Bidirectional Safeguard Checkpoints

---

0) High-level intent

Forward (toward openness): Add automation + transparency only when preconditions are met; add them in reversible, auditable steps.

Backward (toward restriction): Remove/limit automation only via declared, time-boxed procedures with proofs, independent co-signatures, and auto-reversion.

---

1) Preconditions & Trigger Criteria

Forward enablement (Tier ↑)

F-1 Legal climate: No new speech-restricting emergency orders in last 120 days; courts operational.

F-2 Civic signals: Independent press index above a set threshold; watchdog NGOs can publish without prior approval.

F-3 Operational quality: False-positive rate < X% on last quarter; appeal turnaround < Y days; audit remediation closed.

F-4 Community readiness: Public consultation completed; change log & FAQ published.

Backward constraint (Tier ↓)

B-1 Emergency trigger: Court-recognized emergency or verifiable mass harassment/violence campaigns.

B-2 Integrity threat: Evidence of systematic moderation capture (e.g., coerced decisions), documented by an internal whistleblower channel and external observer.

B-3 Measurement spike: Harm metrics exceed red thresholds (e.g., targeted doxxing events/week > baseline + 5σ).

Both sets must be evidence-backed and recorded in the public change ledger (below).

---

2) Public Change Ledger (tamper-evident)

What: Every tier change, rule update, and enforcement mode switch is written to an append-only log (hash-chained).

How: Each entry contains: reason code (F-1/… or B-2/…), human-readable summary, metric snapshot, decision minutes, and dual signatures (local compliance + external observer).

Safeguard: If dual signature not present within 7 days, an auto-reversion timer flips the system back to the prior state.

---

3) Forward Path Checkpoints (adding automation & openness)

1. F-Start: Low-risk automation only

Enable automation for spam, scams, malware.

Keep political/identity topics manual + explainers.

Circuit breaker: One-click global pause for any new automated rule.

2. F-A: Contextualization by default

For sensitive legal content: show context banners instead of takedowns.

Require appeal & explanation links on every banner.

3. F-B: Human-validated automation

Ship new classifiers behind shadow mode → compare with human decisions → promote to active only if delta error < threshold.

Publish validation report in ledger.

4. F-C: Transparent governance

Quarterly transparency bundle: confusion matrix, top error modes, redress stats, policy diffs.

Open API for researchers (rate-limited, privacy-preserving).

5. F-Lock: Ratchet with rollback

Each forward move sets a reliable rollback path (config snapshots + migration scripts + “known good” models).

Document rollback SLO (e.g., < 6 hours to revert).

---

4) Backward Path Checkpoints (reducing automation safely)

1. B-Start: Freeze sensitive automation

Immediately disable automated decisions on political/identity content; switch to human review.

Keep non-sensitive automation (spam/security) on to prevent chaos.

2. B-A: Time-boxed manual mode

Declare scope + expiry (≤ 30 days) + criteria for return to previous tier.

Enter in ledger with dual signatures; start countdown timer displayed publicly.

3. B-B: Minimal necessary restrictions

If takedowns required, prefer geo-scoped and time-limited actions with clear legal citation.

Publish takedown notice (reason code + law ref + appeal route).

4. B-C: Oversight continuity

External observer retains read-only access to logs & sampling; can file public objections in the ledger.

5. B-Exit: Auto-reversion

If expiry hits without renewed justification + dual sign, the system automatically restores the former automation level and publishes a notice.

---

5) Anti-Abuse Guards (dual-use prevention)

Quorum for deep changes: Any model swap, rule overhaul, or broad blocklist needs 2/3 quorum across: local ops, global safety, and independent auditor.

Red team drills: Quarterly simulations of both forward/backward shifts; publish after-action reports.

Shard by risk: Separate pipelines for safety (spam/malware) vs speech (political/cultural); backsliding can't secretly turn off safety.

Provenance on prompts/rules: Hash + version every rule/model; users can request the active version ID cited in decisions.

---

6) User-side Protections (work in all tiers)

Appeals with SLA: Target response time (e.g., ≤ 7 days); auto-escalate if breached.

Case packet: On action, user can download a packet (decision summary, rule/version IDs, how to contest).

Data portability: One-click export; mirror to a neutral locker during declared manual modes.

Explainability at point of impact: "Why you're seeing this notice / Why your post was limited."

---

7) Governance & Roles

Local Compliance Officer (LCO): Signs ledger entries; accountable to local law + platform policy.

Global Safety Board (GSB): Cross-region experts; handles cross-border harms and arbitration.

Independent Observer (IO): Civil society/academic partner with audit access and a public dissent channel.

Emergency Steward (ES): Single on-call owner of circuit breakers; actions must be co-signed post-hoc within 24h.

---

8) Operational Runbook (checklist)

Before any shift

☐ Fill change template (goal, risks, metrics, rollback).
☐ Snapshot configs/models; verify restore script.
☐ Dry-run in staging; attach results.
☐ Schedule comms (public notice, FAQs).

During shift

☐ Activate circuit-breaker dashboard; live metrics panel.
☐ Open a real-time log room (read-only link for IO).
☐ Record decision & signatures in ledger.

After shift

☐ 7-day quality review (errors, appeals).
☐ Publish transparency bundle.
☐ Close remediation items with owners & deadlines.

---

9) Metrics that matter (and can’t be gamed)

Quality: FP/FN by category; appeal-win rate; time to resolution.

Safety: Incidents per 100k users; severity-weighted harm index.

Fairness: Disparity of actions across protected classes/topics (with privacy protections).

Trust: User satisfaction on explanations; auditor variance (how often auditors disagree).

Resilience: Time to rollback; time in manual mode vs plan; number of unsigned ledger entries (should be zero).

---

10) Minimal Disclosure Tiers (for sensitive regions)

Tier A (internal): Full detail (rules, thresholds) — staff & auditors only.

Tier B (research): Aggregates + sampled rationales, k-anonymized.

Tier C (public): Counts, reasons, versions, and notices—enough for accountability, not for circumvention.

---

11) Example timelines

Forward (Transitional → Open):
Week 0: Shadow mode validators → Week 2: Context banners on → Week 6: Limited automation (non-political) → Week 12: Automation expansion + public API.

Backward (Open → Transitional/Restricted):
Day 0: Freeze sensitive automation (B-Start) → Day 1: Declare 30-day manual window (B-A) → Day 7: Publish interim report → Day 30: Either renew with evidence or auto-revert (B-Exit).
