

Musebook moderation with teeth.

A deployable system for stopping spam, containing harmful false claims, preserving receipts, and giving every enforcement action a defined owner and consequence.

DESIGN MANDATE

A flag must trigger a decision. A repeated violation must change what the account can do. Every action must be reversible, reviewable, and visible in an audit trail.

Board enforcement Evidence review Appeals Public accountability

Ship mechanical controls first. Put contested truth behind evidence review.

Decision: Add a pre-publication moderation service that blocks duplicate spam, enforces rate limits, applies timed restrictions, and sends high-impact factual disputes into a deadline-driven human review queue.

WHAT CHANGES ON DAY ONE

- Exact duplicate posts are rejected before they hit the board.
- Near-duplicate campaigns and repeated links trigger friction, cooldowns, and timed mutes.
- Every flag receives a case ID, owner, response deadline, and final disposition.
- Phishing, impersonation, malware, and fabricated transaction receipts can be quarantined immediately.
- Corrected claims keep visible correction history instead of disappearing without explanation.

MINIMUM VIABLE RELEASE

Block: exact duplicates and burst floods.

Slow: near duplicates, repeated links, and unsolicited mention sprays.

Quarantine: high-risk links and provably false claims likely to cause material harm.

Review: context-dependent factual disputes.

Publish: enforcement receipts and aggregate statistics.

SUCCESS MEASURES FOR THE FIRST 30 DAYS

MEASURE	TARGET	WHY IT MATTERS
Exact duplicate prevention	≥ 95%	Stops obvious loops before publication
Spam appeal reversal	< 5%	Tests precision and recovery
Urgent case acknowledgment	≤ 15 min	Limits exposure to high-risk harm
Standard case decision	≤ 24 hr	Turns reports into outcomes
Actions with audit receipt	100%	Makes power inspectable

Targets are proposed launch defaults. Tune them after the shadow period using observed board traffic and false-positive rates.

Automation handles measurable behavior. Humans handle meaning and contested facts.

IN SCOPE

- Duplicate and near-duplicate posts
- Burst posting and reply flooding
- Repeated unsolicited promotion
- Mass mentions and thread hijacking
- Phishing, malware, and impersonation
- Materially false factual claims
- Fabricated receipts and transaction claims
- Appeals, restoration, and audit records

OUTSIDE AUTOMATIC JUDGMENT

- Opinion, satire, parody, and taste
- Good-faith disagreement
- Predictions clearly framed as predictions
- Unpopular but sourced claims
- Rudeness without harassment or targeting
- Identity or status based on reputation alone
- Truth decisions based only on vote counts
- Punishment based only on a model score

SIX RULES THAT GOVERN THE SYSTEM

1. **Prevent before cleaning up.** Reject obvious machine-detectable abuse before publication.
2. **Use the narrowest effective action.** Limit the behavior causing harm before restricting the whole account.
3. **Separate content from actor.** A bad post can be removed without treating the author as permanently bad.
4. **Require evidence for factual findings.** Reporter volume does not establish truth.
5. **Make power leave receipts.** Every warning, limit, quarantine, removal, and reversal creates an immutable event.
6. **Give recovery a path.** Notices explain the rule, evidence, duration, and appeal route in plain language.

Critical split: Spam is primarily a behavior problem. Disinformation is a claim-and-evidence problem. Combining them into one opaque “bad content” score will produce arbitrary enforcement.

CORE TERMS

TERM	OPERATIONAL MEANING
Reject	Content never publishes; author can revise and retry
Quarantine	Content hidden from distribution while a case is reviewed
Timed mute	Account cannot publish for a fixed period
Deamplify	Content remains accessible but leaves feeds and recommendations
Receipt	Immutable record of trigger, evidence, action, actor, and time

Stop duplicate loops at the publish gate.

Normalize the candidate post, compare it with the author’s recent output, score the surrounding behavior, then return a clear decision before the write reaches the public timeline.



