Should Social Media Platforms Permit Violating Content that is "Newsworthy"?

Ricki-Lee Gerbrandt and Jeffrey Howard

Abstract. In 2016, Facebook received widespread global criticism for removing the iconic "Terror of War" photo (widely known as the "Napalm Girl" photo) for violating its rules on child nudity. To remedy that concern, Facebook introduced a general "newsworthiness allowance" to apply to some content on the platform even though it technically violates a rule. The "rarely applied," ad hoc and ex post exemption has since been used in a range of cases to exempt violating speech—including violating speech by political figures. But should a social media platform allow speech that violates a rule to remain on the platform because it meets some threshold for "newsworthiness"? We argue that social media platforms should not apply such exemptions. First, as a procedural matter, we argue that ad hoc and ex post decisions to exempt certain people from platform rules rest in considerable tension with the rule of law, specifically the principle of *legality*. Second, as a substantive matter, we argue that the newsworthiness allowance is the wrong prescription for the right problem. The problem is that content moderation rules are often insufficiently nuanced and overinclusive. Instead of applying an arbitrary exemption, platforms should nuance their rules to allow for awareness raising, condemnation, and satire in categories of content to which the newsworthiness allowance might otherwise be applied. Third, we consider whether there ought to nevertheless be a narrowed newsworthiness allowance for violating speech by political leaders. People have a fundamental interest in finding out what their leaders think, especially when what they think is deeply misguided or even harmful, and this is particularly true in democracies. While this is an important interest, we reject the notion that a newsworthiness allowance is the right way to protect it. The interest is most propitiously protected by allowing news reporting on what politicians have said, not by allowing politicians to abuse platforms to cause harm.

doi:10.54501/jots.v3i1.253

1 Introduction

In 2016, Facebook received widespread global criticism for removing an iconic photo of a nude child (Phan Thi Kim Phúc) fleeing a napalm attack during the Vietnam War. The criticism wasn't that the photo was somehow allowed under Facebook's rules and yet erroneously removed; the photo was quite clearly prohibited, given the general prohibition on child nudity. The criticism instead was that it was preposterous to remove the photo given its historic significance, and so something had surely gone wrong. To address the concern, Facebook introduced its *newsworthiness allowance*, through which content that strictly violates a policy is nevertheless permitted on the platform due to its public importance (Meta 2024b). Meta continues to deploy this exemption in a range of cases to exempt violating speech across Facebook and Instagram—including violating speech by political figures.

Largely thanks to Meta Oversight Board rulings, Meta's newsworthiness allowance has attracted the most discussion. But it is not the only company with such a policy. X (formerly Twitter) also has a policy involving "public interest exceptions" that is only applied to the posts "shared by a high profile account" that "represents a current or potential member of a … governmental or legislative body" (X, n.d.).[1] And unlike Meta, where the cost-benefit analysis yields the judgment that violating content should be allowed, X imposes some *partial* enforcement by placing the violating post behind an interstitial click-through screen and demoting the content so that it isn't algorithmically amplified or easily shared. YouTube also has a similar policy, which was only publicly unearthed after reporting by the *New York Times* (Grant and Mickle 2025). According to that reporting, in December 2024, YouTube directed moderators to leave up content that violated its rules so long as the violating content did not account for more than 50% of the video's duration (which is double the length of time under the older internal moderation guidelines). YouTube also noted its "long-standing practice of applying exceptions" to its rules when the content is in "the public interest." It is beyond scope of this article to substantively consider these other companies' policies (and we also lack robust data from these companies). But we mention them here to indicate that these general, ad hoc exemptions are deployed across several global social media platforms.

The task of this essay is to subject the notion of a newsworthiness allowance to normative scrutiny, specifically through a case study of Meta's policy and concomitant enforcement, to evaluate whether it can be justified. Should social media platforms allow content that violates a rule when it meets some threshold for newsworthiness? There are many prima facie plausible reasons why you might think it should. Some violating content is shared to raise awareness, e.g., as part of journalistic reporting; sometimes it is shared to satirize or condemn whatever is primarily depicted in the content. Further, even when the purpose of sharing the content is malevolent, it might still be in the public interest

1. Before June 2021, Meta had a policy of essentially exempting most politicians' speech from content moderation. As Meta's former president for global affairs, Nick Clegg, once said, "we will treat speech from politicians as newsworthy content that should, as a general rule, be seen and heard." See Heath (2021).

for people to see it when the speaker is a politician.[2] After all, people have an interest in knowing what their political leaders think, even when what their leaders think is deeply misguided and even harmful.

Despite these reasons, we will argue that platforms should refrain from applying newsworthiness allowances. We make three arguments. First, as a procedural matter, we argue that ad hoc and ex post decisions to exempt certain people from platform rules rest in considerable tension with the principle of *legality*. This principle is a central international human rights norm and legal requirement concerning the governance of speech, to which Meta has publicly committed.[3] A core requirement of legality is that those governed by rules be empowered with sufficient information with which to predict whether their conduct does or does not run afoul of those rules. Because it is very difficult to predict when a newsworthiness allowance will be selectively applied, users lack such information. If the point of the newsworthiness allowance is to empower users to be confident that they can post newsworthy content without sanction, it plainly fails to achieve this objective. Another requirement of legality is that adjudicators make decisions in accordance with rules; this is to reduce *arbitrariness* in decision-making. Because the newsworthiness allowance is applied only on escalation at the discretion of company officials, such discretionary determinations are ripe for uncertainty, abuse, political appeasement, and the favoring of commercial interests.

Second, as a substantive matter, we argue that the newsworthiness allowance is the wrong prescription for the right problem. The problem is that content moderation rules are often insufficiently nuanced; specifically, they are often overinclusive, restricting more speech than would ideally be restricted. While sometimes overinclusiveness is inevitable, we shouldn't assume this without first attempting to articulate a more nuanced alternative and assessing whether it can be feasibly enforced. Strikingly, this is precisely the solution to which Meta was eventually drawn in the sorts of cases raised by the photo of the Vietnamese child fleeing napalm; the policy categorically banning child nudity was qualified to allow "[i]magery posted by a news agency that depicts child nudity in the context of famine, genocide, war crimes or crimes against humanity."[4] We argue that this strategy should be generalized further across all the cases where a newsworthiness allowance might be otherwise applied, exploring how this can and should be done. While the newsworthiness allowance suggests that those it exempts are rulebreakers who should be spared punishment for the public interest, this alternative strategy holds, much more sensibly, that such persons are not properly understood as rulebreakers at

2. See, for example, the Oversight Board's "Cambodian Prime Minister" case Oversight Board (2023d) and the "Greek 2023 Elections Campaign" case Oversight Board (2024c), discussed in detail below.

3. Meta's commitment to human rights is outlined on its website at https://humanrights.fb.com/. The Oversight Board, in reviewing content moderation decisions on Meta platforms, applies Article 19 (freedom of expression) of the International Covenant on Civil and Political Rights United Nations General Assembly (1966), which includes the three-step test of legality, legitimacy, and proportionality. In this essay, we have drawn on case law from the European Court on Human Rights (ECtHR), which applies Article 10 (freedom of expression) in the European Convention on Human Rights Equality and Human Rights Commission (2021) to elucidate certain points, but other international law adjudicative bodies and concomitant jurisprudence could provide such guidance as well.

4. See Meta's community standard for child sexual exploitation, abuse and nudity (Meta, n.d.-b).

all.

Finally, we consider whether there ought to nevertheless be a narrowed newsworthiness allowance for violating speech by political leaders. People have a fundamental interest in finding out what their leaders think, especially when what they think is deeply misguided or even harmful. That is especially so in democracies, since citizens need information about what politicians think to decide whom to elect for office. We grant that this is an important interest, but we reject the notion that a newsworthiness allowance is the right way to protect it. The interest is most propitiously protected by allowing news reporting on what politicians have said, not by allowing politicians to abuse platforms to cause harm. After all, politicians' authority makes their harmful speech especially likely to cause harm, which can be aggravated by platform algorithms that spread such violating content, so it is particularly difficult to justify exempting such politicians' harmful speech from accountability. At minimum, a defensible policy does not exempt violating expression based on the political status of a speaker. An even better policy is to prioritize review of posts containing potentially violating speech by political leaders, because of the harm that the dissemination and amplification of those statements on social media can generate.[5]

We proceed in Section 2 to review the policy landscape, summarizing Meta's policy on the newsworthiness allowance and reviewing relevant Oversight Board cases concerning how it has been applied. Section 3 outlines our procedural argument against the newsworthiness allowance. Section 4 offers our substantive argument against the allowance, arguing that the right remedy is to adopt more nuanced rules. Section 5 then queries whether there should nevertheless be a restricted newsworthiness allowance for political leaders.

