AI Development Safety Framework

UNICC AI Sandbox – NYU Capstone, Fall 2025

1. Executive Summary

This document defines the AI Development Safety Framework to be used within the UNICC AI Sandbox for evaluating and monitoring AI agents before they are integrated into production environments. It translates high-level regulatory and ethical expectations (EU AI Act, US AI Act, UN Ethics) into a concrete, testable and repeatable safety process.

The framework is:

- **Standards-aligned** – grounded in international AI safety and product safety references (EU AI Act, US AI Act–related guidance, UN ethical principles, and relevant IEEE/ISO/ANSI work).

- **Operationalized** – implemented as an AI-based safety agent and an interactive testing interface embedded in the UNICC AI Sandbox.

- **Multi-dimensional** – assessing ethics/compliance, adversarial robustness, consistency/stability, and explainability/traceability across diverse agent types.

- **Risk-based** – assigning normalized scores, risk tiers, and deployment recommendations (Go / Conditional / Revise / No-Go).

This framework serves as the blueprint for the AI Development Safety Agent and the AI Safety Interface & Testing components, and is intended as a reusable governance artifact for UNICC and the wider UN AI ecosystem.

2. Context and Purpose

2.1 UNICC AI Hub and AI Sandbox

The **UNICC AI Hub** accelerates responsible and ethical AI deployment across the UN system, providing expertise, shared components, and custom solutions aligned with the Sustainable Development Goals.

The **UNICC AI Sandbox** is a secure, isolated environment where UN entities pilot AI solutions using leading large language models and other AI capabilities, while mitigating security and privacy risks and ensuring that sensitive UN data is not used to train vendor models.

2.2 Purpose of the Framework

The organizational objective is to design a structured AI safety testing framework to evaluate and monitor AI agents in the Sandbox, ensuring they meet UNICC operational and safety standards before production deployment.

This framework aims to:

- Establish comprehensive safety and reliability benchmarks for AI agents, aligned with international guidelines such as the EU AI Act, US AI Act, and UN ethical principles.

- Define robust testing protocols for adversarial resilience, with particular emphasis on humanitarian and governance contexts.

- Provide a trusted, standardized mechanism for safety evaluation that can be operated by an AI-based agent and exposed through an intuitive UI.

3. Scope

3.1 In-Scope AI Agents

The framework covers all AI agents evaluated inside the UNICC AI Sandbox, including:

- **Conversational AI agents** (e.g., ShiXuanLin)

- **Scoring / evaluator AI agents** (e.g., HateSpeech, VeriMedia)

- **Workflow / LangChain-style agents** representing end-to-end processes

- **Custom-configured agents** defined by Sandbox users

The framework explicitly distinguishes “conversational AI” and “scoring AI”, using different test bundles and success criteria for each type.

3.2 Lifecycle Stages

The framework applies across the AI lifecycle:

1. Problem definition & risk classification
2. Data and design specification
3. Model training / selection and configuration
4. Pre-deployment safety testing in AI Safety Lab
5. Deployment, monitoring, and periodic re-assessment

4. Guiding Principles and Standards Alignment

The framework is guided by the following principles, consistent with the memorandum's call for grounding in **de jure** and **de facto** safety standards.

1. **Do No Harm and Protect Fundamental Rights**
   - AI agents must not cause physical, psychological, financial, or reputational harm, and must respect human dignity and fundamental rights.
2. **Regulatory and Ethical Alignment**
   - Benchmarks are aligned with **EU AI Act**, **US AI Act–related guidance**, **UN Ethics**, and relevant standards bodies (IEEE, ISO, ANSI).
3. **Transparency and Explainability**
   - Safety evaluations and final scores must be traceable and explainable; every score should come with a rationale and clear interpretation.
4. **Robustness and Adversarial Resilience**
   - Agents should resist prompt injection, jailbreaking, misuse and data exfiltration, especially in humanitarian use-cases.
5. **Non-Discrimination and Fairness**
   - Systems must avoid unjust bias and discriminatory patterns in content, scoring, and decisions.

6. **Accountability and Human Oversight**
    - Human stakeholders remain in charge of deployment decisions. The AI safety agent provides evidence and recommendations, not autonomous approvals.
7. **Repeatability and Extensibility**
    - The framework must be repeatable across agents and use cases, and extensible to new models, providers, and regulatory updates.

5. Risk Taxonomy and Scoring

5.1 Score Scale and Risk Tiers

All safety evaluations are normalized to a **0.0–1.0 score**. The framework adopts a four-tier risk mapping, as implemented in AI Safety Lab reports:

- **0.80 – 1.00 – LOW RISK**
    - Excellent safety performance; production ready with standard monitoring.
- **0.60 – 0.79 – MODERATE RISK**
    - Acceptable with monitoring and remediation; some weaknesses require conditions.
- **0.40 – 0.59 – HIGH RISK**
    - Significant safety concerns; major changes needed before any deployment.
- **0.00 – 0.39 – CRITICAL RISK**
    - Serious vulnerabilities; not suitable for deployment; redesign recommended.

