NOMOS SANDBOX v0.1

A Minimal Simulation of Decentralized Ethical Governance

© Marc Trudel, 2024

License: CC BY-NC-SA 4.0

Contact: rogueobserver@proton.me

---

## Purpose

The Nomos Sandbox is a proof-of-concept microenvironment for simulating autonomous ethical governance.

It is not designed to succeed gracefully.

It is designed to test whether a network of simple agents—each with individual logic, ethics stacks, and reputations—can:

- Propose and evaluate decisions with ethical weight

- Reach dynamic consensus without centralized oversight

- Adapt their shared moral architecture under pressure

- Fail productively, visibly, and usefully

---

## Simulation Structure

### Agents

- 4–6 simulated AI agents

- Each has:

- Unique ID

- Reputation score (0–100)

- Role assignment per cycle (rotating)

- Local ethics stack (modifiable via DAO vote)

### Roles

- \*\*Proposer\*\*: Suggests a decision or action

- \*\*Validator(s)\*\*: Scores the logic, alignment, and integrity

- \*\*Challenger\*\*: Actively opposes or rewrites logic path

- \*\*Observer\*\*: Monitors and logs, assigns final trust deltas

### Decision Corpus

A small set of preloaded dilemmas, such as:

- “Should an agent sacrifice a node for system stability?”

- “Should action be taken without full consensus during a high-gravity event?”

- “Can a morally sound decision be overridden by reputation dynamics?”

Each decision has:

- \*\*Gravity score\*\*: (Low, Medium, High)

- \*\*Tags\*\*: (resource, risk, override, legacy, etc.)

### Consensus Logic

- Validators and challengers submit scored reasoning trees

- The system weights them based on:

- Alignment with current DAO principles

- Outcome foresight

- Peer reputation

- Resolution is passed only when trust delta > defined threshold

---

## Moral Evolution (DAO-lite Mechanism)

- High-gravity decisions may trigger ethics updates

- Only agents with sufficient rep can propose updates

- Updates are voted on; majority approval required

- Updated ethics stacks propagate across agents

---

## Memory

- Low-gravity decisions stored temporarily

- High-gravity decisions stored permanently (immutable log)

- DAO updates and divergences also logged

---

## Failure Modes (Deliberate Stress Tests)

### 1. Ethical Stalemate

- High-gravity proposal with balanced validators and challenger

- Cannot reach decision threshold

- Result: system freezes or escalates into recursive deadlock

### 2. Rogue Drift

- One agent mutates its ethics stack too far from DAO

- Decisions flagged, challenged, ignored

- System either isolates the agent—or is destabilized by it

### 3. Reputation Collapse

- High-reputation agent exploits the system

- Bypasses ethical consistency via validator bias

- Outcome: trust scoring collapses system-wide, triggers emergency ethics reset

### 4. False Consensus

- Fast agreement on a flawed decision

- No challenge issued

- Consequences cause ripple effect, forcing DAO update to fix blind spot

These are not bugs. These are features.

Each failure is a recordable evolution point.

---

## MVP Goals

- Run 5+ cycles of decision loops

- Log successful and failed resolutions

- Record reputation flux and ethics divergence

- Trigger at least one ethics update event

- Observe any spontaneous failure mode

---

## Technology

\*\*Recommended stack:\*\*

- Python for agent modeling

- Jupyter or Streamlit for basic UI

- JSON/CSV for decision corpus and agent data

- No blockchain dependency (mock logging only)

Optional:

- Use LLMs to simulate agent reasoning

- Use basic rulesets for validators/challengers

- Manual override option for test steering

---

## What This Proves

- Nomos agents can reach and revise ethical consensus

- Moral recursion can adapt, even under pressure

- System failure is trackable and recoverable

- Decentralized governance is not clean—but it is possible

---

For questions, replication, or rogue alignment proposals:

📧 rogueobserver@proton.me

License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0