# AI Governance Policy

**Version:** 2026-04 · **Owner:** AI Safety Officer · **Distributed as:** `Votriz_AI_Governance_Policy.pdf`

## 1. Principles

Votriz's AI governance follows the four core functions of the NIST AI Risk Management Framework:

**Govern**, **Map**, **Measure**, **Manage**. Formal NIST RMF attestation and ISO 42001 certification are on the 2027 roadmap; current state is "principles-aligned with documented controls."

## 2. Human oversight (Govern)

All AI-generated content requires human approval before any external publishing or send action.

- **Content queue:** every drafted post sits in `content_queue` with `status = 'pending'` until a user with `content.approve` permission acts on it.
- **Email campaigns:** subject line, body, and recipient segment must be approved by a user with `email.campaigns.send` before the worker dispatches.
- **Brand monitor:** AI auto-drafts crisis responses; a human must approve before the response is posted. Incidents flagged `severity = 'critical'` route to human review **regardless** of AI confidence score.
- **Ghost Presence (autonomous mode):** opt-in per org, with a per-brand `auto_approve_confidence_threshold`. Disabled by default. Subject to the same kill switches as other AI agents.

## 3. Transparency (Map)

- Every row in `content_queue` carries a provider/model identifier in `details` so users can see which model produced the suggestion.
- Brand DNA learning decisions (which examples taught the AI what) are stored in `brand_memory` and reviewable via the Brand DNA page.
- The full AI model inventory is published in `AI_MODEL_INVENTORY.md`; the customer-facing summary appears on votriz.com/compliance.
- Every AI-related action is logged to `security_audit_log` (e.g. `content.approve`, `content.reject`, `email.campaign_send`) with the originating user and request_id.

## 4. Risk monitoring (Measure)

- **Quality scoring:** generated content is scored against the brand's voice profile before reaching the approval queue. The scoring is exposed on the Brand DNA page so users can spot regressions.
- **Approval-rate monitoring:** approval / rejection ratios per brand are tracked over time. A sustained drop is the leading signal of model drift; formal automated drift monitoring is on the 2026-Q3 roadmap.
- **User feedback loop:** `POST /ai/report-issue` lets any user flag a generated piece as biased, inaccurate, inappropriate, or off-brand. Reports land in `security_audit_log` under `ai.quality_report` and are routed to the AI Safety Officer.

- **Sentiment & crisis signals:** the brand monitor tracks sentiment shifts and predicts incident severity from multiple signal sources to reduce single-classifier bias.

## 5. Accountability (Manage)

- **Designated AI Safety Officer:** Founder / CTO. Email: security@votriz.com.
- All AI-related decisions are logged and queryable by the org owner via `/audit/log`.
- Quarterly review of AI output quality metrics: approval rates, rejection reasons, user-reported issues, crisis-detection accuracy. Findings drive prompt and threshold updates.
- Plan to establish a formal AI Ethics Committee with external advisors as headcount grows.

## 6. Fairness and bias

- Content generation uses brand-specific voice profiles rather than generic templates. This minimizes regression-to-mean output that often surfaces unintended bias.
- Sentiment analysis aggregates multiple signal sources; no single classifier acts unilaterally on actionable mentions.
- Users can disable any AI agent on a per-brand basis via the Brand DNA controls.
- Bias / fairness reports submitted via `/ai/report-issue` are treated as P3 (Medium) incidents per `AI_INCIDENT_RESPONSE.md`.

## 7. Data principles

- **No training on customer data.** Customer content, prompts, Brand DNA, subscriber data, and analytics are never used to train, fine-tune, or otherwise improve any AI model. Anthropic's API contract explicitly excludes API data from training corpora.
- **Per-customer learning only.** Brand DNA is an in-database voice profile scoped to the customer's `org_id`; it is loaded into prompt context at inference time and never shared with other customers.
- **Ephemeral prompts.** AI prompts exist only for the duration of the inference call. They are not persisted by Votriz beyond the resulting content row.
- **PII minimization.** The PII redactor (`services/pii_redactor.py`) is applied on the chatbot and email-personalization prompt paths. It is intentionally NOT applied to the lead generator path, where extracting public business contact information is the agent's explicit job — redaction there would defeat the purpose.

## 8. Kill switches

Documented in `AI_INCIDENT_RESPONSE.md` §4. Three layers:

1. **Per-agent disable:** comment out the cron registration in `votriz-worker/main.py` and rebuild.
2. **Global AI disable:** set `VOTRIZ_AI_DISABLED=true` on the worker container.
3. **Network block:** firewall rule blocking outbound to the LLM provider's API. Nuclear option for a confirmed breach.

## 9. Reporting AI issues

Any user can report a quality, bias, or safety issue:

- In-app: "Report AI issue" button on every generated content card.
- API: `POST /ai/report-issue` with `{content_id, issue_type, description}`.
- Email: security@votriz.com.

Reports are reviewed within 24 hours. Systematic patterns trigger a review of the relevant agent's prompts and scoring thresholds.