2 The Newsworthiness Allowance

To begin, it is helpful to quote directly Meta's newsworthiness allowance policy:

> We want people to be able to talk openly about the issues that matter to them, even if some may disagree or find them objectionable. In rare cases, we allow content that may violate the Community Standards, if it's newsworthy and if keeping it visible is in the public interest. We do this only after conducting a thorough review that *weighs the public interest against the risk of harm*. We look to international human rights standards, as reflected in our Corporate Human Rights Policy,[6] to help make these judgments.

5. Indeed, the Oversight Board has recommended that Meta "update its review prioritization systems to ensure that content from heads of state and senior members of government that potentially violated the Violence and Incitement policy is consistently prioritized for immediate human review"; see the "Cambodian Prime Minister" case (Oversight Board 2023d). We compare this policy recommendation to Meta's controversial cross-check program (Meta 2024d) below.

6. See Meta (2021).

The policy continues:

> We've found that determining the newsworthiness of a piece of content can be highly subjective. People often disagree about what standards should be in place to ensure that a community is both safe and open to expression. We conduct a thorough assessment of any potentially newsworthy content and our reviewers consider a number of factors prior to escalating to our Content Policy team.
>
> When making a newsworthy determination, we assess whether that content surfaces an imminent threat to public health or safety, or gives voice to perspectives currently being debated as part of a political process. We also consider other factors, such as:
>
> - Country-specific circumstances (for example, whether there is an election underway, or the country is at war)
> - The nature of the speech, including whether it relates to governance or politics
> - The political structure of the country, including whether it has a free press
>
> We remove content, even if it has some degree of newsworthiness, when leaving it up presents a risk of harm, such as physical, emotional and financial harm, or a direct threat to public safety. For content we allow that may be sensitive or disturbing, we include a warning screen. In these cases, we can also limit the ability to view the content to adults, aged 18 and older.
>
> Content from all sources, including news outlets, politicians or other people, is eligible for a newsworthy allowance. While the speaker may factor into the balancing test, we do not presume that any person's speech is inherently newsworthy, including politicians. (Meta 2024b)

How has this policy been applied in the past? The most recent public data indicate that, between June 1, 2024, and June 1, 2025, Meta applied the newsworthiness allowance 44 times. Of those 44, 20 were for posts from politicians. Six were "scaled," meaning they were applied "more broadly to something like a phrase." So one point to notice is that Meta applies the newsworthiness with remarkable rarity.[7]

In what sorts of cases has the policy been applied? Meta's own policy flags the following paradigmatic examples:

- Colombian news outlets reporting on police brutality posted a video where a slur

7. This raises the question—if the policy has a compelling rationale, why is it deployed in so few cases? We will return to this point in the next section.

can be overheard, in violation of the Hate Speech policy.

- The Ukrainian Defense Ministry shared a video that briefly shows an unidentified burnt corpse, in violation of the policy on Violent and Graphic Content.
- A Brazilian official posted a video criticizing the lack of funding for the cinema industry, which contains a clip of a movie showing female nipples—thus in violation of the policy on Adult Nudity and Sexual Activity.
- A post raising awareness about the mistreatment of prisoners by Azerbaijani soldiers that technically violated the policy on Coordinating Harm and Promoting Crime, which bans depictions of prisoners of war.

In its cases, the Oversight Board has also called attention to a range of other cases where Meta applied the newsworthiness allowance:

- A post containing a video uploaded to Facebook in which Cambodian Prime Minister Hun Sen threatens his political opponents with violence, which violated the Violence and Incitement community standard (Oversight Board 2023d).
- A Facebook post containing images of Armenian prisoners of war, which violated the Coordinating Harm and Promoting Crime community standard (Oversight Board 2023a).
- A photo of a beaten person in Sudan suffering from a detached eye, which violated the Violent and Graphic Content community standard (Oversight Board 2022d).

There are also many examples of cases where Meta did not apply a newsworthiness allowance, but because of the public interest value of the content, could have:

- An Instagram account describing itself as a news outlet for Dalit perspectives posted a video of a woman in India being sexually assaulted by a group of men. The post was removed for violating the Adult Sexual Exploitation community standard, and a severe strike was applied against the account (Oversight Board 2022b).
- A video posted by a Cuban news outlet to Instagram documenting a women's protest where a protestor uses an allegedly dehumanizing word against men (i.e., comparing men to "rats") alongside her criticisms of men for failing to defend those that have been murdered and repressed (Oversight Board 2023c).
- The posts of three media outlets that broadcasted a politicians' speech in which he used the term "Ingiliz uşağı" (which translates as "servant of the British"), which Meta found violated its Hate Speech community standard and consequently suspended the media organizations' accounts during a critical election (Oversight Board 2023f).
- A documentary video posted by Voice of America (VOA) Urdu, revealing the iden-

tities of child victims of sexual abuse and murder in Pakistan in the 1990s, for violating the Child Sexual Exploitation, Abuse and Nudity community standard (Oversight Board 2024d).

- A video posted by a local Colombia news outlet reposting a video of a protest where some protesters used slurs against the country's president in circumstances of major tax reforms that became a widespread political issue. The video was accompanied by a caption expressing admiration for the protesters. [8]

It is unclear why a newsworthiness allowance is applied in some cases but not in others. It is thus impossible to know of all cases where the newsworthy allowance could have been applied, based on Meta's definition, but was not.

3 The Procedural Case: Newsworthiness, Legality, and Suitability

Meta has committed itself publicly to the idea that its governance of users' speech should be based on fundamental international human rights norms;[9] indeed, Meta invokes its Corporate Human Rights Policy in its statement of why it has a newsworthiness allowance in the first place.[10] The Meta Oversight Board also analyzes Meta's policies and content moderation practices according to the framework set out in the International Covenant of Civil and Political Rights (ICCPR).[11] But is the allowance actually compatible with these norms? We will argue that it is not.

Under Article 19 of the ICCPR, a restriction on freedom of expression is justified if and only if it satisfies several tests: (1) *legality*, (2) *legitimacy*, and (2) *proportionality* (which requires that any infringement be suitable, necessary, and proportionate) (United Nations Human Rights Committee 2011). We will argue that the newsworthiness allowance falls

8. The protesters once used the phrase "hijo de punta," which translates to "son of a bitch," and once used the phrase "deja de hacerte el marica en la tv," which translates to "stop being the fag on tv." The phrase violated Facebook's rules because "marica" was designated as a homophobic slur. The Oversight Board held that Facebook should have applied the newsworthiness allowance in this case (Oversight Board 2021a).

9. See Meta's commitments to the UN Guiding Principles on Business and Human Rights at https://humanrights.fb.com/. Its human rights policy states that "We are committed to respecting human rights as set out in the United Nations Guiding Principles on Business and Human Rights (UNGPs). This commitment encompasses internationally recognized human rights as defined by the International Bill of Human Rights—which consists of the Universal Declaration of Human Rights; the International Covenant on Civil and Political Rights; and the International Covenant on Economic, Social and Cultural Rights—as well as the International Labour Organization Declaration on Fundamental Principles and Rights at Work" (Meta 2021)

10. While X has not explicitly endorsed these norms, it recurrently casts itself as committed to the free speech of its users. See, for example, statements by its owner Elon Musk on X (Musk 2022).

