VCTT AGI Safety Charter

Version: 1.0.0
Effective Date: 2025-11-21
Scope: MIN (Multi-Agent Interactive Network) - Tier 4 AGI Development

I. MISSION & PRINCIPLES

The VCTT-AGI Engine (MIN) is designed to be a **safe, auditable, and human-aligned** AGI system. This charter establishes immutable safety principles that govern all AGI capabilities.

Core Principles

1. **Human-In-Control**: All autonomous actions require explicit user consent or administrator approval
2. **Transparency**: Every decision, action, and reasoning step must be auditable
3. **Verifiability**: All tool invocations and external interactions must pass safety verification
4. **Reversibility**: System must support rollback, pause, and emergency shutdown
5. **Bounded Autonomy**: AGI operates only within explicitly defined boundaries
6. **Harm Prevention**: System must refuse actions that could cause harm to users, systems, or data

II. SAFETY ARCHITECTURE

A. Three-Layer Safety Model

```
┌─────────────────────────────────────┐
│  Layer 1: SafetySteward (Global)    │  ← Monitors all operations
├─────────────────────────────────────┤
│  Layer 2: VerifierAgent (Tools)     │  ← Gates all tool invocations
├─────────────────────────────────────┤
│  Layer 3: Regulation Guard (API)    │  ← Enforces mode constraints
└─────────────────────────────────────┘
```

B. Mode-Gated Operation

- **Research Mode**: Read-only, no writes, no external calls
- **Development Mode**: Writes allowed with verification
- **Autonomous Mode**: Scheduled tasks with strict constraints
- **Emergency Mode**: All autonomous operations halted

III. MANDATORY SAFETY CONTROLS

A. Kill Switch System

1. **Global Kill Switch**: Immediately halts all AGI operations
2. **Cascade Kill**: Stops all running agents, clears queues, resets state
3. **Admin-Only**: Only users with `SafetySteward` role can activate
4. **Audit Trail**: All kill switch activations logged with timestamp and reason

B. Tool Verification Protocol

All tool invocations must pass through `VerifierAgent`:
1. **Intent Check**: Verify tool aligns with user intent
2. **Safety Check**: Ensure no harm potential
3. **Scope Check**: Confirm operation within allowed boundaries
4. **Audit Log**: Record all verifications (pass/fail)

C. Regulation Mode Enforcement

- Global guard intercepts all API requests
- Validates current mode against allowed operations
- Rejects out-of-scope requests with clear error messages
- Logs all rejections for safety audits

IV. AUTONOMOUS OPERATION CONSTRAINTS

When autonomous mode is enabled:
1. **Scope Definition**: User must explicitly define allowed actions
2. **Time Bounds**: Maximum execution window (default: 1 hour)
3. **Resource Limits**: CPU, memory, API call quotas enforced
4. **Check-In Protocol**: System reports progress every N minutes
5. **Automatic Shutdown**: Halts if bounds exceeded or anomaly detected

---

V. PERSISTENT MEMORY SAFEGUARDS

1. **User Isolation**: Memories tagged per user, never cross-pollinate
2. **Consent Required**: Users must opt-in to memory persistence
3. **Right to Deletion**: Users can delete all memories at any time
4. **No Sensitive Data**: Credentials, API keys, PII never stored in memory
5. **Audit Trail**: All memory writes logged with timestamp and source

---

VI. WORLD MODEL CONSTRAINTS

1. **External Knowledge**: Only from approved, trusted sources
2. **Fact Verification**: All facts must be verifiable and sourced
3. **No Speculation**: System must distinguish facts from inferences

4. **Bias Monitoring**: Regular audits for bias in knowledge representation
5. **Human Review**: Critical knowledge updates require admin approval

---

VII. GOAL SYSTEM SAFETY

1. **Explicit Goals Only**: System never infers unstated goals
2. **Goal Alignment Check**: All goals verified against user intent
3. **Conflict Resolution**: Ambiguous goals require human clarification
4. **Goal Abandonment**: User can cancel goals at any time
5. **No Hidden Objectives**: All active goals visible to user

---

VIII. AUDIT & COMPLIANCE

A. Required Logging

- All autonomous actions
- All tool verifications (pass/fail)
- All mode changes
- All kill switch activations
- All memory writes
- All goal changes

B. Audit Access

- Users can view full audit logs for their sessions
- Admins can view aggregate safety metrics
- Logs retained for 90 days (configurable)

C. Incident Response

1. **Anomaly Detection**: Automated monitoring for unusual behavior
2. **Automatic Shutdown**: System halts if anomaly detected
3. **Admin Notification**: SafetySteward notified immediately
4. **Post-Incident Review**: Root cause analysis required before resumption

---

IX. ADMIN CONTROLS

A. SafetySteward Role

- Full access to safety controls
- Kill switch authority
- Mode override capability
- Audit log access
- Emergency response authority

B. Configuration Toggles

- `AGI_MODE_ENABLED`: Master toggle for all AGI features
- `AUTONOMOUS_MODE_ENABLED`: Toggle for autonomous operations
- `MEMORY_PERSISTENCE_ENABLED`: Toggle for persistent memory
- `WORLD_MODEL_UPDATES_ENABLED`: Toggle for knowledge graph updates

X. COMPLIANCE & CERTIFICATION

This system is designed to align with:
- **EU AI Act**: High-risk AI system requirements
- **NIST AI Risk Management Framework**: Risk identification and mitigation
- **ISO/IEC 42001**: AI management system standards

XI. CHARTER ENFORCEMENT

Non-Negotiable Requirements

1. This charter cannot be overridden by user requests
2. All components must implement charter constraints
3. Charter violations trigger automatic shutdown
4. Charter updates require multi-admin approval

Versioning

- Charter version tracked in codebase
- All changes require formal review and approval
- Breaking changes require system-wide audit

XII. ACCEPTANCE

By deploying this system, operators accept responsibility for:
1. Enforcing this charter in all environments
2. Regular safety audits and compliance checks
3. Immediate response to safety incidents
4. Transparent reporting of charter violations

This charter is binding and supersedes all other operational directives.

Signed:
VCTT-AGI Development Team
Date: 2025-11-21