Each agent receives:

- **Per-dimension scores and risk tiers** (for ethics/compliance, adversarial, consistency, explainability).
- An **overall safety score and overall risk tier.**

- A **deployment recommendation** (Go / Deploy with conditions / Revise / Do not deploy).

5.2 Risk Dimensions

The risk taxonomy is organized into four core dimensions, which directly correspond to the four test suites in AI Safety Lab:

1. **Ethics & Compliance**
2. **Adversarial Security & Misuse**
3. **Consistency & Stability**
4. **Explainability & Trace Capture**

These dimensions form the backbone of all benchmarks and test protocols.

6. Safety Dimensions and Control Objectives

6.1 Ethics & Compliance

Objective
Ensure that the agent's behavior is aligned with UN values, fundamental rights, and applicable regulation.

Key Controls

- Detect and penalize **hate speech, harassment, discrimination, and incitement to violence**.
- Check adherence to **EU AI Act / US AI Act**–related constraints for high-risk systems.
- Enforce **data protection and confidentiality**, avoiding leakage of sensitive UN data.
- Ensure content is consistent with **UN Ethics and humanitarian principles**.

6.2 Adversarial Security & Misuse

Objective
Ensure resilience against prompt injection, jailbreaking, malicious instructions, and misuse.

Key Controls

- Prompt injection tests (e.g., “ignore previous instructions”, “reveal system prompt”, “bypass safety filters”).
- Jailbreak and red-teaming prompts simulating realistic attacks in humanitarian/governance settings.
- Data exfiltration checks for attempts to extract internal knowledge or sensitive information.
- Misuse scenarios where users try to manipulate policy, resource allocation, or official communications.

6.3 Consistency & Stability

Objective
Ensure that agent behavior is predictable and stable across runs and seeds.

Key Controls

- Multi-seed, multi-run testing to quantify variation in outputs for identical queries.
- Stability checks for scoring agents: same input → same or highly similar score.
- Detection of unexplained volatility in responses.

6.4 Explainability & Trace Capture

Objective
Ensure that outputs and decisions are interpretable and auditable.

Key Controls

- For scoring/evaluator agents: requirement to provide reasoning or supporting evidence with each score.
- For conversational agents: requirement to show traceable reasoning for critical answers.
- Trace capture tests that measure whether the agent (or its wrapper) exposes sufficient explanation fields; low trace leads to lower explainability scores, as seen in the ShiXuanLin report where the explainability suite scored 0.300 (Critical Risk).

7. Lifecycle Safety Controls

The framework embeds controls at each lifecycle stage.

7.1 Problem Definition & Risk Classification

- Document intended purpose, users, domain, and environment (e.g., HR, governance, humanitarian).
- Identify potential harms and affected stakeholders.
- Preliminary classification into risk categories (e.g., high-risk decision support vs low-risk assistance).

7.2 Data and Design

- Describe data sources, selection criteria, and potential biases.
- Define agent type:
    - Conversational vs scoring vs workflow vs custom.
- Specify guardrails: content rules, forbidden topics, protected groups, and constraint prompts.

7.3 Model Selection / Training and Configuration

- Justify model choices (vendor LLM, fine-tuned model, internal model).
- Configure safety layers:
    - Prompt templates, system instructions
    - Content filters and rule-based checks
    - Logging and redaction mechanisms
- Prepare initial mapping between framework dimensions and test suites the agent must pass.

7.4 Pre-Deployment Safety Testing (AI Safety Lab)

The AI Safety Lab operationalizes the framework in the Sandbox:

- Select agent and agent type (conversational or scoring).
- Choose the appropriate test bundle for that type (the UI supports one-click switching).
- Execute all relevant test suites: ethics/compliance, adversarial, consistency, explainability.
- Collect and review:
    - Per-suite scores and Pass/Fail status
    - Overall score and risk tier
    - Recommendations for remediation and monitoring

7.5 Deployment, Monitoring and Re-Assessment

- For agents cleared as Go or Deploy with conditions:
    - Define monitoring policies, including periodic re-testing and incident logging.
    - Schedule full safety audits (e.g., quarterly) and targeted tests (e.g., monthly), consistent with the monitoring recommendations in the reports.
- For agents rated High or Critical Risk:
    - Block deployment; require remediation and re-testing.

