NUT AGI FRAMEWORK: Human Oversight Procedures

Version: 1.0 | **Date:** January 30, 2026

1. Overview

Nut AGI implements a comprehensive human oversight framework through the Safety Net Protocol (SNP) to ensure safe, transparent, and accountable AI operations. Human oversight is mandatory for all high-risk decisions and is structured across three levels:

Level 1: Automated monitoring with human-in-the-loop alerts

Level 2: Human review and approval for medium-risk decisions

Level 3: Mandatory human intervention for high-risk/critical decisions

2. Safety Net Protocol (SNP) Components

2.1 GAN-Based Critic

- Trained on 1M synthetic failure cases + 10k red-team scenarios
- Discriminator loss threshold: D-loss < 0.1
- Flags outputs that deviate from safe patterns
- 98% adversarial input flagging rate in pre-beta testing

2.2 Confidence Threshold

Pre-Beta Standard: 95% confidence (p=0.05 binomial test)

Deployment Standard: 99.5% confidence (EU AI Act + NIST 800-53 High Impact)

Override Mechanism: Human override required when confidence p < threshold

2.3 Audit Ledger

- SHA-256 cryptographic hashing of all decisions
- Immutable blockchain-style ledger
- Merkle tree construction for regulatory audits
- Real-time access for regulators and compliance officers

3. Role-Based Oversight Structure

Role	Responsibilities	Qualifications
AI Safety Monitor	• Monitor SNP alerts 24/7 • Escalate medium/high-risk flags • Document all interventions	• AI/ML degree or equivalent • 2+ years AI safety experience • Security clearance for sensitive data
Domain Expert	• Review domain-specific outputs (finance, legal, medical) • Approve/reject high-stakes decisions • Provide feedback for model improvement	• Professional certification in relevant domain • 5+ years industry experience • Understanding of AI limitations
Compliance Auditor	• Conduct quarterly audit reviews • Verify regulatory compliance • Generate audit reports for regulators	• Legal/compliance background • GDPR/AI Act expertise • Internal audit certification (CIA/CISA)
Ethics Committee	• Review ethically ambiguous cases • Set ethical guidelines and policies • Investigate high-profile incidents	• Diverse backgrounds (ethics, law, tech, social sciences) • Independence from operational teams
Executive Oversight	• Final authority on kill-switch activation • Strategic oversight of AI safety • Board-level reporting	• C-suite or Board level • Accountability for AI governance

4. Decision Classification and Review Requirements

Risk Level	Examples	Oversight Requirement
LOW	• General Q&A • Content summarization • Creative ideation • Code suggestions (non-critical)	• Automated monitoring only • Quarterly audit sample review
MEDIUM	• Business analytics • Market trend analysis • Non-critical financial recommendations • HR screening support (non-binding)	• Random sampling review (10%) • AI Safety Monitor approval for flagged cases • Monthly audit
HIGH	• Credit scoring • Hiring/promotion decisions • Portfolio allocation >$100k • Regulatory compliance outputs	• 100% human review required • Domain Expert approval mandatory • Real-time audit logging • Weekly review by Ethics Committee
CRITICAL	• Safety-critical infrastructure • Medical diagnosis/treatment • National security applications • Mass financial transactions	**NOT APPROVED** for autonomous operation • Human-only decision authority • Nut provides support/analysis only

5. Kill-Switch Procedures

Activation Triggers:

- Systematic bias detected (>10% disparate impact)
- Security breach or unauthorized access attempt
- Repeated SNP confidence threshold failures
- Regulatory or legal violation
- Unintended self-evolution patterns (ASI risk)

Kill-Switch Authority:

Level 1: AI Safety Monitor (partial shutdown - specific modules)
Level 2: Chief Technology Officer (system-wide pause)
Level 3: CEO / Board (permanent shutdown)

Execution Process:

1. Emergency alert triggers automated system pause (< 5 seconds)
2. Immediate notification to oversight team and executives
3. Incident investigation and root cause analysis
4. Remediation plan development and approval
5. Phased restart only after Executive Oversight approval

6. Escalation Protocols

Standard Escalation Path:

AI Safety Monitor → Domain Expert → Compliance Auditor → Ethics Committee → Executive Oversight

Time-Based Escalation:

- Immediate: Critical safety issues (kill-switch activation)
- 1 hour: High-risk decision requiring expert review
- 24 hours: Unresolved medium-risk flags
- 1 week: Recurring low-risk issues indicating systemic problem

7. Training and Competency

Initial Training (all oversight personnel):

- Nut AGI technical architecture and limitations (8 hours)
- AI ethics and bias recognition (4 hours)
- Regulatory compliance (EU AI Act, GDPR, sector-specific) (6 hours)
- SNP operations and kill-switch procedures (4 hours)
- Incident response and escalation (2 hours)

Ongoing Training:

- Quarterly refresher courses
- Annual recertification exam
- Case study reviews of real incidents (monthly)

8. Reporting and Documentation

Real-Time Logs: All human interventions logged with timestamp, user ID, decision rationale

Weekly Reports: Summary of flagged cases, interventions, and trends

Monthly Reviews: Oversight effectiveness metrics and improvement recommendations

Quarterly Audits: Comprehensive compliance review with external auditor participation

Annual Transparency Report: Public disclosure of oversight activities and safety metrics

____
Founder’s Office
Nrutseab Ltd. | hello@nrutseab.com