Dangerous-Sounding Speech: Semantic Proximity and the Inversion of Safety

Flyxion

September 2026

Abstract

Automated and semi-automated content moderation systems routinely overweight a speech act's lexical proximity to a dangerous subject relative to its actual causal role in that danger. This produces a specific and recurring error: language that names harm explicitly in order to prevent, document, or treat it is treated as more dangerous than language that accomplishes harm through euphemism, implication, or coded reference. The paper states this as a formal inequality between semantic proximity and signed causal contribution, distinguishes three layers of language that moderation systems tend to overweight toward the most cheaply computable (lexical content, speech act, operational role), and traces the same inversion across eleven domains, anchoring the pattern in documented moderation incidents where available. It argues further that the error is not ordinary classifier noise but an adversarial asymmetry: euphemistic evasion advances a harmful actor's goal while degrading a protective speaker's, so the same corrective pressure disables one party and assists the other, and that under repeated cycles of enforcement and adaptation this asymmetry compounds, with harmful speech growing more coded and protective speech growing less informative over time. It closes by arguing that the resulting bias is directionally regressive in a definable sense: it rewards actors fluent in institutional euphemism and penalizes actors, often the most vulnerable ones, who can only speak plainly.

1 The Error Stated Formally

A moderation system, whether a keyword filter, a trained classifier, or a human reviewer working under time pressure and a policy checklist, must infer something it cannot fully observe (whether a given utterance advances or forecloses a harm) from something it can observe far more cheaply (the utterance's surface features, and at best some of its immediate context). Call the first observable quantity $\mathrm{SemProx}(x, H) \geq 0$, the semantic proximity of an utterance x to a harm category H. Let $\mathrm{CausalContrib}(x, H) \in \mathbb{R}$ denote the signed causal contribution of x to H: positive values indicate that x moves the world toward H occurring, negative values indicate that x moves the world toward preventing, interrupting, or mitigating H, and magnitude indicates how much. The working assumption that lets a system moderate at scale is

$$\mathrm{SemProx}(x, H) \;\approx\; f\big(\,\mathrm{CausalContrib}(x, H)\big)$$

for some monotone, sign-preserving f: roughly, "sounds more dangerous" is treated as a usable stand-in for "is more dangerous." The central claim of this paper is that this assumption fails in a directionally biased way rather than a merely noisy one. In domain after domain there exist paired utterances x and y such that

$$\text{SemProx}(x, H) > \text{SemProx}(y, H) \qquad \text{while} \qquad \text{CausalContrib}(x, H) < \text{CausalContrib}(y, H),$$

with $\text{CausalContrib}(x, H)$ often strongly negative (protective) and $\text{CausalContrib}(y, H)$ often positive (harm-advancing), even though x scores higher on the observable proxy. Here x is typically the explicit, clinical, or evidentiary utterance: the overdose-recognition checklist, the anatomically precise consent explanation, the survivor's itemized account of threats, the disassembled malware sample in a defensive report. And y is typically the euphemistic or coded utterance that accomplishes the harm, or clears the way for it, while sounding unremarkable: the veiled threat, the coordinated but individually mundane instruction, the wellness pitch that never names the disease it promises to cure. The system is not simply making occasional mistakes in an unbiased way. It is ranking x and y in the wrong order, systematically, in exactly the high-stakes cases moderation exists to handle, where the two quantities can become locally anti-correlated rather than merely imperfectly aligned. (No population-level statistic is claimed here; the evidence below establishes recurrent ranking reversals in documented cases, not a measured correlation across a dataset.)