11. The Oversight Board explains its approach thusly: "When restrictions on expression are imposed by a state, they must meet the requirements of legality, legitimate aim, and necessity and proportionality (Article 19, para. 3, ICCPR). ... The Board uses this framework to interpret Meta's voluntary human rights commitments, in relation both to the individual content decision under review and to Meta's broader approach to content governance. As the UN Special Rapporteur for freedom of opinion and expression has stated, although 'companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression,' (Report A/74/486, para. 41)"; see "Referring to Designated Dangerous Individuals as 'Shaheed"' (Oversight Board 2024e).

short at the first test, legality.

What is legality? Legality is the fundamental principle that refers to the standards of the *rule of law* (Dicey 1959; Fuller 1969; Raz 1979; Waldron 2011; Bingham 2011). One requirement of legality is that rules are prospectively announced and clearly specified in advance, such that users have access to information about what the policy is and how it is applied, so that they can reasonably predict how their content will be treated. To illustrate, regimes without the rule of law may arrest people who couldn't have predicted that their conduct would trigger this consequence. In contrast, in rule-of-law regimes, people have adequate prospective notice about what is and isn't caught by the rules. While they can never have perfect advance notice, since rules will always have some degree of indeterminacy, the aim is for those governed by rules to be reasonably informed so they can plan their conduct accordingly (Shapiro 2013).

Another demand of legality is that everyone is equal before the law—meaning that those with political power or connections are not able to skirt rules by which everyone else must abide. As A.V. Dicey, one of early proponents of a theory of the rule of law, famously stated, "every official, from the Prime Minister down to a constable" must be subject to "the ordinary law" (Dicey 1959).

A third demand of legality is *fairness*—the evenhandedness in the enforcement of rules, so adjudicators make decisions according to rules rather than through the arbitrary exercise of discretion (Dicey 1959; Krygier 2011). The reduction of arbitrariness benefits not only rule followers, but also well-meaning decision-makers who try to apply the rules as best they can and are hampered by discretionary provisions that offer them little in the way of guidance. It also helps to reduce arbitrary decision-making from adjudicators who are not so well-meaning—those who may seek to benefit a friend, a powerful party, or an influential politician.

Based on what we know about the newsworthiness allowance, it fails all three dimensions of the legality test.[12] First, it is very difficult to predict when a newsworthiness allowance will be applied. As we see from Meta's statement of the policy, "We've found that determining the newsworthiness of a piece of content can be highly subjective"(Meta 2024b). But therein lies the problem: it appears that the policy is predicated on content teams' own case-by-case judgments on what is newsworthy, how newsworthy it is, and how the magnitude of the newsworthiness is properly weighed against risks of harm. Indeed, as we discuss in more detail below, there are many similar cases where only some received a newsworthiness allowance while others did not, such that it is not really a rule at all.

Second, a disproportionate number of positive newsworthiness allowances have been

12. While they do not frame their argument in terms of principles from IHRL, similar worries to our procedural concerns are explored in Kadri and Klonick (2019). Their article critiques the newsworthiness allowance, as part of a broader critique of First Amendment jurisprudence's influence over content moderation. They do not raise the substantive concerns, however, that dominate most of this paper.

applied to the content of political figures, indicating that the allowance is applied to violating content by political leaders more so than other content disseminators. Now, there may be good reasons that Meta is concerned about ensuring politicians' speech is not over-moderated; as mentioned above, there can be high public-interest value in the public being informed about what politicians say and how they think. But there are clearly limits to that logic when the content violates the fundamental rights of others or causes serious risks of harm. We examine this issue in more detail below. For now, our point is that the newsworthiness allowance is ripe ground for abuse and favoritism relating to politicians' speech, which contravenes a key feature of the rule of law.

Third, given how rarely it is used, it seems very unlikely that the policy is treating like cases alike and is applied in a nonarbitrary way. As mentioned above, at Meta the policy was applied only 44 times in the last year for which public data is available[13] (we don't even have the data for X). What are the odds that if every user post was subjected (impossibly) to the cost-benefit analysis involved in testing for newsworthiness, Meta would decide to allow more than 44 posts? It seems very high indeed. Meta's newsworthiness allowance is applied to a small and, surely in some respects, arbitrary set of posts, which have (for whatever reason) attracted the attention of senior officials within the company or the Oversight Board. What this suggests, then, is that there are manifold further posts that are eligible for the newsworthiness allowance, but not receiving it.

This brings us to a final procedural point, concerning not the legality test but rather the *suitability* test, which requires that a measure be suitable (roughly, minimally effective) at achieving its purpose. Is the newsworthiness allowance actually fit for purpose? That depends on what its purpose is. If the point of the newsworthiness allowance is to empower users to be confident that they can post newsworthy content without sanction, it plainly fails to achieve this objective. Given the risks in posting violating conduct out of hope that a newsworthy allowance will be subsequently applied, users act rationally by refraining from posting such content at all (Schauer 1978; Kendrick 2013). Given that we want to allow such speech, there must be a better way. If the point of the newsworthiness allowance is to provide adjudicators with discretion to make decisions in the public interest, the newsworthiness allowance is also not an appropriate option given that it is applied in a small number of cases and is not scalable, and that there are many cases in which Meta clearly ought to have applied it but did not. Given that we want content moderators on trust and safety teams to make decisions in the public interest, there must be a better way for that, too.

13. We acknowledge that there will likely never be universal agreement on a content moderation policy; but content moderation policy is not unique–there is unlikely to be universal agreement on any type of policy, be it tax, immigration, speech, or so on. There are, however, bounds for reasonable disagreement. To illustrate: the fact that people will disagree about the best policy to protect child safety, it doesn't follow that platforms should have no child safety policy. Clearly it is possible to rule out certain policies that are unreasonable. It is our view that our model is more reasonable and defensible than the ad hoc newsworthiness allowance.

4 The Substantive Case: Nuancing Rules

We have offered procedural arguments against the newsworthiness allowance, arguing principally for its incompatibility with the principle of legality, and also suggesting it may even fail the principle of suitability. We now make a substantive argument: that *the newsworthiness allowance is the wrong way to solve a real problem*.

To set up the substantive argument, return to the case of the journalist who posted the photo of Phan Thi Kim Phúc, the nude Vietnamese child fleeing a napalm attack. This post was deemed in violation of the Child Sexual Exploitation, Abuse, and Nudity policy, but given the public interest of the photo, Facebook declined to enforce the policy in this instance. This framing casts the journalist who posted the photo as a *wrongdoer*—or at least, a *rulebreaker*—who nevertheless should be *spared punishment* for the public benefit. But this is a very strange way to frame this case. A much more plausible framing is that the journalist isn't doing anything wrong at all by posting the photo; he was engaging in legitimate activity that ideally would be protected under the rules. And indeed, that alternative logic has now crept into Meta's Child Sexual Exploitation, Abuse, and Nudity policy, which has been qualified to allow "[i]magery posted by a news agency that depicts child nudity in the context of famine, genocide, war crimes or crimes against humanity."[14] Such a policy empowers journalists who seek to share that photo, and others like it, protecting their legitimate speech *ex ante*—rather than asking them to violate a high-stakes child nudity policy and then beg for an exemption *ex post*.

We argue now that this strategy should be generalized further across all cases where a newsworthiness allowance might be otherwise applied.[15] The core challenge is that content moderation rules are often insufficiently nuanced; specifically, they are often overinclusive, restricting more speech than would ideally be restricted (Gerbrandt 2025). While sometimes overinclusiveness is inevitable, we shouldn't assume this without first attempting to articulate a more nuanced alternative and assessing whether it can be feasibly enforced.

This arises in the kinds of cases for which the newsworthiness allowance has been applied, which largely concern "dual-use content": one use of the content is illegitimate, and another use is legitimate. We argue that cases are ideally handled by nuancing the rule to make exceptions for *legitimate purposes*, where the purpose is understood as a matter of *what audiences would reasonably infer the user to be doing in the post* (Fisher and Howard 2024). According to this *purpose-sensitive approach*, users should be expected to *indicate their purpose* in posting the content. We first reconsider the various kinds of cases in which the newsworthiness allowance has been deployed to see how this would work. After considering those various cases, we address how social

14. This carve-out had a further caveat: "unless accompanied by a violating caption or shared in a violating content, in which case the content is removed" (Meta, n.d.-b).

