Journal of Ethics and Social Philosophy
VOL. 28, NO. 2 · AUGUST 2024
https://doi.org/10.26556/jesp.v28i2.3359

AMBIGUOUS THREATS

"DEATH TO" STATEMENTS AND THE MODERATION OF ONLINE SPEECH ACTS

Sarah A. Fisher and Jeffrey W. Howard

In July 2022, a Facebook user in Iran posted the phrase "death to Khamenei" ("marg bar Khamenei" in the original Farsi).[1] The content was posted in a public group that described itself as supporting freedom for Iran. The post shared a cartoon (dating from a 2011 blog post) of the Iranian supreme leader, Ayatollah Khamenei, his beard forming a fist, which grasped a woman wearing a hijab, a blindfold, and a chain around her ankles. A speech bubble next to the cartoon stated that being a woman is forbidden. The user's caption, accompanying the cartoon, included the text "death to the anti-women Islamic government and its filthy leader Khamenei" alongside wider criticism of the Iranian regime for its treatment of women. The post appeared shortly before Iran's National Day of Hijab and Chastity, an occasion used by critics of the government to protest against the mandatory hijab and other illiberal policies.

Another Facebook user complained about this post to Meta's content moderation teams, which govern the speech of users on the social media platforms Facebook and Instagram. In response, the post was removed, having been deemed to violate Facebook's Violence and Incitement Community Standard.

Did Meta make the right decision by removing the post? This single question, it turns out, illuminates a litany of deeper philosophical issues concerning what sort of speech is properly targeted by platforms' content moderation systems. Upon review, both Meta and its Oversight Board concluded that the platform had erred in removing the post. But the arguments that brought them to this conclusion were fundamentally in conflict and involved crucial confusions about the criteria for removing speech. This article pinpoints an important source of instability in how speech is currently governed online—namely, a failure to distinguish the illocutionary and perlocutionary dimensions of speech acts (very roughly, their communicative force versus their downstream effects).

1 The details of this case are taken from the Meta Oversight Board's decision 2022-013-FB-UA, available at https://www.oversightboard.com/decision/FB-ZT6AJS4X/.

By offering a solution, our goal is to provide broader normative guidance for the proper design and enforcement of platform rules.

In section 1 we describe further details of the case. In section 2 we map the conclusions of Meta and the Oversight Board to a philosophical distinction between illocution and perlocution. Section 3 identifies confusions in both parties' reasoning about the case and traces these back to Meta's policy. In section 4 we show why the confusions are objectionable and set out three possible solutions, whereby content moderation targets (1) the probability of the speech having harmful (perlocutionary) effects, (2) the probability of its (illocutionary) force belonging to a prohibited category, or (3) both of the above aspects, using a systematic combinatory procedure. We conclude by defending an "illocutionary-first" version of target 3.

1. CASE BACKGROUND

1.1. At-Scale Moderation Decision

Like all administrators of social media, Meta specifies what speech is and is not allowed on its platforms via a suite of "Community Standards." The policies therein cover topics ranging from privacy, nudity, and sexual exploitation to hate speech, misinformation, and incitement, among many others. They are enforced through proactive and reactive moderation, undertaken by a combination of automated systems and human reviewers. Meta's content moderation activity (and that of large social media companies in general) often proves controversial, given its impacts on billions of active users.[2] The "death to Khamenei" post involves one such controversy.

Although the complainant reported the post as hate speech, Meta's at-scale moderation team did not judge it to violate Facebook's Hate Speech Community Standard; it was removed for violating the platform's Violence and Incitement Community Standard. That policy stated:

> We aim to prevent potential offline harm that may be related to content on Facebook. While we understand that people commonly express disdain or disagreement by threatening or calling for violence in non-serious ways, we remove language that incites or facilitates serious violence. . . . We also try to consider the language and context in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety. In determining whether a threat is

2 As of December 31, 2022, Facebook had 2.96 billion monthly active users, according to Meta's Q4 2022 Earnings Report, available at https://investor.fb.com/home/default.aspx.

> credible, we may also consider additional information such as a person's public visibility and the risks to their physical safety.[3]

Subsequent investigation by Meta's Oversight Board shed light on the moderators' decision-making. On one hand, "death to *X*" statements are usually allowed on the platform, in that they typically express a mere wish or hope that *X* dies rather than a credible threat or call for lethal violence. On the other hand, Meta removes "death to *X*" statements wherever *X* is a member of a high-risk category. The reason why "death to Khamenei" triggered enforcement action was the fact that the target was a head of state whose safety was therefore deemed to be at elevated risk.[4] Accordingly, the post was classified as a violation of the Violence and Incitement Community Standard, resulting in enforcement action.[5]

1.2. Appeal and Review

The moderation decision was immediately appealed by the user. However, that appeal was not prioritized by Meta's automated systems (which take account of signals concerning the type, virality, severity, and recency of the content), and it was subsequently closed without further review. The user then appealed to the Oversight Board, an independent body established by Meta in 2020 to review and adjudicate content rules and decisions on Facebook and Instagram.

While the Oversight Board was deliberating about the case, Meta conducted its own internal review of the decision. The company continued to maintain that the user had indeed violated the Violence and Incitement Community Standard. However, it nevertheless concluded that the post should have been allowed to stand on the grounds of its "newsworthiness." Meta controversially grants "newsworthiness allowances" to disallowed speech if it is in the public interest for people to see it. To decide whether violating speech merits a

3 The Community Standard has been available in Farsi since February 2022. Note that the Violence and Incitement Community Standard encompasses both *threats* of violence and *incitements* to violence, even though these are distinct categories causing different harms and subject to differing normative analyses in free-speech literature. For the full statement of the current policy and the relevant change log, see https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/.

4 According to the Oversight Board's report, other categories of high-risk targets include former heads of state, candidates and former candidates for head of state, candidates in national and supranational elections for up to thirty days after election if not elected, people with a history of assassination attempts, activists, and journalists.

