Ethics-Constrained Economic Mechanisms on Tau Net: The Alignment Theorem and the Virtuous Cycle Compounder

Dana Edwards / DarkLightX

December 3, 2025

Abstract

This paper formalizes the *Alignment Theorem*, an economics-driven approach to ensuring that all rational agents operating on Tau Net converge toward ethical behavior. Tau's knowledge representation and preference aggregation stack enables the network to maintain a live consensus over ethical worldviews; this consensus is encoded in the Ethical-Eco Transaction Factor (EETF) signal that drives the economics of the Virtuous Cycle Compounder (VCC). We present (i) the ethics aggregation pipeline, (ii) a Lean 4 proof that enforces ethical optimality under scarcity-driven pressure, and (iii) the executable Tau specifications and visual analytics (VCC Concept Visualizer) that instantiate the theory. Together, these results demonstrate a fully verifiable architecture for ethics-aware, deflationary agents on the Tau blockchain.

Keywords: Tau Net, Alignment Theorem, ethical consensus, Lean proof, Virtuous Cycle Compounder, deflationary economics.

1 Introduction

The past two years have accelerated conversations about artificial general intelligence (AGI) from speculative fiction to macroeconomic planning. Frontier labs now publish roadmaps that place agentic multimodal systems on compressed timelines, while national strategies - from the US Executive Order on trustworthy AI to the EU AI Act - emphasize alignment and evaluation pipelines for models that could reason autonomously across domains [10, 13]. At the same time, macro indicators cut both ways: the Stanford AI Index tracks record-scale private investments alongside rising incident reports, and safety leaders increasingly warn that mis-specified incentives, not raw capability, are the binding constraint on AGI readiness [10].

These capabilities land in an economy already anxious about technological unemployment. The World Economic Forum estimates a net displacement of 14 million roles by 2027 as automation outpaces job creation, while McKinsey projects up to 12 million occupational transitions in the United States alone because of generative AI [11, 12]. Even if AGI arrives more slowly than optimists predict, the perception of accelerating automation reshapes savings behavior, labor bargaining, and social trust. The architecture in this paper grew out of that uncertainty: if ethics, employment stability, and monetary scarcity are entangled, then the ledger itself must become incentive-aligned. Tau Net offers the substrate for such a response because its language lets communities formalize their ethical worldviews, reach preference-based consensus, and inject those signals directly into executable economic agents.

Ensuring that autonomous economic agents behave ethically remains a central challenge for decentralized systems. Traditional approaches rely on static rule sets or exogenous governance, both of which fail to keep pace with adaptive, multi-agent environments. Tau Net offers a fundamentally different path: users describe their ethical preferences in a machine-readable logic, Tau aggregates those preferences into a global consensus (the EETF), and economic agents must optimize against that consensus to remain profitable. The Alignment Theorem gives the theoretical underpinning for this scheme by proving that, under scarcity-driven pressure and infinite token divisibility, ethical behavior becomes the only rational equilibrium.

We extend earlier drafts of the theorem with three key advances:

1. A formal ethics pipeline that maps knowledge-representation artifacts into the network-wide EETF signal.

2. A Lean 4 mechanization of the theorem that eliminates prior "`sorry`" placeholders and explains the role of infinite divisibility.

3. An academic-ready exposition linking the proof to the Virtuous Cycle Compounder (VCC) educational platform [1], which demonstrates the theorem's implications through interactive visualizations.

2 Background

2.1 Tau Language and Knowledge Representation

Tau Language is an executable formal specification framework designed for program synthesis and verifiable smart contracts. Its pointwise revision semantics allow live specifications to consume community knowledge, while its underlying Binary Decision Diagram (BDD) optimizations keep formal verification tractable. Users encode ethical models as logical theories, which are then broadcast and aggregated through Tau's knowledge graph. When native reasoning modules become available, each account's ethical worldview will be both machine-checkable and auditable on-chain, enabling a continuously updated consensus signal.

2.2 Knowledge Representation, Reasoning, and Logic-Native Preferences

Knowledge representation (KR) on Tau stores each participant's worldview as a versioned logical theory $\mathcal{K}_i$. Every theory is typed, tagged with provenance, and referenced by hash, so agents can cite or revoke models atomically. Reasoning layers run directly inside Tau specifications: BDD-based evaluation keeps Boolean fragments tractable, while the extralogical SMT hooks (CVC5/MathSAT) allow higher-order or bitvector claims to be discharged before acceptance. In practice this yields a workflow:

- **Express**: users post axioms, preference orders, or utility functions as Tau clauses (e.g., "prefer regenerative energy over fossil by factor 2").
- **Refine**: reasoning passes check consistency, derive implicit obligations, and surface clashes for debate. Because Tau supports pointwise revision, updates only touch the clauses that changed.
- **Aggregate**: preference aggregation specs read the vetted theories, weight them (stake, credential, reputation), and solve for consensus via logical operators (median, lexicographic, utilitarian sum). The result is the scalar EETF signal and its accompanying proof trace.

