# Kleidi: An Incentive-Aligned Fraud Prevention Network

Elliot Friedman

May 6, 2025

## *Abstract*

Infrastructure providers across the cryptocurrency ecosystem lack direct economic incentives to prevent fraudulent transactions. This paper proposes Kleidi, a decentralized Fraud Prevention Network for account abstraction wallets. Validators stake collateral, earn rewards for blocking malicious transactions and incur penalties for reporting false positives. This design aligns infrastructure providers' economic incentives with user security outcomes and has the potential to reduce losses from hacks and scams.

# Introduction

Statistics from the U.S. Federal Reserve shows merchants and credit card issuers split 80% of fraud losses in 2021[1]. This creates incentives for both of these parties to mitigate losses. Fraudulent credit card charges can be and often are reversed if a consumer disputes a payment. Contrast this with cryptocurrency transactions, which are final, and have no such parties with incentive structures to protect users. Wallets, node providers, and validators earn fees regardless of a transaction's legitimacy. Account abstraction allows wallets to become programmable, which is a significant leap in UX and security[2]. We propose a decentralized Fraud Prevention Network for these wallets, where staked validators earn rewards for blocking attacks in real time and incur penalties for false positives. This aligns a new piece of infrastructure with users by adding a party to transactions with incentives to protect them.

# The Incentive Gap

Cryptocurrency transactions are push-only and final[3]. If a hacker tricks a user into authorizing a transfer, there is no equivalent of a chargeback. Importantly, no intermediary suffers financially when a theft occurs. Consider the following infrastructure providers and their incentives in a typical transaction:

| Infrastructure Provider | Incentive |
|---|---|
| Software Wallets | Secure key management in browser, transaction broadcasting and basic scam avoidance |
| Hardware Wallets | Secure key management and signing |
| RPC Providers | Provide high uptime access to the chain |
| Blockchain Miners and Validators | Order and execute transactions, receive transaction fees |
| Transaction Simulators and Analyzers | Paid per use, provide accurate simulation and analysis |
| Blockchain Data Providers | Provide accurate real time information on known threat actors |

Each provider secures their own piece of the transaction supply chain, but the *user's intent* is never validated. No actor is both empowered and rewarded to prevent a suspicious transaction from executing. Users losing funds are often the least equipped to detect the sophisticated methods that cause these losses, while the actors best positioned to detect fraud, the infrastructure providers, often lack an ability to do so. Kleidi introduces a new participant in the transaction supply chain whose *sole role* is protecting users.

## A Fraud Prevention Network for Account Abstraction Wallets

Kleidi introduces a new layer of security by economically incentivizing specialized validators to protect users from fraudulent transactions. Transactions are optimistically approved, meaning users can still broadcast transactions and quickly opt out if needed.

### User Payments

Users opt into the network by pre-approving payment of fees that are automatically charged when fraudulent transactions are blocked. This ensures validators are compensated when they save funds.

### Incentive Structure

Validators who correctly identify and flag a fraudulent transaction receive a financial reward. Rewards are sourced from a percentage of the funds saved by preventing fraud, which users pay. Conversely, if a validator flags a legitimate transaction as malicious, they are penalized by losing a portion of their stake. This financial disincentive is key in preventing network abuse and ensuring validators strive for high accuracy. Validators are motivated to develop and deploy sophisticated fraud detection methods, ranging from algorithmic analysis and AI-driven behavior modeling to leveraging known malicious address databases. Those with superior detection capabilities are likely to profit, while less effective validators will face losses. This competitive market dynamic naturally selects and rewards the most effective performers, regardless of their detection methodology.

## Reputation

Validator reputation allows the network to adaptively prune and respond to its participants. Reputation is publicly available and shows a validator's tally of correctly and incorrectly reported incidents as well as instances where they acted maliciously. Incidents where malicious validator behavior is discovered by the review council results in an increase to their malicious incidents score.

Reputation can be taken into account by the review council when determining slashing events. Higher reputation actors may receive lower slashing amounts than lower reputation peers. This works to incentivize long term behavior from validators.

## Slashing

Validators stake capital, in the form of restaked BTC or ETH [12]. Entering the network requires a minimum bond of *$1,000,000* worth of BTC or ETH. Validators are allowed to stake over the minimum amount, increasing their share of rewards and capital at risk. Governance can adjust these values over time to ensure the ability for new nodes to enter the network.

With this capital at stake, validators can be slashed for incorrectly flagging transactions as malicious. Slashing is performed automatically if a validator flags a transaction and the network does not reach consensus on that transaction being malicious. This slashed capital is then locked into a dispute window, during which the validator can appeal for an additional bond if the categorization is incorrect. If the categorization is found to be correct, the validator loses their dispute bond and slashed amount. A portion of the slashed amount will go towards the aggrieved user, another portion to the review council, and a final amount towards the insurance fund. However, if the categorization is found to be honest by the review council, just not reaching consensus, the slashed amount will be refunded. In this case, the review council's fee will be paid from the insurance fund.

Security guarantees of the network rely on three baseline premises: (i) a majority of staked validators acting honestly, (ii) the bonded review council rendering impartial rulings, and (iii) TDX enclaves correctly simulating and scrubbing pre-execution traces without leaking data [11]. Economic incentives and slashing mechanisms reinforce each other.

## Insurance Fund

The insurance fund is capitalized from a fixed percentage of slashed collateral seized from validators and serves to underwrite dispute‑related council fees. When a user or validator challenges a ruling and the council ultimately decides in the challenger's favor, and no other party is unambiguously liable, the fund covers the cost of the review bond and associated fees. By shouldering these expenses only in cases where the outcome vindicates the challenger, the insurance fund ensures equal access to dispute resolution without exposing honest participants to prohibitive costs.

## Network Entry Point Nodes

Network Entry Point Nodes will start out as a trusted entity, without defining consensus rules for how these are elected and churn in or out. Further into the network's life, consensus rules will be defined for how these entities enter and exit the network beyond the original permissioned set.

## Validator Entry and Exit

As validators enter and exit the network, their weighted stake changes along an exponential curve. This disincentivizes validator churn as the majority of their capital remains idle throughout this process.

## Transaction Review Quorum

All transactions that are blocked must reach an internal quorum on the network. A percentage of the staked capital must agree that a transaction should be blocked. Currently this is undefined as further economic modeling will be needed to find correct values and ranges. This parameter can be adjusted by a governance proposal to lower or raise the threshold depending on network performance. Each validator that signals a transaction veto will sign a signature of the transaction hash to block. Once a veto reaches quorum, the transaction to block will bundle up these signatures and submit them on chain.

## Review Council

Over the course of the network, disputes will arise. The review council will judge challenges from both validators and users. All reviews will require both a bond and fee from the challenging party. The fee is distributed among the council members for their efforts reviewing, and the bond is refunded if the challenging party prevails. This imposes a large cost on a challenger griefing or submitting a challenge if they do not prevail.

# Prior Art and Comparison

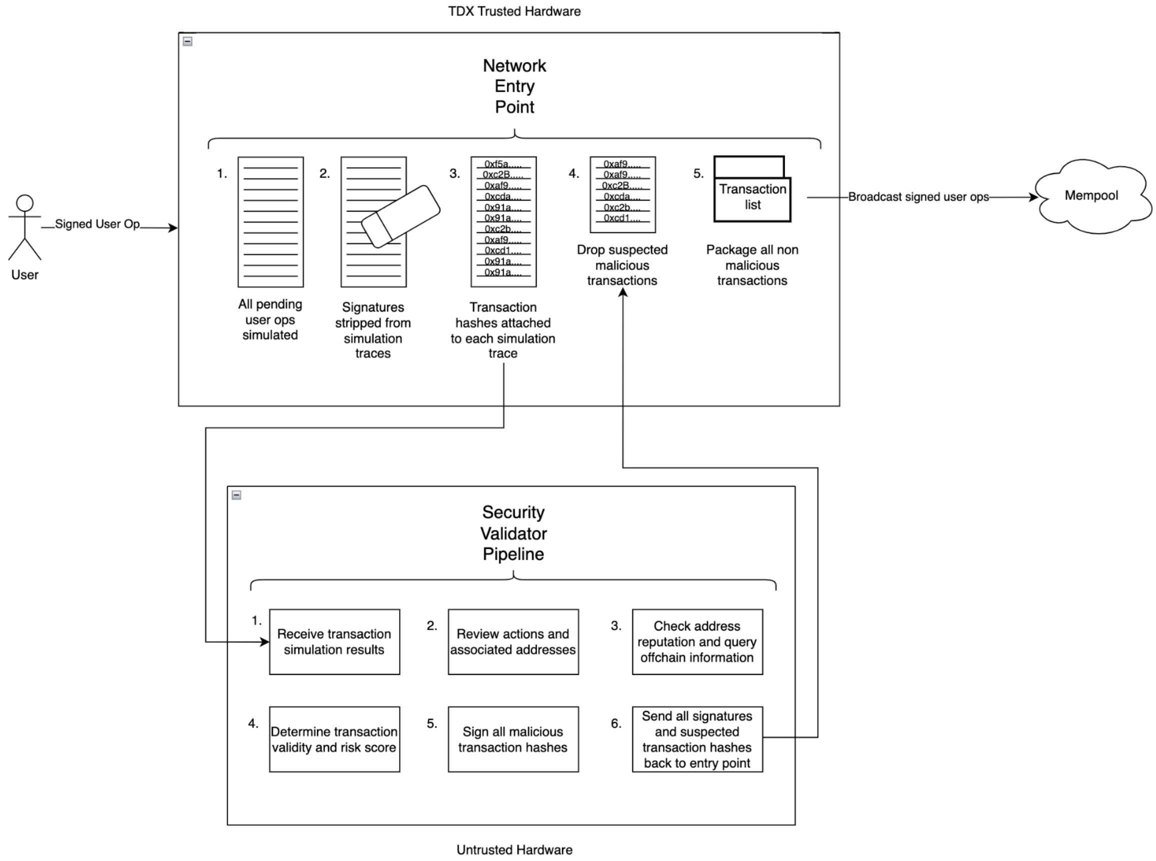
The concept of an incentive-aligned security layer draws inspiration from several existing projects. Each project implements a piece required to make Kleidi work in its own way. Kerberus, Scam Sniffer and Revoke Cash all actively review transaction calldata before signing.

| Project | Mechanism | Limitation |
|---|---|---|
| Lossless Protocol [4] | Token-level freeze-and-reverse mechanism; reporters stake LSS to flag suspicious transfers | Requires token-level opt-in; operates post-attack with 24-hour freeze period |
| Forta network [5] | Decentralized detection bots scan mempool; emit alerts; bot | Focused on protocol security, not end users. Only provides warnings without enforcement; |

| | | |
|---|---|---|
| | authors earn fees from paid feeds | no direct reward/penalty per alert |
| Kleros [6] | Staked jurors vote on disputes; rewards majority voters; slashes dissenters | Multi-day judging process too slow for stopping live attacks |
| Sentinel Protocol [7] | Crowdsources threat intel; community reports scam addresses; experts verify and publish to database | No slashing for bad reports; users must manually check database |
| Kerberus [8],<br>Scam Sniffer [9],<br>Revoke Cash [10] | Reviews transaction calldata, page contents and site URL | Requires users trust the extension to read and write all browser pages |

Kleidi combines and ties together the best of these approaches, transaction reviews, economic incentives, real-time monitoring, and dispute resolution into a unified service.

# Example Transaction Flow


TDX Trusted Hardware
Network Entry Point
User
Signed User Op
1.
All pending user ops simulated
2.
Signatures stripped from simulation traces
3.
0xf5a.....
0xc2B.....
0xaf9.....
0xcda....
0x91a....
0x91a....
0xc2b....
0xaf9.....
0xcd1....
0x91a....
0x91a....
Transaction hashes attached to each simulation trace
4.
0xaf9.....
0xaf9.....
0xc2B.....
0xcda....
0xc2b....
0xcd1....
Drop suspected malicious transactions
5.
Transaction list
Package all non malicious transactions
Broadcast signed user ops
Mempool
Security Validator Pipeline
1.
Receive transaction simulation results
2.
Review actions and associated addresses
3.
Check address reputation and query offchain information
4.
Determine transaction validity and risk score
5.
Sign all malicious transaction hashes
6.
Send all signatures and suspected transaction hashes back to entry point
Untrusted Hardware

Users send their signed user op into the network entry point. This entry point is running trusted hardware, obfuscating the transaction contents from the operator. The following phases occur inside the network:

1. **Simulation:** All pending user ops are simulated inside a TDX trusted hardware entry point. This produces a detailed stack trace for each transaction.
2. **Anonymization**: All signatures and cryptographic material are stripped from simulation traces to prevent a validator from broadcasting a malicious transaction. All amounts and token identifiers from trades are stripped to prevent front-running.
3. **Transaction Bundling and Association**: Transactions are combined together, with transaction hashes associated with their scrubbed stack traces. Transactions are then broadcast to the validators.
4. **Validator Review**: Validators now analyze the transactions, searching for fraudulent transactions or behavior. Any transactions the validator deems fraudulent are signed, signaling their intent to block the transaction.
5. **Validator Broadcast to Entry Point**: Validators send only their signed results to the entry point.

6. **Entry Point Review**: Signed validator transactions are reviewed by the network entry point. Transactions that reach quorum from validators are excluded from the bundle.
7. **Fee Settlement**: Users with dropped transactions pay fees for their blocked transactions. Users without dropped transactions do not receive payments

### Collusion and Sybil Attacks

If enough nodes on the network are compromised, they can ignore a malicious transaction they created or have knowledge of, resulting in slashing for the honest validators. The review council can overturn the result of a slashing, refunding the amount lost to the honest validator.

### Frontrunning Opportunities

Validators have access to transactions before they reach the chain, however they do not have the ability to order transactions. Their ability to see sensitive trade details is mitigated as token amounts are scrubbed by the network entry point. Validators may be able to guess token amounts based on the user sending the transaction within their TDX machine. At the start of the network, validator entry will be permissioned and require legal agreements to bind validators from attempting front running.

### Censorship and Denial of Service

Malicious validators could collude to censor a particular user’s transactions. However, penalties for malicious censorship are high, costing a large amount of the validator’s stake. This allows for compensation for the affected user if censored. Economic damages can be repaid by slashing stakes of the offending party, with the proceeds going to the recipient of the damages.

## Conclusion

A significant gap exists between sophisticated offchain fraud intelligence and onchain transaction execution. Many parties possess advanced threat detection systems, machine learning models, and comprehensive databases with relationships between fund movements and known threat actors. However, these actors currently lack both the ability and economic incentive to intervene on users' behalf during live transactions. This disconnect results in preventable losses as valuable threat intelligence remains disconnected from end users. Kleidi bridges these two worlds by creating the first real-time fraud prevention network. This network changes fraud from a costly externality to a core economic component in the way transactions are reviewed and blocks are built.

# References

[1] **Board of Governors of the Federal Reserve System.** "2021 Interchange Fee and Payment Card Statistics." https://www.federalreserve.gov/paymentsystems/2021-Interchange-Fee.htm (accessed 6 May 2025).

[2] **Crypto for Innovation.** "What Is Account Abstraction and Why Is It Important?" https://cryptoforinnovation.org/what-is-account-abstraction-and-why-is-it-important/ (accessed 6 May 2025).

[3] **Investopedia.** "Blockchain." https://www.investopedia.com/terms/b/blockchain.asp (accessed 6 May 2025).

[4] **Lossless Team.** *Lossless Protocol Whitepaper.* https://lossless.io/whitepaper.pdf (accessed 6 May 2025).

[5] **Forta Foundation.** "Forta Network Documentation." https://docs.forta.network/ (accessed 6 May 2025).

[6] **Kleros Cooperative.** *Kleros: Short Paper v2.* https://kleros.io/static/whitepaper_en-8bd3a0480b45c39899787e17049ded26.pdf (accessed 6 May 2025).

[7] **Uppsala Security.** *Sentinel Protocol Whitepaper.* https://sentinelprotocol.github.io/whitepaper/ (accessed 6 May 2025).

[8] **Kerberus Labs.** "Kerberus Wallet Security Extension." https://kerberus.app/ (accessed 6 May 2025).

[9] **Scam Sniffer.** "Scam Sniffer Documentation." https://scamsniffer.io/docs (accessed 6 May 2025).

[10] **Revoke.cash.** "Revoke.cash – Open-Source Repository." https://github.com/RevokeCash/revoke.cash (accessed 6 May 2025).

[11] **Intel Corporation.** "Intel® Trust Domain Extensions (TDX) Technology Overview." https://www.intel.com/content/www/us/en/architecture-and-technology/tdx.html (accessed 6 May 2025).

[12] **EigenLayer.** "Restaking: Extending Cryptoeconomic Security." https://docs.eigenlayer.xyz/ (accessed 6 May 2025).