5 In addition to the post being removed, the user who posted it received one "strike" against their account (where accumulating more strikes leads to greater restrictions on one's ability to create content on the platform) and two "feature limits," meaning that they could not post or comment in groups for the next thirty days and could not post or comment on other Facebook surfaces (except Messenger) for seven days.

newsworthiness allowance, the company putatively weighs the public interest in seeing the speech against the risk of harm caused by the speech. In this case, the company deemed the risk of offline harm to the Iranian supreme leader to be outweighed by the significant public interest in hearing the antiregime perspective.[6] That perspective was considered especially newsworthy given the timing of the post (when protests were being organized in the run-up to the National Day of Hijab and Chastity) and its contribution to political and religious discourse (in criticizing Iran's mandatory hijab laws and the government's overall stance towards women). The argument for allowing the speech was further compounded by the Iranian government's history of suppressing free expression, including online (with Facebook having been banned in Iran since 2009 and only accessible via technical workarounds). As a result, the post was reinstated on the platform in August 2022.[7]

1.3. Oversight Board Ruling

In January 2023, the Oversight Board published the results of its own investigation into the case. Although the board similarly determined that the post should be allowed to stand, it marshaled quite different reasons and argumentation in support of that conclusion. Specifically, the Oversight Board ruled that "death to Khamenei" here was better interpreted as "down with Khamenei," a "clearly rhetorical" expression of criticism, disdain, or disgust for Khamenei, rather than a genuine threat or call for his assassination. In other words, it ruled that the speech was not a violation of the Violence and Incitement Community Standard at all.

6 At the time of this article's publication, Meta's description of the newsworthiness allowance (available at https://transparency.fb.com/en-gb/features/approach-to-newsworthy-content/) states:

> 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.

On this occasion, a *narrow* newsworthiness allowance was granted, not a *scaled* one, meaning that other posts of "death to Khamenei" were still subject to removal.

7 The strike against the user's account was also removed. However, it was impossible to reverse the temporary feature restrictions, which had already run their course.

The Oversight Board's interpretation of the user's speech was based on testimony from Farsi speakers familiar with Iranian current affairs (including language experts and members of the public who submitted comments on the case), who attested to a common rhetorical use of "death to Khamenei" as a political slogan in contemporary public discourse. Deployed in this way, the board ruled, the phrase does not threaten or call for violence against Khamenei but instead falls within the Community Standard's exemption for speech that expresses disdain or disagreement by threatening or calling for violence in *non-serious* ways.[8] The board judged that the author of the post was using "death to Khamenei" in its rhetorical sense, merely criticizing rather than threatening or calling for violence. As such, the post did not violate the Community Standard, nor did it require any special allowance to appear on the platform.[9]

In light of its findings, the Oversight Board issued a series of recommendations to Meta, including to revise the public-facing Violence and Incitement Community Standard and the internal guidance for moderators, so as to implement a more contextually nuanced treatment of seemingly threatening or inciting speech.

Meta has accepted the Oversight Board's guidance that "death to Khamenei" is a political slogan that is integral to the ongoing protests in Iran. The phrase is now allowed on the platform in that context (and previous enforcement actions are being reversed). At the time of writing, Meta is still considering several of the other recommendations made by the Oversight Board, including those to revise the Violence and Incitement Community Standard and internal guidance for moderators.[10]

How exactly should Meta proceed? Drawing on longstanding distinctions in the philosophy of language, we argue that the answer depends on whether platform rules should primarily target an utterance's illocutionary force or instead its perlocutionary effects. The analysis we provide in what follows is

8 In fact, the Community Standard states only that Meta "understands that" people commonly express disdain or disagreement by threatening or calling for violence in nonserious ways, not that such speech is permitted. This would seem to leave room for a stricter interpretation than the Oversight Board's presumption of an exemption, although we will not pursue the point here. We assume the policy contains a deliberate carveout for "nonserious" speech.

9 This view is reminiscent of the position adopted by the US Supreme Court, which has held that many statements that may *appear* to threaten are in fact mere hyperbole and do not constitute what it calls "true threats." Such cases of hyperbole are ubiquitous in political discourse, which is often "vituperative, abusive, and inexact." See *Watts v. United States* 394 U.S. 705 (1969).

10 Details of Meta's response are taken from https://transparency.fb.com/en-gb/oversight/oversight-board-cases/caricature-of-ayatollah-ali-khamenei (accessed July 12, 2023).

intended to support the development of a principled, fair, and appropriate approach in content moderation policy and enforcement.

2. THE ILLOCUTION/PERLOCUTION DISTINCTION

It is striking how the different reasoning applied by Meta and its Oversight Board tracks a longstanding distinction in speech act theory. Yet this distinction is currently obscured in online speech governance, leading to inconsistent—and, we will argue, unfair—results for users. By demonstrating the normative significance of the distinction, we aim to offer guidance on how content rules should be interpreted and enforced by platforms. Given that large online platforms are now arguably the most important fora for the exercise of our communicative liberties, it is vital that they be governed by clear and defensible principles.

What exactly is the distinction we have in mind? Austin famously described several different kinds of acts we perform when speaking—or things we *do* with words.[11] Most relevant to the current discussion are his categories of "illocution" and "perlocution." We will present each in turn and explain how they pertain to the "death to Khamenei" case, before proceeding to offer our central normative argument.[12]

2.1. Illocution

We perform illocutionary acts in using words with a particular *force*, such as telling, instructing, warning, promising, or thanking. Consider the sentence "The window is open." Even after accounting for the meanings of its constituent words and the way they are combined, there is still scope for this sentence to be uttered with differing force. In one scenario, a speaker could simply be informing her audience that the window is open. On a different occasion, though, uttering "The window is open" could be a request or a command to close it (say, where the speaker manages a building in which it is forbidden to open the windows, and she has just walked in on a tenant who is breaking the rules). In yet another kind of context, "The window is open" could be a warning (say,

11 Austin, *How to Do Things with Words*.

12 The claim that distinctions within Austinian speech act theory can shed light on pressing normative issues is by now familiar in social philosophy. See, for example, Langton, "Speech Acts and Unspeakable Acts" and "Blocking as Counter-speech"; Langton and Hornsby, "Free Speech and Illocution"; Hornsby, "Disempowered Speech"; McGowan, "Conversational Exercitives" and "Oppressive Speech"; Maitra, "Silencing Speech"; Dotson, "Tracking Epistemic Violence, Tracking Practices of Silencing"; Kukla, "Performative Force, Convention, and Discursive Injustice"; Hesni, "Illocutionary Frustration"; and Schiller, "Illocutionary Harm."

to the parent of a small child who is entering a fourth-floor apartment with floor-to-ceiling windows).