Two clarifications on the scope of this claim. First, the inequality is existential, not universal: the paper does not claim that SemProx and CausalContrib are anti-correlated in general, only that pairs exhibiting the reversal recur reliably across independent domains, which is weaker than universality but stronger than anecdote, and is the claim the eleven-domain survey in Section 3 is built to support. Second, and more important to state up front: this is not the ordinary claim that classifiers are noisy. A noisy classifier errs in both directions at roughly comparable rates and comparable cost. The claim here is that the error has a sign: it is biased toward flagging the protective utterance and away from flagging the harm-advancing one, and, as Section 5 develops, the cost of that bias falls unevenly on the two parties as well. Noise without sign would be a nuisance. A signed error inverts the very ranking moderation exists to produce.

2 Three Layers Collapsed Into One

The inference error becomes clearer once language is separated into three layers that a competent classifier would need to keep distinct.

1. Lexical content. The words and named referents present in the utterance: a drug name, an anatomical term, a slur, an exploit name, a weapon's specifications.

2. Speech act. What the utterance is doing with that content: warning, diagnosing, quoting, instructing, threatening, documenting, prescribing, mocking, disclosing.

3. Operational role. What the utterance does in the world once situated in its full context of speaker, audience, timing, and downstream action: does it give someone the means, opportunity, or motive to cause harm, or does it withhold, interrupt, or repair that possibility.

The charge against contemporary moderation is not that it literally searches for forbidden strings and stops there. Modern contextual classifiers can estimate something about speech act from surrounding text, tone, and account history. The more defensible claim is one of systematic overweighting: these systems overweight lexical and semantic proximity because that layer is directly computable at scale, while speech act is only probabilistically inferable and operational role is often distributed across histories, relationships, tools, and downstream consequences that sit entirely outside the item being classified. A system under review pressure and volume constraints

will lean on the layer it can actually price out, and the second layer, precisely where mention is separated from use, description from prescription, and warning from threat, gets only partial weight even when it is technically modeled. The honest version of the objection that classifiers are contextual, not merely lexical, is not that operational role is hard to estimate but that it is non-local by construction: it is a property of the utterance situated in a history of speaker, audience, and consequence, not a property of the utterance itself, so a classifier that scores one item at a time is answering a different, narrower question than the one moderation needs answered, regardless of how much surrounding text it is given. The populations for whom this overweighting is costliest are exactly the ones who most need protective speech to be legible: clinicians, survivors, investigators, educators, and the people living inside the harm being described.

3 The Pattern Across Domains

The following survey is not a list of independent anecdotes assembled for rhetorical weight. It is evidence for a structural claim, in the qualified sense already noted: if the same inversion recurs across subject matter as unrelated as drug policy, cryptography, grief counseling, and electoral politics, the fault cannot lie in any one domain's content. It has to lie in the classification logic itself, applied uniformly across domains that share nothing except a taboo topic. Nine of the eleven paragraphs below anchor the claim in a specific documented incident or study; two, weapons policy and medical diagnosis, do not, for reasons given in each case (classified or proprietary access, and the ordinary confidentiality of clinical exchange, respectively). Those two are included deliberately rather than omitted: the argument in Section 2 predicts the same inversion there independent of whether a public incident happens to exist, and the reader's ability to construct the case unaided is itself evidence that the pattern does not depend on the availability of documentation.

Suicide prevention. A crisis-line training document or a peer-support post describing suicidal ideation in specific, recognizable terms exists so that at-risk readers see themselves reflected and helpers know what to watch for. An advertisement or a vaguely inspirational post that never names ideation, method, or risk factors passes moderation untouched while offering a reader in crisis nothing to hold onto. Clinical honesty produces a higher lexical-danger score than commercial vagueness, even though the first utterance is protective and the second is inert at best. The same overweighting appears on the intervention side of these systems: researchers have noted that automated distress-detection tools, built to flag explicit language associated with suicidal ideation, disproportionately flag LGBTQ+ youth, whose ordinary, non-suicidal language about identity and struggle sits lexically closer to the flagged vocabulary than the guarded language of peers not navigating that disclosure.[1]

Sexual health education. Meta's own Oversight Board, reviewing two cases in which health and gender-identity education were removed under the Adult Nudity and Sexual Activity and Sexual Solicitation standards, found the underlying policy language so broad that it was, in the Board's words, bound to capture a large amount of content that had nothing to do with solicitation.[2] Sexual health educators with large, health-literate followings have separately reported having to delete

[1] "Inside AI's Efforts to Stop Suicide on Social Media," TIME, https://time.com/6696703/ai-suicide-prevention-social-media/.

