Open Source, Open Season: Why AI Has Broken DeFi's Risk Premium

---

DeFi's foundational premise was that smart contract risk was knowable and manageable — that auditable, deterministic code created a risk profile investors could estimate, price, and hold against a yield. That premise was already thin before 2026. AI has broken it. The adversarial infrastructure required to exploit every deployed contract on every major blockchain, at scale and near-zero cost, has been in place for years. It was waiting for a capability upgrade. That upgrade has arrived.

In 2020, Paradigm researchers Dan Robinson and Georgios Konstantopoulos published "Ethereum is a Dark Forest," describing an on-chain environment where generalized frontrunning bots monitor every pending transaction in the mempool and automatically copy any profitable move — replacing addresses, executing the same call, and extracting value before the original transaction lands. Their closing observation: "the future is only going to get scarier." They were describing a world of automated bots operating on deterministic rules. The AI escalation is not a different threat. It is the next layer of the same one. The yield premium over traditional finance is thinner than most participants acknowledge. And the risk side of the equation has been permanently altered by a development the market has not yet priced: AI has fundamentally changed who can find exploits, how fast, and at what cost.

This is not an argument that DeFi is worthless or that every protocol will be drained. It is a more specific claim: the risk-adjusted return on DeFi capital deployment and DeFi token ownership is worse than it appears, the yield premium does not currently compensate for a new and unquantifiable attack variable, and the market is pricing DeFi as if AI capability is static. It is not.

The Yield Premium Was Already Thin

Before introducing AI, the baseline case for DeFi capital deployment was already weaker than its proponents admitted.

US Treasuries currently yield approximately 4–4.5%. This is the true risk-free rate against which DeFi yields should be measured. Supplying USDC to Aave on Ethereum in normal market conditions generates approximately 3–4% APY. In periods of low utilisation — which describe the majority of calendar time — the DeFi yield is at or below the Treasury yield. The premium over risk-free is near zero, and sometimes negative, while the risk being taken is considerably higher.

DeFi capital deployment carries a risk stack that has no equivalent in Treasury ownership: smart contract vulnerability (code that may contain bugs that cannot be patched post-deployment), oracle manipulation risk (price feeds that can be manipulated to trigger incorrect liquidations or borrows), governance risk (token holders who can vote to change protocol parameters, including upgrading contracts in ways that harm depositors), stablecoin peg risk (the asset you're supplying may de-peg, as USDC briefly did in March 2023), and liquidity risk (in stress conditions, withdrawal queues form and capital is locked). Each of these is a non-trivial risk. In combination, they represent a risk profile that would demand a substantial premium over risk-free in any rational risk-adjusted framework.

Yields do spike — Aave USDC supply hit approximately 9–10% on Ethereum during periods of peak borrowing demand in mid-2026. But these periods are episodic, unpredictable, and brief. A depositor cannot reliably earn the spike yield; they earn the time-weighted average, which in practice sits close to or below Treasuries for much of the year. The existence of spike yields does not solve the risk-premium problem; it shows that compensation is episodic while exploit risk is continuous. The case for simple DeFi capital deployment over Treasuries, on a risk-adjusted basis, was already marginal before introducing AI.

The hack data makes the risk side of this calculation concrete. In the first five months of 2026, DeFi protocols lost over $840 million to exploits — a 70% year-over-year increase in losses over the same period in 2025, with more than 50 incidents recorded. April 2026 was especially severe, with over $600 million drained in a single month. The frequency of attacks has increased 68% year-over-year. These numbers describe the environment *before* AI-augmented exploit capability reaches general availability.

AI Adds an Unquantifiable Variable to an Already Thin Premium

The standard framing for DeFi risk was that smart contract risk could be bounded and estimated. Audits exist. Historical hack rates give a base rate. TVL at risk is public. An informed depositor could form a view on expected annual loss probability and compare it against yield.

AI collapses that framework.

Historically, the constraint on exploit discovery was human expertise and time. Finding a critical vulnerability in a complex protocol required skilled security researchers spending days or weeks reviewing code, simulating edge cases, and modelling economic attack paths. That constraint is gone. Large language models trained on code can audit smart contracts at a scale and speed no human team can match. The cost of scanning a protocol's entire codebase for vulnerability classes drops toward zero. The time required to identify novel attack paths across composable protocol interactions — the kind of multi-step exploits that caused the largest recent hacks — compresses from weeks to hours or minutes.

This creates a new attack surface dynamic that the DeFi risk premium has not been recalibrated to reflect. Before AI, you could estimate the population of humans capable of finding and executing sophisticated exploits and form a rough view on attack probability. That population is now effectively unbounded. Anyone with access to frontier code models and economic motivation can attempt exploit discovery at scale. The base rate of sophisticated attack attempts increases. The probability that any given critical vulnerability is found and exploited before it can be patched — if patching were even possible, which it usually is not — increases with it.

The evidence that this transition is already underway is now quantified. In 2025, Anthropic published the Smart Contract Exploitation benchmark (SCONE-bench) — a dataset of 405 real-world smart contracts successfully exploited between 2020 and 2025. Frontier models at the time successfully exploited over half of them, generating $550 million in simulated revenue. The 2026 follow-on evaluation used a harder set of contracts exploited after the models' knowledge cutoff dates; on that test, Claude Mythos Preview led materially over all other evaluated models. The evidence points in the same direction across both evaluations: earlier frontier models already exploited a large share of historical DeFi targets, while the 2026 follow-on showed Mythos materially ahead of the next-closest model. Simulated exploit revenue has been doubling every 0.7 months — faster than the 1.1-month doubling time recorded in the prior benchmark cohort — and Anthropic explicitly stated they have not

yet reached a plateau. Their conclusion is direct: “the knowledge and expertise required to develop exploits will drop significantly as Mythos-level capabilities become more widely available,” with Mythos-level models expected to reach general availability within 6–12 months, making this class of exploit capability “increasingly commoditized.”

This is not speculative framing from outside observers. It is Anthropic’s own assessment of their own models’ capabilities, published alongside a decision to withhold Claude Mythos Preview from general release specifically because of its exploit development abilities — deploying it instead through a controlled programme, Project Glasswing, precisely because its capabilities in this domain were judged too dangerous for general access. And eventually, models at this capability level will reach general availability.

The evidence that AI is reaching below the smart contract layer arrived the same month. Security researcher Taylor Hornby, using Anthropic’s Claude Opus 4.8, discovered a critical vulnerability in Zcash’s Orchard privacy pool — a flaw that had existed undetected for four years despite human auditing. The flaw — a circuit constraint written too loosely, causing the proof engine to accept fraudulent transactions as legitimate — would have allowed an attacker to create unlimited counterfeit ZEC within the shielded transaction system with no on-chain trace. An emergency hard fork patched it before exploitation. When the vulnerability was publicly disclosed, ZEC fell approximately 45% in a single day. Whether the flaw had actually been exploited was, and remains, structurally unknowable: a separate supply-accounting mechanism could confirm the total ZEC supply had not inflated, but Orchard’s privacy properties mean there is no way to cryptographically prove or disprove that an individual forged transaction occurred within the shielded pool before the patch. The market sold off into that uncertainty, not despite it.

Two things about this incident matter for the risk premium argument. First, the vulnerability was not a smart contract bug — it was a flaw in the underlying cryptographic circuit governing Zcash’s privacy layer. AI is not merely finding application-layer code bugs; it is capable of identifying errors in the mathematical foundations that protocols are built on, at a depth and speed that four years of human review missed. Second, the model that found it is publicly accessible. The same capability that Taylor Hornby used for responsible disclosure is available to any actor with different intentions.

Critically, this is a variable whose magnitude is unknown, non-static, and already trending in one direction. Exploit capability on SCONE-bench has been doubling every 0.7 months — faster than the previous 1.1-month doubling time — and Anthropic explicitly states they have not reached a plateau. A capability doubling every three weeks, with no ceiling in sight, cannot be incorporated into a yield premium. The traditional DeFi risk framework asks: what is the probability that a sophisticated attacker finds a critical vulnerability this year? AI turns that question into a moving target. The answer changes every month, and it changes upward. A yield calibrated against last year’s attack probability is not calibrated against next month’s. The market is treating AI exploit risk as background noise. It should be treating it as a structural repricing event.

The Attacker Only Needs to Be Right Once

The fundamental asymmetry in DeFi security is not new, but AI makes it more severe.

An attacker needs to find one exploitable vulnerability in one protocol at one point in time. The reward for success is immediate access to some or all of the protocol’s TVL — potentially hundreds of millions of dollars, transferred in a single transaction, with no recourse

mechanism. The attacker's cost is the time and compute required to find the vulnerability, execute the transaction, and launder the proceeds. AI has dramatically reduced the first of those costs.

The defender — the protocol team, the auditors, the community — faces the inverse problem. They must maintain perfect security across the entire codebase, indefinitely, against an attack population that is growing in capability. One mistake, at any point in time, in any part of the code, is sufficient for total loss. The cost of maintaining that security — audit fees, bug bounties, operational overhead, ongoing monitoring — is paid out of protocol fee revenue, which is a small percentage of TVL. As TVL grows, the attack prize scales linearly while the defence budget scales only with fees, which grow slowly and incompletely relative to TVL.

The Euler case makes the asymmetry concrete. Six independent auditors reviewed the protocol. The vulnerable function passed all of them. Sherlock, a smart contract insurance and audit platform, paid a $4.5 million claim for the miss. The attacker needed to find one flaw that six professional teams missed; the defenders needed to find every flaw before the attacker found any one of them. AI shifts this ratio further: it increases the number of effective searchers on the attack side without meaningfully constraining the defender's requirement for completeness.

This means the economic incentive structure is structurally inverted. Consider a protocol with $500 million TVL. At typical utilisation rates, the protocol earns $1–3 million annually in protocol fees — from which it must fund audits, bug bounties, ongoing monitoring, and security infrastructure. The attacker's prize for finding a single critical flaw is up to $500 million, realised in one transaction, with no counterparty and no recourse. The ratio is not close. AI widens it further: it drives the attacker's cost of discovery toward zero while leaving the defender's requirement unchanged — be right across the entire codebase, simultaneously, forever. The winning move for the attacker is one error. The winning condition for the defender requires none.

It is worth noting that AI can also assist defenders — automated auditing tools, formal verification, real-time anomaly detection. The relevant question is not whether AI helps defenders. It does. The question is whether it changes the attacker/defender balance. In DeFi specifically, it does not change it in the defender's favour. In traditional finance, code is closed-source, which limits the attacker's surface area. In DeFi, all protocol code is public and permanently available for inspection. The defender cannot obscure the attack surface. AI is equally accessible to both sides, but the attacker's problem — find one flaw — is structurally easier than the defender's problem — have no flaws — and AI's ability to exhaustively scan a known codebase benefits the attacker more.

The 2026 hack data introduces a further dimension: the majority of DeFi losses this year came not from smart contract bugs but from key theft, credential compromise, and social engineering. The two largest single incidents — Drift Protocol ($285M, April 2026) and KelpDAO ($292M) — involved no code vulnerabilities at all. Drift was preceded by a six-month social engineering campaign by Lazarus Group operatives who patiently built relationships with team members before extracting privileged access credentials. KelpDAO was drained through compromised RPC nodes feeding false data to a bridge running a single-verifier configuration. This might seem to cut against the code-vulnerability argument made above. It does not — it broadens it. AI augments these vectors too — AI-generated deepfakes lower the cost of impersonation, AI-powered phishing personalises social engineering at scale, and AI can automate the reconnaissance required to identify which team members hold privileged access. The argument is not solely that AI will find more smart contract bugs. It is that AI expands the effective attack surface across every vector simultaneously: code, credentials,

and human operators alike.

Composability Is a Systemic Risk Amplifier

DeFi's core design principle — that protocols can compose with each other, building more complex financial products from simpler building blocks — is also its primary systemic risk vector, and AI makes this worse in a specific way.

A single-protocol exploit has bounded damage. If Aave is exploited, Aave depositors lose funds. If Compound is exploited, Compound depositors lose funds. The damage stays within the exploited system.

DeFi composability breaks this containment. Protocols share oracle feeds, liquidity pools, collateral assets, and liquidation mechanisms. An exploit in Protocol A creates price dislocations that trigger forced liquidations in Protocol B. Liquidity withdrawals from Protocol A reduce pool depth in Protocol C, which relies on shared liquidity. Bad debt in Protocol A, if collateralised by an asset that Protocols D and E accept, creates cascading solvency questions across the ecosystem.

The Euler Finance hack in March 2023 demonstrated this precisely. Approximately $197 million was drained from Euler through a flaw in a little-used donateToReserves function — a function that had passed review by six separate auditors, including Sherlock, which subsequently paid a $4.5 million insurance claim for missing it. But the damage did not stop at Euler's contracts. Angle lost over $17 million in agEUR collateral and Balancer lost $11.9 million; several others, including Temple DAO, Idle, Yield Protocol, Yearn, and Inverse, were also affected. None of these protocols had any vulnerability in their own code. All of them took losses because they held positions in or integrated with Euler. The composability that made DeFi useful made Euler's failure everyone's problem.

AI extends this risk in a specific direction. Multi-step, cross-protocol exploits — where an attacker borrows from Protocol A to manipulate prices in Protocol B to trigger a liquidation that drains Protocol C — require sophisticated economic modelling to design and execute. This is precisely the category of reasoning where AI excels. The most complex and damaging exploit architectures, which previously required rare combinations of smart contract expertise and quantitative finance knowledge, become accessible to a broader attacker population. The containment assumption in DeFi risk assessment — that your exposure is limited to the protocols you interact with — is already false in practice, and AI makes the cross-protocol attack surface larger and more accessible.

The Attack Surface Doesn't Disappear When Protocols Do

There is a dimension of DeFi's security problem that receives almost no attention in risk frameworks: deployed code is permanent. When a protocol upgrades, deprecates, or shuts down, its old smart contracts remain on-chain indefinitely, immutably, often still holding user funds who never migrated. They cannot be taken offline. They cannot be patched. And they are still publicly readable by anyone — including AI systems scanning for vulnerabilities.

In June 2026, Aztec Connect — a privacy bridge that had been officially deprecated and shut down in March 2023 — was exploited for $2.1 million. The root cause was a mismatch between the contract's verification and settlement logic: the proof verification function checked

only part of the submitted proof data, leaving the parameters governing token transfers unverified. An attacker found this mismatch three years after the protocol was abandoned and drained approximately 909 ETH, 270,000 DAI, and 167 wstETH in a single transaction. Three days later, a second attacker exploited a different deprecated Aztec contract through the same fundamental class of flaw, taking another $2 million. Over $4 million was extracted from contracts that had been considered dead for three years.

Aztec Labs confirmed that the exploited contracts had no connection to the current Aztec Network. That distinction, while legally and reputationally accurate, is cold comfort for the users whose funds were held in contracts the team could no longer protect and could not patch.

This is the legacy code graveyard problem. DeFi's history of deployment has created an ever-growing inventory of unmaintained, immutable contracts still holding real value. Human auditors are resourced to review new deployments — not to continuously re-audit years of accumulated legacy code. AI has no such constraint. It can scan every contract ever deployed on every public blockchain simultaneously, at trivial cost, for any known or novel vulnerability class. The relevant attack surface is not "currently maintained DeFi protocols." It is the entire history of deployed code on every chain, most of which is no longer being watched by anyone.