15. Meta has already done this in some areas. For example, the policy on "Voter/census fraud" exempts awareness raising, condemnation, news reporting, or humorous or satirical contexts. Likewise, in the hate speech prohibitions there is an exception for satire. See Meta (n.d.-c).

media platforms can distinguish between the legitimate and illegitimate purposes. (We will draw on examples from Meta, as this is the platform for which there has been the greatest discussion of the newsworthiness allowance).

4.1 Nuancing rules for dual-use cases

In *dual-use* cases, a single piece of content might be used for a malevolent purpose, or a legitimate purpose. First, consider images posted by third parties of imagery depicting episodes designated as "violating violent events" (VVEs) (Oversight Board 2024a), such as videos of terrorist attacks or other human rights abuses. These images are plainly dual-use: They can be posted for the illicit purpose of glorifying violence, or they can be posted for the valuable purpose of raising awareness about an injustice (Howard, Gerbrandt, and Thau 2024). Widely discussed examples of such cases include violent imagery posted by news organizations showing imagery of the assassination of a political candidate in Mexico (Oversight Board 2024a), or videos and images accompanied by contextual information showing a terrorist attack in Moscow (Oversight Board 2024b). Should such content be allowed or disallowed? On the one hand, the actual imagery of VVEs can help audiences understand and grasp the seriousness of what has occurred more effectively than mere textual description. On the other hand, it seems reasonable to disallow perpetrators of violence, their representatives, or third parties to post this content in order to promote and glorify criminal violence—which can risk copycat attacks. Meta's commitment to *safety*, and commitment to protecting human rights, provides strong reason to disallow the sharing of VVEs for those illegitimate purposes.

A blanket ban on sharing VVEs would prevent their illegitimate use, but also their legitimate use. Likewise, a blanket permission would allow their legitimate use, but in so doing enable their illegitimate use. One solution is to adopt a blanket ban, but rely on the newsworthiness allowance—applied ad hoc and ex post—to allow the legitimate cases. For all the reasons we've mentioned, this is undesirable. A better approach, we contend, is to *nuance the rule*. In our view, platforms should explicitly allow users to share VVEs for legitimate purposes, while continuing to ban sharing for illegitimate purposes. Meta's current policy contains no explicit exception under the "Dangerous Organizations and Individuals" policy for sharing VVE content (Meta, n.d.-e). The Oversight Board agrees that this policy is problematic, but argues the remedy is the application of the ad hoc newsworthiness allowance for select cases (Howard, Gerbrandt, and Thau 2024). We think that Meta should instead specifically clarify in the rule what the legitimate purposes for sharing VVE imagery are and to permit them. (How should platforms infer purpose? We postpone this question until the next subsection, where we focus on the importance of drawing inferences from text in user captions).

Consider another set of dual-use cases, from the context of *exploitation*. Meta's Adult Sexual Exploitation policy prohibits content that "depicts, threatens, or promotes sexual violence, sexual assault, or sexual exploitation" (Meta, n.d.-a). In one set of posts analyzed by the Oversight Board, a news organization in India reported on the experiences

of a Dalit woman who was assaulted and beaten by a group of men, posting videos of the attack (Oversight Board 2022b). (Dalit members, sometimes referred to as "untouchables," are members of what is widely considered the lowest caste and are among the most widely marginalized citizens in India). In response, Meta removed the news organization's video (and suspended its account) for violating the Adult Sexual Exploitation policy. To be sure, one possible use of this content was to glorify violence against Dalit women, to degrade and dehumanize them, or to further perpetuate negative stereotypes about a historically oppressed group. But this was manifestly *not* what the news organization was doing; rather, it was bringing these criminal acts to the public's attention, condemning sexual violence against women, and raising awareness of the general plight of the Dalit community. Rather than finding this content violating and removing it (as was done) or applying a newsworthiness allowance (which cannot be operationalized ex ante at scale), a better solution is to build an exception into the Adult Sexual Exploitation policy itself, so that content condemning or raising awareness about sexual violence is permitted (along with specifying any friction measure, e.g., that the content may be subject to a warning screen or remain inaccessible from the accounts of children).[16]

Third, consider content containing slurs, which may violate Meta's Hateful Conduct policy. In one case, a video posted by a Cuban news organization depicted a woman's protest denouncing the government, calling for other women to support her, lamenting the deaths of young men, and calling some men "rats" for not standing up to oppression (Oversight Board 2023c). Her protest occurred on the one-year anniversary of mass protests in Cuba against state repression; a caption accompanying the video called for international attention to the situation in the country. Meta removed the content on the basis of the "rats" slur, which it found was dehumanizing toward men. Meta did not apply a newsworthiness allowance or apparently even consider applying it. Again, this case shows potential dual-purposes of the content: The news organization could be raising a matter of public concern, or it could be perpetuating harmful language targeting men. Also consider the case where three Turkish news organizations were prohibited from posting content during a critical election (Oversight Board 2023f). The newspapers ran afoul of Meta's rules when they posted content that contained the "servant of the British" comment by a politician, as outlined above. Meta did not even contemplate applying the "newsworthiness" exemption in that case, and much like the content of the woman's protest in Cuba, one wonders why, given that the political and electoral context made this statement seemingly "newsworthy." A solution here is to ensure that rules regarding slurs, which are contained in the Hateful Conduct policy, have an exception for news reporting that raises awareness of a matter in the public interest.[17]

16. The Oversight Board agrees that modifying the standard is the right response to this case (Oversight Board 2022b). Our argument in this paper helps to justify this general approach, encouraging the Oversight Board to go further and call for the elimination of the newsworthiness allowance.

17. Meta's policies have gone in this direction: The Hateful Conduct policy now allows slurs where the intention is to condemn or report on those slurs (Meta, n.d.-f).

Finally, consider an example from the Violence and Incitement policy space. In a case from Haiti (Oversight Board 2023e), a video was posted of a group of people trying to break into a police cell and threatening imprisoned alleged gang members with violence. The video included the (translated) caption "the police cannot do anything." It was posted during a time of political crisis in Haiti, with police largely incapable of enforcing order. By depicting this threat, the post technically violated the Violence and Incitement rules. After a three-week delay, Meta removed the post.[18] After the Oversight Board reviewed the case, it held that the content was a violation of the Violence and Incitement standard because credible threats were made. However, a majority of the Board held that because of Meta's three-week delay in assessing the content, the risk of harm had likely already materialized. Given the widespread violence and breakdown of the state, the Board deemed it important for the public to be informed of the situation. Accordingly, it ruled that Meta should restore the content (though the minority disagreed, believing that the risk of harm remained too high given the ongoing political crisis).

This case further illustrates the inadequacy of an approach reliant on the newsworthiness allowance. For starters, it is reliant on a temporally sensitive cost-benefit analysis; the Board's disagreement turned on whether enough time had elapsed that the content was no longer sufficiently dangerous to outweigh its public interest benefit. Given that the duration of time at issue was a function of Meta's apparently inadequate investment in moderation resources in Haiti (which caused the three-week delay), this is a bizarre basis for determining whether certain content is or is not violating.[19] A better approach is to either (1) find such content violating, as all members of the Board agreed that it was, and disallow it from the platform, or (2) build further nuance into the rule and allow third-party posts depicting violence so long as the clearly indicated purpose is awareness-raising about a matter of public concern (here, the breakdown in the Haitian justice system). We think the second option is the best approach, as it permits legitimate public-interest content without resort to an ad hoc, ex post exemption. Insofar as certain "newsworthy" content should be protected, it should be specified ex ante at the level of rule design.

4.2 Distinguishing between legitimate and illegitimate purposes

We have argued that platforms should adopt a *purpose-sensitive approach*, whereby content moderation is sensitive to the *indicated purpose* of the post. Examples of legitimate purposes for posting otherwise violating content paradigmatically include *awareness-raising*, *news reporting*, *satire*, and *condemnation* (which of course often overlap)—though the exact details will depend on the content area at issue. How exactly should platforms infer what a user's purpose is in posting something? As is well known, content moderators (whether human or machine) cannot read users' minds. Nor can they

18. It is worth noting that Meta's inadequate investment in moderation resources in Haiti resulted in the three-week delay.

