MARKET RESOURCE

Prohibiting Harmful Conduct Without Limiting Speech

How companies can address threats, harassment, and other genuine harms through clear rules that protect every user equally.

7-minute read • Market • Updated July 2026

WHY IT MATTERS

Policies against threats, harassment, and other harmful conduct serve an important purpose. But they can become unequal or suppress lawful expression when they protect only selected groups or rely on subjective labels such as “hateful,” “intolerant,” or “offensive.”

The Core Principle

Prohibit the harmful behavior itself, define it clearly, and protect every customer, user, seller, creator, or developer from it. A rule against credible threats should protect everyone. A rule against targeted harassment should protect everyone. The company should not need to classify a person by race, sex, religion, sexual orientation, gender identity, political belief, or another trait before deciding whether the conduct is prohibited.

This approach allows companies to address genuine harm while reducing the risk that ordinary religious, political, or social disagreement will be treated as misconduct.

What This Resource Covers

This resource applies to operative policies governing a company’s core products or services, including:

- Terms of service and acceptable-use policies
- Community standards and content-moderation rules
- Seller, creator, advertising, and developer policies
- Payment, banking, hosting, cloud, and platform-use agreements

SCOPE NOTE

Standard workplace nondiscrimination policies and incidental comment sections are generally outside this resource unless the rule also governs customers or users of the company’s core service.

Two Risks to Avoid

1. Protection limited to selected groups

A policy may prohibit threats, harassment, or abuse only when the target belongs to a listed group. That leaves other users without the same protection and makes identity part of the enforcement decision.

2. Rules based on subjective offense

Terms such as “hateful,” “demeaning,” “intolerant,” or “harmful” may be used to restrict lawful expression when they are undefined or depend primarily on how a listener reacts. Religious teachings, political arguments, satire, criticism, and discussion of contested issues can all be described with those labels.

What Strong Policy Language Looks Like

Stronger approach	Problematic approach
Defines threats, harassment, stalking, fraud, or other conduct through objective elements.	Uses “hate,” “bigotry,” “intolerance,” or “offense” without clear limits.
Protects every user from the prohibited behavior.	Protects only users with listed characteristics.
Distinguishes criticism of ideas from conduct directed at a person.	Treats disagreement with an idea as an attack on a person.
Accounts for context such as religious teaching, political debate, news reporting, and satire.	Requires users to affirm contested terminology or beliefs.
Explains what evidence, duration, repetition, or intent is required.	Reserves broad discretion without meaningful standards.

Context Matters

Words such as “threaten,” “violence,” “harassment,” and “discrimination” are not automatically improper. They become higher risk when the policy does not define them, limits protection to selected groups, or allows ordinary disagreement to qualify.

- A threat rule should require a serious expression of intent to cause unlawful harm, evaluated in context.
- A harassment rule should identify targeted, repeated, or severe conduct rather than isolated disagreement.
- A discrimination rule should address unequal access or treatment in a transaction, not merely an unpopular opinion.
- A violence rule should distinguish advocacy, reporting, historical discussion, and figurative language from incitement or credible threats.

Before-and-After Examples

Harassment

PROBLEMATIC	Users may not harass or demean members of protected groups.
MORE PRECISE	Users may not direct repeated, unwanted communications at another person when the conduct is intended to intimidate or substantially disrupt that person’s use of the service.

Threats

PROBLEMATIC	Users may not post hateful or threatening content.
MORE PRECISE	Users may not communicate a serious expression of intent to commit unlawful violence against an identifiable person or group.

Discrimination

PROBLEMATIC	Content that discriminates against protected groups is prohibited.
MORE PRECISE	Sellers may not refuse a transaction or provide materially different terms because of a customer's protected status, except where required by law.

What Unequal or Overbroad Policies Look Like in Practice

The examples below illustrate different drafting risks. They do not establish that a company enforced the provision against a particular person or viewpoint.

Airbnb

POLICY LANGUAGE	Airbnb prohibits language that "demeans, insults, stereotypes, or seeks to convey a person's inferiority" based on protected characteristics and expressly includes deadnaming, misgendering, microaggressions, and "all other forms of hateful speech."
CORE CONCERN	Identity-based speech rules
WHY IT MATTERS	Subjective terms may reach ordinary disagreement, and specified speech rules may require users to adopt the company's position on contested questions of sex and gender.
CLEARER APPROACH	Define targeted harassment through repetition, severity, intent, and concrete interference without requiring users to affirm a contested belief.

View evidence

Akamai Technologies

POLICY LANGUAGE	The Linode community policy states that it prioritizes "marginalized people's safety over privileged people's comfort" and will not act on complaints involving "reverse racism," "reverse sexism," or "cisphobia."
CORE CONCERN	Unequal complaint handling
WHY IT MATTERS	The policy announces in advance that similar complaints will receive different treatment depending on how the parties are classified.
CLEARER APPROACH	Apply one conduct standard to every complainant and target, regardless of group identity.

View evidence

Block, Inc.

POLICY LANGUAGE	Square states that promotion of historically excluded or disadvantaged protected classes "may not constitute discrimination."
CORE CONCERN	Asymmetric discrimination standard
WHY IT MATTERS	The same treatment may be classified differently depending on which group benefits, rather than whether the conduct itself is unfair or exclusionary.
CLEARER APPROACH	Define discrimination as materially unequal access or treatment and apply that definition consistently.

View evidence

Morgan Stanley

POLICY LANGUAGE	Morgan Stanley's bill-pay terms restrict communications and payments involving "hate," "racial intolerance," "bigoted," "hateful," "racially offensive," "indecent," or "discourteous" content.
CORE CONCERN	Subjective restrictions on financial activity
WHY IT MATTERS	These terms do not identify a concrete financial, legal, security, or operational harm and could reach lawful books, events, advocacy, or religious materials.
CLEARER APPROACH	Limit payment restrictions to defined illegality, fraud, sanctions, threats, or other identifiable financial and compliance risks.

View evidence

A Six-Step Policy Review

1. Inventory operative policies. Review every agreement and incorporated rule governing the company's core products or services.

2. Flag unequal protections. Identify rules that protect only listed groups or expressly disregard complaints from other users.

3. Flag subjective terms. Search for "hateful," "bigoted," "demeaning," "intolerant," "harmful," "deadnaming," "misgendering," and equivalent language.

4. Define the actual harm. Specify the conduct, target, repetition, intent, context, and measurable harm required for enforcement.

5. Test competing viewpoints. Apply the rule to comparable religious, political, and ideological scenarios to check for even treatment.

6. Publish and train. Make the controlling policy accessible and train reviewers to enforce the defined standard consistently.

Common High-Risk Terms

These terms are not automatically improper, but they require careful definition and context when used to regulate access or expression:

Subjective judgments	Identity and status	Harm labels	Enforcement terms
Bigotry	Deadnaming	Hate speech	Threaten
Demeaning	Misgendering	Hateful conduct	Violence
Degrading	Discriminatory	Harm	Abuse
Derogatory	Direct attack	Bullying	Offensive
Intolerant		Harassment	Unsafe

What the Viewpoint Diversity Score Evaluates

The Viewpoint Diversity Score examines harmful-conduct policies governing core products and services. It asks whether the company avoids restricting speech merely because someone may find it offensive and whether genuinely harmful conduct is prohibited equally for every user.

A USEFUL DRAFTING RULE

If conduct is harmful enough to prohibit, define the conduct and protect everyone from it. Do not make a user's identity or viewpoint the condition for protection.

Source note: Company examples are drawn from current 2026 Viewpoint Diversity Score research records; links lead to the underlying public evidence.

Back to Resources