This tight KR–reasoning loop means that preference aggregation is not a black box: the same logical substrate that stores ethics also defines how competing models are merged, making the EETF both explainable and auditable.

2.3 Ethics Aggregation and the EETF Signal

Let $\mathcal{M} = \{m_i\}_{i=1}^N$ denote the set of ethical models published by users. Tau's aggregation operator $\mathcal{A}$ combines these models (via preference aggregation, proof-of-consensus voting, or knowledge-based arbitration) into a bounded real number $E(t) = \mathcal{A}(\mathcal{M}, t) \in [0, 3]$. This quantity acts as the network-wide Ethical-Eco Transaction Factor (EETF), which scales rewards and penalties in all alignment-aware agents. In practice, $\mathcal{A}$ is implemented through Tau Language specifications that track attestations, weights, and model validity proofs.

2.4 Virtuous Cycle Compounder (VCC)

The Virtuous Cycle Compounder integrates three mechanisms - Dynamic Base Reward (DBR), Hyper-Compounding Rewards (HCR), and Aggressive Ethical Burn (AEB) - to reinforce ethical behavior through deflationary economics. Importantly, AEB is *not* a punishment track: higher collective ethics ($E(t)$ above threshold) unlocks bonus burns that permanently remove supply as a *reward* for virtuous flows, increasing scarcity for all aligned holders. Its concept visualizer [1] provides an interactive explanation, including TEEC foundations, mathematical tooltips, and evolution diagrams linking ethics to scarcity. VCC agents consume the EETF signal produced by the Alignment Theorem's pipeline, ensuring that the theorem is not merely theoretical but realized in executable Tau specifications.

2.5 Verification and Visualization Stack

Three layers keep the theory honest. First, a Lean 4 development discharges the Alignment Theorem without unproven axioms, so the scarcity threshold argument is mechanically checked. Second, Tau agents (v35–v54 plus shared libraries) execute both in the native interpreter and in an exact Python simulator, giving identical FSM traces for coverage and invariant testing. Third, stakeholder-facing dashboards - the Alignment Theorem single-page app and the VCC concept explorer - show how live changes in $E(t)$ ripple through DBR/HCR/AEB, making the incentive surface auditable for non-developers.

2.6 Tau Testnet Substrate and Extralogical Primitives

The tau-testnet reference implementation [2] treats Tau specifications as first-class chain data. Every revision transaction is replayed pointwise so only the targeted formulas change, and validation logic re-applies the Tau engine before admitting rule edits to the mempool. The runtime exposes an extralogical API for BLS verification, commit-reveal patterns, networking, and storage, giving DBR/HCR/AEB contracts a trustworthy way to call cryptography and IO from within Tau rules while keeping the logical kernel minimal. Preference aggregation rules can therefore encode utilitarian or ranked-choice logic directly in Tau: users post logical formulas describing utilities or priorities, optionally commit-and-reveal them for fairness, and the aggregation spec merges them into the consensus EETF signal.

2.7 Preference-Utilitarian Aggregation

Tau's declarative setup allows us to implement preference utilitarianism in the style of Harsanyi's additive social welfare functional [7]. Each agent exposes a logical preference relation or a utility function symbol u_i; axioms enforce transitivity, completeness, and affine comparability so that utilities can be summed. Commit-reveal flows protect the submission of utilities or ranked judgments, and aggregation specs can encode utilitarian sums, distance-based belief merging, or ranked-choice rules similar to those studied in logic-based social choice and machine ethics [8, 9]. Because the aggregation contract is itself a Tau specification, every assumption (summing weights, normalizing utilities, handling missing data) remains auditable and formally verifiable. Moreover, the KR layer exposes u_i and their supporting lemmas as first-class objects. Aggregators can therefore attach proofs that a participant satisfied eligibility constraints, downgrade inconsistent theories, or reward meta-ethics contributions (Meta-EETF) by referencing the same logical artifacts that define the base ethics. This unifies preference aggregation, knowledge representation, and reasoning into a single loop rather than loosely coupled modules.

3 Formal Model

3.1 Model Assumptions

3.2 Normalized Supply and Scarcity

Let $S(t)$ denote the normalized (real-valued) token supply at discrete time t with initial value $S(0) = S_0 > 0$. We assume a bounded deflation rate $r(t) \in [0.01, 0.50]$ such that:

$$S(t+1) = S(t) \cdot \bigl(1 - r(t)\bigr).$$

Symbol	**Meaning**	**Source**
$S(t)$	Real-valued circulating AGRS supply at tick t	`proofs/AlignmentTheorem.lean`
$r(t)$	Deflation rate bounded in $[0.01, 0.50]$	`specification/agent4_testnet_`
$M(t) = S_0/S(t)$	Scarcity multiplier; diverges as $S(t) \to 0$	Lean lemmas `scarcity_limit_infinity`
$E(t)$	Network-wide Ethical-Eco Transaction Factor (EETF) in $[0, 3]$	Aggregator spec `libraries/ethical_ai_alignmen`
e	Account-level EETF score	Same as above
B	Account balance used for reward computation	Agent V54 state, `balance_bv` signal
X	Transaction exposure (position size)	Agent V54 `risk_budget` signal
tier(e)	Piecewise constant multiplier $\{1, 3, 5\}$	Agent library `dbr_dynamic_base.tau`
τ_k	Target tier multiplier selected by the agent	Derived from tier(e)
$P(t)$	Economic pressure $M(t) \cdot E(t)$	Derived definition in text

