

Addressing “Lawful but Awful” Content

What are the best ways to manage online content that is “lawful but awful”?

CNTI’s Assessment

The “lawful but awful” dilemma describes a tension between legality and morality: systems, practices and speech may comply with legal standards but still cause profound social harm. Platforms struggle to apply fragmented, often U.S.-centric “marketplace of ideas” frameworks across diverse contexts. Is the goal to reduce harm, preserve speech, redistribute visibility, or make platform power transparent and contestable — goals that may conflict, with governance approaches implicitly prioritizing one over others?

Policy Deliberation

Policymakers face a tradeoff between free expression and limiting systemic risk, with no universal definition of “harmful” content and little agreement on when harm justifies intervention. This plays out most visibly in the divide between the U.S.’s Section 230, which protects platforms’ editorial discretion, and the EU’s Digital Services Act, which requires transparency and risk assessment — and whichever direction policy moves, it concentrates authority somewhere, either with private platforms or with governments, each carrying distinct risks. Underlying all of this is a hard practical truth: consistent, context-sensitive moderation at scale may not be achievable.

Journalistic Concerns

Journalists face uneven legal protections and often must share graphic material to document abuses — essential archives that also risk audience harm. As journalists and content creators blur together, who qualifies for protection grows murkier, while algorithms frequently misread political speech or protest language as violations, revealing the need for human, context-aware review.

Technology Governance

Tech firms must build transparent, democratic systems that resist abuse, despite algorithms’ poor grasp of nuance like satire. Multi-stakeholder groups face generative AI’s dual role as detector and source of harm, compounded by cultural bias and platforms’ incentives to amplify controversy — pointing toward one emerging direction: shifting from content removal to user-controlled filtering.

The Issue

“Lawful but awful” describes a tension between legality and morality in which systems, practices and speech comply with legal standards but cause profound social harm, especially content available to the public on digital platforms.

Debates around “lawful but awful” speech predate the digital era. Years before the 1994 genocide, licensed Rwandan radio stations broadcast [dehumanizing propaganda](#), such as calling Tutsis “cockroaches.” This propaganda, while legal, helped make the genocide “socially acceptable.”

Platformization increases the reach of harmful content and obscures who is legally responsible, since distribution is shared among original creators, hosting platforms and account holders who share content for [various reasons](#).

The “lawful but awful” dilemma is often framed as how to moderate content, obscuring a more fundamental question: What should the goal of intervention be? The challenge is not improving moderation tools but defining the desired end state of the information environment — reducing harm, maximizing free speech, redistributing control over visibility, or making platform power transparent and contestable. This primer’s core argument is that harm prevention and free speech cannot be definitively reconciled, meaning solutions must focus on procedures over specific outcomes.

Much of the current debate takes a platform-centric view, focusing on tools like removal, downranking and account sanctions. But these questions extend to spaces where such tools don’t easily translate, such as private messaging, email and live public events. Forcing public-square policies onto private correspondence risks significant privacy violations. Digital rights groups like the Electronic Frontier Foundation [note that](#) policy mandates designed for public feeds often fail to account for the unique protections of private communications.

FULL ISSUE PRIMER →



Addressing “Lawful but Awful” Content

What are the best ways to manage online content that is “lawful but awful”?

What Makes This Complex

I. There is no universal standard for defining or governing

“awful,” harmful speech. This tension between legal protection and social harm is compounded by cultural and legal divergence worldwide, though major platforms have generally applied a U.S.-centric [“marketplace of ideas”](#) framework globally.

II. Meaning is context-dependent. Automated systems often can’t interpret meaning. For example, [satire and political speech](#) require contextual analysis to avoid wrongful takedowns, yet fully manual review is impossible at scale.

III. The same content can have simultaneous social harms and social benefits. Platforms serve as archives documenting abuse and injustice, while also exposing audiences to [psychological harm](#) through exposure to violent material.

IV. The platform economy creates financial incentives to amplify controversial content. The [financial incentives](#) of the platform economy encourage platforms, advertisers and users to share emotionally evocative content, fueling the spread of propaganda and extremism.

V. Current approaches to “lawful but awful” content disproportionately affect marginalized communities. For example, Meta’s lower AI confidence thresholds for Arabic versus Hebrew content led to the [removal of over 1,000 documented human rights reports](#) in Palestine.

VI. There are many methods of content moderation, which makes it hard to measure their consistency, impact and effectiveness. [Shadowbanning](#), for instance, quietly reduces visibility without notice, leaving creators uncertain why.

State of Research

Current research overwhelmingly frames ‘lawful but awful’ as a content moderation and platform governance problem, overlooking the infrastructural and societal factors at play. Among this research, a strong [consensus](#) has emerged around the need for a balanced approach to safety and speech. [Researchers and journalists conclude](#) that global moderation policies are Euro- and U.S.-centric, leaving the rest of the world without resources for culturally relevant oversight. There is also broad consensus that platforms are “black boxes,” and that current practices often moderate marginalized voices while amplifying [those](#) promoting controversial content for monetary gain.

State of Research (cont).

While there is consensus on exclusionary, opaque policies, deep disagreements persist over platforms’ fundamental legal responsibility for filtering content. Scholars remain divided over whether platforms should act as neutral common carriers or active stakeholders curating the online environment. This highlights a clear gap: there is a distinct lack of technical frameworks ensuring these systems are democratic, transparent and genuinely [sensitive to context](#).

State of Legislation

A transatlantic dispute over accountability for “lawful but awful” content defines today’s regulatory landscape. Under this [“duty of care”](#) shift, policies increasingly require platforms to proactively address risks rather than simply prohibit content. The U.S. relies mainly on litigation, deferring to First Amendment protections and [Section 230](#), which shields platforms from liability for third-party content while letting them moderate freely. This has let platforms disclaim editorial responsibility while wielding editorial power — though plaintiffs and state attorneys general increasingly target platforms’ own design choices and algorithms, testing Section 230’s limits.

The EU takes a different approach: the [Digital Services Act](#) imposes [due diligence obligations](#) — risk assessments, audits, algorithmic transparency — rather than mandating removal of lawful content. The European Media Freedom Act adds procedural protections for recognized media, raising concerns about unequal treatment among journalists, creators and users. Section 230 faces bipartisan reform pressure, while the DSA has sparked geopolitical tension, with the U.S. calling it a [“foreign censorship threat.”](#) Elsewhere, regulation remains fragmented and evolving.

Notable Legislation

Germany: [Germany’s Network Enforcement Act \(NetzDG\)](#) requires large platform companies to remove content tied to 22 criminal categories — such as hate speech, defamation and incitement — within 24 hours if “clearly illegal” and within seven days otherwise, while maintaining transparency through reporting obligations. However, steep fines create strong incentives for platform companies to over-enforce, often removing content that is technically lawful, thereby suppressing legitimate discourse.

