SOP-AI-2025 — Handling AI Misuse

Purpose: Define protocols for detecting, responding to, and preventing harmful AI-generated content. **Scope:** Applies to contributors, moderators, and governance teams across all Guardian Override platforms.

Workflow:

1. **Identification:**
   - Flag AI-generated outputs that are harmful (misinformation, harassment, impersonation, manipulative content).
   - Contributors must disclose AI tool + version used.
2. **Classification:**
   - Minor: harmless but undisclosed AI use → advisory notice.
   - Moderate: misleading or manipulative AI content → restriction.
   - Critical: harmful, harassing, or impersonation content → suspension/ban.
3. **Response:**
   - Immediate suspension for Critical cases.
   - Escalation to Conduct Officers within 24h.
   - Documentation in MEI case log.
4. **Restorative Pathways:**
   - First-time offenders may undergo training on AI disclosure and ethics.
5. **Audit:**
   - Quarterly review of flagged AI cases.
   - Transparency report includes AI misuse statistics.