader Abdullah Öcalan (2021-006-IG-UA), and a case that involves the mentioning of Al-Qassam Brigades, the military wing of the Palestinian group Hamas (2021-009-FB-UA). In these cases, the designated dangerous individuals and groups were mentioned to satirically comment on current politics, to raise awareness about prison rights, and to republish news. The content was removed in all these cases even though the users did not praise or support the individuals or organisations. These interpretive approaches and method of handling indeterminacies have impact beyond the current case,

[44] Which may be understood as "human-developed algorithmic systems that analyse data and develop solutions in specific domains". United Nations General Assembly, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, , 29 August 2018. Like all existing AI technology, automated detection is based on narrow AI, and will remain so in the foreseeable future.

[45] S. RUSSELL, P. NORVIG, *Artificial Intelligence: A Modern Approach*, 2020, pp. 823–878.

[46] "State of the art" attempts to apply machine learning in identifying satire rely on semantic rather than pragmatic cues; successful "detection" occurs within a limited set of test items. V. RUBIN et al., *Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News*, in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, San Diego, California: Association for Computational Linguistics, 2016, pp. 7–17, https://doi.org/10.18653/v1/W16-0802 (detecting satire by looking for reference to unexpected entities such as people, location and places in the last line of the article); O. LEVI et al., *Identifying Nuances in Fake News vs. Satire: Using Semantic and Linguistic Cues*, in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 2019, pp. 31–35, https://doi.org/10.18653/v1/D19-5004 (distinguishing fake news and satire articles by comparing their semantic and syntactic features).

[47] S. T. ROBERTS, *op. cit.*, pp. 1–266.

[48] J. KOETSIER, *Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day*, in *Forbes*, 9 June 2020, https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/.

as case decisions are used in classifier training that is supposed to improve the accuracy of the automated systems. Errors could therefore perpetuate through automation if not discovered and removed.

Meta's content moderation practices are similar to the industry approach to multilingual management in the digital society. Gramling (2020) uses the term supralingualism to describe a structural ideology and an aggressive industrial effort observed in applied research by global commercial enterprises to manage multilingualism in online settings for practical purposes.[49] Driven by a client agenda that is indifferent to nationalism and partisanship, these companies pursue technologies in cross-linguistic information retrieval and machine translation that work with translingually controlled meanings, which are now highly valued commodities, at the expense of other variations of meanings. Literalization and decontextualization are among features of supralingualism that Gramling has identified. Literalization involves preferring literal over non-literal meanings, or as Gramling explains, "[m]odes of meaning-making that rely on silence, implicature, inuendo, and subtlety are … dispreferred as data sources in supralingualism, where explicit propositional content is the primary source of meaning-making potential" (p. 143). In terms of decontextualization, since computational approaches to language understands context in terms of textual proximity and frequency, they tend to ignore the social nature of speech and "lack the depth of genre, aesthetics, pragmatics, and polysemy that inhere in the usage" (ibid). Applying a similar logic, modalities that are easier to process are preferred over other modalities, and a monomodal orientation optimizes content moderation processes by controlling the scope of meaning that is accessed.

5. Conclusion: meeting the shortfalls

As the volume of its content grew, Meta's loose standards in content moderation ("Feel bad? Take it down") hardened into an elaborate set of internal rules around 2009.[50] According to Meta's executives Dave Willner and Jud Hoffman, Meta formulated objective rules to ensure consistency and uniformity. Willner considered the distillation from standards to rules "a form of technical writing". Part of the concern was that human moderators with diverse backgrounds would bring in their cultural values and norms instead of applying Meta's. According to Sasha Rosse, who was involved in training the first content moderation team in Hyderabad:

> I liked to say that our goal was [to have a training system and rules set] so I could go into the deepest of the Amazon, but if I had developed parameters that were clear enough I could teach someone that had no exposure to anything outside of their village how to do this job.[51]

Universality comes at the cost of context, which informs speaker intentionality. An "objective" algorithm that filters by keywords without analysing context cannot tell the difference between a racist post and a post that calls out racial injustice. This is why when a black mother went on Facebook to vent about a white man uttering racist slurs to her children in a supermarket, her post was removed as violating content.[52]

Meta's algorithmic struggle with context impacts its users disproportionately, contrary to the claim that algorithmic detection of violating speech based on objective rules is consistent and unbiased. Algorithms work better in languages that are frequently used, and less well in minority languages.

---

[49] D. GRAMLING, *Supralingualism and the Translatability Industry*, in *Applied Linguistics*, 41, n. 1, 2020, pp. 129–47, https://doi.org/10.1093/applin/amz023.