Returning to our case, the phrase "death to Khamenei" similarly underdetermines the illocutionary act being performed. What act, exactly, did the user in question perform in posting it? Were they threatening the life of the Iranian leader or calling for others to assassinate him?[13] Were they criticizing the leader and his regime? Plainly, this question is at the heart of the disagreement between Meta and the Oversight Board. Whereas Meta took itself to be dealing with a threat or call for violence against Khamenei (later deciding that the speech should be allowed anyway), the Oversight Board interpreted the utterance as an act of mere criticism.[14]

How should the dispute be adjudicated? This depends in part on a contested question in speech act theory concerning which criteria determine illocutionary force. Broadly speaking, theorists have appealed to three conditions for performing illocutionary acts: (i) the speaker's *intending* to perform that act (in this case, intending to threaten or intending to criticize); (ii) the audience's "uptake" of the act (whether they *interpreted* the speaker to be either threatening or criticizing); and/or (iii) the prevailing social conventions for performing the act being met (including the speaker's having used a particular kind of linguistic formulation, in a particular kind of context, with the requisite degree of authority). As is to be expected, philosophers of language disagree as to which of conditions i–iii are necessary or sufficient for the performance of any given illocutionary act.[15]

While we cannot resolve that debate here, we will need to make *some* minimal assumptions in order to apply the notion of illocution in our context. We assume that what is relevant for online speech governance is *how the illocutionary force of a post would reasonably be interpreted by its audience*. In standard cases, what audiences would reasonably interpret the speech to be doing will be sensitive to shared evidence (available to audience members, speakers, and platform adjudicators) about the relevant language, speaker, and wider context.

13 Following Searle's classification in "A Taxonomy of Illocutionary Acts," threats are standardly treated as illocutionary acts.

14 Those familiar with Austin's work might worry that rhetorical uses of language simply fall outside the scope of the theory, being an instance of nonliteral or nonserious speech (akin to acting in a play, making a joke, or writing a poem, which Austin explicitly excluded from consideration). However, even if one did not believe that any felicitous illocutionary act was performed with "death to Khamenei," that would still be sufficient to undercut Meta's claim that the user was threatening or calling for violence, and that is the important point for our purposes here.

15 For relevant discussion, see Strawson, "Intention and Convention in Speech Acts"; Searle, *Speech Acts*; and Sbisà, "How to Read Austin." For an overview of the contemporary literature on speech act theory, see Fogal, Harris, and Moss, *New Work on Speech Acts*.

The proposed approach is broadly friendly to those who emphasize the importance of the speaker's intention, insofar as reasonable interpretations of speech are those that attempt to recover what the speaker was trying to do. The approach also recognizes a role for uptake, insofar as it is the audience's interpretation that is being estimated. Finally, our approach respects the importance of social conventions, insofar as reasonable interpretations rely on what a speaker of a particular kind would normally be doing with those words in that context.[16] What we end up with then is an illocutionary category tailored to the specific case of speech governance that tracks the speaker's (likely) intentions, the (reasonable) audience's uptake, and the surrounding conventions (other things being equal) without giving priority to any one criterion in its pure form.[17]

2.2. Perlocution

Perlocutionary acts are things we do *by means of* our utterances—the effects our speech achieves in the world. For example, by saying "The window is open" I might cause you to believe that fact; alternatively, I might cause you to close the window; or I might cause you to take additional care to avoid accidents. Across these cases, the focus is on the *consequences* of speech. Likewise, in the "death to Khamenei" case, the perlocutionary question concerns what effects the speaker was (likely to be) producing by means of their speech. Were they causing Khamenei to experience intimidation, or inducing audience members to attempt an assassination? Or alternatively, would the words simply have the effect of raising awareness, causing audiences to reflect on the oppressive government and perhaps take up protest against it?

16 There are complications across all three dimensions. First, a clumsy or misguided speaker might intend to do one thing but reasonably be interpreted as doing another. In some cases, this will be due to negligence or recklessness on the speaker's part, as will be discussed in section 4. Further, we note the possibility that an unreasonable audience might interpret the speaker as doing one thing while a reasonable audience interprets them as doing another. In section 4 we will discuss why we think it is appropriate for content moderation to track the latter. Finally, we note the possibility that available evidence about a particular user or context might affect what interpretation is most reasonable, all things considered, even if that diverges from convention.

17 Even so, the argument below would apply *mutatis mutandis* if we were to adopt (implausibly, we suspect) a single necessary-and-sufficient criterion of illocutionary force. For example, a hardcore intentionalist version would require content moderators to focus on only what the speaker was trying to say, whereas an uptake-centric version would require them to focus on how audiences actually interpret the speech.

There is an obvious and close connection between the illocutionary and perlocutionary aspects of an utterance.[18] For instance, telling *S* that *p* tends to result in *S* coming to believe that *p*; requesting *S* to ϕ tends to result in *S* ϕ-ing; and so on. More precisely, illocutionary acts have the *normal function* of producing the corresponding perlocutionary effects. Yet they will not produce those effects on every occasion, meaning that illocutionary force and perlocutionary effects can diverge in practice. For example, I might tell you that the window is open, but you, seeing it clearly closed, may refuse to believe me. By the same token, a particular threat or call for violence might fail to result in harm.[19] Conceptually, then, the illocutionary force of an utterance can be held apart from its perlocutionary effects, even if the two normally go hand in hand. We will see later why this is so important for thinking about the governance of speech.

With the illocution/perlocution distinction in place, we can rationally reconstruct the positions of Meta and the Oversight Board. For Meta, the author of the "death to Khamenei" post performed a particular illocutionary act—the act of threatening or calling for the killing of Khamenei. The perlocutionary effects of this act included some risk of actual harm, but they also included beneficial awareness raising. The newsworthiness allowance was applied following a cost-benefit analysis of these perlocutionary effects, which determined that the benefits of the speech outweighed the costs. Yet that determination did not alter the illocutionary status of the speech as a threat or a call for murder; the speech constituted a violation, even though it was exempted due to its instrumental benefits.

In contrast, the Oversight Board interpreted the same speech as an illocutionary act of criticism, which was legitimate and protected by Facebook policy, independent of whatever downstream perlocutionary effects (good or bad) it might have had in the situation at hand.

All of that said, it turns out that neither party managed to hold apart illocutionary and perlocutionary issues in a fully consistent way. By analyzing those

18 There can also be disagreements over whether a particular verb is illocutionary or perlocutionary. Indeed, "inciting" seems to be one of these; for relevant discussion, see Kurzon, "The Speech Act Status of Incitement." Because of that, we prefer to contrast threatening or calling for violence (illocutionary) with instigating violence (perlocutionary).

19 It remains a contested issue in speech act theory whether perlocutionary effects should include only those that the speaker intended or also any unintended effects of their speech. For the purposes of applying the theory to content moderation, we wish to include all downstream effects, whether intended or unintended. Those who object to this use of "perlocution" are free to substitute a different label.

inconsistencies in the next section, we will then be in position to recommend our preferred systematic approach.

3. Conflations, Complications, and Confusions

Platforms must decide whether a given utterance violates or does not violate their rules. When deciding this, they might focus on the illocutionary force of the utterance—i.e., whether it constitutes a speech act belonging to a prohibited category, such as an act of threatening or an act of calling for murder. Alternatively, they might focus on the perlocutionary force of the utterance—i.e., its (likely) causal effects in the world. Or they might opt for some principled combination of the two. Soon we shall explain what we think the right position on that is. Before doing so, we will linger on our core case slightly longer to show why platforms' current thinking on this issue is unhelpfully muddled.

We start with the Oversight Board's conflation of illocutionary and perlocutionary issues. The board notes, quite reasonably, the importance of context in determining the force of "death to *X*" utterances. However, it then considers a hypothetical case in which the target is Salman Rushdie rather than the Ayatollah Khamenei. The board argues that this case would pose a much more significant risk of harm (due to the fatwa against Rushdie, the recent attempt on his life, and ongoing concerns for his safety) and would therefore need to be taken more seriously. While it is no doubt true that a threat against Rushdie would be more credible than one against Khamenei, this is a fact about the downstream dangerousness of the threat *once issued*. It does not tell us whether such a speech act (the act of threatening) was performed in the first place. The risks facing Rushdie do not themselves make "death to Rushdie" more likely to be an illocutionary act of threatening or calling for violence than "death to Khamenei." On the contrary, the Oversight Board's decision on the "death to Khamenei" case rests on the claim that "death to" here has a rhetorical use equivalent to "down with." The fact that the "death to Khamenei" post was deemed to instantiate this rhetorical use was what exempted the post from enforcement action. Presumably, then, moderators must rule out the possibility of any other "death to" statement being merely rhetorical before taking enforcement action under the Violence and Incitement Community Standard. This applies equally to a post of "death to Rushdie" as to a post of "death to Khamenei." In sum, changing the identity of the target cannot settle what the (violating or nonviolating) illocutionary force of a "death to" statement is.[20]

20 Perhaps the board would ultimately want to say that the rhetorical use of "death to" is possible only for particular targets, say those who hold political office or those who feature

It seems then that the Oversight Board *does* want moderators to consider perlocutionary effects, despite its claim that acts of criticism should be excluded from consideration. In its discussion of Rushdie, the board implies that even rhetorical uses of "death to *X*" might sometimes be appropriate targets of enforcement, so long as the risks to *X*'s safety are sufficiently high. If this is the right interpretation, the board's reasoning is unclear. Should Meta take a view on the illocutionary force of a "death to" speech act before deciding whether to take enforcement action? Or should it first assess the likely risk of harm to the target? Or should it do both concurrently? The answer is not clear.

That confusion, visible at the level of enforcement and review, is, we think, a direct result of the conflation of illocutionary force and perlocutionary effects in Meta's underlying policy. On one hand, the overarching rationale for moderating content (under the Violence and Incitement Community Standard and other policy provisions) is to prevent harm, i.e., damaging perlocutionary effects of speech. In line with this, Meta states that it will take enforcement action where speech "facilitates serious violence," where there is believed to be a "genuine risk of physical harm or direct threats to public safety," and where these threats are considered "credible." On the other hand, there are carve-outs for users to "express disdain or disagreement by threatening or calling for violence in non-serious ways"—speech that does not in fact threaten or call for violence but has some other illocutionary force.

One might think that nonserious threats are permitted precisely because they tend not to cause harm. As we saw in section 2.2, though, the connection between illocutionary and perlocutionary acts is not one that is guaranteed to hold on every occasion, being contingent on a range of ad hoc contextual factors. There could, for example, be particular illocutionary acts of criticism that are in fact likely to lead to violence; on the flipside, there could be particular illocutionary acts of threatening that are unlikely to do so.

Imagine, first, a celebrity who has a coterie of zealous fans. The celebrity posts something that is intended as a mere criticism (say, "down with Khamenei," in the context of a peaceful political protest movement) and is universally interpreted that way. Yet the zeal of her hardcore fans is such that they will take it upon themselves to attempt serious harm on anyone their idol disagrees with. We can even imagine that this fact is known to the platform's moderation team, making the offline harm entirely foreseeable to them. In this case, the post is likely to instigate violence without having threatened or called for it. Should such speech acts be protected, regardless of their potentially harmful effects?