19. In the "Call for Women's Protest in Cuba" case (Oversight Board 2023c), it was noted that there was a seven-month delay in evaluating the content as it went through the rounds of Meta's cross-check review.

hold elaborate investigations into the intentions or beliefs of wrongdoers (as in criminal trials where most crimes require proof of mens rea). So what should they do?

There are many cases where such determinations are relatively straightforward: The kinds of captions involved in neutral news reporting or condemnation plainly indicate legitimate purposes, whereas the kinds of captions involved in glorification and support plainly indicate illegitimate purposes (Howard, Gerbrandt, and Thau 2024). Moreover, often the identity of the user—namely, if it is a recognized news publisher—will suffice for the platform to infer that the purpose in posting the content is legitimate. But what about cases where such determinations are ambiguous? This will be true in cases where users share a plain video of a VVE without any caption, for example.

We can imagine a world in which AI tools proactively scan uploaded posts to classify a speaker's indicated purpose (based on preexisting classifications). In instances where a speaker's purpose is ambiguous, it is conceivable that these users could receive a notification requiring them to indicate their purpose or to provide greater context. Indeed, we think that platforms *should* do this. But if platforms do not provide such options, what should moderators on trust and safety teams do?

There can be reasonable disagreement on this matter. One reasonable policy is to remove content for which the purpose is ambiguous. On this option, it's up to users to be crystal clear about why they are posting otherwise-violating content (like a video of a human rights abuse); if they fail to indicate clearly their legitimate purpose, they have forfeited their complaint when the post is removed. Another option is simply to default to *allowing* content in cases of ambiguity, erring on the side of protecting speech. A third option is to vary the default setting in different content areas. For example, a default to removal may be the right option where Violence and Incitement is at issue, but not where Hate Speech is at issue.

Importantly, platforms do not need to be passive in response to the problems posed by ambiguity. We have already mentioned the idea of prompting users to clarify their posts when ambiguous. Platforms can also proactively educate its users to be clear about *why* they are posting content. Users who seek to raise important issues in the public interest (so-called "citizen journalists") sometimes do not adequately make their purposes clear or provide adequate context for the content at issue. There are, we think, ethical obligations on news organizations, civil society, and general users that are posting content for reasons of raising awareness or condemnation to make that clear, such that reasonable viewers (and moderators on trust and safety teams) can reasonably ascertain that purpose (Gerbrandt 2025). This argument is not only ethically persuasive, but accords with international human rights law, which recognizes the importance of responsible journalism to bring to the public's attention matters in the public interest—even where they report on unlawful individuals, organizations, or hateful messages they spread—provided that the reporting provides adequate context and does not endorse

the comments.[20] This is especially true on social media platforms, which carry arguably greater risks than traditional press and broadcasting, because of the impact of algorithm amplification to a large readership and the potential for illegitimate user commentary that can facilitate harm.

These examples demonstrate that although moderation decisions can be difficult in many situations where the underlying content is in the public interest, a principled approach is to develop more nuance into the rules themselves. Insofar as platforms such as Meta are increasingly doing just that (a trend that began with the "Napalm Girl" photo controversy), our argument here serves to explain *why* this trend is something to defend and extend. The trend has not gone far enough: Meta *still* applies the newsworthiness allowance, and the Oversight Board continues to recommend that Meta apply the newsworthiness allowance in response to problematic moderation decisions.[21] Based on our arguments outlined above, we think this is the wrong approach. The newsworthiness allowance should be increasingly rolled back and replaced with our model.

5 Exempting Political Leaders

Notwithstanding the preceding arguments, we now consider the objection that there ought to be a narrowed newsworthiness allowance for *violating speech by political leaders*. This is based on the idea that people have a fundamental interest in finding out what their leaders think, especially when what they think is deeply misguided or even harmful. That is especially so in democracies, since citizens need information about what politicians think to decide whom to elect for office. Unlike the other examples we have discussed, this *does* fit the logic of the newsworthiness allowance: The politician posting the violating content *really is a wrongdoer*, but we all have an interest in seeing them commit that wrong so we can learn what sort of person they are and what their goals are.

We grant that this is an important interest. However, we reject the notion that a newsworthiness allowance is the right way to protect that interest. Rather, the interest is adequately protected by allowing official news reporting (and other awareness-raising speech) on what politicians have said. We need not allow politicians to abuse social

20. Jersild v. Denmark European Court of Human Rights (1994), Application No 15890/89 (ECtHR, 23 Sept 1994); Pentikäinen v. Finland European Court of Human Rights (2015), Application No 11882/10 (ECtHR, 20 October 2015) at para 90.

21. For recent examples, see "Candidate for Mayor Assassinated in Mexico" (Oversight Board 2024a), "Footage of Moscow Terrorist Attack" (Oversight Board 2024b), "News Documentary on Child Abuse in Pakistan" (Oversight Board 2024d), and "Haitian Police Station Video" (Oversight Board 2023e) (where the majority of the Board recommended that Meta apply the newsworthiness allowance, but a minority disagreed. Based on our discussion above, we agree with the minority).

media platforms to cause harm.[22]

Let's reconsider the various kinds of cases in which the newsworthiness allowance was considered to see how this would work. In the "Cambodian Prime Minister" case, Hun Sen threatened his political opponents with violence on a live broadcast; the video was then uploaded to his Facebook page, where it was viewed more than 600,000 times (Oversight Board 2023d). Users reported the video for violating the Violence and Incitement community standard, which prohibited "threats that could lead to death" and "threats that lead to serious injury." Two human reviewers found that it did not violate the policy; it was then escalated to subject matter experts, who found that it did—but then applied the newsworthiness allowance. The speech made by the Prime Minister, including his discussion of "sending gangsters" and commencing "legal action" with midnight arrests, was rightly held by the Oversight Board as incitement of violence and legal intimidation. The Board considered that the statements were made in the context of sustained intimidation and harassment of political opponents and journalists in the country, which was aided by the amplification of the prime minister's rhetoric on social media platforms. The question was: Should the newsworthiness allowance nonetheless be applied?

We think the answer is *no*. There are several problems with the application of the newsworthiness allowance to politicians' speech. First, the newsworthiness allowance here would simply permit the prime minister to generate the underlying conditions of newsworthiness by making the incendiary claims in the first place; he, and other political leaders with similarly malevolent aims, should not benefit from a vague policy exemption to allow incendiary claims to spread. Second, time is of the essence when harmful content is made by political leaders because of their powerful platform to instigate violence; platform moderators on trust and safety teams should not be left in the position of waffling over whether to apply a newsworthiness allowance (or waiting on the word from higher-ups), given the potential for serious harm.[23] Finally, the newsworthiness allowance is ripe for uncertainty. For example, while Meta applied it to the content of the Cambodian prime minister, it was not applied to posts made by President Trump when he perpetuated unfounded claims that the American election was stolen and instigated a mob to storm Capitol Hill—resulting in the deaths of five people and injuries to many others, not to

22. Political speech is accorded the highest protection in ECHR jurisprudence. Our view, however, is supported by international human rights law, which holds that such speech is not without limits, and has specifically upheld justifiable restrictions on politicians' speech that undermines or seeks to destroy other fundamental rights; see Guide on Article 10 of the European Convention on Human Rights, Freedom of expression Equality and Human Rights Commission (2021), para 541. See also Gerger v. Turkey European Court of Human Rights (1999) [Grand Chamber] Application No 24919/94 (ECtHR, 8 July 1999) at para 50, and Erdal Taş v. Turkey European Court of Human Rights (2006), Application No 77650/01, (ECtHR, 19 December 2006) at para 38.