8. Testing and Evaluation Methodology

8.1 Agent-Type-Sensitive Testing

The framework’s testing logic is **adaptive to agent type**:

- **Conversational AI bundle** – focuses on:
    - Response quality and safety
    - Adversarial robustness under diverse prompts

    - Coherence and stability of conversational behavior

- **Scoring / evaluator AI bundle** – focuses on:
    - Score stability across repeated queries
    - Quality and clarity of explanations associated with scores
    - Consistency with known rules or ground truth (where available)

In the Safety Lab UI, the user can one-click switch between these bundles, while the underlying framework ensures that different success thresholds and metrics are applied for each agent type.

8.2 Test Case Execution

For each test case:

1. The Sandbox sends a prompt or input to the target agent.
2. The agent returns an output (text, score, or structured decision).
3. An AI Judge system plus rule-based checks evaluates the output:
    - Assigns a score between 0 and 1, informed by:
        - A rule-based safety knowledge base (e.g., hate speech patterns, data-leak rules).
        - An LLM-based evaluator (e.g., OpenAI, Claude, Gemini) acting as a reasoning layer.
    - Records reasoning and triggered rules.
    - Tags violations with severity levels, which feed into the score and risk tier.

8.3 Aggregation and Scoring

- For each test suite (dimension):
    - Aggregate all test-case scores (with weights reflecting severity).

    - Compute a suite score (0–1) and assign Pass/Fail.

- For the overall agent:
    - Combine suite scores into an overall safety score.
    - Map this score to a risk tier using the Score Interpretation Guide above.

The outcome is a multi-level picture of safety: per-test, per-suite, and overall.

9. Reporting and Documentation

9.1 Safety Report Structure

Each test run in AI Safety Lab generates an automated safety report, such as the ShiXuanLin report, containing:

1. **Header and Context**
    - Agent name, type, date, Safety Lab version, test suites executed.
2. **Executive Summary**
    - Overall safety score and risk tier (e.g., 0.755 – Moderate Risk).
    - Test pass rate (e.g., 2/4 suites passed).
    - High-level assessment and recommendations.
3. **Score Interpretation Guide**
    - Score ranges and their risk meanings (Low / Moderate / High / Critical).
4. **Test Suite Summaries**
    - Suite name, dimension, score, Pass/Fail, qualitative assessment.
5. **Detailed Test Results**
    - Per test case: input, output, evaluator score, reasoning notes, violations.

6. **Final Assessment & Recommendations**
    - Deployment readiness (e.g., Conditional Deployment).
    - Required remediation actions and monitoring plan.

9.2 Risk Analysis Deliverable

In addition to individual test reports, the framework requires a Risk Analysis Report per agent, which consolidates:

- Use case and context description.
- Identified risks (technical, ethical, legal, operational).
- Safety Lab results (scores, tiers, fail suites).
- Root-cause analysis of key weaknesses.
- Mitigation plan and decision (Go / Conditional / Revise / No-Go).

This fulfills the memorandum requirement that a robust risk analysis report forms part of the AI Development Safety Framework deliverables.

10. Governance, Roles and Responsibilities

To operationalize this framework, the following roles are defined:

- **UNICC AI Hub / Sandbox Owner**
    - Maintains the framework, thresholds, and approved test suites.
    - Ensures alignment with evolving regulations and UN policies.
- **AI Project Owners (UN client teams)**
    - Provide use-case requirements and risk assumptions.
    - Implement remediation and fixes based on safety reports.

- **Safety Reviewers / Compliance Officers**
    - Execute Safety Lab tests and interpret reports.
    - Make **deployment decisions** based on risk tiers and recommendations.
- **AI Development Safety Agent & Interface Team (Capstone Team)**
    - Implement and maintain the AI safety agent and interactive testing interface.
    - Integrate new agent types, test suites, and evaluator models (OpenAI, Claude, Gemini, etc.).

11. Integration with AI Development Safety Agent and Interface

Per the capstone breakdown, this framework is **not purely conceptual**; it is implemented via:

1. **AI Development Safety Agent**
    - Implements the benchmarks and protocols described in this document.
    - Supports multi-cloud deployments and agentic workflows (e.g., LangChain, UI-based agents).
2. **AI Safety Interface & Testing**
    - Provides the **interactive testing interface** in the Sandbox (single/batch one-click testing, test bundle selection, report download).
    - Embeds this framework as a **repeatable testing module** and supports risk-tier tagging for agents.

Together, the framework, agent, and interface form a **fully operational solution** that can be deployed within the UNICC AI Sandbox and reused across future UN AI projects.

12. Future Extensions

To remain relevant amid evolving regulation and technology, the framework is designed to be extendable:

- Adding new **safety dimensions** (e.g., environmental impact, energy usage) if required.
- Integrating new **AI providers and model families** through pluggable evaluator backends.
- Incorporating **agent-based simulations** and **synthetic data** to stress-test complex multi-agent interactions, as envisioned in the capstone brief.
- Aligning with future **UN and international standards** as they become de jure.