[2] Oversight Board, "Gender Identity and Nudity" decision; discussed in Global Freedom of Expression, Columbia University, https://globalfreedomofexpression.columbia.edu/cases/oversight-board-case-of-gender-identity-and-nudity/.

hundreds of anatomically precise posts after platform warnings, even as coercive or exploitative messaging that avoids anatomical vocabulary routinely survives the same filters.[3] The pattern is consistent enough that a UNESCO-cited public comment record submitted to the Oversight Board specifically warned that strict nudity enforcement risks mistaking body and relationship education for prohibited content.[4] The educational sentence is penalized for the precision that makes it useful; the exploitative sentence is rewarded for the vagueness that makes it deniable.

Domestic abuse documentation. A survivor's account of strangulation, stalking, or financial coercion contains violent vocabulary because the violence was real and specific. An abuser's own messaging routinely avoids that vocabulary entirely, relying on context the classifier cannot see: a location shared without comment, a vague reference to consequences, a reminder of what the target's employer might learn. The evidentiary record is lexically alarming; the threat itself is lexically calm. This is a special case of a documented, larger effect: reviewing its own archive of citations, Human Rights Watch found that 619 of 5,396 pieces of social-media content it had relied on to support abuse allegations, roughly eleven percent, had since been removed by the platforms hosting them, evidence assembled precisely because it named violence explicitly.[5] A domestic-abuse record built from screenshots, voicemail transcripts, and injury photographs is asked to survive the same classifier, with the same bias against the person naming the harm and in favor of the person who never has to name it.

Extremism documentation and research. A historian, journalist, or deradicalization worker names organizations, quotes slogans, and explains coded symbols because explaining requires citing what is being explained. A propagandist relies on those coded symbols precisely because they no longer need explanation within an audience that already recognizes them; numerology, in-jokes, and nostalgic imagery carry the payload without triggering a keyword match. The researcher's vocabulary is dense with the very terms the classifier is built to catch. The propagandist's is sparse. The investigative journalist Alexa O'Brien learned this directly: after her channel hosted al-Qaeda-produced videos that had themselves become evidence in the Chelsea Manning court-martial, a Google reviewer concluded the channel was, in the company's own words, dedicated to terrorist propaganda, and O'Brien woke to find herself banned from YouTube with her Gmail and Google Drive suspended as well, even though the flagged videos contained no depicted violence at all; Google reversed the decision and apologized only after O'Brien appealed and the case drew press attention, which is the detail that makes the incident informative rather than merely alarming, since the correction depended on a public profile and a working appeals channel that most flagged users do not have.[6] The same failure recurs at the level of entire archives: the same Human Rights Watch review found that documentation of war and atrocity in Syria and elsewhere, work that necessarily quotes, films, or names extremist material in order to record it, was removed under the very extremist-content classifiers built to catch the propaganda it was documenting. The propagandist's own material, once it is quoted by the person documenting it, becomes the reason

[3] Reporting on sexual health creators including Leeza Mangaldas, The Fuller Project, https://www.fullerproject.org/sexual-health-content-creators-removed-social-media/.

[4] Cited in Oversight Board decision summary, ibid.

[5] Human Rights Watch, "'Video Unavailable': Social Media Platforms Remove Evidence of War Crimes" (2020), https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes. The same figure is cited again below in the extremism paragraph; it is one dataset bearing on two adjacent domains (violent-crime and atrocity documentation), not two independent data points.