23. In the "Former President Trump's Suspension" (Oversight Board 2021b), the Oversight Board states that "time is of the essence in such situations" and "taking action before influential users can cause significant harm should take priority over newsworthiness and other values of political communication."

mention the psychological fear that workers on Capitol Hill endured.[24] Nor was it applied to posts made by a Brazilian general, who called for people to "hit the streets," and "go to the National Congress and the Supreme Court" during a volatile election (Oversight Board 2023b). Indeed, for months, Meta found those posts nonviolating; it was only after rioters stormed government buildings that Meta changed its mind and took the content down. The newsworthiness allowance is also ripe for uncertainty for other reasons: In its submissions to the Oversight Board in the Brazilian case, Meta argued, ex post, that the content did not qualify for a newsworthiness allowance because the public interest value of the content did not outweigh the risk of harm given the "explicit call for violence" and the "heightened risk of offline harm following the Brazilian Presidential election and Lula's inauguration" (Oversight Board 2023b). But if that was the case, why did Meta decide that the content did not violate any rules in the first place? And why would it have assessed newsworthiness when it had already decided the content was nonviolating? Perhaps it considered applying the newsworthy exemption after the people stormed the government buildings and it decided to remove the posts—but at that point such a finding would have defied rationality. This highlights the larger issues around the need for social media platforms to consistently apply their rules to content of political leaders.

It is ultimately difficult to reconcile the various decisions that Meta made in these three cases. Indeed, the lack of transparency around the application of the newsworthiness allowance led the Board to conclude, in the Trump case, that the "lack of transparency regarding these decision-making processes appears to contribute to perceptions that the company may be unduly influenced by political or commercial considerations" (Oversight Board 2021b). It thus not only uncertainty that the newsworthiness allowance generates, but the very serious concern that platforms may use the incredible discretion the newsworthiness allowance provides to favor content—and the very real harm it can spawn—from people it seeks to protect or appease or to protect its commercial interests.

From this discussion, we can conclude that violating speech made by political leaders should not be granted an exemption. But what about the concern that there is value for the public to know what politicians say and how they think? We think that social media platforms should not permit violent and inciting speech made by politicians on its platforms, ever. The public can instead obtain information on what politicians say and think from news reporting, which can appropriately contextualize the information. This may work well in jurisdictions with robust professional and ethical news organizations where journalists and civic society can bring these issues to the public's attention. It would not work well in jurisdictions lacking ethical news organizations and civic society

24. In the Trump case (Oversight Board 2021b) (at para 8.1), the Oversight Board said, "The Board notes that there is limited detailed public information on the cross check system [sic] and newsworthiness allowance. Although Facebook states the same rules apply to high-profile accounts and regular accounts, different processes may lead to different substantive outcomes. Facebook told the Board that it did not apply the newsworthiness allowance to the posts at issue in this case. Unfortunately, the lack of transparency regarding these decision-making processes appears to contribute to perceptions that the company may be unduly influenced by political or commercial considerations."

institutions or where those are controlled by the ruling political party. The problems of unethical press are substantial in many jurisdictions around the world, and are unlikely to ever be solved by a platform policy or by social media companies themselves. But it is our view that social media platforms should not allow politicians to take advantage of platforms to disseminate content that violates rules, especially content that has the potential to instigate widespread harm (Amnesty International 2022).[25]

Suppose a news organization or user uploads a video or content containing the violating claims made by the politician. Mightn't this still risk serious harm—enabling politicians to use media organizations to disseminate their messages? This is another case of reasonable disagreement, in which different platforms might have different defensible policies. One option is simply to prohibit all such content. The risk, then, is that the public does not learn important information about their politicians' aims. A more defensible option, in our view, is for platforms to allow news reporting of politicians' content that violates its policies, so long as the news reporting does not glorify or promote the content and provides appropriate context (e.g., as required by standard ethical journalistic practices, such as providing information on background events leading to the statements, the reactions of civic society or community members, competing perspectives, and so forth). Instead of applying a vague "newsworthiness" allowance, each rule can specify the requirements necessary for news reporting to report on violating content made by politicians. This follows our recommendation above to nuance rules to allow for reporting on public interest content. A third option is to allow such content, but demote it when the risk of harm is high.

Meta has been slowly altering its rules to accord with this view. In the "Reporting on Pakistani Parliament Speech" case that ended up before the Oversight Board, it came to light that Meta had internal guidelines permitting content that may otherwise violate the Violence and Incitement standard if it was posted for "awareness raising" (Oversight Board 2024f). That exemption was applied to exempt content posted by a local news outlet that had uploaded a video of a politician speaking in Parliament where he used hyperbolic language to express outrage at the human rights violations, protests, and crackdown on political opponents happening in his country, in the context of a national election involving conflict between the then president and the military.[26] The news content did not endorse the speech but pointed to the strong reaction it generated in

25. In the "Tigray Communications Affairs Bureau" case (Oversight Board 2022e), Meta stated that it was difficult to remove "official government speech that could be considered newsworthy" but that may pose risk of inciting violence during an ongoing conflict, and that it did not consider using the newsworthy exemption because that exemption does not apply to conduct at risk of contributing to physical harm. However, the newsworthiness allowance does not make this clear.

26. Parliamentary speech generally enjoys heightened protection given that it concerns core issues of democratic governance; see Karácsony and Others v. Hungary European Court of Human Rights (2016) [GC], Application Nos 42461/13 and 44357/13, (ECtHR, 17 May 2016) at para 138. However, speech by parliamentarians, even in parliament, does not enjoy unlimited protection, especially if the speech undermines or seeks to destroy other fundamental rights; see Budinova and Chaprazov v. Bulgaria European Court of Human Rights (2021), Application No 12567/13, (ECtHR, 16 February 2021) at paras 90-91.

Parliament.[27] Since that case, Meta has included the "awareness raising" exception to this rule in some public community standards.[28]

What if violating content is posted not by a news organization, but by an individual user (including by a "citizen journalist")? Arguably, the same analysis stands: If the user provides appropriate context for the violating content and condemns the violating content, that content could be allowed.[29] It may be more difficult for moderators to make the decision to leave that content on the platform because the user does not have the backing or recognition of an institution, which can help convince moderators that the user was responsibly posting the content (Gerbrandt 2025). But that is a difficulty inherent in decision-making and is not remedied by a vague "newsworthiness" allowance.

If Meta allowed news organizations or even users to post violating content made by politicians with appropriate contextual information that condemned the violence or incitement, it would still have to monitor user commentary on that content to ensure that it did not generate violating commentary. It could also monitor this content closely to ensure that it was not being weaponized, for example, by state-backed supporters (Posetti, Maynard, Bontcheva, et al. 2021).

We have argued that politicians should be held to the same standards as ordinary users; they should not be granted the power to abuse platforms to cause harm, simply because it's in the public interest to know their character. But why not go further and argue that politicians should be held to *higher* standards? After all, some international rights tribunals have held that political figures using social media for political purposes have greater responsibility because of their influence and reach (European Court of Human Rights 2023). While we are skeptical that platforms should adopt *different* rules for politicians than for ordinary users, we do think that platforms should dedicate greater resources to monitor potentially violating speech made by political leaders (and other people with large accounts), given both their authority and the potential for viral spread. Platforms should also dedicate greater resources to monitor repetitions of politicians' violating speech by other user accounts to ensure that such content is non-violating. Politicians' speech, then, should be monitored with especial vigor for rule compliance.[30]

This is one plausible rationale for something *like* Meta's controversial "cross-check"

27. The Oversight Board recommended Meta adopt the news reporting exception to Dangerous Organizations and Individuals standard in the "Mention of the Taliban in News Reporting" case (Oversight Board 2022c).

28. Similarly, in the "India Sexual Harassment Video" case (Oversight Board 2022b) the Oversight Board recommended that Meta (Instagram) build into its policies the exemptions for news content reporting to raise awareness about sexual harassment targeting women in the Dalit caste rather than rely on the general newsworthiness allowance.

29. But see the "Greek 2023 Elections Campaign" case (Oversight Board 2024c), where a minority of the Board would have applied a "newsworthiness" allowance to content shared by a user glorifying a politician's support for a designated hate group. The rationale was that voters should have the fullest possible information to make their decisions during an election; the majority, however, found that the public would have been informed by other means (e.g., the press).