[50] Community Standards are a simplified version of the rules that are publicised to its users. K. KLONICK, *The New Governors: The People, Rules, and Processes Governing Online Speech*, cit., pp. 1631–35.

[51] K. KLONICK, *The New Governors: The People, Rules, and Processes Governing Online Speech*, cit., p. 1642.

[52] T. JAN, E. DWOSKIN, *A White Man Called Her Kids the N-Word. Facebook Stopped Her from Sharing It*, in *Washington Post*, 31 July 2017, https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html.

Moreover, Meta allocates unequal resources and prioritises attention to cases based on the urgency to control bad press. Unwarranted censorship experienced by powerful users, such as famous authors, political leaders, and newspaper editors, are often rectified quickly, while the average user's wrongfully removed content may never be reinstated.[53] Until recently, politicians are given a free pass for posting violating content because of the "newsworthiness" of their speech.[54] On the other hand, false negatives affecting communities with less political power[55] and those speaking less popular languages[56] are tolerated for much longer. Former Meta employees complain about content moderation rules not being applied equally across geopolitical spaces.[57]

Despite the lack of pragmatic competence in current AI technologies, social media companies perpetuate the myth that AI is now assessing content "holistically" and analysing it "deeply"[58]. Meta's response to the Board's recommendations about enforcement is technochauvinistic[59], promoting the belief that technology is always the solution. For example, after its automated system failed to identify a breast cancer awareness campaign as a policy exception to the Nudity and Sexual Activity standard, Meta launched "keyword-based improvements" and "a new predictive model that will contribute more detail to the original system" [60]. It is impossible for the public to know how these promises of improvement translates into more accurate enforcement. Meta perpetuates a rhetoric of improvement through its software updates and enforcement reports. Given the frequency of "glitches" that occur, it is easy to forget that they have been moderating content based on internal rules for more than a decade now. But the algorithms that are now responsible for most of the content moderation decisions are still perpetually in training. The impact of existing content moderation practices on people's lives today cannot await future technology.

Pragmatic deficiency is routinely normalized, justified as an inevitable trade-off for objectivity, cultural neutrality, scalability, and efficiency. While the system design of content moderation no doubt has to take into account competing considerations, the public do not know how these are weighed against one another. Meta has been known to use the normalization of deviance as a strategy—the idea that a problem has become so accepted that it is no longer seen as problematic. After the personal information of more than half a billion Meta users had been "scraped", a leaked internal memo reveals that a long-term strategy for the company is to "normalize the fact that this activity happens regularly".[61]

Since pragmatic competence, which allows us to draw inferences, identify non-literal meaning, and deduce intent, is something that human possesses but AI does not have, it is tempting to suggest that human review of every post is the solution to the problem. For a company with an annual net income of

---

[53] K. KLONICK, *The New Governors: The People, Rules, and Processes Governing Online Speech*, cit., pp. 1654–1655.

[54] A. HEATH, *Facebook to End Special Treatment for Politicians after Trump Ban*, in *The Verge*, 3 June 2021, https://www.theverge.com/2021/6/3/22474738/facebook-ending-political-figure-exemption-moderation-policy.

[55] S. STECKLOW, *Why Facebook Is Losing the War on Hate Speech in Myanmar*, in *Reuters*, 15 August 2018, https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

[56] AVAAZ, *How Facebook Can Flatten the Curve of the Coronavirus Infodemic*, 15 April 2020, pp. 1–21.

[57] C. BUNI, S. CHEMALY, *The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech*, in *The Verge*, 13 April 2016, https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech.

[58] M. SCHROEPFER, *How AI Is Learning to See the Bigger Picture*, in *Facebook Technology* (blog), 19 May 2021, https://tech.fb.com/how-ai-is-learning-to-see-the-bigger-picture/.

[59] M. BROUSSARD, *Artificial Unintelligence: How Computers Misunderstand the World*, Cambridge (MA), 2019, pp. 1–200.

[60] Meta, *Meta Q2 + Q3 2021 Quarterly Update on the Oversight Board*, November 2021, pp. 1–37, https://about.fb.com/wp-content/uploads/2021/11/Meta-Q2-and-Q3-2021-Quarterly-Update-on-the-Oversight-Board.pdf .

[61] S. HALPERN, *Facebook and the Normalization of Deviance*, in *The New Yorker*, 2 May 2021, https://www.newyorker.com/news/daily-comment/facebook-and-the-normalization-of-deviance.

