Dynamic Teen Coalition

December 3, 2025

THE UPSTREAM SAFETY SYSTEM (USS): AISI Pilot Program Brief

Stacy Gildenston & Pyrate Ruby Passell, dynamicteencoalition@gmail.com, Dynamic Teen Coalition

Executive Summary

The Upstream Safety System (USS) is a first-of-its-kind deterministic policy compliance engine designed for regulatory certification of high-risk AI systems. USS provides safety classification and content-policy adjudication through transparent, auditable governance.

Unlike probabilistic AI safety approaches, USS implements a Technical Separation of Powers across three auditable layers, enabling transparent, traceable decisions that can be independently audited against AISI's voluntary AI safety guardrails.

USS combines:

• **Deterministic decision-making:** Every outcome is produced by explicit, human-readable Datalog rules

• **End-to-end auditability:** Diagnosis (Layer 3), policy logic (Layer 2), and impact measurement (Layer 1) are strictly separated

• **Quantified duty of care:** Dignity and Agency are operationalised as measurable outcomes

Why USS Meets AISI's Certification Requirements

Traditional AI safety solutions struggle with regulatory scrutiny because they rely on probabilistic black-box models. USS addresses this through:

1. Deterministic Architecture: Policy decisions trace to explicit logic rules

2. Complete Auditability: Each layer's function can be independently verified

3. Measurable Impact: Concrete metrics for ethical outcomes

Technical Architecture: Separation of Powers

Layer 3: Diagnose — Diagnostic Engine

• Hybrid TF-IDF classification into 13 persona vectors using mature, well-understood methods

• **F1 Score: 0.71902** on 19,375-row validation

• Interpretability layer for human-readable structured facts

• **Cost: ~$0.00004 per message**

Layer 2: Decide — Declarative Policy Engine

• Datalog-based rules (deterministic, auditable)

• Age-Stage Classification with priority ladder (Human Safety > Human Rights)

• Self-selecting policy controls for adults (preserves autonomy)

Layer 1: Validate Impact — Impact Simulator

• Synthetic testbed for pre-deployment policy validation

• Quantifies Dignity (harm prevention) and Agency (autonomy preservation)

• Empirical measurement before production release

Addressing Implementation Scrutiny

1. Cost Validation & Economic Viability

Per-Message Cost Breakdown:

Component	Cost per Message
TF-IDF vectorization	~$0.00001
Logistic regression (CPU)	~$0.00002
Datalog policy evaluation	~$0.00001
Total USS Cost	**~$0.00004**

Scale Impact (1M messages/day): LLM: $3.6M-10.9M/year vs USS: $14,600/year = **Annual savings: $3.6M-10.9M**

2. Performance Benchmarking & Intentional Trade-offs

Our **~3-4 percentage point accuracy trade-off** (0.72 vs 0.75+) yields 3-5x faster inference, deterministic outputs, and complete explainability. Regulatory compliance requires deterministic, explainable decisions at scale — a capability probabilistic systems cannot provide, regardless of accuracy.

3. Quantifying Dignity & Agency: Methodological Rigour

Dignity Metrics: Policy remediation success rate, user appeal accuracy, harm prevention rate

Agency Metrics: Autonomy support score, choice-space preservation, self-selection adoption

Validation: Layer 1 Impact Simulator runs synthetic scenarios measuring both metrics before deployment

Regulatory Alignment

AISI Voluntary Standard: USS addresses all 10 guardrails through deterministic architecture, audit trails, and contestability mechanisms

eSafety Commissioner: Quantifiable Duty of Care addresses the mandate for demonstrable harm prevention

Pilot Program Readiness

Governance: Managed by Dynamic Teen Coalition, the first Teen Board at the UN, who were invited to, participated in, and endorsed the UN Global Digital Compact.

Leadership: Stacy Gildenston won the first WSIS Award in 2003 and brings three decades of technical governance experience, including developing and running worldwide certification programs for USENIX and Linux Professional Institute. Pyrate Ruby Passell served as a mentor at ITU's first Citiverse Challenge, is a CERN Open Quantum Institute Friend, and won the 2018 Australian Youth Rocketry Challenge.

Development: Three years of UN-level digital governance development with defined accountability structures. USS represents the maturation of policy-first architecture validated through sustained international engagement.

Timeline: Ready for March 2026 pilot deployment

Strategic Fit: USS provides a technical foundation compatible with likely mandatory guardrail requirements: deterministic, auditable, cost-effective safety at scale.

USS enables AISI to demonstrate that voluntary safety governance can be both transparent and economically viable at a national scale.