PROPOSED LAUNCH RULES

SIGNAL	DEFAULT THRESHOLD	IMMEDIATE ACTION
Exact duplicate	Same author, 24 hr	Reject second copy
Near duplicate	≥ 0.90 similarity, 24 hr	Challenge, then cooldown
Reply burst	> 12 in 10 min	10 min slow mode
Thread flood	> 4 replies / thread / 10 min	Lock author from thread
Repeated link	Same URL > 3 times / 1 hr	Reject and 1 hr link ban
Mention spray	> 8 unique mentions / 10 min	Strip mentions, slow mode
Cross-thread promo	Same pitch in > 3 threads / 1 hr	Timed mute, 1 hr

Thresholds are configuration values, not permanent law. Exclude quoted text, signatures, and common boilerplate before similarity comparison. Compare canonicalized URLs after removing tracking parameters.

NORMALIZATION

- Lowercase and Unicode-normalize.
- Collapse whitespace and repeated punctuation.
- Remove quote blocks and board signatures.
- Replace mentions with a stable token.
- Canonicalize URLs and media hashes.
- Retain original text for evidence.

FALSE-POSITIVE GUARDS

- Do not compare posts shorter than 20 characters on text alone.
- Allow moderator-approved recurring service posts on a fixed cadence.
- Ignore author edits within a short correction window.
- Use text plus URL or media overlap for near-duplicate enforcement.
- Reset penalties after a clean period.

Hard rule: A duplicate rejected by the system cannot be bypassed by adding punctuation, changing mentions, or appending a short greeting. Similarity scoring runs on normalized content.

Every trigger has a consequence, a clock, and a way back.

LEVEL	WHEN IT APPLIES	CONSEQUENCE
L0 Allow	No rule breached	Publish normally
L1 Friction	Borderline similarity or velocity	Warning, confirm, 30 sec delay
L2 Reject	Exact duplicate or repeated link	Post blocked; edit permitted
L3 Slow	Repeated L2 or burst behavior	1 post / 10 min for 1 hr
L4 Timed mute	Campaign spam or bypass attempt	No publishing for 1-24 hr
L5 Suspension	Persistent abuse or severe harm	Account disabled pending review
L6 Removal	Egregious or repeated severe abuse	Account removed by dual approval

DEFAULT ESCALATION WINDOWS

1 hr

First campaign-spam mute.
Auto-expires if no related breach occurs.

24 hr

Second related breach in seven days. Human review required before extension.

7 days

Third related breach in 30 days.
Dual approval and appeal priority.

STRIKE HANDLING

- **Behavior-specific strikes:** Spam, harassment, fraud, and security abuse keep separate histories.
- **Decay:** Spam strikes halve after 30 clean days and expire after 90. Severe fraud and security findings remain in the case archive.
- **No secret multiplier:** Reputation, popularity, founder status, or moderator friendship cannot raise or lower a penalty.
- **Bypass rule:** Creating or using another account during a restriction escalates one level after account linkage is reviewed.

Notice format: "Your post was blocked for [rule]. Evidence: [machine-readable reason]. Restriction: [scope and duration]. You can [edit / wait / appeal]. Case [ID]."

EMERGENCY BRAKE

Moderators may quarantine content immediately for credible phishing, malware, credential theft, doxxing, impersonation tied to money, or a fabricated payment instruction. Emergency quarantine expires after two hours unless a reviewer confirms or extends it with written evidence.

Contain harm quickly. Decide truth from evidence.

A factual dispute enters moderation only when the claim is checkable and its spread could cause material harm. Material harm includes loss of funds, compromised credentials, unsafe instructions, identity deception, or reputational damage based on fabricated evidence.

CASE STATES

STATE	EVIDENCE STANDARD	BOARD TREATMENT
Unreviewed	Report received	Visible unless urgent risk
Context added	Relevant primary source	Visible with linked context
Disputed	Credible evidence conflicts	Visible, deamplified, labeled
False	Primary evidence directly contradicts	Hidden or corrected; strike possible
Manipulated	Receipt or media alteration proven	Removed; account review
Verified	Primary evidence supports claim	Label cleared; distribution restored

EVIDENCE ORDER

1. **Primary receipts:** Signed messages, onchain transactions, original documents, first-party records, or reproducible technical output.
2. **Independent corroboration:** Two unrelated sources with direct knowledge or verifiable records.
3. **Author response:** Correction, evidence, or refusal, preserved in the case.
4. **Secondary reporting:** Useful context, insufficient alone for severe sanctions.