39.37 billion US dollars (in 2021)[62], one may say that there is room for more resources to be devoted to content moderation. However, even for Meta, this solution is impractical due to the sheer quantity of content shared on the platform. Some have proposed mandatory human review just for appeal cases. The Board itself has recommended that human content moderator be assigned to appeals on algorithmic decisions, at least for certain types of alleged violations (such as adult nudity, see 2020-004-IG-UA). This is a recommendation that Meta was not willing to accept. In fact, mandatory human review will be even harder to achieve for smaller social media platforms and could stifle competition. From the perspective of due process, prioritising human review may also reduce the speed of content moderation, which may also affect perception of fairness. There is a sense in which the decisions made by the Board, which resulted in reinstating a breast cancer awareness post months after the campaign ended or a political post after a conflict subsided, are not useful remedy to the users, because they are not timely enough.

A further constraint is that the Board can only review a miniscule fraction of appeal cases that come before them. Like appeal courts, it arrives at its conclusions through thorough contextual analyses. It urges Meta to pay more attention to context, but also acknowledges the challenge of content moderation at scale, as contextual analysis is labour intensive. Meta could not possibly replicate the Board's approach to case analyses, other than in selected, high-profile cases. The Board's impact is less likely exerted through its leaving up versus taking down decisions, than through the use of soft power—its policy recommendations. By pressuring Meta in its policy recommendations, the Board has had some success in requiring Meta to reveal what would otherwise be opaque processes to its users and in drawing attention to quality assurance issues in content moderation processes.

Although the Board's decisions are binding and have precedential value, unlike courtroom litigation, Meta has little stake in winning or losing cases. There is no penalty for wrongful removal of content. An outcome-based focus is therefore most economical for Meta. Meta has repeatedly urged the Board to focus on outcomes rather than processes, but accountability in processes is clearly critical to ensure the equity of outcomes. The enforcement and communication errors highlighted in Section 3.2 do not inspire confidence in Meta's operational processes, but the Board has faced obstacles in getting to the bottom of these problems. When the Board asked Meta for details that are needed to assess the accuracy of enforcement, such as error rates by individual rules, and by moderators versus automation, Meta turned down the requests on the basis that "the information is not reasonably required for decision-making in accordance with the intent of the Charter" (2021-006-IG-UA). The limited information provided in Meta's transparency reports and its reluctance to share more information with the Board make it difficult to check for and correct any bias in Meta's decision-making processes.

The cases examined in this paper demonstrate the urgency for external oversight on not only Meta's content moderation outcomes but also its processes, such as whether it is following its own rules, whether it is treating different user groups fairly, and whether its content moderation processes achieve a reasonable balance among competing demands. In other words, questions about system design and quality assurance that have impact beyond individual cases. This would be especially important for people and communities who are impacted by speech disseminated on Meta's platforms, but are not platform users themselves. It may be that the pragmatic deficits identified in this paper are necessary trade-offs against other pressing concerns, but the public currently have no way of knowing. Given the positive reputation that it has built, the Board is well positioned to expand its oversight on Meta's content moderation processes. There are inherent limits to thinking about the Board as a Supreme Court rather than as an authority that has general oversight on Meta's decision-making over speech and online safety.[63]

[62] Available at https://investor.fb.com/investor-news/press-release-details/2022/Meta-Reports-Fourth-Quarter-and-Full-Year-2021-Results/default.aspx .

[63] In a similar vein, Douek argues that the focus on correcting erroneous decisions through appeal mechanisms is misplaced. Since content moderation resembles an administrative system more than a legal system, it should strive to have aggregate accountability rather than alignment with rule of law values.

Bibliography

AVAAZ, *How Facebook Can Flatten the Curve of the Coronavirus Infodemic*, 15 April 2020.

BARRETT P. M., *Who Moderates the Social Media Giants? A Call to End Outsourcing*, New York, June 2020.

BBC News, *Ville de Bitche: Facebook Mistakenly Removes French Town's Page*, 13 April 2021. https://www.bbc.com/news/world-europe-56731027.

BROUSSARD M., *Artificial Unintelligence: How Computers Misunderstand the World*, Cambridge (MA), 2019.

BUNI C., CHEMALY S., *The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech*. *The Verge*, 13 April 2016, https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech.

COBIAN J., SCURATO C., CASTILLO B. V., *Facebook and the Disinformation Targeting Latinx Communities*. *Colorlines*, 19 March 2021, https://www.colorlines.com/articles/op-ed-facebook-and-disinformation-targeting-latinx-communities.

DOUEK, E., *The Limits of International Law in Content Moderation*, in *UC Irvine Journal of*

International, Transnational, and Comparative Law, 6(1), 2021, pp. 37–76.

DOUEK, E., *Content Moderation as Systems Thinking*, in *Harvard Law Review*, vol. 136, number2, 2022, pp. 526-607.

