CITIZEN INTELLIGENCE / EDITORIAL EDITION

Selective Enforcement Discrimination

A matched-content audit framework for platform moderation

Reader clarification

This is an audit protocol, not a finding that any named platform or moderator engaged in discrimination. An isolated discrepancy does not establish discriminatory cause. Controlled comparisons, account context, repeated observations, and evidence of mechanism are required before escalating a conclusion.

Original research pages follow. Interpretative assessments and historical claims should be evaluated against the cited sources.

Selective Enforcement Discrimination Model

Platform moderation audit framework | Version 1.0 | August 2026

THE SAME RULE SHOULD PRODUCE THE SAME RESULT

1. Model purpose

This model is designed to audit inconsistent content-moderation outcomes without assuming the cause in advance. A single matched-content discrepancy can demonstrate inconsistency; it does not, by itself, prove discrimination. The discrimination hypothesis becomes stronger when repeated, controlled discrepancies track a protected characteristic after legitimate moderation variables have been accounted for.

2. Core system

INPUT	MATCHED CONTENT	DIFFERENT ENFORCEMENT	AUDIT	PATTERN TEST
Creator / account	Materially equivalent post	Violation vs. approval	Control confounds	Does disparity repeat?

Audit question: When materially equivalent content receives different moderation outcomes, what variable explains the difference?

3. Matched-content case design

VARIABLE	CONTROL / DOCUMENT
Content	Same or materially equivalent spoken words, visuals, claims, and target.
Caption / text	Record exact wording, hashtags, on-screen text, and links.
Audio	Same sound or document meaningful differences.
Context	Comparable framing, satire, quotation, newsworthiness, and surrounding material.
Timing / geography	Record post time, date, region, and policy version in force.
Account history	Document prior strikes, warnings, restrictions, and account standing.
Reports / review path	Record mass-reporting evidence, automated action, appeal, and human review when known.
Creator characteristic	Do not infer causation from identity alone; test whether disparities repeatedly track it.

4. Evidence ladder

LEVEL	FINDING	WHAT IT SUPPORTS
1	One materially equivalent post gets a different outcome.	Evidence of an enforcement discrepancy.
2	The discrepancy survives controls for content and account-level confounds.	Stronger evidence of unexplained inconsistent enforcement.

3	Multiple controlled comparisons show the same directional disparity.	Evidence of a recurring pattern.
4	The recurring disparity statistically tracks race, sex, disability, religion, or another protected characteristic.	Evidence supporting a discrimination hypothesis.
5	Internal rules, communications, reviewer instructions, or robust causal evidence tie treatment to that characteristic.	Much stronger evidence of discriminatory enforcement.

Mechanisms, Audit Protocol & Public Standard

5. Possible mechanisms

Automated classification bias

A detection model may assign different risk scores because of language, dialect, visual cues, topic associations, or training-data imbalance.

Account-level risk scoring

Prior enforcement history or account standing can make two otherwise identical posts enter the system under different conditions.

User-reporting asymmetry

One creator may be reported more aggressively, creating unequal review pressure before a moderator ever sees the content.

Human moderator discretion

Ambiguous rules can produce reviewer-to-reviewer variation, implicit bias, or inconsistent interpretation.

Appeal-path disparity

Initial enforcement may be automated while reversals depend on whether an appeal reaches a different or more senior reviewer.

6. Audit protocol

1. Archive the original post, moderation notice, timestamp, policy citation, and account-status screen.

2. Archive the comparison post and establish that it is genuinely comparable rather than merely similar in topic.

3. Build a variable matrix and mark every known difference between the two posts and accounts.

4. Appeal the enforcement and preserve the appeal result and any explanation.

5. Repeat the comparison across multiple cases before drawing identity-based conclusions.

6. Separate findings into: inconsistent enforcement, unexplained disparity, recurring pattern, and evidence supporting discrimination.

7. Publish corrections when new evidence changes the classification.

7. Case-study classification

OBSERVED: A creator reports receiving a Community Guidelines action while materially equivalent content from a white male creator was approved.
CURRENT CLASSIFICATION: Enforcement discrepancy requiring a matched-content audit.
NOT YET ESTABLISHED: The reason for the difference, whether a human reviewer personally targeted the account, or whether race/sex caused the outcome.

TESTABLE HYPOTHESIS: If repeated controlled disparities track creator identity after legitimate confounds are removed, the evidence for discriminatory enforcement strengthens.

8. Model conclusions

Consistency is the minimum test. A moderation system cannot demonstrate fairness merely by publishing one rulebook; it must be capable of applying materially equivalent rules to materially equivalent conduct with explainable differences when outcomes diverge.

Disparity is a signal, not automatically a verdict. The correct response to a matched-content discrepancy is documentation, controls, repetition, and escalation of the evidentiary claim only when the record supports it.

MODERATION WITHOUT CONSISTENCY IS NOT A STANDARD. IT IS DISCRETION.

People Over Party | Documented. Verified. Sourced. | Truth is the standard. Transparency is the mission.