---

in particular protest slogans. However, even if such an argument could be sustained, it is not provided in the case ruling.

On the flipside, a post that is intended—and widely understood to be—a call for violence (say "One of you should kill Khamenei when he leaves the Beit Rahbari compound at 10 AM tomorrow") may nevertheless carry a very low probability of causing harm to the target. We can imagine that the user has very few followers, none of whom is in a position to attempt an attack, and the compound in question is extremely well protected. In this case, illocutionary and perlocutionary aspects come apart in the opposite direction: the speech has the illocutionary force of a threat or call for violence but no prospect of generating harmful perlocutionary effects. Should such a post be removed despite its extremely low risk?

These two limit cases illustrate the need to clarify whether illocutionary or perlocutionary factors are driving content moderation. Even though illocutionary force and perlocutionary effects are closely connected, we have seen that they can and do diverge. Thus, the question of whether a post constitutes a call for violence is simply not the same as the question of whether it is likely to cause offline harm. By conflating both aspects, policies like Facebook's Violence and Incitement Community Standard fail to give us clear principles for handling the limit cases sketched above and a vast spectrum of intermediate cases that arise at enormous scale and frequency on social media platforms. As we have seen, this confusion at the policy level destabilizes the reasoning of Meta and the Oversight Board during enforcement and review.

4. THREE MODELS OF CONTENT MODERATION

This brings us to the foundational question raised through our case study: Should content moderation on social media platforms target the illocutionary or perlocutionary aspects of speech? For example, when specifying what counts as a forbidden "threat," should platforms emphasize the illocutionary aspect, the perlocutionary aspect, or some combination thereof?[21] At a minimum, we submit that platforms should take a consistent and transparent line in their policy and enforcement. This is both for the sake of moderators, who need a procedure for resolving cases where illocutionary force and perlocutionary effects diverge, and for the sake of users, who have an interest in knowing how their speech will be handled. Indeed, given the important role played by platforms like Facebook in facilitating the speech of billions of people, the

21 Specifying what should count as a forbidden threat for the purposes of online content moderation (or, relatedly but distinctly, what should count as a forbidden threat for the purposes of criminal law) is different from the question of what counts as a threat for the purpose of conceptual or linguistic analysis. Among other reasons, concerns of free speech are relevant for specifying the former but not the latter.

current lack of a clear moderation principle arguably impairs their enjoyment of an intrinsic good—i.e., their ability to express themselves. Meanwhile, at the social level, we are denied a proper balance between the good of free expression in increasingly important online environments and the bad of potentially harmful effects. Accordingly, it is a moral imperative for platforms to get clear on whether their content moderation targets illocutionary force or perlocutionary effects.[22]

In what follows, we contrast three possible approaches to content moderation that focus respectively on perlocutionary effects, illocutionary force, and (as we prefer) a systematic combination of the two.

4.1. Moderating Perlocution

Given that the end goal of content moderation is to minimize the harmful effects of speech, a natural suggestion would be to adopt a purely perlocutionary strategy whereby each utterance is moderated based on its likely resulting harm. On such an approach, enforcement action would be taken against posts that exceed some threshold of risk, as a function of the probability of the harmful effect occurring and its degree of harmfulness. Meanwhile, posts deemed unlikely to cause harm would be left to stand and spread.

Such a strategy immediately runs into two central difficulties. First, it fails to offer clear normative guidance about what exactly platforms ought to do. That is because our ability to predict the effects of particular utterances is highly limited. But second, a purely perlocutionary approach would lead to objectionable overmoderation, licensing the removal of potentially large swaths of legitimate speech. If all that matters is whether an utterance has a contingent effect of causing some harmful result downstream, plenty of legitimate speech that merits protection will be vulnerable to censorship.

22 It might be objected at this point that because platforms are private companies, they do not wrong their users when they remove legitimate speech. Even if a state would wrong them by removing such speech, a *company* does not wrong them by removing it. In reply, though, we note that platforms have themselves committed to respect the value of users' expressive and communicative interests. Meta has expressly committed itself to respecting "voice" as a paramount value. In this way, platforms have voluntarily taken on such a moral obligation. While this modest point is sufficient to establish that they can wrong their users by removing legitimate speech, we are further tempted by a stronger thesis: that platforms exercise a kind of governance power over the public discourse and, in virtue of this power, are bound by similar principles to respect free speech as states are. That stronger thesis underpins the Oversight Board's provocative contention that principles of international human rights law should govern content moderation decisions by the large platforms. For discussion, see Howard, "The Ethics of Social Media."

To see the problem, consider again the limit case raised above involving a celebrity with overly zealous fans. We saw there how an entirely innocent illocutionary act, intended and interpreted as a mere criticism, could nevertheless have foreseeably harmful effects due to the unreasonableness of a small number of audience members. Giving this unreasonable minority the normative power to silence the speaker, rendering her speech eligible for removal, would entirely misallocate the burdens: the unreasonable fans, not the celebrity, are chiefly responsible for any violence that ensues.[23] Thus, a purely consequentialist focus on harm leaves platform users' speech rights hostage to factors that are intuitively irrelevant, such as how many unreasonable people follow them or how much security the subject of their critical speech happens to have.

By the same token, ignoring illocutionary force also disregards the expressive interests of speakers. Suppose that it is true that by posting "death to Khamenei," the user was engaging in speech that was both intended and (more importantly) likely to be understood as vociferous protest against an authoritarian regime. Such expression is at the heart of what the right to free speech protects. It is not plausible to suggest that Iranian users on Facebook have a moral duty to refrain from such expression.[24]

On the flipside, a purely perlocutionary approach would also lead to objectionable undermoderation of illegitimate speech. For instance, if a user issues a threat or call to violence, it is much less plausible to think that they have a moral right to engage in such speech. There is a moral duty to refrain from issuing death threats and encouragements to murder; such activity familiarly falls outside the protective scope of free speech. Therefore it is plausibly an abuse of a platform to engage in such speech, which, while having the normal function of instigating harmful behavior, does not depend for its wrongness and unprotected status on causing harm in every instance.[25]

23 This is compatible with the claim that speakers have moral duties to look out for ways in which unreasonable listeners might misinterpret their speech and to rearticulate or clarify their messages in ways that reduce such risks. Just because a speaker has a right to express a certain message does not mean that there are no moral considerations that bear on *how* she ought to express that message. But we are skeptical that such highly nuanced issues arising from occasionally foreseeing the unreasonable responses of deranged audience members should limit what people are allowed to say. Rather, they bear on how speakers might use discretion when exercising their speech rights.

24 The intuitive force of this claim is doubtlessly bolstered by the fact that the Iranian regime is seriously unjust. But even if the regime were in fact just, free speech theories familiarly protect citizens' general right to protest regardless of the moral standing of their state.

25 See Howard, "Dangerous Speech." Some might think assassination of tyrants is legitimate, such that the illocutionary act of threatening or calling for such assassination is also legitimate. It is possible that people's intuitions on this issue may influence their reactions to

These considerations lead us to reject a purely consequentialist strategy of moderating the perlocutionary effects of speech (whether under Facebook's Violence and Incitement Community Standard or any other platform policy): such an approach unsurprisingly fails to attend to the nonconsequentialist significance of users' rights and duties, which militate in favor of centering illocutionary force.

4.2. Moderating Illocution

At the other extreme, we might be tempted to focus entirely on the illocutionary act performed by a piece of online speech. Thus, a post would be removed if deemed to be a violating illocutionary act, such as threatening or calling for violence (even if it is in fact unlikely to lead to harm in a given case). Meanwhile, legitimate illocutionary acts would be left alone (regardless of any harmful effects they might have).

On our view, this approach is nearly correct. It respects users' freedom of speech by protecting utterances that have valuable or innocuous illocutionary force; speakers have no moral duty to refrain from such speech and indeed have a right against its restriction. In contrast, speakers do not enjoy rights to achieve perlocutionary effects—there is no plausible right, for example, to convince, persuade, or instigate a particular chain of events, whereas there is a right to assert, argue, criticize, and so on. The right to free expression sits at the illocutionary level; and speech enjoys a blanket protection wherever it is reasonably interpreted as having innocent illocutionary force (even if it could end up being harmful for reasons beyond the speaker's control).

At the same time, the illocutionary approach enables platforms to target categories of speech acts—like wrongful threats and calls for violence—that normally function to inflict serious wrongful harms on others. Partly in virtue of the harms that such utterances are calculated (and often tend) to produce, they lack value *qua* free expression since (on the standard view) such unprotected speech is utterly disconnected from the values (such as autonomous expression, democratic citizenship, and the search for truth) that justify free speech in the first place.[26] This is especially clear in cases where the user *intends*

---

the case under discussion. We set aside this complication here, as it plays no explicit role in the official reasoning about the case. Meta disallows calls for assassination of political leaders, even seriously autocratic ones.

26 We assume that freedom of speech is a moral principle that makes it very difficult to justify restrictions on communication by public (and some private) institutions, especially restrictions that silence particular viewpoints. We further assume that this principle is justified by a plurality of interests (of both speakers and audiences), including autonomous self-expression and self-development, education, and democratic self-government. Such

to engage in the relevant speech act; she is not contributing to public discourse by sharing her opinion on matters of public concern but rather endeavoring to cause harm. Such speech is indeed a violation of duties she owes to others. But even in cases where there is no such intention, if the reasonable construal of the utterance is that it constitutes a speech act of threatening or calling for violence, the user may be violating her duty through recklessness or negligence. Such a user cannot stand on her free-speech rights to immunize her speech from interference; she too violates duties she owes to others not to perpetrate wrongful utterances.[27]

Notwithstanding its virtues, a purely illocutionary strategy depends upon a simplifying assumption: that any given utterance admits of only one reasonable interpretation. Yet this is plainly false. In reality, a single utterance can often have multiple competing illocutionary interpretations, each of which may be deemed similarly plausible given the available evidence about the language, the speaker, and the wider context. In fact, the "death to Khamenei" post seems to exhibit just this kind of ambiguity, despite the Oversight Board's confident assertion that it was mere criticism.

The complexities of establishing an utterance's illocutionary force are evident from ongoing debates in speech act theory and should in no way be underestimated. Across a wide range of cases, there are likely to be genuine difficulties in determining which illocutionary act a speaker has performed. After all, Austin's core insight (discussed in section 2.1) was that a speaker's form of words commonly underdetermines what is done in uttering them. This immediately creates room for uncertainty and disagreement about illocutionary force.

Moreover, the context for a social media post is often so sparse that homing in on a unique illocutionary act is simply impossible. Accordingly, we will often need to hold open different possibilities.[28] This points to the need to nuance

---

a pluralist view of the grounds of free speech is widely held in the scholarly literature and avoids the challenge of having to decide what the *real* justification for free speech is. For one pluralist view, see Cohen, "Freedom of Expression."

27 See Howard, "Dangerous Speech," for an argument along these lines in the context of speech calling for wrongful violence.

28 For some speech act theorists, the question of which illocutionary act is performed may even be more deeply and metaphysically indeterminate. Some argue, for example, that a single utterance can have multiple illocutionary forces—and not just because of a divergence between the primary and secondary speech act (as when a speaker asks a question in order to make a request) but also because of different interpretations at one of those levels of analysis. For relevant discussion, see Sbisà, "Some Remarks about Speech Act Pluralism"; Johnson, "Investigating Illocutionary Monism," "Mansplaining and Illocutionary Force," and "Illocutionary Relativism"; and Lewiński, "Illocutionary Pluralism." Even

the illocutionary approach in its application to online content moderation: a pure version of the approach can tell us only what to do with speech deemed to perform violating or nonviolating illocutionary acts; it has nothing to say about the (potentially many) cases where there is genuine uncertainty about which type of act was performed.

4.3. A Hybrid Approach

We suggest that a hybrid approach to content moderation will be most defensible. Such an approach, we argue, should synthesize illocutionary and perlocutionary considerations in a systematic and predictable way rather than haphazardly conflating the two, as in current platform policy. There are potentially several different ways to pursue a hybrid approach. For example, one might treat violating illocutionary force and harmful perlocutionary effects as individually necessary and jointly sufficient conditions for enforcement action to be taken (such that a post must be reasonably interpreted as a threat or call to violence *and* likely to lead to violence, in order to be a legitimate target for moderation). Or one might treat each condition as individually sufficient (such that a post must be reasonably interpreted as a threat or call to violence *or* likely to lead to violence, in order to be a legitimate target for moderation).

The arguments in the previous two subsections, however, point to an important asymmetry between illocutionary and perlocutionary aspects of speech: illocutionary force is far more closely linked to speakers' rights and duties and thus to predictable, justifiable restrictions of their speech, whereas perlocutionary effects are more arbitrary. Accordingly, we propose an "illocutionary-first" moderation strategy. The strategy is comprised of three rules:

1. If an utterance performs a legitimate illocutionary act (such as a criticism), it should not be moderated.
2. If an utterance performs an illegitimate illocutionary act (such as a threat or a call for violence), it should be moderated.
3. If an utterance's illocutionary status is ambiguous between legitimate and illegitimate acts, it should be moderated only if it has net harmful perlocutionary effects.

In effect, we propose to supplement the illocutionary approach described in section 4.2 with a procedure for resolving cases of uncertainty—namely, to

---

if the pluralist view is correct, we believe our proposed approach remains the right one: where multiple interpretations are available (regardless of whether they are understood in terms of illocutionary ambiguity or illocutionary pluralism), content moderators will need to decide what to do about the potentially violating speech; and this is when they should look to perlocutionary effects.

conduct a cost-benefit analysis, weighing the probability of harm against the probability of benefit (roughly along the lines of Meta's "newsworthiness" test).

In principle, other resolution procedures are available at step iii. Most obviously, one could simply apply a blanket approach in cases of uncertainty, either subjecting them all to moderation (on the basis that the speakers should have been clearer about their illocutionary intents) or exempting them all (and giving speakers the benefit of the doubt). However, we believe these blanket approaches fail to track speakers' rights and responsibilities sufficiently closely.

First, speakers have substantial interests in expressing themselves, even if they are not always perfectly clear about what they are saying or doing with their words. Given the scope for performing different illocutionary acts with the same words, even in similar contexts, and given humans' bounded capacities to optimize language choice (considering, for example, our imperfect knowledge, finite vocabularies, performance frailties, and so on), it would be unreasonable to expect all utterances to have clear illocutionary force. A blanket policy of shutting down all *potentially* violating speech would therefore involve unacceptable costs for speakers' rights.

On the other side, a blanket exemption for such speech would be too risky. In certain cases, ambiguous speech is dangerous. Consider President Donald Trump's exhortation to his followers to "fight like hell."[29] Was this an illocutionary act of calling for his supporters to attack the Capitol? Or was it merely advocating vigorous but peaceful protest? It was ambiguous. Trump, we suggest, had a moral duty to be clearer about what precisely he meant in that case; it is this duty that explains why Trump would not have been wronged had the ambiguous speech been taken down in such a case. Here the crucial distinction is between *innocuous ambiguity*—where the illocutionary force is ambiguous, but it is no big deal because perlocutionary risks are low—and *dangerous ambiguity*—where the illocutionary force is ambiguous, and there is indeed a serious risk of harm. In the latter cases (but not the former), speakers have a duty to clarify what they mean. If they do not, they are liable to have their speech moderated. (One could imagine a mechanism whereby ambiguous posts trigger an auto-prompt, encouraging speakers to clarify the meaning of their post.) This is the rationale for appealing to perlocutionary effects as the arbitrating factor: *when the stakes are high, users must take pains to clarify the legitimate illocutionary function of their speech.*

What does this mean for the "death to Khamenei" case? If the Oversight Board was right that this speech, in context, was clearly a mere criticism, the first step of our illocutionary-first approach requires that the post be left

29 See Naylor, "Read Trump's Jan. 6 Speech."

unmoderated. However, if, as we suspect, it was genuinely unclear whether the post was threatening or calling for violence, the third step in our procedure turns to the likely perlocutionary effects. At this point, we agree with Meta that the risk of harm to Khamenei was negligible and outweighed by the opportunity to raise awareness and foster political resistance.

Our illocutionary-first strategy nicely reflects the underlying purpose of content moderation. Ultimately, platforms seek to restrict speech because of its connection to wrongful harm. The reason why illocutionary acts are appropriate targets for moderation is because of the perlocutionary effects they normally function to produce without being subject to all sorts of contingent factors affecting whether or not those effects are produced on any given occasion. In other words, the focus on illocution never takes harmfulness out of the picture but seeks to counter it in a way that best respects a speaker's freedom of expression. This reveals the illocutionary-first strategy to be thoroughly in the business of finding the appropriate balance between free speech and the avoidance of harm—and makes the appeal to perlocutionary effects in cases of uncertainty a very natural one.

Further, our approach avoids counterintuitive implications of a purely perlocutionary approach. Wherever ambiguous speech is removed under an illocutionary-first approach, the user can typically rearticulate their post so as to produce an illocutionary act that is more clearly permissible (and thus exempt from moderation). In contrast, under a purely perlocutionary strategy, this option would not be available in the same way, since no illocutionary acts would be automatically protected (instead requiring case-by-case assessment of perlocutionary effects).