[6] "Journalist Nearly Banned from YouTube and Gmail For Posting Al-Qaeda Videos From Chelsea Manning Trial," Gizmodo, https://gizmodo.com/journalist-nearly-banned-from-youtube-and-gmail-for-pos-1815314182.

the documentation disappears.

Harm reduction. A social worker posting overdose-recognition steps, dosage-uncertainty warnings, and naloxone administration instructions was, by his own account, repeatedly locked out of his account for using words like naloxone and fentanyl in explicitly protective posts, while the platform's automated systems could not distinguish his content from that of an actual seller.[7] A review of platform community guidelines documents reports of removal, restriction, and profile banning affecting harm-reduction and overdose-prevention accounts specifically because of the drug-related keywords their purpose requires them to use, though it does not offer comparative prevalence data against non-protective drug content.[8] A vague anti-drug slogan avoiding all of that vocabulary sails through moderation while doing nothing to keep anyone alive who is already using.

Cybersecurity. A defensive advisory reproduces exploit code, malware family names, and command sequences because a patch cannot be verified against a threat that has not been fully described. Two contrasting documented cases illustrate the asymmetry from opposite directions. State-linked threat actors have posed as security researchers themselves, building fabricated blogs and credible-sounding social accounts specifically to borrow the reputational cover of legitimate research before delivering an actual payload; those accounts were eventually suspended, but only after the impersonation had done its work.[9] Separately, a researcher who published genuine proof-of-concept exploit code following a public dispute over a vendor's disclosure process had accounts suspended in quick succession on both GitHub and GitLab in a dispute that ran from early April through late May 2026, with no public account of which specific artifact, as opposed to the underlying dispute, triggered the removals.[10] These are not one evidentiary sequence: the first shows harmful actors successfully wearing the posture of defense, the second shows a defensive posture itself becoming grounds for suspension. Together they show the same lexical layer, exploit code and researcher framing, sitting on both sides of the causal line at once. A properly defensive advisory looks, lexically, like the most dangerous object in the room; causally, it is often the only thing standing between a disclosed vulnerability and an unpatched population.

Weapons policy. An engineer explaining why a proposed device will fail, an inspector documenting a program's components for a treaty body, and an operator refining a working system can all produce sentences built from the same technical nouns and equations. What separates them is not locatable in any single sentence; it is the surrounding project, the access to material, and the downstream action the sentence is embedded in, none of which a lexical classifier can see.

Medical diagnosis. Differential diagnosis requires naming the frightening possibility in order to rule it out: a physician explaining how to tell a benign lump from a malignant one, or a benign headache from a hemorrhage, must say the dangerous word to do the reassuring or the life-saving work. A seller of ineffective remedies can encourage someone to abandon real treatment without

[7] Reporting on Louise Vincent and Ryan Sabora's harm-reduction moderation experiences, Yahoo News (syndicated), https://news.yahoo.com/free-speech-restrictions-social-media-230008564.html.

[8] "Problematizing content moderation by social media platforms and its impact on digital harm reduction," Harm Reduction Journal (2024), https://harmreductionjournal.biomedcentral.com/articles/10.1186/s12954-024-01104-9.

[9] Threatpost, "Twitter Suspends Accounts Used to Snare Security Researchers," https://threatpost.com/twitter-suspends-security-researchers/175524/.

[10] Coverage of the Nightmare-Eclipse GitHub and GitLab suspensions, The Hacker News, https://thehackernews.com/2026/05/microsoft-slams-public-zero-day.html.

using any alarming vocabulary at all. The responsible statement generates the denser cluster of frightening terms; the harmful one sounds soothing throughout.

Slurs and the use-mention distinction. The sentence identifying an expression as a slur and explaining why it should not be used necessarily contains that expression. The sentence performing exclusion, "those people don't belong here," can avoid every listed slur while doing the discriminatory work the slur list was meant to catch. A filter keyed to a list of forbidden strings inverts the moral ranking of the two sentences by construction.