DWOSKIN E., TIKU N., KELLY H., *Facebook to Start Policing Anti-Black Hate Speech More Aggressively than Anti-White Comments, Documents Show*, in *Washington Post*, 3 December 2020, https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/.

DYNEL M., *Irony, Deception and Humour: Seeking the Truth about Overt and Covert Untruthfulness*, 1st ed., Boston/Berlin, 2018.

FRANKS M. A., T*he Cult of the Constitution*, Stanford, 2020.

GRAMLING D., *Supralingualism and the Translatability Industry*, in *Applied Linguistics*, 41, n. 1, 2020, pp. 129–47, https://doi.org/10.1093/applin/amz023.

HALEVY A., FERRER C. C., MA H., OZERTEM U., PANTEL P., SAEIDI M., SILVESTRI F., STOYANOV V., *Preserving Integrity in Online Social Networks*. *Proceedings of Facebook AI*, n. ACM, New York, 2020, http://arxiv.org/abs/2009.10311.

HALPERN S., *Facebook and the Normalization of Deviance*, in *The New Yorker*, 2 May 2021, https://www.newyorker.com/news/daily-comment/facebook-and-the-normalization-of-deviance.

HEATH A., *Facebook to End Special Treatment for Politicians after Trump Ban*, in *The Verge*, 3 June 2021, https://www.theverge.com/2021/6/3/22474738/facebook-ending-political-figure-exemption-moderation-policy.

HELFT M., *Facebook Wrestles With Free Speech and Civility*. *The New York Times*, 13 December 2010, https://www.nytimes.com/2010/12/13/technology/13facebook.html.

JAN T., DWOSKIN E., *A White Man Called Her Kids the N-Word. Facebook Stopped Her from Sharing It.*, in *Washington Post*, 31 July 2017, https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html.

KLEIN E., *Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next*, in *Vox*, 2 April 2018. https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge.

KLONICK K., *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, in *The Yale Law Journal*, 129, 2020, pp. 2418–99.

KLONICK K., *The New Governors: The People, Rules, and Processes Governing Online Speech*, in *Harvard Law Review*, 131, 2018, pp. 1958–1670.

KOSTSIER J., *Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day*, in *Forbes*, 9 June 2020, https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/.

LEUNG J. H. C., *The Audience Problem in Online Speech Crimes*, in *Journal of International Media & Entertainment Law*, 9, n. 2, 2021, pp. 189–234.

LEVI O., Hosseini P., Diab M., Broniatowski D. A., *Identifying Nuances in Fake News vs. Satire: Using Semantic and Linguistic Cues*, in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 2019, pp. 31–35. https://doi.org/10.18653/v1/D19-5004.

Meta, *Meta Q2 + Q3 2021 Quarterly Update on the Oversight Board*, November 2021.

ROBERTS S. T., *Behind the Screen: Content Moderation in the Shadows of Social Media*, New Haven, 2019.

RUBIN V., CONROY N., CHEN Y., CORNWELL S., *Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News*, in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, San Diego, California: Association for Computational Linguistics, 2016, pp. 7–17, https://doi.org/10.18653/v1/W16-0802.

RUSSELL S., Norvig P., *Artificial Intelligence: A Modern Approach*, 4th ed., Hoboken, 2020.

SCHROEPFER M., *How AI Is Learning to See the Bigger Picture*. *Facebook Technology* (blog), 19 May 2021, https://tech.fb.com/how-ai-is-learning-to-see-the-bigger-picture/.

SHUY R. W., *Linguistics and Terrorism Cases*, in COULTHARD M., JOHNSON A. (eds.), *Routledge Handbook of Forensic Linguistics*, London, 2010, pp. 558–75.

STECKLOW S., *Why Facebook Is Losing the War on Hate Speech in Myanmar*, in *Reuters*, 15 August 2018, https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

The Copia Institute, *Detecting Sarcasm Is Not Easy (2018), Case Study Series, Trust & Safety Foundation Project* (blog), 29 July 2020, https://www.tsf.foundation/blog/detecting-sarcasm-is-not-easy-2018.

The Oversight Board, *Oversight Board Transparency Reports Q4 2020, Q1 & Q2 2021*, October 2021.

United Nations General Assembly, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, 29 August 2018.

VAN DIJK T. A., *Discourse and Context: A Sociocognitive Approach*. Cambridge, 2008, https://doi.org/10.1017/CBO9780511481499.

VINCENT J. *AI Won't Relieve the Misery of Facebook's Human Moderators*, in *The Verge*, 27 February 2019, https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms.