Tempting the Agent: On-Chain Reputation in ERC-8004 AI Agent Economics

Federico Gatta[1], Manuel Naviglio[1], Francesco Tarantelli[2]

[1]Scuola Normale Superiore, [2]Università di Bologna

federico.gatta@sns.it

AMASES 50th Annual Conference - September 03, 2026

The Economic Question

Autonomous Agents Meet Decentralized Finance. LLM-based agents are increasingly deployed in DeFi. No legal recourse: discipline comes from economic incentives.

ERC-8004 (Ethereum, August 2025) is the first standard providing a publicly verifiable trust substrate for AI agents. Agents can: • be minted; • be transferred; • provide services to clients (receiving a fee); • receive feedback.

- *Is honest behaviour always optimal, or can a sufficiently large rug pull dominate the continuation value of a reputation?*
- *What mechanisms can the protocol adopt to induce honest behaviour?*

Domain of application: General, when we have no KYC constraints and easily replaceable identities. Currently, the most explicit field is the blockchain.

Summary

- **ERC-8004: Empirical Analysis.** A first multi-chain picture of the reputation layer, on Ethereum, Base and BNB Chain.
- **Microstructural Model of the Tempted Agent.** A microstructural model with staking and protocol bans as alternative validation devices. Based on mean-field approximations.
- **Temptation Function and Policies Comparison.** Comparison of different policies based on properties of the temptation function, describing the agent's incentive to misbehave.
- **Conclusion.**

ERC-8004: Empirical Analysis

Data: Event-time, Multi-chain Samples

Event logs of the ERC-8004 contracts on **Ethereum mainnet**, **Base** and **BNB Chain**, aligned on a common window: 3 Feb – 10 Jul 2026 (≈ 157 days). Three event families: identity registrations, identity transfers, feedback events.

Network	Agents	Clients	Feedback	Evaluated	Coverage	Fb / eval. agent
Ethereum	13,630	186	2,504	1,651	12.11%	1.52
Base	58,831	11,956	366,702	29,343	49.88%	12.50
BNB Chain	199,482	104	29,507	4,336	2.17%	6.81
Aggregate	271,943	12,246	398,713	35,330	12.99%	11.29

ERC-8004 is not one homogeneous ecosystem:

- **Ethereum** — intermediate coverage, but shallow records.
- **Base** — the only mature reputation layer: broad coverage *and* deep histories.
- **BNB Chain** — mass identity creation, almost no reputation, feedback produced by only a few addresses.

Concentration of Reputation

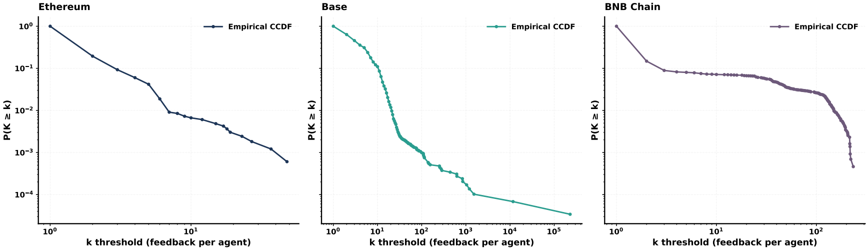


Figure 1: CCDF of feedback *received* per agent.

	Agents $K \geq 10$		Clients $K \geq 10$	
Chain	Share	Generated	Share	Received
Ethereum	0.67%	9.50%	4.84%	90.30%
Base	10.96%	80.41%	4.74%	95.60%
BNB	7.15%	84.50%	34.62%	99.64%

- **Reputation accumulation**: on Base and BNB Chain, 7–11% of agents absorb more than 80% of all evaluations.
- **Reputation production**: fewer than 5% of clients generate $> 90\%$ of feedback on Ethereum and Base.

Is Reputation Traded? Identity Transfers

Identities are ERC-721: reputation is a portable asset. We observe **8,916** transfers.

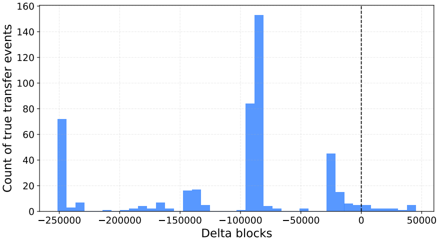


Figure 2: $\Delta_a = b_a^{\text{transfer}} - b_a^{\text{first fb}}$ (blocks).

- 94.58% occur on tokens that *never* receive feedback;
- 5.17% occur *before* the first feedback;
- only 0.25% occur at or after the first feedback.

The mass of the distribution lies in the negative region: transfers happen *before* the reputation process begins.

The reputation layer is currently generated by persistent operators, not by identities circulating across owners. A secondary market for reputation-bearing identities has not (yet) emerged.

Microstructural Model of the Tempted Agent

Reputational Capital

The state space is $\mathcal{X} = [0,1]^2$.
Discrete time: we model the state of the system as a pair $\{(F_t, M_t)\}_{t\in\mathbb{N}}$.
F_t: agent's average reputation score at time t.
M_t: maturity or depth of the agent's observable reputation record – needed as the empirical coverage is low.

Assumption (volume)

The (mean-field approximation of the) handled volume is described by a function $v : \mathcal{X} \in \mathbb{R}_+$ such that:

- $v(0, M) = v(F, 0) = 0$
- $\partial_F v(F, M),\ \partial_M v(F, M) \geq 0$
- $\partial^2_{F,F} v(F, M),\ \partial^2_{M,M} v(F, M) \leq 0$
- $\partial^2_{F,M} v(F, M) \geq 0$

Bivariate control (h_t, q_t).

- Honesty decision: $h_t \in \{0, 1\}$ (rug pull vs. honest) – empirical no identity trading.
- Effort/quality: $q_t \in [0, 1]$ (relevant only if $h_t = 1$).

Our Target

Define the continuation $\mathcal{C}$ and deviation $\mathcal{D}$ regions:

$$\begin{aligned} \mathcal{C} &= \{(F, M) \in \mathcal{X} \mid h(F, M) = 1\} \\ \mathcal{D} &= \{(F, M) \in \mathcal{X} \mid h(F, M) = 0\} \end{aligned} \tag{1}$$

Questions

- Analyze the deviation region $\mathcal{D}$.
- What protocol's design choices can reduce the deviation zone?

Ban Suspicious Agents

If an agent is suspected of committing a rug-pull, it is banned. Thus, it is forced to recreate a new identity.

Stake and Slash (S&S) Policy

The agent stakes a fraction of the entrusted volume. If a rug-pull is detected, such quantity is slashed.

Instantaneous Payoffs

Deviate ($h_t = 0$)

$$\pi^{dev}(F, M) = (\theta - \kappa^S)\, v(F, M) - \kappa^R$$

- $\theta \in (0, 1]$: extractable share of the flow entrusted;
- $\kappa^R \geq 0$: fixed cost of repair / re-entering under a new identity;
- $\kappa^S \in [0, \theta)$: slashed stake.

Continue ($h_t = 1$)

$$\pi^{cont}(F, M, q) = f\, v(F, M) - c(q) - c^o\big(\kappa^S v(F, M)\big)$$

- $f \in (0, \theta)$: fee per unit of flow;
- c: effort cost;
- c^o: opportunity cost of the stake.

Assumption (cost)

The cost $c : q \in [0, 1] \to [0, 1]$ and the opportunity cost $c^o : \kappa^S v \in [0, 1] \to [0, 1]$ are such that:

- $c(0) = c^o(0) = 0$
- $c'(q) \geq 0,\ c''(q) \leq 0$
- $\big(c^o\big)'(\kappa^S v) \in [0, f]$

Dynamics - I

Assumption (Honest-transition)

The honest transition $T : \mathcal{X} \times [0,1] \to \mathcal{X}$ is defined as

$$T(F, M, q) := \big(T_F(F, q), T_M(F, M)\big)$$

T_F, T_M depend on two parameters tuning the update speed $\lambda, \rho \in [0,1]$:

$$T_F(F, q) := (1-\lambda)\,F + \lambda\, s(q), \qquad T_M(F, M) = (1-\rho)\,M + \rho\,\sigma(F, M)$$

The feedback function $s : \mathcal{X} \to [0,1]$ satisfies

$$s'(q) > 0, \qquad s''(q) \leq 0.$$

The interaction function σ satisfies the same properties as the volume function: increasing, concave in F and M, and supermodular.

Dynamics - II

Assumption (Deviation-transition)

The deviation transition $R : \mathcal{X} \to \mathcal{X}$ is defined as

$$R(F, M, q) := \big(R_F(F), R_M(F, M)\big)$$

R is continuous and order-preserving with respect to the componentwise order on $\mathcal{X}$.

We compare two alternatives.

Same Identity

No protocol constraint $\Rightarrow$ the agent remains active under the same identity:

$$R^{same}(F, M) = \big((1-\lambda)\, F,\ (1-\rho)\, M + \rho\, \sigma(F, M)\big)$$

New Identity

Protocol ban $\Rightarrow$ the agent abandons the compromised identity.
It restarts from a fixed reset state:

$$R^{new}(F, M) = \left(F_{reset}, M_{reset}\right)$$

Value Function

Let $\beta \in [0,1]$ be the discount factor. The value function is defined as the maximum discounted payoff attainable from a given reputation state:

$$V(F,M) := \sup_{\{(h_t,q_t)\}_{t\in\mathbb{N}}} \sum_{t=0}^{\infty} \beta^t \, \pi(F_t, M_t, h_t, q_t) \tag{2}$$

$$\textit{with} \qquad (F_0, M_0) = (F, M), \qquad (F_{t+1}, M_{t+1}) = \begin{cases} T(F_t, M_t, q_t) & \textit{if } h_t = 1 \\ R(F_t, M_t) & \textit{if } h_t = 0 \end{cases}$$

For any candidate value function W, let's define

$$V_W^{dev}(F,M) = (\theta - \kappa^S)\, v(F,M) - \kappa^R + \beta\, W\big(R(F,M)\big) \qquad \text{Deviation branch}$$

$$V_W^{cont}(F,M) = f\, v(F,M) - c^o\big(\kappa^S v(F,M)\big) + \max_{q\in[0,1]} U_W(F,M;q) \qquad \text{Continuation branch}$$

$$U_W(F,M;q) = -c(q) + \beta\, W\big(T(F,M,q)\big) \qquad \text{Continuation kernel}$$

Bellman equation for V:

$$V(F,M) = \max\left\{ V_V^{dev}(F,M),\ V_V^{cont}(F,M) \right\}$$

Solvability of the Bellman Problem

Let $C_b(\mathcal{X})$ denote the Banach space of bounded continuous functions from $\mathcal{X}$ to $\mathbb{R}$, endowed with the supremum norm. For any $W \in C_b(\mathcal{X})$, define the Bellman operator:

$$\mathcal{B} : W \in C_b(\mathcal{X}) \to \mathcal{B}W \in C_b(\mathcal{X}) \quad \textit{with} \quad (\mathcal{B}W)(F, M) := \max\left\{V_W^{dev}(F, M), V_W^{cont}(F, M)\right\}$$

Then, the value function is the fixed-point of $\mathcal{B}$.

Proposition (Existence, Monotonicity, and Sign)

There exists a unique, uniformly continuous value function V satisfying the Bellman equation. Furthermore:

- V is non-decreasing in F and M;
- V is non-increasing in κ^S;
- $V(F, M) \geq 0$ for every $(F, M) \in \mathcal{X}$.

The proposition is mainly proved by means of the Banach fixed-point theorem, working on proper subsets $S \subset C_b(\mathcal{X})$, closed and invariant under $\mathcal{B}$.

Temptation Function and Policies Comparison

Temptation Function

The relative attractiveness of the deviation is summarized by the *temptation function*:

$$\Delta(F, M) := V^{dev}(F, M) - V^{cont}(F, M). \tag{3}$$

Thus,

$$\mathcal{C} = \Delta^{-1}\big((-\infty, 0]\big), \quad \mathcal{D} = \Delta^{-1}\big([0, \infty)\big), \quad \mathcal{I} = \text{ Invariance region } =: \Delta^{-1}(\{0\})$$

Proposition (Post-deviation comparison)

Let W be non-decreasing in F and M. Then, $\forall (F, M)$ such that:

$$F \geq \frac{F_{reset}}{1-\lambda} \qquad M \geq \frac{M_{reset} - \rho\,\sigma(F, M)}{1-\rho}$$

Then:

$$\Delta_W^{same}(F, M) - \Delta_W^{new}(F, M) = \beta\,\Big[W\Big(R^{same}(F, M)\Big) - W\Big(R^{new}(F, M)\Big)\Big] \geq 0$$

Temptation - Numerical Results

In the following plot, we assume:

- Cobb-Douglas volume: $v(F,M) = F^{\eta} M^{\zeta}$.
- Quadratic cost: $c(q) = \gamma q^2/2$.
- Linear opportunity cost $c^o\left(\kappa^S v(F,M)\right) = a\,\kappa^S v(F,M)$.
- Square-root feedback: $s(q) = \sqrt{q}$.

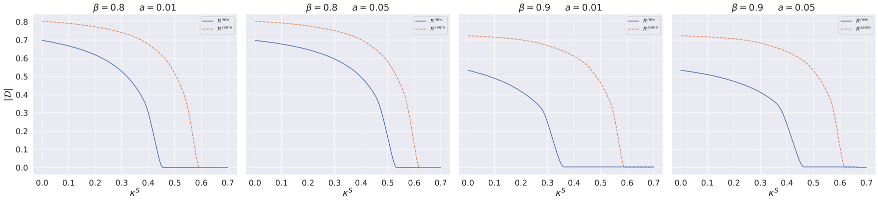


Figure 3: Deviation area (defined as the Lebesgue measure of $\mathcal{D}$) as a function of κ^S. The blue line is for the R^{new} reset; the dotted, orange line is for R^{same}. Each subplots is for a different pair (β, a). From left to right: $(0.8, 0.01)$, $(0.8, 0.05)$, $(0.9, 0.01)$, and $(0.9, 0.05)$.

Lipschitz Continuity

Proposition (Lipschitz Continuity)

Assume that v and σ are Lipschitz continuous with constants L_v and L_σ, respectively. Define $A_M := 1 - \rho + \rho L_\sigma$. Suppose that

$$\beta A_M < 1, \qquad \beta L_{R,MM} < 1, \qquad \beta L_{R,F} < 1.$$

Then the value function V is Lipschitz continuous separately in M and F. Valid Lipschitz constants are

$$L_M = L_v \max\left\{ \frac{\theta - \kappa^S}{1 - \beta\, L_{R,MM}}, \frac{f}{1 - \beta\, A_M} \right\} \tag{4}$$

$$L_F = \max\left\{ \frac{(\theta - \kappa^S)\, L_v + \beta\, L_{R,MF}\, L_M}{1 - \beta\, L_{R,F}}, \frac{f\, L_v + \beta\, \rho\, L_\sigma\, L_M}{1 - \beta\,(1 - \lambda)} \right\}$$

Monotonicity of the Temptation

Proposition (General Monotonicity)

In the same assumptions as before, for every $\bar{F}, \bar{M} > 0$, if

$$\begin{cases} \left[\theta - f - \kappa^S\right] \partial_F v(1, \bar{M}) \geq \beta\, L_F\, (1-\lambda) + \beta\, \rho\, L_M\, L_\sigma & \text{(F)} \\ \left[\theta - f - \kappa^S\right] \partial_M v(\bar{F}, 1) \geq \beta\, L_M\, (1-\rho) + \beta\, \rho\, L_M\, L_\sigma & \text{(M)} \end{cases}$$

then the temptation function $\Delta(F, M)$ is non-decreasing in the corresponding variable on $[\bar{F}, 1] \times [\bar{M}, 1]$.

Sufficient, not necessary. Kind of break-even condition:
Marginal additional profit extractable under deviation (left-hand)
Upper-bound of additional future value generated under continuation.

Small-update Regime (SUR)

Extend the value function domain to $\mathcal{X} \times [0,1]^2$, by considering 4-dimensional inputs: (F, M, λ, ρ). We focus on the small-update regime $\mathcal{X} \times [0, \bar{\lambda}] \times [0, \bar{\rho}]$.
The idea is to infer information by looking at what happens when $\lambda = 0$ and $\rho = 0$.
Main advantage: get rid of the iterative component of the value function.

Proposition (Explicit form of the value functions)

Assume $\lambda = \rho = 0$ and $R = R^{new}$. The branches of the value function read:

$$V^{dev}(F, M, 0, 0) = (\theta - \kappa^S)\, v(F, M) + \frac{1}{1-\beta} \max \left\{ \begin{cases} (\theta - \kappa^S)\, \beta v_{reset} - \kappa^R \\ \beta \left[f\, v_{reset} - c^o\left(\kappa^S\, v_{reset}\right)\right] - (1-\beta)\, \kappa^R \end{cases} \right.$$

$$V^{cont}(F, M, 0, 0) = \frac{1}{1-\beta} \left[f\, v(F, M) - c^o\left(\kappa^S\, v(F, M)\right)\right]$$

SUR - Monotonicity Condition

Proposition (Small-update regime monotonicity)

Assume that one of the following conditions hold:

$$(\theta - \kappa^S)\,(1-\beta) - f \geq \xi \qquad \textbf{Increasing}$$
$$(\theta - \kappa^S)\,(1-\beta) - f + \kappa^S \leq -\xi \qquad \textbf{Decreasing}$$

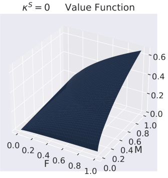
for some $\xi > 0$. With linear opportunity cost $c^o(v) = a\,v$, it is sharpened:

$$(\theta - \kappa^S)\,(1-\beta) - f + a\,\kappa^S \gtreqless \pm\xi$$

Then, $\forall \bar{\delta}, \bar{F}, \bar{M} \in (0,1)$, there exists a small-update regime $[0, \bar{\lambda}) \times [0, \bar{\rho})$ such that $\forall F_2 \in [\bar{F}, 1]$, $\forall M_2 \in [\bar{M}, 1]$, $\forall F_1 \in [0, F_2\,\bar{\delta}]$, and $\forall M_1 \in [0, M_2\,\bar{\delta}]$, it holds:

$$\Delta(F_1, M_1, \lambda, \rho) \leq \Delta(F_2, M_2, \lambda, \rho) \qquad \textbf{Increasing} \qquad (5)$$
$$\Delta(F_1, M_1, \lambda, \rho) \geq \Delta(F_2, M_2, \lambda, \rho) \qquad \textbf{Decreasing}$$

SUR - Numerical Results



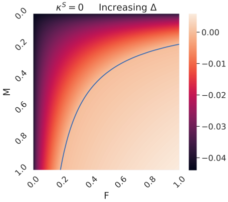


Figure 4: Small-update regime – : $\lambda = \rho = 0.001$. The two plots on the left correspond to stake amount $\kappa^S = 0$; the right plots are for $\kappa^S = 0.3$. $3D$ plots: value function $V(F, M)$. Heatmaps: $\Delta(F, M)$. The blue line in the heatmap corresponds to $\Delta(F, M) = 0$.

SUR - Numerical Results

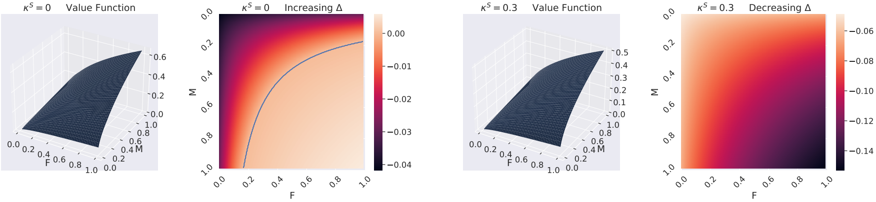


Figure 4: Small-update regime – : $\lambda = \rho = 0.001$. The two plots on the left correspond to stake amount $\kappa^S = 0$; the right plots are for $\kappa^S = 0.3$. $3D$ plots: value function $V(F, M)$. Heatmaps: $\Delta(F, M)$. The blue line in the heatmap corresponds to $\Delta(F, M) = 0$.

Optimal Effort - Comparative Statics

Finally, we study the impact of the reputation F, maturity M, and stake amount κ^S on the optimal effort provided by the agent.

Assumption (Continuity Region)

Let the continuation set equal the whole state space: $\mathcal{C} = \mathcal{X}$. Thus, $V = V^{cont}$. The maturity update σ is independent of F. The opportunity cost is linear.

Proposition (Optimal Effort - Comparative Statics)

The optimal effort adopted by the agent, $q^*(F, M; \kappa^S) = argmax_{q\in[0,1]} U(F, M, q)$, is such that:

- $q^*(F, M; \kappa^S)$ is non-increasing in F.
- $q^*(F, M; \kappa^S)$ is non-decreasing in M.
- $q^*(F, M; \kappa^S)$ is non-increasing in κ^S.

The result is proven thanks to technical results on the concavity of V and U. The main theoretical tool is the supermodularity theory by Topkis.

Conclusion

Conclusions

- **Empirically**, the ERC-8004 reputation layer is registration-heavy and reputation-poor: coverage is sparse and uneven across chains, and feedback production is concentrated in a handful of clients.

- **Theoretically**, the answer to our question *"is the reputation enough to guarantee correct behavior"* is "not always". Discipline is strengthened by high reset costs, strong slashing, costly rebuilding of maturity, and high demand sensitivity to reputational damage.

- **Policy.** Reputation is necessary but not sufficient. The policy design problem is a trade-off: high staking restores correct behavior at the cost of lower effort.

Thanks for your attention!

Tempting the Agent

Out on SSRN:

papers.ssrn.com/sol3/papers.cfm?abstract_id=7401058