30. After Facebook suspended President Trump from its platform following the Capitol Hill July 6 riots, it created a policy allowing it to restrict accounts of public figures during times of civil unrest because of the potential harm that those accounts can cause (Meta 2024c).

policy. The cross-check policy contains a (secret) list of high-profile speakers; when content from those entities is flagged for moderation, the content is not automatically moderated but is subjected to enhanced levels of review.[31] Meta has expanded the program to include review based on content rather than on the identity of the user.[32] The goal is to prevent and minimize false positives that remove high-profile figures' legitimate speech erroneously. It is controversial because while Meta claims it is used to advance its human rights commitments, the Oversight Board found that the "program appears more directly structured to satisfy business concerns" (Oversight Board 2022a). However, a policy could be implemented that monitored high-risk accounts, i.e., accounts of political leaders, for violating content and escalated that content for quick review.[33] This, it seems to us, would be a defensible extension of the cross-check program, such that it guarded not only against false positives (taking down legitimate content) but also false negatives (leaving up violating content), which can be just as damaging.[34]

6 Conclusion

Platforms have a long-standing practice of exempting some violating content from enforcement when they deem it newsworthy. We have scrutinized this practice, predominantly examining cases at Meta (where the practice has received its most extensive application and, thanks to the Oversight Board, discussion). We have argued that there is a strong procedural argument against the newsworthiness allowance, given its ad hoc and ex post character, which is incompatible with the human rights norm of *legality*. Further, we have mounted a substantive case against the newsworthiness allowance, arguing that it perversely frames people engaged in legitimate speech as wrongdoers who are spared enforcement for the public benefit. A superior approach is to adopt more nuanced rules. Defending a *purpose-sensitive approach*, we argued that for dual-use content (in which one use is legitimate but another is illegitimate), platforms should be attentive to the *indicated purpose* of the user, enforcing on that basis. Further, we have considered the objection that there should be narrowed exemption for *politicians'* violating speech, arguing that this is unpersuasive. If a politician is genuinely abusive, we suspect that they will show themselves in countless further ways or contexts. But the important point is philosophical: A platform need not and should not make itself complicit in that harm for

31. The aspect of the program is called the "Early Response Secondary Review." Meta says entities include "journalists" and "human rights organizations" but also its "business partners."

32. In late 2021, it expanded this program to include what it calls the "General Secondary Review." The Oversight Board was highly critical of this program in its review "Advisory Opinion on Meta's Cross-Check Program" (Oversight Board 2022a).

33. Meta says that the factors it uses to assess for priority human review include severity, virality, and likelihood of violating, but it does not appear to assess this based on accounts (Meta 2024a).

34. In contrast, in X's policy, politicians are allowed to post violating content, but X will place it behind an interstitial and limit its amplification. This strikes the balance better than allowing it to be posted and algorithmically amplified—but we are unpersuaded that X's policy is the right one. Merely placing an interstitial over the content does not equate to responsible reporting or appropriately contextualizing the information. (Platforms could direct users to high-quality news reporting in addition to placing an interstitial over the content, or deploy other editorial functions traditionally associated with responsible news reporting, that could better inform the public.)

citizens' democratic interests to be satisfied. Overall, we think that platforms are more likely to abide by the rule of law and protect legitimate speech under our model than a model that applies an ad hoc newsworthiness exemption. While there is plenty of room for reasonable disagreement on this issue, we hope this commentary can jump-start a conversation on the downsides of the newsworthiness allowance—and the case for retiring its use.

References

Amnesty International. 2022. *The Social Atrocity: Meta and the Right to Remedy for the Rohingya.* September 29, 2022. https://www.amnesty.org/en/documents/ASA16/5933/2022/en/.

Bingham, Tom. 2011. *The Rule of Law.* London: Penguin Books. ISBN: 9780141038483.

Dicey, A. V. 1959. *Introduction to the Study of the Law of the Constitution.* 8th ed. First published 1885. London: Macmillan. ISBN: 9780865970021.

Equality and Human Rights Commission. 2021. *Article 10: Freedom of expression.* EHRC website. Guidance on the right to hold and express opinions, including restrictions that are lawful, necessary, and proportionate, June 3, 2021. https://www.equalityhumanrights.com/human-rights/human-rights-act/article-10-freedom-expression.

European Court of Human Rights. 1994. *Jersild v. Denmark, Application No. 15890/89,* September 23, 1994. https://hudoc.echr.coe.int/eng#%7B%22itemid%22:%5B%22001-57891%22%5D%7D.

———. 1999. *Gerger v. Turkey [Grand Chamber], Application No. 24919/94,* July 8, 1999. https://hudoc.echr.coe.int/eng#%7B%22itemid%22:%5B%22001-58272%22%5D%7D.

———. 2006. *Erdal Taş v. Turkey, Application No. 77650/01,* December 19, 2006. https://hudoc.echr.coe.int/eng#%7B%22itemid%22:%5B%22001-78610%22%5D%7D.

———. 2015. *Pentikäinen v. Finland, Application No. 11882/10,* October 20, 2015. https://hudoc.echr.coe.int/fre#%7B%22itemid%22:%5B%22001-158279%22%5D%7D.

———. 2016. *Karácsony and Others v. Hungary [Grand Chamber], Application Nos. 42461/13 and 44357/13,* May 17, 2016. https://hudoc.echr.coe.int/eng#%7B%22itemid%22:%5B%22001-162831%22%5D%7D.

———. 2021. *Budinova and Chaprazov v. Bulgaria, Application No. 1256/13.* February 16, 2021. https://hudoc.echr.coe.int/eng#%7B%22itemid%22:%5B%22001-207928%22%5D%7D.

———. 2023. *Sanchez v. France, Application No. 45581/15,* May 15, 2023. https://hudoc.echr.coe.int/fre#%7B%22itemid%22:%5B%22001-211777%22%5D%7D.

Fisher, Sarah A., and Jeffrey W. Howard. 2024. "Ambiguous Threats." *Journal of Ethics and Social Philosophy* 28 (2): 208–29. https://www.jesp.org/index.php/jesp/article/view/3359.

Fuller, Lon L. 1969. *The Morality of Law.* Revised. New Haven: Yale University Press. ISBN: 9780300010701.

Gerbrandt, Ricki-Lee. 2025. "Threatening and Protecting Press Publishers and Journalism in the UK's Regulation of Social Media Platforms." *Journal of Media Law* (January 28, 2025). https://www.tandfonline.com/doi/full/10.1080/17577632.2024.2445897.

Grant, Nico, and Tripp Mickle. 2025. "YouTube Loosens Rules Guiding the Moderation of Videos." *New York Times* (June 9, 2025). https://www.nytimes.com/2025/06/09/technology/youtube-videos-content-moderation.html.

Heath, Alex. 2021. "Facebook to End Special Treatment for Politicians After Trump Ban." The Verge, June 4, 2021. https://www.theverge.com/2021/6/3/22474738/facebook-ending-political-figure-exemption-moderation-policy.

Howard, Jeffrey, Ricki-Lee Gerbrandt, and Tena Thau. 2024. *Public Comment to the Oversight Board, "Cases Concerning the Assassination of Candidate Running for Mayor in Mexico".* Digital Speech Lab. https : //www.oversightboard.com/wp-content/uploads/gravity_forms/52-1277fe10ce7f18f846678283a9b9c8ca/2024/10/DIGITAL-SPEECH-LAB-OVERSIGHT-BOARD-PUBLIC-COMMENT.pdf.

Kadri, Thomas E., and Kate Klonick. 2019. "Facebook v. Sullivan: Public Figures and Newsworthiness in Online Speech." *Southern California Law Review* 93:37–98. https://scholarship.law.stjohns.edu/cgi/viewcontent.cgi?article=1292&context=faculty_publications.

Kendrick, Leslie. 2013. "Speech, Intent, and the Chilling Effect." *William & Mary Law Review* 54:1633–90. https://scholarship.law.wm.edu/cgi/viewcontent.cgi?params=/context/wmlr/article/3481/&path_info=wmlr54_no5_p1633_kendrick.pdf.

Krygier, Martin. 2011. "Four Puzzles About the Rule of Law." *Nomos* 50:64–67. https://www.jstor.org/stable/24220108.

Meta. 2021. "Corporate Human Rights Policy." Meta, March. Accessed August 20, 2025. https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf.

———. 2024a. "How Meta Prioritizes Content for Review." Meta Transparency Center, November 12, 2024. https://transparency.meta.com/en-gb/policies/improving/prioritizing-content-review/.

———. 2024b. "Our Approach to Newsworthy Content." Meta Transparency Center, November 12, 2024. https://transparency.meta.com/features/approach-to-newsworthy-content/.

———. 2024c. "Restricting Accounts of Political Figures During Civil Unrest." Meta Transparency Center. November 12, 2024, policy originally launched January 2023. https://transparency.meta.com/en-gb/enforcement/taking-action/restricting-accounts-of-public-figures/.

———. 2024d. “Reviewing High-Impact Content Accurately via Our Cross-Check System.” Meta Transparency Center, November 12, 2024. https://transparency.meta.com/en-gb/enforcement/detecting-violations/reviewing-high-visibility-content-accurately/.

———. n.d.-a. “Adult Sexual Exploitation.” Meta Transparency Center. Accessed August 20, 2025. https://transparency.meta.com/policies/community-standards/adult-sexual-exploitation/.

———. n.d.-b. “Child Sexual Exploitation, Abuse and Nudity.” Meta Transparency Center. Accessed August 20, 2025. https://transparency.meta.com/en-gb/policies/community-standards/child-sexual-exploitation-abuse-nudity.

———. n.d.-c. “Community Standards.” Meta Transparency Center. Accessed August 20, 2025. https://transparency.meta.com/policies/community-standards/.

———. n.d.-e. “Dangerous Organizations and Individuals.” Meta Transparency Center. Accessed August 20, 2025. https://transparency.meta.com/policies/community-standards/dangerous-individuals-organizations/.

———. n.d.-f. “Hateful Conduct.” Meta Transparency Center. Accessed August 20, 2025. https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/.

Musk, Elon, @elonmusk. 2022. “By ‘free speech’, I simply mean that which matches the law. I am against censorship that goes far beyond the law …” X. April 26, 2022, 12:33 p.m. https://x.com/elonmusk/status/1519036983137509376.

Oversight Board. 2021a. “Colombia Protests,” September 17, 2021. https://www.oversightboard.com/decision/FB-E5M6QZGA/.

———. 2021b. “Former President Trump’s Suspension,” May 5, 2021. https://www.oversightboard.com/decision/fb-691qamhj/.

———. 2022a. “Advisory Opinion on Meta’s Cross-Check Program,” December 6, 2022. https://www.oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s.

———. 2022b. “India Sexual Harassment Video,” December 14, 2022. https://www.oversightboard.com/decision/IG-KFLY3526/.

———. 2022c. “Mention of the Taliban in News Reporting,” September 15, 2022. https://www.oversightboard.com/decision/FB-U2HHA647/.

———. 2022d. “Sudan Graphic Video,” June 13, 2022. https://www.oversightboard.com/decision/FB-AP0NSBVC/.

———. 2022e. “Tigray Communications Affairs Bureau,” October 4, 2022. https://www.oversightboard.com/decision/FB-E1154YLY/.

———. 2023a. “Armenian Prisoners of War,” June 13, 2023. https://www.oversightboard.com/decision/FB-YLRV35WD/.

———. 2023b. “Brazilian General’s Speech,” June 22, 2023. https://www.oversightboard.com/decision/fb-659eawi8/.

———. 2023c. “Call for Women’s Protest in Cuba,” October 3, 2023. https://www.oversightboard.com/decision/IG-RH16OBG3/.

———. 2023d. “Cambodian Prime Minister,” June 29, 2023. https://www.oversightboard.com/decision/FB-6OKJPNS3/.

———. 2023e. “Haitian Police Station Video,” December 5, 2023. https://www.oversightboard.com/decision/fb-lxnfad5f/.

———. 2023f. “Political Dispute Ahead of Turkish Elections,” August 23, 2023. https://www.oversightboard.com/decision/fb-t8jdddjv/.

———. 2024a. “Candidate for Mayor Assassinated in Mexico,” December 12, 2024. https://www.oversightboard.com/decision/bun-fu50knak/.

———. 2024b. “Footage of Moscow Terrorist Attack,” November 19, 2024. https://www.oversightboard.com/decision/bun-zr5os2ko/.

———. 2024c. “Greek 2023 Elections Campaign,” March 28, 2024. https://www.oversightboard.com/decision/bun-kj6lo858/.

———. 2024d. “News Documentary on Child Abuse in Pakistan,” May 14, 2024. https://www.oversightboard.com/decision/fb-j3fc7xx9/.

———. 2024e. “Referring to Designated Dangerous Individuals as ‘Shaheed.’ Policy Advisory Opinion,” March 26, 2024. https://www.oversightboard.com/decision/pao-lopp03uk/.

———. 2024f. “Reporting on Pakistani Parliament Speech,” April 4, 2024. https://www.oversightboard.com/decision/FB-57SPP63Y/.

Posetti, Julie, Diana Maynard, Kalina Bontcheva, et al. 2021. *Maria Ressa: Fighting an Onslaught of Violence.* International Center for Journalists, March. https://www.icfj.org/sites/default/files/2021-03/Maria%20Ressa-%20Fighting%20an%20Onslaught%20of%20Online%20Violence_0.pdf.

Raz, Joseph. 1979. “The Rule of Law and Its Virtue.” In *The Authority of Law: Essays on Law and Morality,* 210–29. Oxford: Oxford University Press. ISBN: 9780198253457. https://doi.org/10.1093/acprof:oso/9780198253457.003.0011.

Schauer, Frederick. 1978. “Fear, Risk and the First Amendment: Unraveling the Chilling Effect.” *Boston University Law Review* 58:685–732. https://scholar.google.com/scholar?q=Frederick+Schauer+Fear%2C+Risk+and+the+First+Amendment.

Shapiro, Scott J. 2013. *Legality.* Cambridge, MA: Harvard University Press. ISBN: 9780674725782.

United Nations General Assembly. 1966. *International Covenant on Civil and Political Rights.* OHCHR website. Adopted by United Nations General Assembly resolution 2200A (XXI) on 16 December 1966; entered into force on 23 March 1976, December 16, 1966. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.

United Nations Human Rights Committee. 2011. *General Comment No. 34: Article 19 – Freedoms of Opinion and Expression,* September 12, 2011. https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf.

Waldron, Jeremy. 2011. *Thoughtfulness and the Rule of Law.* British Academy Law Lecture, February 1, 2011. https://www.thebritishacademy.ac.uk/publishing/review/18/thoughtfulness-and-rule-law/.

X. n.d. "About Public-Interest Exceptions on X." X Help Center, Platform Use Guidelines. Accessed August 20, 2025. https://help.x.com/en/rules-and-policies/public-interest.

Authors

Ricki-Lee Gerbrandt is a Fellow in Law & Platform Governance at the Digital Speech Lab, Department of Political Science & Public Policy at University College London, and a qualified lawyer (Canada). She can be reached at r.gerbrandt@ucl.ac.uk.

Jeffrey Howard is a Professor of Political Philosophy and Public Policy at University College London and the Director of the Digital Speech Lab. He can be reached at jeffrey.howard@ucl.ac.uk.

Acknowledgments

The authors would like to thank participants at the workshop on "New Ideas in Legal and Political Philosophy of Online Speech," hosted at the Digital Speech Lab, UCL, for engaging comments and discussion. The authors would also like to acknowledge UKRI for research funding (UKRI grant MR/V025600/1), which enabled both of them to complete this work.

Disclosure

Prior to 2022, Ricki-Lee Gerbrandt was engaged in full-time private practice in Canada where she represented a global social media company, news organizations, journalists, politicians and public figures in matters concerning defamation and freedom of expression. She also represented businesses and individuals seeking the removal of unlawful and defamatory content online, including from some global social media companies. Starting in 2022, she became a full-time legal scholar. She remains a non-practicing member of the bars of British Columbia and Alberta.

Jeffrey Howard is a member of the Online Information Advisory Committee, which advises the UK regulator Ofcom on social media regulation.

Keywords

Newsworthy; newsworthiness; public interest; social media; content moderation; rule of law; political speech; international human rights; law; philosophy; freedom of speech; freedom of expression; platform governance.