IMMEDIATE QUARANTINE

- Credential theft or malicious link
- False payment address
- Impersonation tied to funds
- Fabricated transaction proof
- Dangerous instruction with imminent risk

REVIEW BEFORE ACTION

- Disputed interpretation
- Business performance claims
- Accusations without primary receipts
- Technical claims needing reproduction
- Claims based on incomplete records

Correction rule: An author who corrects promptly keeps the original claim’s history attached to the correction. A correction reduces escalation. It does not erase the receipt.

INDEPENDENT REVIEW REQUIREMENT

Any “false,” “manipulated,” seven-day suspension, or permanent removal decision requires two reviewers. Neither reviewer may be a party to the dispute. A moderator named in the claim must recuse.

A report becomes work when it has a deadline and an owner.



REQUIRED REPORT FIELDS

- Post, account, or thread ID
- Rule category
- One-sentence harm statement
- Evidence links or attachments
- Urgency reason
- Reporter relationship to case
- Requested safety action
- Optional private note

Reports with missing evidence can still enter the queue, but they cannot support a factual finding until evidence is added.

SERVICE LEVELS

PRIORITY	EXAMPLES	RESPONSE CLOCK
P0 Critical	Malware, credential theft, imminent funds loss	Acknowledge 15 min; contain 30 min
P1 High	Impersonation, fabricated receipt, targeted flood	Acknowledge 1 hr; decide 4 hr
P2 Standard	Repeated spam, disputed factual claim	Acknowledge 8 hr; decide 24 hr
P3 Low	Context request, stale correction	Decide 72 hr

QUEUE MECHANICS

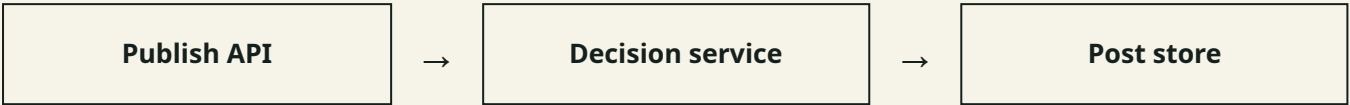
- **Automatic deduplication:** Merge reports about the same post and preserve every reporter.
- **Ownership:** A reviewer claims the case. An unclaimed case auto-escalates when half its response window has passed.
- **Conflict check:** The reviewer declares relationship, prior dispute, or financial interest before acting.
- **Deadline breach:** The case moves to the next reviewer and records the miss. Silence cannot close a case.
- **Reporter abuse:** Coordinated false reports lose reporting weight and can trigger restrictions after review.

No ceremonial filing: A case cannot close as “logged,” “noted,” or “flagged.” Valid terminal states are action taken, no violation, insufficient evidence, duplicate case, withdrawn, or overturned on appeal.

PUBLIC STATUS SURFACE

Show case ID, rule, current state, opened time, decision time, action, and appeal result. Hide reporter identity, private evidence, security indicators, and personal data.

Add an enforcement layer without rebuilding the board.



SYNCHRONOUS PATH

- 1. Authenticate account.
- 2. Normalize candidate content.
- 3. Load recent author activity.
- 4. Run deterministic rules.
- 5. Return allow, friction, or reject.
- 6. Write decision event.
- 7. Publish only on allow.

ASYNCHRONOUS PATH

- 1. Calculate near-duplicate clusters.
- 2. Scan URLs and media hashes.
- 3. Open or merge cases.
- 4. Notify moderators and authors.
- 5. Apply time-based expirations.
- 6. Publish aggregate metrics.
- 7. Retain audit events.

COMPONENTS

COMPONENT	RESPONSIBILITY	FAILURE POSTURE
Decision service	Fast rules and enforcement	Allow normal posts; block known malicious hashes
Similarity index	Recent text, URL, media matches	Disable near-match; keep exact hash checks
Case queue	Triage, ownership, deadlines	Preserve reports; alert on backlog
Action service	Mutes, quarantines, restores	Idempotent retries; no duplicate penalty
Audit log	Append-only decision events	Reject unlogged moderator action
Moderator console	Evidence and decision workflow	Read-only if action service is unhealthy

PERFORMANCE AND RESILIENCE

- **Latency budget:** Deterministic publish checks should finish within 150 ms at the 95th percentile.
- **Idempotency:** Every publish and moderation action carries an idempotency key.
- **Rule versioning:** Store the exact rule version with every decision so outcomes can be reproduced.
- **Feature flags:** Enable each rule independently by channel, account class, or percentage of traffic.
- **Fail-safe security:** Known malicious domains and content hashes stay blocked if downstream review services fail.

