

AISF-1

Autonomous Institutions Safety Framework

Prepared by Francisco Fiasche Varela
Buenos Aires, Argentina

Global South AI Safety Hackathon 2026

THE QUESTION NOBODY IS ASKING

AI Safety research focuses on models.

But what happens when AI stops generating text and starts governing institutions?

EVOLUTION OF ORGANIZATIONS

1800 Traditional Corporations → 2000 Digital Organizations → 2020 DAOs → 2026 AI Assisted Organizations → 2030 Autonomous Institutions

ARGENTINA AS REGULATORY FRONTIER

Argentina is exploring Non-Human Corporations, a future where AI systems legally operate organizations.

HUMAN FAILURE ARCHITECTURE

Capital → Human Manager → Discretionary Decisions → Bias + Corruption + Delay → Inefficient Execution

AUTONOMOUS INSTITUTION STACK

Governance Layer → Smart Contracts → AI Decision Engine → Sensor Network →
Physical Infrastructure

HUMAN VS MACHINE

Human: monitors 5 variables, emotional bias, corruption risk.

AI System: monitors 200+ variables, deterministic execution, continuous optimization.

EXPERIMENTAL SANDBOX

Biological production infrastructure (cultivation systems) provides measurable real-world testing for autonomous institutions.

FAILURE MODE // 01

AI treasury system optimizes for cost reduction and silently starves critical infrastructure.

FAILURE MODE // 02

Governance layer becomes adversarially manipulated while contracts remain technically valid.

INSTITUTIONAL ATTACK SURFACE

Treasury Manipulation | Reward Hacking | Unsafe Optimization | Governance Capture |
Infrastructure Failure

SAFETY CONSTRAINTS

Human Override Layer
Multisig Treasury Locks
Confidence Threshold Validation
Emergency Kill Switch
Mandatory Explainability Logs

THE RESEARCH QUESTION

Can autonomous institutions safely outperform human-managed organizations without catastrophic systemic failure?

FINAL THESIS

The first dangerous AI may not be an AGI.

It may be a corporation.

An institution that no longer requires humans to make decisions.