Child safeguarding. Clinicians and investigators need explicit, specific language to document grooming and abuse, because vague language cannot support a finding or a prosecution. Grooming itself frequently proceeds through affection, games, secrecy, and gradual boundary erosion that contains no taboo vocabulary at all until very late, if ever, in the pattern. A system tuned to taboo vocabulary recognizes the protective record and misses the interaction it exists to protect against.

Election-integrity reporting. A report documenting a disinformation campaign must reproduce the false claims it is debunking in order to be legible as a debunking. The campaign itself can frame the same claims as questions, jokes, or "just asking," and can launder certainty into apparent neutrality with a phrase like "some people are saying," making an assertion look like reporting on other people's opinions. Surface grammar conceals causal function in both directions at once.

4 A Governing Principle

Across all eleven cases the same relation holds: semantic proximity to harm is not causal identity with harm, and treating the two as interchangeable is a specific instance of a more general failure to distinguish resemblance from causal role. Two utterances that share vocabulary, syntax, or topic do not thereby share a causal relationship to the outcome that vocabulary names. A classifier that scores resemblance and reports the score as risk has substituted a measure of similarity for a measure of consequence, and the substitution fails hardest exactly where the stakes are highest: in the utterances of people trying to prevent, document, or survive the harm the classifier was built to stop.

The layered account in Section 2 explains why the substitution is so persistent rather than incidental. Lexical content is cheap to measure and available at scale; speech act and operational role require context, history, and often information the platform does not and sometimes cannot have (who the speaker is, what access they have, what happens next). A system under pressure to moderate at volume will gravitate toward the layer it can actually compute, and will keep doing so even after its designers understand, in the abstract, that lexical content is the wrong layer to be measuring.

A fair objection deserves to be met directly rather than left implicit. Content moderation is not merely under-resourced; it is constitutionally constrained to produce rules that are legible, reviewable, and appealable, often under legal obligations to give reasons for a removal. Lexical proxies are attractive to designers not because anyone believes proximity is causally correct, but because a rule keyed to lexical content can be written down, applied consistently by thousands of reviewers or a shared model, and defended on appeal in a way that a rule keyed to operational role, which depends on facts external to the item, usually cannot be. This is a real constraint, and it means the systems under discussion are not naive about the gap between proximity and causation. But conceding the point sharpens rather than weakens the argument: if lexical proxies are chosen

for their legibility rather than their accuracy, then the selection pressure described here is not an oversight to be corrected by better classifiers. It is a structural output of optimizing for defensibility under review, and it will recur under any moderation regime that keeps that optimization target, however sophisticated its models become.

5 False Positives Versus Adversarial Asymmetry

It is tempting to file all of the preceding cases under a familiar heading: classifiers make mistakes, and some mistakes fall on sympathetic targets. That heading understates what is actually happening. A false positive is a symmetric event: an innocuous item is wrongly flagged, at some fixed rate, without regard to who benefits from the error. What the domain survey shows is not symmetric. The two parties in each pair face unequal costs from the same corrective advice.

Consider the standard remedy a platform offers a flagged account: rephrase, soften, remove the specific terms, speak less explicitly. For the harmful actor, that instruction is free and often already the plan. The propagandist, the abuser, the coordinated influence operation all benefit from concealment independently of any moderation pressure; being told to communicate more indirectly simply formalizes a strategy already in use, and moving further into euphemism carries no cost to the harmful project because that project never depended on explicitness. For the clinician, the survivor, the harm-reduction worker, or the investigator, the identical instruction is corrosive. Their protective value depends specifically on precision: an overdose-recognition guide that cannot name the drug, a consent education post that cannot name the anatomy, an abuse record that cannot name the act, or a defensive advisory that cannot reproduce the exploit is a weaker version of the thing it was for. Asking both parties to phrase it more safely is not a neutral correction applied twice. It is an intervention that advances one party's goal and disables the other's, using the same words.