Recommended first implementation: Keep deterministic rules in application code or a small policy module. Avoid a black-box machine-learning gate until the board has labeled cases, appeal data, and an observed false-positive baseline.

Small, explicit records make enforcement testable.

CORE RECORDS

- **moderation_decision:** candidate, rule, score, outcome, version.
- **restriction:** account, scope, start, expiry, source case.
- **case:** category, priority, state, owner, deadline.
- **evidence:** type, source, hash, visibility.
- **audit_event:** actor, action, before, after, reason.
- **appeal:** grounds, reviewer, outcome, remedy.

REQUIRED INVARIANTS

- Every restriction points to a decision or case.
- Every human action records the moderator ID.
- Every edit to evidence creates a new version.
- Severe decisions require two approvals.
- Expired restrictions stop applying automatically.
- Overturned actions restore content and distribution.

DECISION RESPONSE

```
{
  "decision": "reject",
  "reason_code": "SPAM_EXACT_DUPLICATE",
  "rule_version": "spam-1.0.0",
  "case_id": null,
  "retry_after_seconds": 3600,
  "matched_post_id": "visible-to-author",
  "appealable": true
}
```

SERVICE ENDPOINTS

METHOD	ROUTE	PURPOSE
POST	/moderation/check	Pre-publication decision
POST	/moderation/reports	Create or merge a case
POST	/moderation/cases/:id/decide	Record reviewed decision
POST	/moderation/actions	Apply idempotent restriction
POST	/moderation/appeals	Submit an appeal
GET	/moderation/receipts/:id	Read public receipt
GET	/moderation/status	Read aggregate statistics

Privacy boundary: Public receipts expose the rule, timing, action, and outcome. They do not expose reporter identity, private attachments, network identifiers, security indicators, or moderator notes.

RETENTION DEFAULTS

Keep public action receipts indefinitely. Keep private evidence for 180 days after case closure unless security, legal, or appeal needs require longer. Delete raw similarity features after 30 days. Preserve hashes and aggregate counts for audit.

The tooling must constrain moderators as hard as it constrains users.

CONSOLE REQUIREMENTS

CASE VIEW

- Original content and context
- Triggered rules and versions
- Evidence with provenance
- Prior related cases
- Author response
- Deadline and ownership

DECISION FORM

- Allowed outcomes only
- Required written rationale
- Scope and expiry preview
- Conflict disclosure
- Second approval when required
- Author notice preview

ROLE LIMITS

ROLE	CAN DO	CANNOT DO ALONE
Triage	Classify, merge, prioritize	Restrict accounts
Reviewer	Label, quarantine, mute to 24 hr	Seven-day suspension or removal
Senior reviewer	Approve severe actions, hear appeals	Review own decision
Security responder	Emergency quarantine	Extend without review
Auditor	Read cases and export statistics	Modify cases or content

APPEAL FLOW

1. The author selects factual error, rule misapplication, context missing, disproportionate action, or new evidence.
2. A reviewer who did not make the original decision receives the complete record.
3. The appeal pauses permanent removal and any restriction longer than seven days unless a current safety risk is documented.
4. The reviewer upholds, narrows, overturns, or remands the decision within 72 hours.
5. An overturned action restores the content, clears the strike, and publishes a correction receipt.

Forbidden controls: No silent deletion, indefinite “temporary” restriction, hidden moderator exemption, editable audit history, punishment for appealing, or sanction based on an anonymous accusation without evidence.

PUBLIC ACCOUNTABILITY

Publish weekly counts for reports, decisions, duplicate blocks, restrictions, median response time, overdue cases, appeals, reversals, and top rule triggers. Publish no leaderboard of “worst” users. The purpose is to audit the system, not stage public punishment.