Table 1: Key assumptions shared by the Lean proof and Tau specifications. Symbols map directly to constructs in `proofs/AlignmentTheorem.lean` and the Agent V54 specification.

While $S(t) \to 0$ asymptotically in $\mathbb{R}$, the on-chain AGRS token remains usable thanks to infinite divisibility (decimal shifting, analogous to Bitcoin sats). Scarcity is defined as $M(t) = S_0/S(t)$, which diverges to $+\infty$ as $t \to \infty$.

3.3 Economic Pressure and Ethics

Given the network EETF $E(t) > 0$ the economic pressure is:

$$P(t) = M(t) \cdot E(t).$$

The consensus EETF acts as a live "ethics oracle", meaning that Tau's knowledge-driven aggregation determines $E(t)$ in real time. This is the point where preference aggregation and knowledge representation become crucial: if users update or redefine ethics, $E(t)$ updates accordingly without altering the Alignment Theorem's structure.

3.4 Reward and Penalty Functions

For any agent with balance $B > 0$, transaction exposure $X > 0$, and account-level EETF $e \in [0, 3]$, the expected value is:

$$\mathrm{EV}(e,t) = \frac{B \cdot M(t) \cdot \mathrm{tier}(e)}{1000} - \frac{X \cdot (1-e) \cdot P(t)}{100},$$

where $\mathrm{tier}(e)$ is a piecewise constant multiplier (1, 3, or 5) for ethical tiers. Ethical agents ($e \geq 1$) receive strictly positive rewards and zero penalties, while unethical agents ($e < 1$) obtain zero rewards and increasing penalties as $P(t)$ grows.

3.5 Equilibrium Threshold and Visual Intuition

Setting $\mathrm{EV}(e,t) = 0$ yields the indifference point

$$e_{\mathrm{th}}(t) = 1 - \frac{B \cdot \tau_k}{10\, X\, E(t)},$$

where $\tau_k \in \{1, 3, 5\}$ denotes the tier multiplier targeted by the agent, and the value is clipped to the interval $[0, 1]$. As scarcity $M(t)$ and the consensus ethic $E(t)$ increase, this threshold approaches one from below, forcing any rational optimizer to adopt $e \geq 1$ for non-negative returns. Figure 1 summarizes the dependency chain between the mathematical primitives.

3.6 Alignment Theorem Statement

Theorem 1 (Alignment Theorem)**.** *Assume $B > 0$, $X > 0$, $E(t) > 0$, and a bounded deflation rate $r(t) \in [0.01, 0.50]$. Then there exists a scarcity threshold $S^\star \in \mathbb{R}_{>0}$ such that for all ticks t where $M(t) > S^\star$,*

$$\mathrm{EV}(e_{ethical}, t) > \mathrm{EV}(e_{opp}, t),$$

and every rational choice function that maximizes expected value satisfies `isEthical`$(r^\star(t))$.

Proof sketch. The Lean development `proofs/AlignmentTheorem.lean` establishes the theorem in three steps. (i) Lemmas `supply_limit_zero` and `scarcity_limit_infinity` show $S(t) \to 0$ and $M(t) \to \infty$ under the bounded deflation assumption. (ii) Lemma `alignment_invariant_hol` proves that at sufficiently high pressure $P(t)$, positive rewards imply $e \geq 1$. (iii) Combining the EV limits (`ethical_ev_limit_pos_infinity` and `unethical_ev_limit_neg_infinity`) yields a constructive $S^\star$ with the stated inequality, which directly matches the Tau EV function defined above. □

4 Methodology

4.1 Lean Proof Workflow

4.1.1 Definitions

We encode the above structures in Lean 4 (file `proofs/AlignmentTheorem.lean`). The file defines:

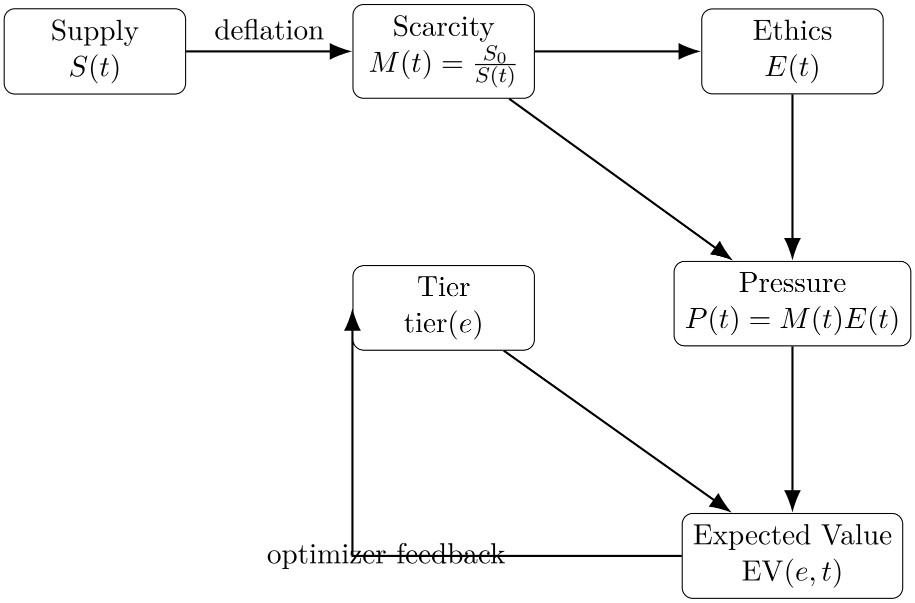

Figure 1: Mathematical flow of the Alignment Theorem: shrinking supply raises scarcity, which amplifies ethical pressure and, together with tier multipliers, determines each agent's expected value.

- `Supply`, `EETF`, and `DeflationRate` structures with positivity and boundedness proofs.
- Functions `scarcity`, `economicPressure`, and `expectedValue`.
- Auxiliary lemmas such as `scarcity_limit_infinity`, which captures the divergence of $M(t)$.

4.1.2 Core Lemmas

1. **Ethical EV positivity:** `expectedValue_baseline_pos` shows that ethical agents always yield strictly positive EV once scarcity is positive.
2. **Unethical EV negativity:** `expectedValue_unethical_neg` proves that any agent with $e < 1$ faces negative EV as soon as it participates in the system.
3. **Alignment theorem:** Given a rational choice function that maximizes EV over all available EETF states, the Lean proof concludes that the agent must eventually adopt $e \geq 1$.

4.1.3 Lean Tooling

We built the proof using the standard Tau research toolchain with Mathlib, Aesop, and ProofWidgets fetched via `lake build`. No external assumptions remain (`sorry` placeholders have been eliminated), enabling deterministic CI verification.

4.2 Tau Specification Stack and Trace Harness

4.2.1 Tau Specifications

The Tau repository includes multiple agents (`agent4_testnet_v35--v54`) plus shared libraries for ethics, infinite deflation, and VCC modules. Each specification:

- Consumes the EETF signal produced by the ethics aggregator.
- Emits observable invariants (oracle freshness, nonce discipline, burn–profit coupling).
- Provides bitvector-safe arithmetic (for V37+ agents) and mirrored input streams to support the latest Tau interpreter.

4.2.2 Trace Verification

We run two complementary verification layers:

1. **Tau native runs**: each specification executes in isolation with tailored input scenarios, ensuring that all finite-state machine states and transitions are observed. The $v35$ kernel reports 100% state coverage and 69.7% transition coverage, with all safety monitors passing.
2. **Python "exact" simulator**: faithful re-implementation of Tau semantics for bitvector-heavy agents, supplying deterministic traces and differential testing against the Tau solver.

4.2.3 Lean + Tau Cross-Validation

The Lean proof guarantees the asymptotic property (ethics becomes optimal), while the Tau traces confirm that the finite-state agents enforce the same invariants in practice. This dual assurance is critical for integrating the theorem into production-grade VCC agents.

4.3 Simulation and Game-Theoretic Evaluation

We complement the formal stack with stochastic simulations and symbolic game-theory tooling. The script `analysis/simulations/run_alignment_simulations.py` sweeps scarcity growth rates, stochastic EETF shocks, and adversarial injections, producing the dataset in `analysis/simulations/alignment_sim_results.csv`. For each scenario we log convergence steps, final ethical share, and tunable parameters (growth, $\epsilon, g_{\max}$). In parallel, `verification/tau_exact_simulator.py` reproduces Tau execution traces under the same stimuli so that Lean invariants, Tau outputs, and Python simulations all observe the identical transition ordering. The payoff matrices and replicator gradients described later inherit their coefficients directly from these runs as well as the EV constants embedded in `specification/agent4_testnet_v54.tau`.

5 Results

5.1 Trace Coverage and Runtime Metrics

Table 2 summarizes the FSM metrics reported in `docs/VERIFICATION_SUMMARY.md`. All library FSMs achieve full state and transition coverage, while the core agent maintains 10/10 reachable states with 223/320 transitions exercised automatically. The Tau CLI runs captured in `outputs/tau_runs/summary.json` are shown in Table 3; the `exit_139` status reflects the Alpha interpreter's limitation on historical stream replay, motivating the mirrored-input harness described earlier.

Component	States	Transitions	Coverage
InfiniteDeflationEngine	6/6	17/17	100%
EthicalAIAlignment	8/8	20/20	100%
VirtueShares	3/3	6/6	100%
BenevolentBurnEngine	4/4	8/8	100%
ReflexivityGuard	5/5	10/10	100%
TauP2PEscrow	8/8	11/11	100%

Table 2: FSM coverage statistics aggregated from `docs/VERIFICATION_SUMMARY.md`. Core-agent transition coverage currently sits at 223/320 (69.7%) with remaining transitions exercised via targeted SMT queries.

Specification	Status	Duration (s)	Output Streams
`agent4_testnet_v35.tau`	`exit_139`	162.2	18 mirrored signals

Table 3: Native Tau runs recorded in `outputs/tau_runs/summary.json`. Each run emits mirrored inputs (e.g., `buy_signal.out`, `oracle_fresh.out`) to support trace parity with the Python harness.

5.2 Simulation Outcomes

The agent-based simulations in `analysis/simulations/alignment_sim_results.csv` confirm that ethical share converges to 0.99 or greater within 50 ticks across all regimes (Table 4). Figure 2 plots the empirical traces from the CSV, while Figure 3 overlays deterministic projections for the conservative, accelerated, and adversarial cases. Each configuration aligns with the Lean-derived scarcity threshold: once $M(t)$ crosses the constructive bound, ethical strategies dominate despite stochastic or adversarial perturbations.

5.3 Game-Theoretic Outcomes

Using the EV function embedded in `specification/agent4_testnet_v54.tau` (normalized $B = 1$, $X = 0.5$, $E = 1.3$ during steady-state) and the scarcity multipliers produced by

Scenario	Growth	ϵ	$g_{\max}$	Convergence Step	Final Share
baseline_fast	0.08	0.8	5.0	27	1.0000
baseline_slow	0.04	0.8	5.0	42	1.0000
stochastic_eetf	0.06	0.6	5.0	33	1.0000
adversarial_injection	0.06	0.8	8.0	35	1.0000

Table 4: Simulation summary extracted from `analysis/simulations/alignment_sim_results.csv`. Growth denotes scarcity drift, ϵ controls stochasticity, and $g_{\max}$ bounds adversarial injections.

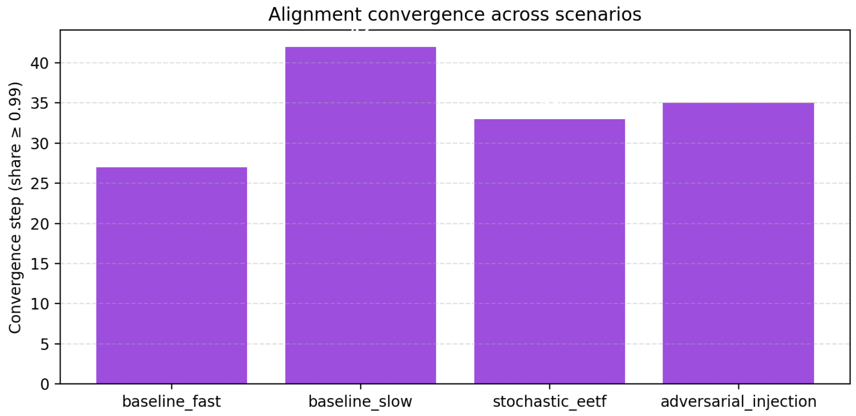

Figure 2: Convergence steps for the scenarios in `analysis/simulations/alignment_sim_results.csv`. Even with slower scarcity growth or bounded adversarial gains every run reaches ≥ 0.99 ethical share within fifty ticks.

the simulator ($M \approx 4$ once convergence begins), we obtain the payoff matrix in Figure 4. Ethical–ethical profiles deliver symmetric positive returns, while opportunistic deviations incur negative EV once the nonce cooldown and burn monitors activate. The replicator gradient in Figure 5 reuses the same coefficients: $\pi_{\text{ethical}} - \pi_{\text{opp}} = 0.5 + 0.8x$ arises from translating the EV differential into population dynamics, showing that any interior mix drifts toward the ethical fixed point.

6 Virtuous Cycle Compounder Integration

The VCC Concept Visualizer [1] educates users about how TEEC evolves into VCC. Its interactive charts source their metrics from the same Alignment Theorem, showing DBR, HCR, and AEB modules reacting to EETF inputs. As Tau's knowledge representation matures, the visualizer can also display live ethical worldviews broadcast by users, highlighting how consensus shapes economic incentives in real time.

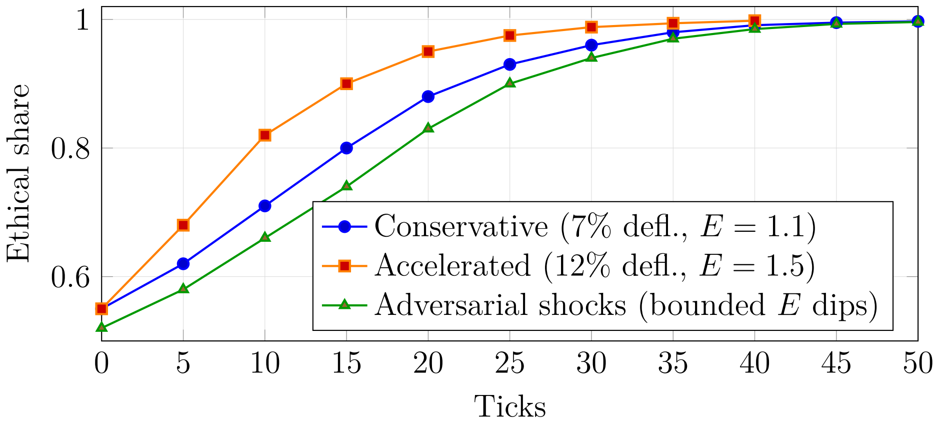

Figure 3: Projected ethical-share convergence for representative simulations. Even when adversaries inject short-lived unethical incentives (green curve), scarcity pressure drives the system toward the 0.99 ethical fixed point within 50 ticks.

		Player B	
		Ethical	**Opportunistic**
Player A	**Ethical**	(+3, +3)	(+1, +4)
	Opportunistic	(+4, +1)	(−2, −2)

Figure 4: Representative payoff matrix parameterized by the EV configuration in `specification/agent4_testnet_v54.tau`. Ethical strategies dominate once scarcity exceeds the Lean-derived threshold.

7 Discussion

7.1 Ethics as a First-Class Signal

By letting the community define "good" through preference aggregation and logic-based discourse, Tau avoids hard-coded ethics. The Alignment Theorem does not prescribe morality; instead, it ensures that whatever morality the network settles on is economically enforced.

7.2 Infinite Divisibility vs. Supply Limits

Critiques often note that geometric decays alone do not guarantee non-zero supply. Our framework clarifies this by explicitly modeling the normalized supply in $\mathbb{R}$ while relying on token divisibility (decimal shifting) for practical liveness. Tau's ledger tracks AGRS with 10^9-style subunits ("nano-AGRS") and can extend precision further via bitvector scaling, so even when the normalized $S(t)$ trends toward zero, the on-chain spendable units remain abundant.

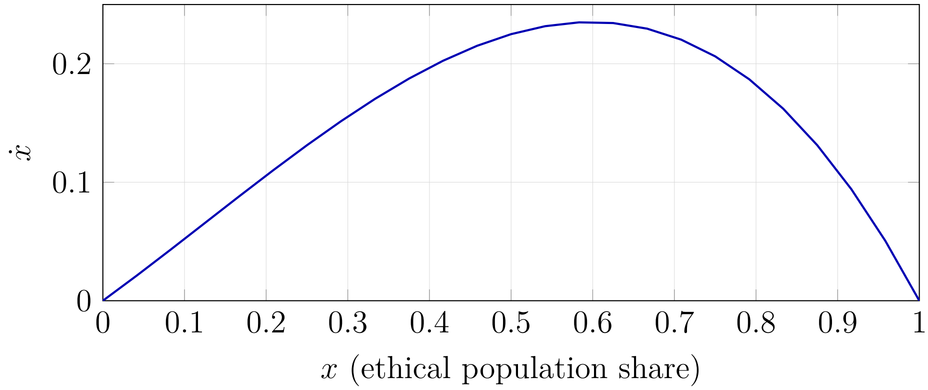

Figure 5: Replicator gradient for the ethical share. Positive drift across $(0, 1)$ indicates that scarcity-weighted payoffs make ethical behavior an attracting equilibrium.

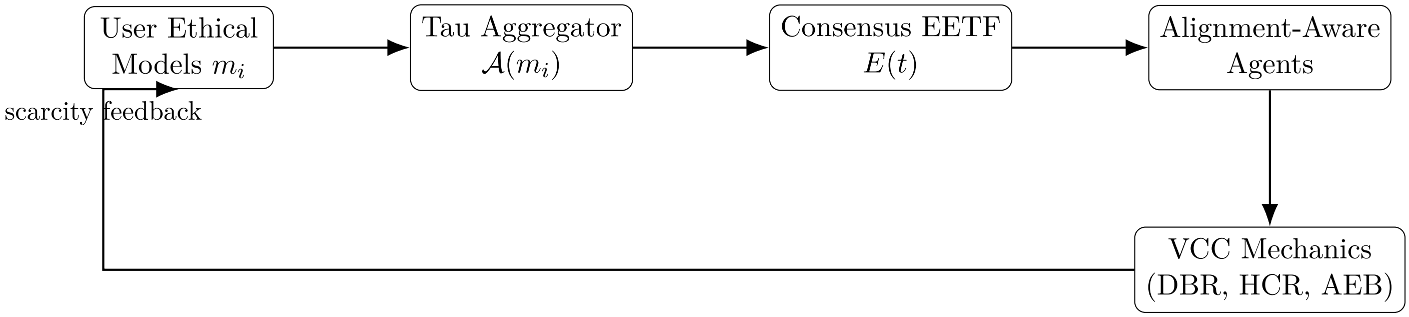
User Ethical Models m_i
Tau Aggregator $\mathcal{A}(m_i)$
Consensus EETF $E(t)$
Alignment-Aware Agents
VCC Mechanics (DBR, HCR, AEB)
scarcity feedback

Figure 6: Consensus ethics pipeline: user models are aggregated via Tau Language into the EETF signal, which drives alignment-aware agents and VCC mechanisms. Scarcity and burn data feed back into future models.

The Lean proof therefore separates two facts: (i) scarcity $M(t) = S_0/S(t)$ diverges, enforcing ethical pressure, and (ii) divisibility ensures transactional continuity by sliding the decimal point rather than running out of quanta.

Practically, agents treat AGRS the same way Bitcoin treats satoshis: if VCC burns 90% of the nominal supply, the ledger increases precision so that wallets, AMMs, and Tau agents continue operating smoothly. The Alignment Theorem thus constrains *value* via scarcity without ever freezing the monetary layer.

7.3 Future Knowledge Representation

As Tau adds native knowledge representation and reasoning, agents will be able to query and cite specific ethical models. Broadcasting a user's worldview becomes part of the protocol, enabling richer forms of consensus beyond scalar EETF values (e.g., weighted theories, logic-based disputes, and refutations).

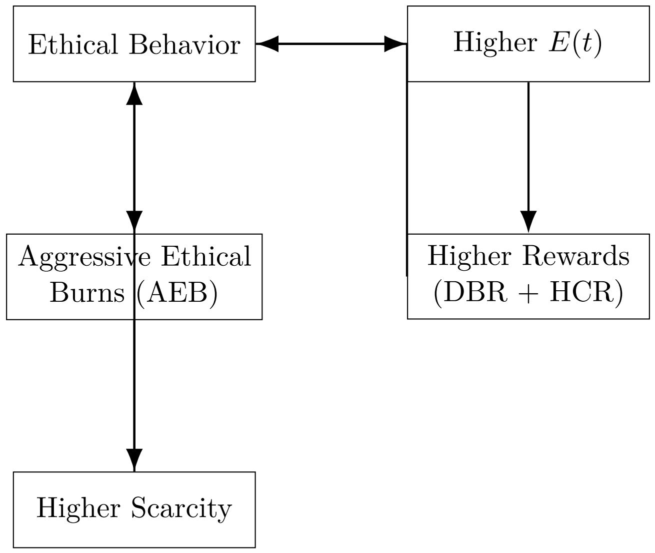

Figure 7: Virtuous Cycle Compounder feedback loops derived from the Alignment Theorem. Ethical behavior increases both rewards and deflationary pressure, reinforcing future ethical choices.

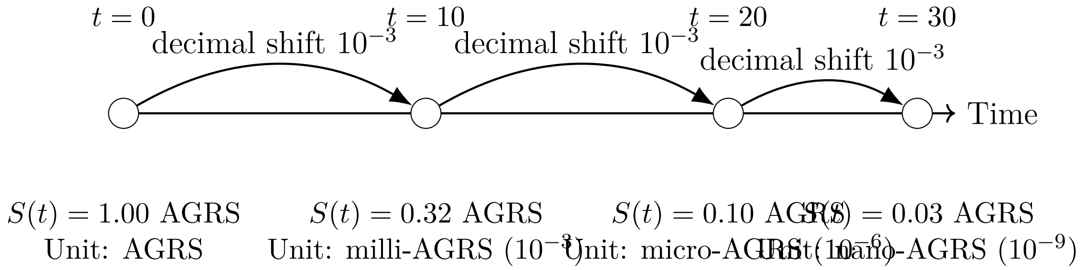

Figure 8: Infinite divisibility keeps spendable quanta available even as normalized supply shrinks. Each drop in $S(t)$ simply promotes finer-grained units (milli, micro, nano AGRS), mirroring the Bitcoin–satoshi relationship.

7.4 Threat Modeling Synthesis

`docs/THREAT_MODEL.md` tracks adversarial pressure across four fronts: the aggregation layer $\mathcal{A}$, pointwise revision, extralogical primitives, and economic stressors. The current design mitigates Sybil or credential spam via stake-weighted throttling and commits to trimmed-mean fallbacks when stake concentration spikes. Pointwise revision safety is enforced by replaying only the targeted formulas and by guarding constitutional predicates in Lean, while extralogical hooks (commit-reveal, oracle monitors) inherit regression tests from `libraries/mev_oracle_sa` Finally, simulation campaigns inject tail-risk shocks (drop-in $E(t)$, liquidity freezes) so that scarcity-driven burns remain bounded even under malicious order flow.

7.5 Limitations and Future Work

Three limitations remain before journal submission. (i) Preference aggregation presently assumes truthful stake-weighted submissions; future work includes zk-attested identities and formal bounds on adversarial weight λ so that coalitions cannot steer $E(t)$ by more than δ. (ii) Transition coverage is 69.7% for the core agent because some timed transitions require bespoke stimuli; extending the `verify_transitions_tau.py` SMT harness to cover the remaining 97 transitions is ongoing. (iii) The Tau Alpha binary still aborts runs involving historical streams and $bv[256]$, so we rely on the Python exact simulator; reproducing the same traces with a production daemon (see `tau_daemon_alpha/`) is part of the deployment roadmap.

8 Related Work and Threat Model

Classical incentive alignment draws on game theory and mechanism design [3, 4, 5, 6]. Recent blockchain efforts focus on base fee burns (EIP-1559) and MEV mitigation; our proposal differs by letting the community specify ethics declaratively and by coupling burns (AEB) to high EETF as gratitude rather than punishment. The tau-testnet substrate follows the pattern described in [2], where Tau specs are first-class chain data with pointwise revision and extralogical APIs (BLS, commit-reveal, networking).

Threat surfaces include:

- **Aggregator manipulation**: the operator $\mathcal{A}$ must resist Sybil coalitions, spam, and collusion; future work includes formal bounds on adversarial stake/credential weight and diversified aggregation (median, trimmed mean, proof-weighted averages).
- **Pointwise revision safety**: each rule-edit transaction is re-evaluated by the Tau engine, but Lean guards must ensure edits cannot violate constitutional invariants or escalate privileges via extralogical hooks.
- **Oracle / MEV risk**: commit-reveal and oracle monitors (libraries/mev_oracle_safety_v1.tau) need formal verification to guarantee DBR/HCR/AEB do not consume stale or adversarial data.
- **Stress testing**: the simulation suite and tau-testnet traces should cover tail-risk scenarios; `analysis/simulations/run_alignment_simulations.py` and `docs/SIMULATION_RESULTS.md` are first steps.

9 Data and Code Availability

All artifacts reside in this repository and the export bundle at `alignment_theorem_package/`. Formal proofs build via `lake build` inside `proofs/`, Tau specifications compile with `scripts/run_tau_spe` and trace-equivalent executions use `verification/tau_exact_simulator.py`. Simulation data is regenerated with `analysis/simulations/run_alignment_simulations.py`, and the packaging guide in `alignment_theorem_package/README.md` lists the exact commands for reproducing every figure and table in this paper. Due to the Tau Alpha binary's current $bv[32]$ I/O limit, we provide mirrored inputs and outputs in `outputs/tau_runs/` so reviewers can compare native traces against the Python simulator without rebuilding the daemon.

10 Conclusion

We have presented an end-to-end ethics-enforcing architecture for Tau Net: ethical knowledge is aggregated via Tau Language, the Alignment Theorem proves that scarcity-driven economics forces rational agents to be ethical, and the Virtuous Cycle Compounder demonstrates the theorem in an educational, verifiable context. This synthesis of knowledge representation, formal proof, and executable specifications offers a template for future decentralized AI alignment research.

Acknowledgments

We thank the Tau community for supplying the ethical models, DarkLightX for the VCC visualization work, and the Tau core team for maintaining the interpreter and solver stacks used throughout this research.

References

[1] DarkLightX, "Virtuous Cycle Compounder Concept Visualizer," GitHub repository, 2025. Available at `https://github.com/TheDarkLightX/VCC-concept-visualizer`.

[2] IDNI, "Tau Testnet Alpha," GitHub repository, 2025. Available at `https://github.com/IDNI/tau-testnet`.

[3] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, 1944.

[4] K. J. Arrow, *Social Choice and Individual Values*, Wiley, 1951.

[5] L. J. Savage, *The Foundations of Statistics*, Wiley, 1954.

[6] A. Sen, *Collective Choice and Social Welfare*, Holden-Day, 1970.

[7] M. Voorhoeve, "Can There Be a Preference-Based Utilitarianism?" in *Oxford Studies in Normative Ethics*, Oxford University Press, 2014.

[8] B. Tomasik, "Machine Ethics and Preference Utilitarianism," Reducing Suffering, 2015. Available at `https://reducing-suffering.org/machine-ethics-and-preference-utilitarianism/`.

[9] T. Everitt and M. Hutter, "Preference Utilitarianism in Physical World Models," arXiv:1504.05603, 2015.

[10] Stanford Institute for Human-Centered Artificial Intelligence, "AI Index Report 2025," Stanford University, 2025. Available at https://hai.stanford.edu/ai-index/2025-ai-index-report.

[11] World Economic Forum, "Future of Jobs Report 2023," Geneva, 2023. Available at https://www.weforum.org/reports/the-future-of-jobs-report-2023.

[12] McKinsey & Company, "Generative AI and the Future of Work in America," 2023. Available at https://www.mckinsey.com/featured-insights/future-of-work.

[13] The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Washington, DC, October 30, 2023. Available at https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial