Kleidi: An Incentive-Aligned Fraud Prevention Network for Account Abstraction Wallets

Elliot Friedman
May 6, 2025

Abstract

Infrastructure providers across the cryptocurrency ecosystem lack direct economic incentives to prevent fraudulent transactions, leaving users exposed. 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

Credit-card networks split over 80 percent of fraud losses between card issuers and merchants, creating incentives for both parties to stop this fraud. These fraudulent charges can even be reversed. Cryptocurrency transactions are final, and no such incentive structures to protect users exist in crypto. Wallets, nodes providers, and validators earn fees regardless of a transaction's legitimacy. Users rely on passive warnings while losses to hacks and scams reached $2.3 billion last year. Account abstraction allows wallets to become programmable, which is a significant leap in UX and security. We propose a decentralized Fraud Prevention Network (FPN) for these wallets, in which staked validators earn rewards for blocking attacks in real time and incur penalties for false alarms. This system aligns a new key piece of infrastructure with that of the users by creating incentives to protect them.

The Incentive Gap in Crypto vs. Traditional Finance

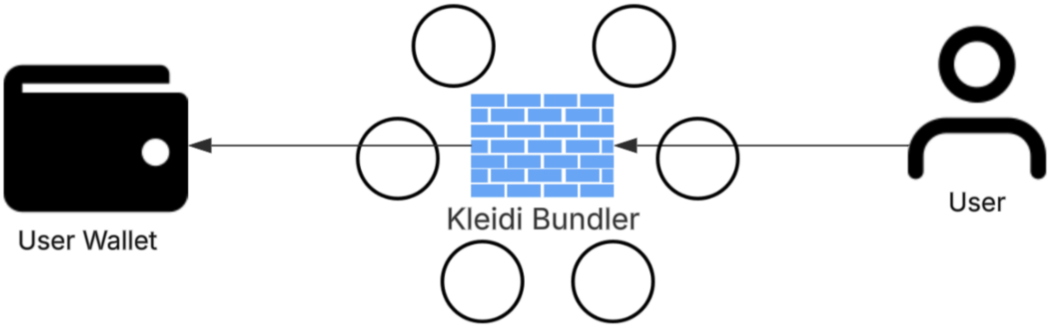
Fraud Prevention in Traditional Finance: Legacy payment systems have strong incentives that shield consumers from fraud risks and motivate intermediaries to prevent it. In credit card networks, if an unauthorized charge occurs, the **liability falls on different parties** depending on the circumstances. For card-present transactions (in-person with chip cards), the card issuer typically bears the loss if fraud occurs, as long as the merchant followed security protocols. This gives banks and issuers a strong incentive to maintain sophisticated fraud detection algorithms to block stolen-card transactions in real time. For online card payments (card-not-present), the liability shifts to merchants. They are usually required to refund customers for fraudulent transactions, meaning the merchant directly loses revenue and their product or service. Consequently, **merchants invest in fraud prevention tools** such as address verification, 3-D Secure verification, and anti-fraud software. This helps avoid chargebacks and losses. The card networks enforce rules (like the EMV chip liability shift) to ensure *someone* in the chain is always accountable and therefore motivated to stop fraud. The result is a layered defense: issuers monitoring spending patterns, networks flagging unusual activity, and merchants screening suspicious orders. Users benefit from this alignment as most fraudulent charges are caught and blocked or reimbursed.

This alignment is fundamentally driven by financial incentives. Statistics from the U.S. Federal Reserve show that in recent years, merchants and issuers have together absorbed roughly 80–90% of fraud losses on debit card transactions, while cardholders themselves absorbed only the remaining 10-20%. In 2019, merchants bore about 55.5% of fraudulent losses and issuers 36.3%, but by 2021 merchants' share had fallen to 47.0% as issuers took on 33.5%. This cost-sharing directly incentivizes both parties to reduce fraud rates. Each fraudulent transaction hurts the bottom line of an issuing bank or a merchant, creating a *business case* for investing in better fraud detection. Over decades, this incentive alignment has led to significant investment in fraud prevention software. The key insight is: **when stakeholders have money at risk, their self interest is to protect users**.

In crypto, transactions are push-only and final. If a hacker tricks a user into authorizing a token transfer, there is no equivalent of a chargeback. The tokens are gone. Importantly, *no intermediary loses money when an individual user is hacked*. Consider the following infrastructure providers and their incentives in a typical cryptocurrency 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 simulation and analysis

Each provider secures their own piece of the transaction supply chain, but the *user's intent* falls through the cracks. No actor is both empowered and rewarded to prevent a suspicious transaction from executing. The users losing money are often the least equipped to detect the sophisticated social engineering and hacking methods that cause these losses, while the actors best positioned to detect fraud, the infrastructure providers, often lack an ability to do so. This is the gap Kleidi aims to fill. It introduces a new participant in the transaction supply chain whose *sole role* is to protect users. This is achieved by providing security experts incentives to prevent this fraud in real time.



An Incentive-Aligned Fraud Prevention Network for Account Abstraction Wallets

This section outlines Kleidi, an FPN designed to operate within the transaction pipeline of account abstraction (AA) wallets. Kleidi introduces a new layer of security by economically incentivizing specialized validators to protect users from fraudulent transactions.

Network Overview

Kleidi functions as a proactive watchdog layer, composed of independent validators who stake collateral. Their mandate is to monitor and analyze pending transactions from participating wallets before these transactions are finalized on-chain. If a transaction is identified as potentially fraudulent or unauthorized, validators can flag it, initiating a process to potentially halt its execution. This system is built on a core incentive mechanism where validators who correctly identify and contribute to stopping a malicious transaction earn rewards. Conversely, validators who incorrectly flag a legitimate transaction as malicious, known as a false positive, incur penalties through the slashing of their staked collateral. This design ensures accurate fraud detection is profitable, while imprecise or malicious flagging is costly. Specifics of arbitration for flagged transactions and the detailed slashing mechanisms are further elaborated in the "Slashing" and "Example Transaction Flows" sections.

Core Operating Principles

Kleidi's operation is guided by key principles that are deeply integrated with existing account abstraction architecture. Firstly, the network employs an optimistic approval model, meaning transactions are assumed to be legitimate and are processed unless explicitly flagged by the network. When a transaction is flagged, its on-chain processing is temporarily paused pending further review or consensus. Secondly, Kleidi utilizes an opt-in integration with account abstraction. Users voluntarily choose to route their transactions through Kleidi for screening. This leverages AA's programmable nature, allowing user operations to be intercepted and reviewed by specialized bundlers that query the FPN before block inclusion. If Kleidi flags an operation, the bundler can prevent its inclusion, effectively canceling the transaction. The technical integration, including the use of functions like `validateUserOp`, underpins this interaction.

Economic Incentive Structure

The efficacy of Kleidi hinges on its economic model, which is designed to make user protection a rational, self-interested act for validators. A foundational aspect of this model is the requirement for validators to deposit significant collateral to participate, an act that gives skin in the game and provides the collateral for potential penalties. For their active participation, validators who correctly identify and flag a fraudulent transaction receive a financial reward. These rewards are sourced from a percentage of the funds saved by preventing the 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 crucial for preventing network abuse and ensuring that validators strive for high accuracy; the process for imposing these penalties and distributing any slashed funds is covered in the "Slashing" section. This comprehensive incentive structure fosters a competitive environment. Validators are motivated to develop and deploy sophisticated fraud detection methods, which may range 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 or reckless validators will face losses, naturally refining the network's overall efficacy over time.

Benefits of Incentive Alignment

By integrating these economic drivers, Kleidi aims to achieve several key benefits for the ecosystem. Primarily, it establishes, perhaps for the first time in a widespread crypto context, a scenario where a key infrastructure participant—the Kleidi validator—has direct financial incentives aligned with preventing user loss from fraud. A validator's profitability becomes directly tied to their success in fraud prevention and their ability to minimize false positives. It is important to note that validators are not penalized for failing to detect a malicious transaction (a false negative), which focuses their efforts on interventions where they have high confidence. Furthermore, the system enhances decentralized trust, as users do not need to rely on a single, centralized entity for this protection. The decentralized nature of the FPN, combined with the requirement for consensus among validators (or a committee review for disputes) and the ever-present risk of slashing for incorrect actions, mitigates concerns about arbitrary censorship or abuse. Finally, the network effectively creates a form of crowdsourced security, functioning as a decentralized, real-time security audit for transactions. It becomes an open marketplace where diverse security expertise and detection strategies can be deployed and competitively validated, thereby enhancing the security posture for all participants. This approach transforms fraud prevention from an externality into a core economic function of the network.

Network Effects

The network is chain-agnostic. It will start protecting Ethereum wallets, and then move to similar smart contract wallets on other chains or rollups. This allows cross-platform scaling of the security layer. The incentive model creates a flywheel while the system is growing. As more wallets join, more transactions flow through, providing more potential rewards for catching fraud. This increased revenue and attention attracts more security experts and researchers to run validators. This positive feedback loop improves security for *all* users without requiring each wallet team separately implementing their own fraud system. Users and projects that value security, e.g. DAO treasuries, institutional users, or retail users holding large balances, would opt in, effectively funding the network when they pay bounties.

Reputation

Validator reputation allows the network to adaptively prune and respond to its participants. Reputation would be measured by the ratio between correctly and incorrectly reported incidents. The formula for scoring would be simple, correctly reported incidents add one to a validator’s score, and incorrectly reported incidents subtract one. Incidents where malicious validator behavior is discovered by the council after a user challenges a ruling can result in multi-point score subtractions, or total erasure of reputation, depending on severity.

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. Entering the network requires a minimum bond of *$1,000,000* worth of the BTC or ETH. Validators are allowed to stake over the minimum amount, increasing their skin in the game. 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.

Insurance Fund

The insurance fund is capitalized from a fixed percentage of slashed collateral 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 equitable access to dispute resolution without exposing honest participants to prohibitive costs. As Kleidi scales, continuous inflows from penalties naturally grow the fund, reinforcing the network's resilience and fairness in adjudicating edge-case disputes.

Validator Hardware

Validators will have access to sensitive and potentially dangerous signed transactions before they arrive on chain. As such, before these validators join the network, they will have to prove they are running their node and detection systems inside of a TDX machine with a trusted program so they cannot broadcast malicious transactions. This keeps the network safe from malicious users, and allows users to trust their transactions will not be sent if they are malicious. TDX is more memory efficient and easier to write programs for than SGX[1], hence the choice of hardware.

Validator Entry and Exit

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

[1] https://www.canarybit.eu/intel-sgx-vs-tdx-what-is-the-difference/

Transaction Review Quorum

All transactions that are blocked must reach an internal quorum on the network, where a certain percentage of the staked capital agrees that a transaction should be blocked. Currently this is intentionally left undefined as further modeling will be needed to find correct values and ranges. This parameter, the percentage of staked capital that must agree to block, can be adjusted by a governance proposal to lower or raise this 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.

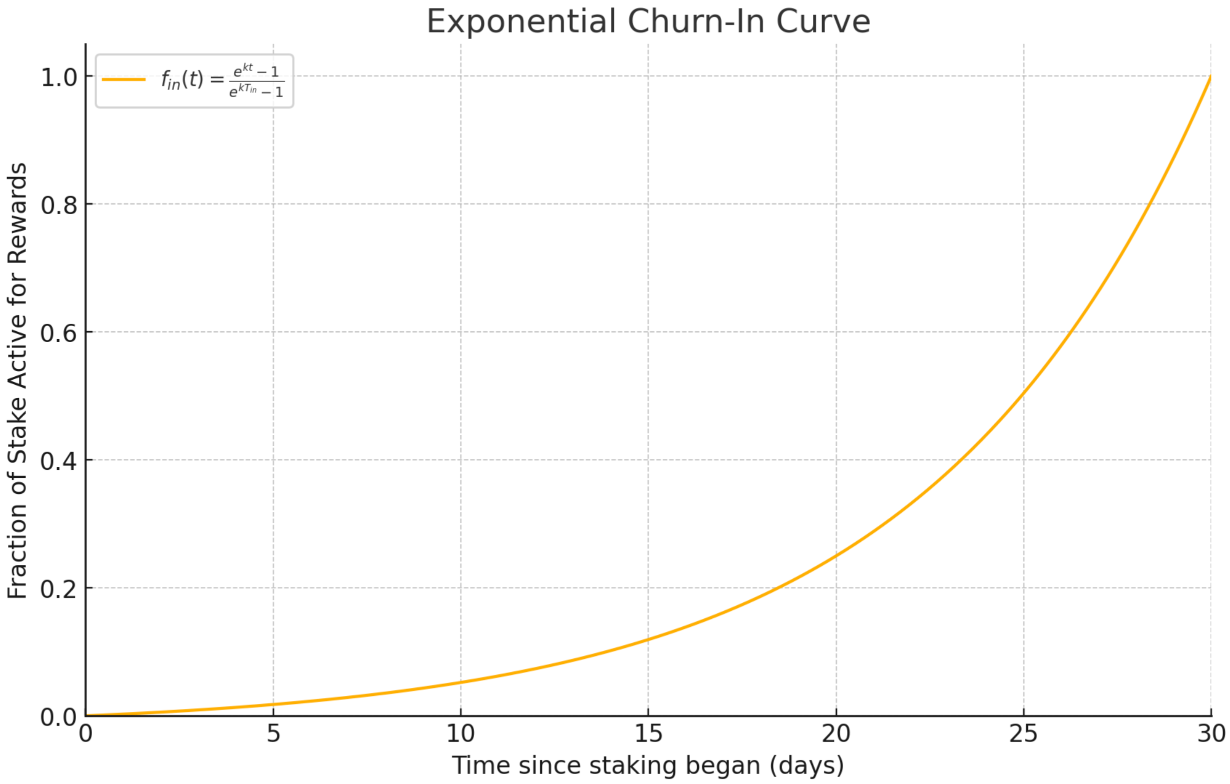
Review Council

Over the course of the network being used, disputes will invariably 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 time reviewing, and the bond is refunded if the challenging party prevails. This aligns incentives and imposes a large cost on a challenger griefing or submitting a challenge if they are in the wrong.

Users could attempt to game the system by submitting a transaction they know will be flagged, i.e. sending funds to North Korea. This transaction will be rightfully blocked by the validators. The user then challenges, paying the fee and posting a bond, claiming this was their intent. This case would go to the council, who would look at the evidence from validators, in this case sending funds to a known hacker address. The council would rule that a user probably did not want to be doing this, force the user to pay, and adjudicate the case in the validator's favor. Users may try this because they get paid once they prove the validator incorrectly blocked their transaction.

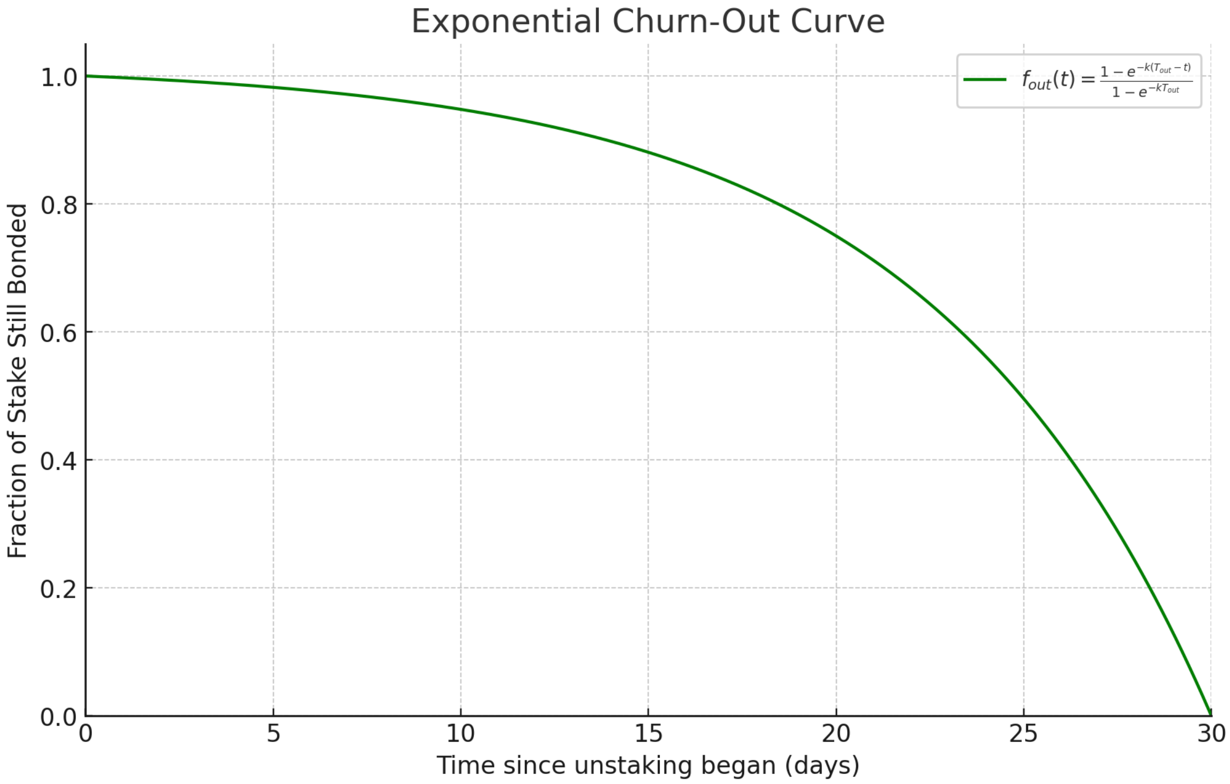
Entry

When a validator enters the network, their staked weight becomes available to them fully after 30 days. During the thirty days, their weight increases to their staked amount.



Exit

When a validator exits the network, their staked weight becomes unbonded to them fully after 30 days. During the thirty days, their weight slowly decreases, preventing a quick exit. This optimizes the network for long term minded validators who will work diligently to increase their reputation and work on behalf of users.



Economic Viability at Scale

Layer two networks process substantial transaction volumes, with Base processing over 6 million daily transactions. This economy provides significant economic opportunity for fraud prevention services. Conservative projections assuming Kleidi captures 5% of these transactions demonstrate strong validator economics.

Funds Saved	Stage	Monthly TX Count	Fraud Rate	Detection Rate	False Positive Rate	Monthly Revenue	Monthly Penalties	Monthly Profit	Annual ROI
$1.6m	Alpha	500,000	.05%	80%	.002%	$10,000	$10,000	$0	0%
$4.2m	Beta	1,000,000	.05%	85%	.001%	$42,500	$10,000	$27,500	33%
$16m	Growth	2,500,000	.06%	87%	.001%	$157,140	$25,000	$125,000	100%
$54m	Expansion	5,000,000	.08%	90%	.0008%	$540,000	$40,000	$450,000	250%

$200m	Scale	10,000,000	0.1%	92%	.0005%	$1.84m	$50,000	$1.5m	~430%

The model shows monthly profits of $27,500+ per validator at modest scale and penetration rates. As the network expands and increases market penetration, validator economics improve dramatically. This creates strong incentives for professional security operators to participate in the network.

Prior Art and Comparison

The concept of an incentive-aligned security layer draws inspiration from several prior efforts in the blockchain space. Here we review four relevant projects, Lossless, Forta, Kleros, and Sentinel Protocol and contrast them with our proposed network:

Project	Mechanism	Limitation
Lossless Protocol	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	Decentralized detection bots scan mempool; emit alerts; bot authors earn fees from paid feeds	Only provides warnings without enforcement; no direct reward/penalty per alert
Kleros	Staked jurors vote on disputes; rewards majority voters; slashes dissenters	Multi-day judging process too slow for stopping live attacks
Sentinel Protocol	Crowdsources threat intel; community reports scam addresses; experts verify and publish to database	No slashing for bad reports; users must manually check database

Lossless allows freezing and intervening in transactions, rewarding those who identify and stop theft. However, it requires projects to use their token standard[2], and imposes delays on transfer, which presents UX challenges. This approach partially breaks composability of the token with DeFi and other external systems. Forta[3] and Sentinel[4] monitor threats in real time, just mostly for projects and not end users. Kleros shows the power of economically motivated courts to adjudicate disputes using a decentralized group of judges and juries[5]. Kleidi is the result of a synthesis of ideas from across these projects. On-chain incentives are combined with off-chain monitoring systems that act proactively on user's behalf. Disputes are handled by decentralized councils that review transactions. Previous solutions addressed individual components of this problem, but did not combine them together. This network is informed by prior art but distinct in that it positions fraud prevention as a generic, chain-wide *service with aligned incentives*, rather than an app-specific security add-on.

Example Transaction Flows

The following two scenarios illustrate how the protocol works. The first is a **successful fraud detection and cancellation** event, showing the network protecting a user in real time. The second is a **false positive case** where a transaction is wrongly flagged and how the system corrects itself and penalizes the mistake.

Scenario 1: Fraudulent Transaction Detected and Stopped

[2] https://docs.lossless.io/protocol/technical-reference/lerc20
[3] https://docs.forta.network/en/latest/scam-detector-bot/
[4] https://www.uppsalasecurity.com/sentinelprotocol/
[5] https://kleros.io/whitepaper.pdf

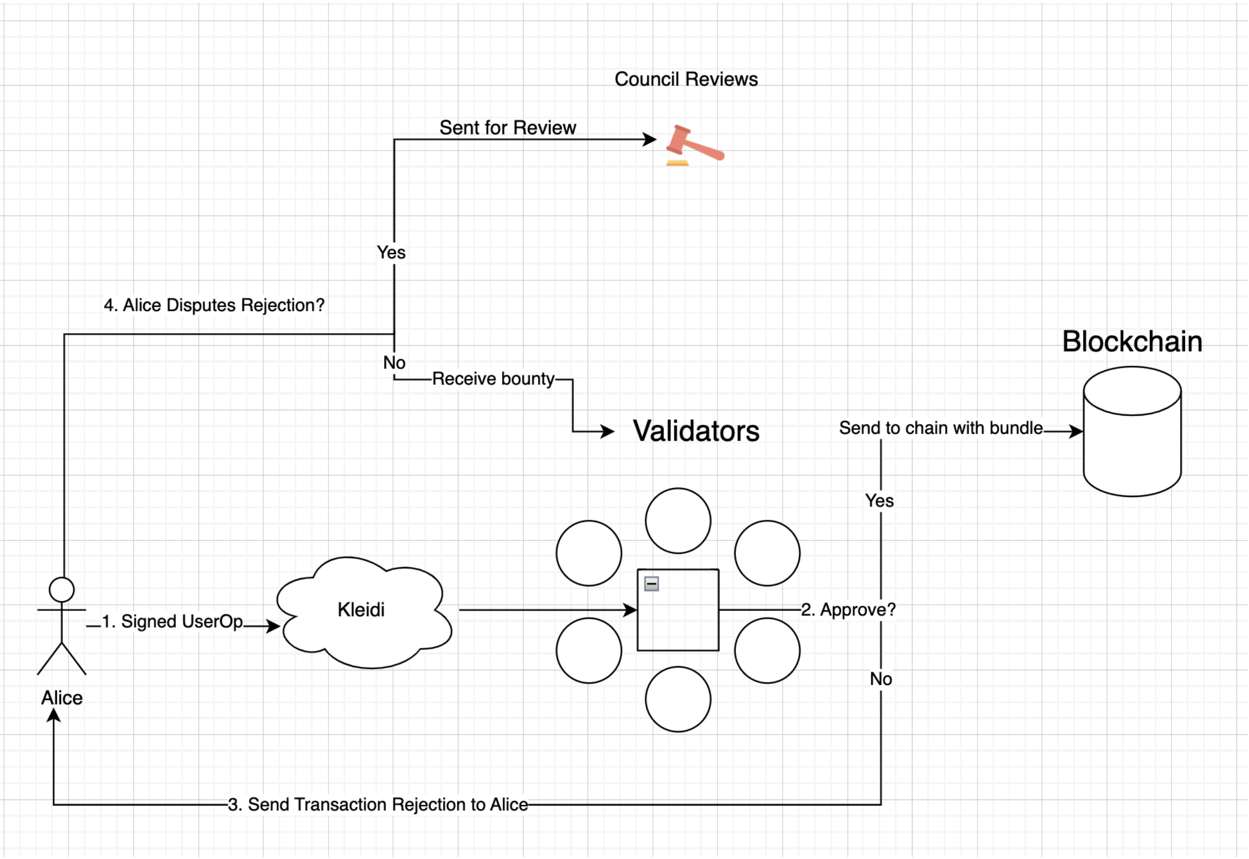


Figure 1: Simplified flow of a fraudulent transaction being intercepted by Kleidi. In this scenario, a user's wallet is about to execute a malicious transaction, but the network identifies the threat and aborts the transaction before it reaches the blockchain, thereby protecting the user's funds and rewarding the validators who intervened.

1. **User Initiates Transaction:** Alice is using an account abstraction wallet to manage her crypto. One day, she interacts with what she believes is a legitimate dApp. However, this is a phishing site, and it crafts a transaction that would transfer all of Alice's tokens to the attacker's address. Upon generation of a signed UserOp, the operation is routed through Kleidi for validator review and inspection rather than immediate broadcast.

2. **Network Analysis and Flagging:** FPN validators receive the pending user operation and simulate or analyze it. They notice several red flags: the recipient address has been reported in Sentinel Protocol's database as a known scam address, and Forta's bots label the transaction as a "likely phishing scam" because it attempts to transfer an unusually large amount of Alice's assets to a single new, unverified contract. Validator Bob, who runs a Kleidi node, immediately **flags the transaction as fraudulent**. In practice, this means Bob's node submits a transaction to a set of contracts on-chain, indicating "Transaction X from Alice looks malicious; recommend rejection." Other validators see the same data. Several other validators independently reach the same conclusion and cast concurring votes to flag the transaction. Once the threshold is met,

the transaction is officially marked as **fraudulent** by the network. At this point, the network instructs the bundler **not to include the transaction** in a block. The wallet itself is designed to wait for an “all clear” from the network, and will not proceed. The transaction is **aborted before execution**. Alice’s funds never leave her wallet.

3. **Outcome and Rewards:** With the transaction stopped, Alice’s wallet notifies her that the transaction was blocked due to suspected fraud. She receives details: “The destination address is flagged as a known scam address. Your transaction was canceled”. The transaction is halted before execution and validators that flagged the malicious transaction receive a reward. Alice pays 1% of the funds saved to Kleidi. Bob and the other flagging validators receive a proportional payout from the reward Alice paid. Their stakes are also safe and not slashed since this was a valid flag. Alice continues using her wallet, now with heightened awareness, and the theft is stopped, at least this time. End-to-end processing completes within three seconds. In this instance, the additional latency is justified by the funds preserved.

4. **Dispute**: Alice has a short window of time after her transaction is blocked and before the reward is distributed to open a dispute on the legitimacy of the block. If disputed, she will have to post a bond as collateral, and a council of humans will review the details of her case. If they side with her that the transaction was legitimate, her bond and reward will be refunded, and if there were damages, it could lead to slashing and reputational damage to both Bob and the validators.

This scenario demonstrates the ideal operation: the user is protected by the collective vigilance of the network. The validators had a clear profit motive to act, they earned a reward, and Alice didn’t have to rely on altruism or after-the-fact intervention.

Scenario 2: False Positive and Validator Slashing

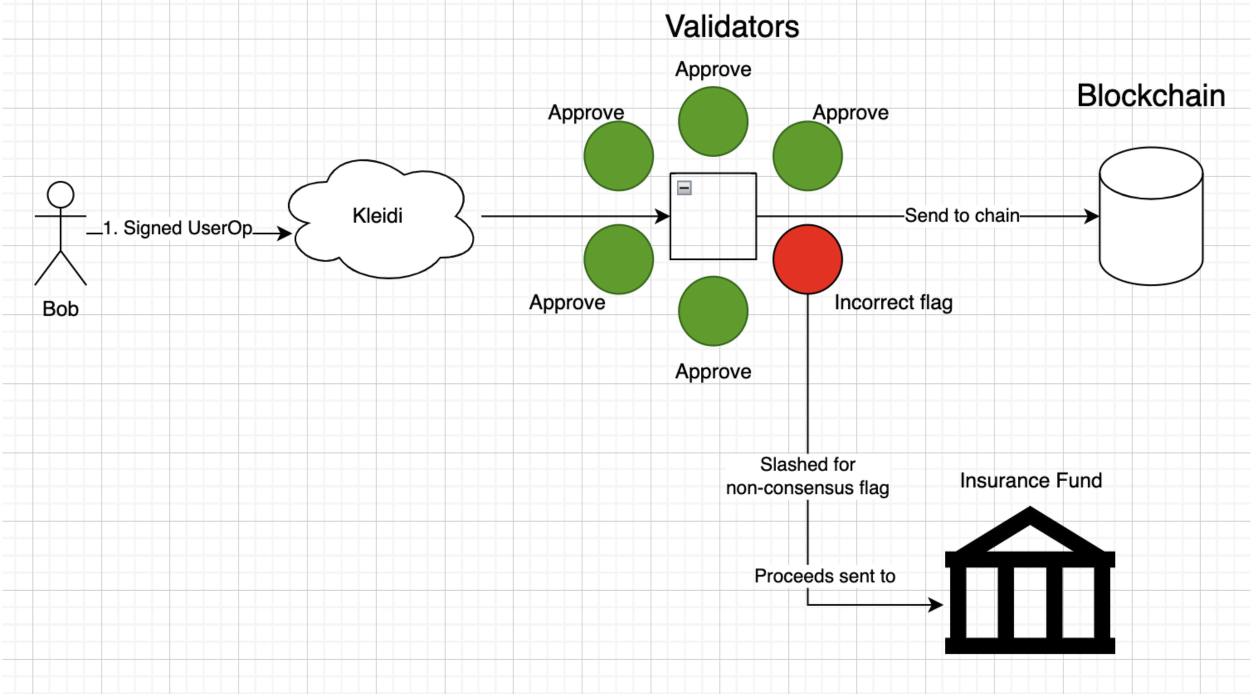


Figure 2: Flow of a false-positive flag. Here a validator wrongly flags a legitimate transaction as fraudulent. All validators review the alert, determine it was an error, and the transaction proceeds. The validator who raised the false alarm is slashed, discouraging incorrect claims.

1. **User Initiates Legitimate Transaction:** Bob, a wallet user is sending 1 ETH to his friend for payment, using his account abstraction wallet with FPN protection. This is a routine transfer to an address Bob has not sent to before. The transaction goes to the FPN for screening as usual.

2. **Validator Flags Suspicion:** Validator Carol sees Bob's transaction. Carol has an aggressive detection algorithm that erroneously marks any first-time outgoing transfer above 0.5 ETH as suspicious, mistaking it for a common scam. Carol flags Bob's transaction as fraud. However, other validators checking the transaction find no red flags: the recipient address is fresh but not on any blacklist, the amount is not large relative to Bob's history, and Bob's wallet behavior doesn't mimic known attack patterns. Carol's flag is a **false positive**.

3. **Slashing:** The protocol immediately slashes Carol, sending $1,000 of her staked assets to the network's insurance fund and subtracting one from her reputation for incorrectly flagging a transaction. The slashed amount is small, reflecting the fact that no economic damage occurred. Had others agreed with Carol, blocking the transaction, and the user disputed the validity of their hold, the damages awarded by a council could have been larger.

4. **Slash Dispute:** Now, Carol faces a decision: she can contest the slash if she believes the other validators were in error and she actually flagged a suspicious transaction. However, disputing is costly, and requires an additional bond to be placed and a council to review. If the council decides Carol acted incorrectly, her bond will be forfeit.

This scenario demonstrates how the network disincentivizes unnecessary interference. By combining automated checks with a human governance layer, the system avoids binary decision making and allows for human judgement when cases are not clear. Slashing not only compensates parties affected by delays, it also acts as a deterrent to prevent malicious actors from abusing the network to censor transactions. An attacker, for instance, would find it expensive to bribe or run validators to falsely flag transactions of others, because each false flag would cost stake and only slow the targeted party down.

Network Security

This network will have extensive checks and balances to mitigate its power over users. Implemented correctly, users will have their transactions checked before inclusion.

Collusion and Sybil Attacks

If enough nodes on the network become compromised, they can ignore a malicious transaction they created, 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 have no ability to order transactions, as the bundle will be passed to Flashbots or another block builder. This would give them the ability to extract MEV by front or back running users. All a validator can do is stop a transaction as their entire simulation and system of checks must occur within their TDX machine.

Censorship and Denial of Service Concerns

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. Additionally, when a user joins the network, they set a veto count. The veto count is the number of times the network can block transactions from their account before they have to re-authorize more blocking transactions. If validators collude over time to continually block transactions from a given user, they will incur large penalties and only succeed in blocking transactions for a short period of time, until the block authorizations run out. In this scenario, the user would contest the blocks in the security council, the council would review the evidence, siding with the user, then proceed to slash the malicious validators involved. This slashed stake would then partially go to the affected user as compensation for the inconvenience and potential funds lost.

Trust Assumptions

Kleidi's security properties rely on specific trust assumptions that differ from traditional cryptocurrency transactions. First, the validator network is assumed to be honest in the majority, with economic incentives designed to reinforce honest behavior. Second, the review council is assumed to provide fair arbitration in disputed cases. Third, TDX secure enclaves are assumed to correctly isolate transaction processing from potential validator manipulation.

These trust assumptions are mitigated through the protocol's decentralized validation, economic penalties for misbehavior, and reputation systems. The security guarantees provided by Kleidi are necessarily different from those of pure self-custody solutions, representing an alternative point in the security-usability spectrum.

Conclusion

By integrating economically motivated validators into the transaction supply chain, user protection becomes aligned with infrastructure provider’s incentives. This stands in contrast to the current status quo, where users are largely on their own and losses to hacks and scams have become a large risk of participating in crypto. Harnessing account abstraction wallets as the integration point, our approach shows how smart contract wallets can be actively safeguarded through programmable consensus with an external network.

Kleidi’s key innovation is its **incentive structure**: rewarding fraud interdiction and penalizing incorrect interference. This forms a mechanism that naturally aligns interests across both users and validators. The system leverages competitive forces among specialized validators, whose diverse detection methodologies collectively enhance the security of the network.

This network aligns economic incentives with user safety, leveraging the strengths of decentralization, global participation and crowdsourcing of detection, while mitigating the lack of built-in consumer protection. As crypto matures, this network could contribute to enhanced protection and user confidence. Empowering validators to help secure users represents a step towards combining the freedoms of cryptocurrency with some of the safeguards of traditional finance.