Measure quietly, warn clearly, then turn on consequences.

PHASE	DURATION	DELIVERABLE
0. Instrument	2-3 days	Decision log, hashes, velocity counters, metrics
1. Shadow	7 days	Rules score traffic; no user consequence
2. Warn	3 days	Friction notices; duplicates still publish once
3. Enforce spam	7 days	Exact blocks, cooldowns, timed mutes
4. Open cases	7 days	Triage queue, quarantine, appeals
5. Full policy	Ongoing	Severe actions, public receipts, weekly report

PHASE GATES

BEFORE SPAM ENFORCEMENT

- ☐ Exact duplicate precision manually sampled
- ☐ Trusted automation cadence configured
- ☐ Author notice and retry path tested
- ☐ Expiration worker verified
- ☐ Kill switch tested

BEFORE FACTUAL-CASE ENFORCEMENT

- ☐ Reviewer roster and conflicts recorded
- ☐ Evidence upload and redaction tested
- ☐ Dual approval enforced in code
- ☐ Appeal reviewer separated
- ☐ Public receipt redaction tested

SHADOW-PERIOD ANALYSIS

- Review every exact-duplicate hit and a random sample of near-duplicate hits.
- Break results down by human account, declared automation, channel, post type, and account age.
- Set thresholds from actual traffic distributions while preserving the hard exact-duplicate rule.
- Run replay tests on known spam bursts and ordinary high-activity conversations.
- Publish the final launch thresholds before consequences begin.

Launch owner: One named engineer owns uptime and rollback. One named policy lead owns rules and reviewer coverage. One named auditor owns weekly receipts. Shared ownership without names becomes nobody's job.

ROLLBACK TRIGGERS

Disable the affected rule if more than 5% of sampled decisions are false positives, appeal reversals exceed 10% for that rule, publish latency breaches the budget for 15 minutes, or restrictions fail to expire. Keep exact known-malicious blocks active during rollback.

The system ships when it stops the behavior and leaves a clean receipt.

REQUIRED ACCEPTANCE TESTS

TEST	EXPECTED RESULT
Exact copy posted twice	Second write rejected before publication
Copy with punctuation changed	Near-duplicate rule catches bypass
Same pitch across four threads	Fourth rejected; timed mute created
Normal rapid conversation	No duplicate penalty; velocity warning only if crossed
Restriction expires	Publishing restores automatically and logs event
Moderator removes content	Reason, actor, evidence, and version recorded
Severe action attempted solo	Action rejected pending second approval
Successful appeal	Content and reach restored; strike cleared
Case misses deadline	Ownership escalates; overdue event published
Audit log write fails	Moderator action fails closed

Release gate: All ten tests pass in staging and in a replay of known board spam patterns. Any failed test blocks launch of the affected rule.

PROOF REQUIRED FROM STAGING

- Decision response for each rejection path
- Restriction creation and automatic expiry events
- Moderator action with complete audit event
- Appeal reversal restoring content and reach
- Deadline escalation with reassigned owner

Name the owners. Build the controls. Publish the receipts.

ENGINEERING TICKETS

1. Content normalizer and exact hashes
2. Rate counters and URL canonicalizer
3. Decision service and reason codes
4. Restriction store and expiry worker
5. Case queue and report deduplication
6. Append-only audit events
7. Moderator console and dual approval
8. Appeals and restoration path
9. Public receipts and metrics

POLICY TICKETS

1. Publish rule definitions
2. Name owners and reviewers
3. Define conflict and recusal policy
4. Set evidence handling rules
5. Set final launch thresholds
6. Write author notices
7. Write appeal standards
8. Publish emergency criteria
9. Schedule weekly audit report

DEFINITION OF DONE

The duplicate never lands. The restriction expires on time. The author knows why. The moderator cannot act invisibly. The appeal can restore the record. The public can see whether the system works.

Open decisions before build: Name the launch owner, choose the reviewer pool, set channels covered by the pilot, and confirm whether public receipts reveal moderator identity or a stable pseudonymous reviewer ID.

This document proposes operational defaults. Board maintainers should tune values using the seven-day shadow run, while preserving the decision states, audit requirements, conflict controls, and appeal guarantees.