In closing, our proposal is decidedly *not* that platforms should start adding more philosophical jargon to their rules. Rarefied Austinian terminology is, we suspect, best left to philosophy journals. The point instead is that platforms should recognize that when enforcing their rule against, for example, threats, they should focus first on whether the speech *constitutes* a threat—something determined by what reasonable audiences would likely infer. If it does not, it should be allowed. If it definitely constitutes a threat, it should be taken down. And if it is unclear, a cost-benefit analysis of likely effects is necessary. Platform employees should be able to understand that order of operations without any fancy nomenclature.

5. CONCLUSION

We have argued for an illocutionary-first approach to online content moderation that primarily enforces against violating illocutionary acts while protecting

those that are nonviolating. The proposed approach has direct implications for moderation practices: human reviewers should first seek to establish the illocutionary force of a piece of online speech (say, threatening, calling for violence, or merely criticizing) and assess its perlocutionary effects (say, instigating violence or peaceful protest) only in cases of genuine illocutionary uncertainty. By the same token, automated moderation systems should not integrate signals relating to perlocutionary effects (such as virality) with signals relating to illocutionary force (such as common rhetorical use in political protest).

While we have arrived at this position by closely examining the specific case of a Facebook user posting "death to Khamenei," it clearly applies well beyond the individual post, the Violence and Incitement Community Standard, and the Facebook platform to policies adopted by social media platforms in general.[30] In this way, we take ourselves to be putting forward a foundational principle that will help ensure that ever larger swaths of speech in the increasingly online world can be justly and robustly supported.[31]

University College London
sarah.a.fisher@ucl.ac.uk
jeffrey.howard@ucl.ac.uk

REFERENCES

Austin, J. L. *How to Do Things with Words*. Oxford: Clarendon Press, 1962.

Cohen, Joshua. "Freedom of Expression." *Philosophy and Public Affairs* 22, no. 3 (Summer 1993): 207–63.

Dotson, Kristie. "Tracking Epistemic Violence, Tracking Practices of Silencing." *Hypatia* 26, no. 2 (Spring 2011): 236–57.

Fogal, Daniel, Daniel W. Harris, and Matt Moss, eds. *New Work on Speech Acts.* Oxford: Oxford University Press, 2018.

Hesni, Samia. "Illocutionary Frustration." *Mind* 127, no. 508 (October 2018): 947–76.

Hornsby, Jennifer. "Disempowered Speech." *Philosophical Topics* 23, no. 2 (Fall

30 As this paper was headed for publication, the Meta Oversight Board published a new decision exhibiting many of the same issues that we identify here. See decision 2023-032-IG-UA, available at https://www.oversightboard.com/decision/IG-6BZ783WQ.

31 We are grateful for feedback from audiences at the 2023 conference of the Society for Applied Philosophy and the 2023 Joint Session of the Aristotelian Society and the Mind Association, as well as from two anonymous reviewers for this journal. The work was funded by UK Research and Innovation (grant reference MR/V025600/1).

1995): 127–48.
Howard, Jeffrey W. "Dangerous Speech." *Philosophy and Public Affairs* 47, no. 2 (2019): 208–54.
———. "The Ethics of Social Media: Why Content Moderation Is a Moral Duty." *Journal of Practical Ethics* (forthcoming).
Johnson, Casey Rebecca. "Illocutionary Relativism." *Synthese* 202, no. 3 (2023): 1–18.
———. "Investigating Illocutionary Monism." *Synthese* 196, no. 3 (March 2019): 1151–65.
———. "Mansplaining and Illocutionary Force." *Feminist Philosophy Quarterly* 6, no. 4 (2020).
Kukla, Rebecca. "Performative Force, Convention, and Discursive Injustice." *Hypatia* 29, no. 2 (Spring 2014): 440–57.
Kurzon, Dennis. "The Speech Act Status of Incitement: Perlocutionary Acts Revisited." *Journal of Pragmatics* 29, no. 5 (May 1998): 571–96.
Langton, Rae. "Blocking as Counter-speech." In Fogal, Harris, and Moss, *New Work on Speech Acts,* 144–64.
———. "Speech Acts and Unspeakable Acts." *Philosophy and Public Affairs* 22, no. 4 (Autumn 1993): 293–330.
Langton, Rae, and Jennifer Hornsby. "Free Speech and Illocution." *Legal Theory* 4, no. 1 (1998): 21–37.
Lewiński, Marcin. "Illocutionary Pluralism." *Synthese* 199, nos. 3–4 (2021): 6687–714.
Maitra, Ishani. "Silencing Speech." *Canadian Journal of Philosophy* 39, no. 2 (June 2009): 309–38.
McGowan, Mary Kate. "Conversational Exercitives: Something Else We Do with Our Words." *Linguistics and Philosophy* 27, no. 1 (February 2004): 93–111.
———. "Oppressive Speech." *Australasian Journal of Philosophy* 87, no. 3 (2009): 389–407.
Naylor, Brian. "Read Trump's Jan 6. Speech, A Key Part of Impeachment Trial." NPR, February 10, 2021. https://www.npr.org/2021/02/10/966396848/read-trumps-jan-6-speech-a-key-part-of-impeachment-trial.
Sbisà, Marina. "How to Read Austin." *Pragmatics* 17, no. 3 (January 2007): 461–73.
———. "Some Remarks about Speech Act Pluralism." In *Perspectives on Pragmatics and Philosophy*, edited by Alessandro Capone, Franco Lo Piparo, and Marco Carapezza, 227–44. New York: Springer, 2013.
Schiller, Henry Ian. "Illocutionary Harm." *Philosophical Studies* 178, no. 5 (May 2021): 1631–46.

Searle, John R. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press, 1969.

———. "A Taxonomy of Illocutionary Acts." In *Language, Mind and Knowledge*, edited by Keith Gunderson, 344–69. Minneapolis: University of Minnesota Press, 1975.

Strawson, P. F. "Intention and Convention in Speech Acts." *Philosophical Review* 73, no. 4 (October 1964): 439–60.