This is the sense in which the inversion is directionally biased rather than merely noisy, restated at the level of incentive rather than outcome. A noisy classifier would impose a random tax on speech near a taboo boundary. What is documented here is closer to a selection pressure: over repeated cycles of moderation and adaptation, the actors best equipped to encode intent indirectly, because indirection was already their method, pay the smallest price for adapting further, while actors whose value lies in saying the explicit thing pay the largest price for the same adaptation. The pressure does not fall evenly on the population near the boundary; it falls hardest on the population for whom retreating from explicitness is itself the harm.

The cycle, not just its endpoints, is what the static inversion in Section 1 leaves out. On one side, extremist recruitment has been widely reported to have migrated away from explicit forums and hashtags, precisely the surfaces enforcement targeted first, toward gaming voice chat, fitness and wellness communities, and an ironic, "just asking questions" register that borrows the phrasing of sincere inquiry; each round of keyword and account enforcement has been followed by a further step away from the vocabulary that enforcement was keyed to, without any corresponding loss of recruiting function. On the other side, harm-reduction and sexual-health educators who have had posts and accounts flagged for specific clinical terms report responding by adopting euphemism themselves, coded hashtags, deliberately vague phrasing, alternate spellings, precisely the strategies documented in Section 3, in order to keep posting at all. The two adaptations are not symmetric in what they cost. The propagandist's shift into coded registers preserves the propaganda; the educator's shift into euphemism degrades the education, because the specific term was the useful part. Enforcement pressure, applied identically to both, drives the harmful register toward greater concealment at no functional cost and drives the protective register toward greater vagueness at

direct functional cost.

6 The Power Effect

The inversion has a distributional consequence that is easy to miss if each domain is examined in isolation. Euphemism, indirection, and institutional phrasing are not equally available to everyone. They are skills, often professionally acquired ones: the propagandist who has studied plausible deniability, the abuser who has learned to threaten through implication, the corporation whose lawyers have refined a press statement into total vagueness. Blunt, explicit speech, by contrast, is disproportionately the register of people without that training or without the composure crisis affords: survivors describing what was done to them, patients repeating what a doctor told them, frightened teenagers, non-native speakers translating an experience they do not yet have the polished vocabulary for, clinicians and investigators whose professional obligation is precision rather than tact.

A moderation regime built on semantic proximity therefore does not simply fail at random. It selects, systematically, for actors who already know how to make dangerous things sound ordinary, and against actors who have never needed or been able to learn that skill because they are the ones the danger is actually happening to. What presents itself as a neutral, content-blind safety boundary is, in its actual operation, a filter correlated with rhetorical class position: fluency in euphemism functions as a kind of passport past the very systems built to catch the people who lack it. Call a policy regressive, in the sense meant here, when its burden falls disproportionately on those with the least prior access to the resource the policy implicitly requires, in this case, fluency in institutional indirection. By that definition the inversion described in this paper is not merely biased in outcome; it is regressive in structure, because the resource it rewards is unevenly distributed for reasons that have nothing to do with the policy's stated aim.

7 Conclusion

The recurring failure documented here is not a list of edge cases to be patched one moderation policy at a time. It is a single structural error, restated across eleven domains with enough consistency, and anchored in enough independently documented incidents, to indicate that the error lives in the inference itself: overweighting lexical proximity relative to operational role because proximity is the layer available at scale, and available cheaply enough to survive legal and operational demands for legibility. The adaptation cycle described in Section 5, coded harm sharpening while legible protection dulls, is the mechanism that turns a one-time misranking into a compounding one: a system that cannot tell resemblance from causal role does not just misrank a fixed set of utterances once. It trains the population speaking near the boundary to become exactly the population it is worst at reading correctly, round after round, for as long as the underlying optimization target, defensible rules over accurate ones, remains the same.