The attack surface extends further still — below the contract level to the compilers and language implementations protocols are built on. In July 2023, $69 million was drained from Curve Finance and four protocols integrated with it through a bug not in Curve's contract logic but in the Vyper programming language itself. A storage slot misalignment in Vyper versions 0.2.15 through 0.3.0 disabled the reentrancy guard — a bug that had existed since 2021 and was patched accidentally in version 0.3.1 with no protective action taken. Contracts written in the vulnerable versions, however, remained permanently deployed and permanently vulnerable. The affected protocols — JPEG'd, Alchemix, Metronome, Curve — had written correct code. The vulnerability was in the tool they wrote it with. AI can scan compiler source code with the same efficiency it applies to contract code. The attack surface in DeFi is not a layer. It is a stack, all the way down.

Two Expressions, Different Risk Profiles

It is important to be precise about what this argument applies to, because DeFi risk manifests differently depending on how you are expressing exposure.

Supplying capital to a DeFi protocol — depositing USDC into a lending market, providing liquidity to a DEX — gives you direct smart contract hack risk on your principal, plus the systemic contagion risk described above, offset by yield. The yield provides some compensation. If yields were to reprice upward significantly — if depositors began demanding 8–10% as a baseline to compensate for AI-augmented attack risk — capital deployment could become justifiable again on a risk-adjusted basis. This is a self-correcting dynamic: rational capital suppliers will eventually demand more yield per unit of hack risk, and protocols that cannot offer it will lose TVL.

Owning a DeFi governance token — being long a lending protocol, DEX, or yield aggregator token — is a materially worse expression of the same risk with less compensation. The token holder takes hack risk (a major exploit destroys token price regardless of whether you personally had capital in the protocol), thin value accrual risk (protocol fee revenue is a small fraction of TVL and governance tokens rarely have hard claims on that revenue), and

TVL compression risk (as rational capital reassesses AI-augmented attack probability, TVL outflows reduce fee revenue, which reduces any fundamental case for the token). These three negative exposures compound, and they trigger simultaneously from the same event. A major exploit collapses token price on news, accelerates TVL outflows as capital flees, and reduces fee revenue in lockstep with TVL — all three strike in the same direction at the same time. There is no self-correction mechanism for token holders equivalent to the yield repricing available to capital suppliers. Depositors can demand more yield. Token holders have no equivalent lever.

The spread between these two expressions widens as AI risk scales. At low AI risk, both are marginal but defensible. At high AI risk, supplying capital to simple, battle-tested protocols with adequate yield might remain defensible; owning governance tokens does not.

The Market Has Not Priced This

DeFi governance token valuations still embed assumptions built in a world of static, human-bounded attack capability. The price-to-fee multiples on major DeFi protocols, the TVL trajectories being extrapolated forward, the narrative frameworks around fee switches and protocol revenue — none of these systematically incorporate an accelerating and open-ended attack variable.

The evidence of mispricing is visible in the disconnect between what DeFi protocols are losing to exploits and how governance tokens are trading through those losses. In the first five months of 2026, DeFi protocols lost over $840 million to exploits — a 70% year-over-year increase. Major DeFi governance tokens largely absorbed these losses as sector-level noise, continuing to trade at multiples of annual protocol revenue that imply TVL growth and stable fee economics. A valuation framework that treats $840 million in losses across five months as noise is not a framework that has priced an accelerating attack variable.

The reason is partially psychological and partially structural. Psychologically, the market has processed AI as a general positive for crypto — AI agents will use DeFi, AI will drive demand for decentralised compute and storage, AI narratives pump token prices. The attack-surface implications of the same AI capability expansion are processed separately, if at all. Structurally, there is no clean mechanism for the market to price an unknown variable. You cannot put a number on “how much better at exploit discovery will AI be in 18 months” and discount it into a token price. So the market ignores it.

There is a third reason, distinct from psychology and structure: markets reprice around discrete, attributable events, not diffuse statistical trends. Subprime mortgage risk had been visible in the data since 2007, but it took Bear Stearns — a major Wall Street institution, not a subprime specialty lender — collapsing in March 2008 for the market to treat the risk as systemic rather than contained. The two largest 2026 DeFi losses, Drift and KelpDAO, were attributed to social engineering and a known state actor, not to an AI model finding a code vulnerability. Until a single event makes that connection unambiguous, the risk stays diffuse, and diffuse risk is what markets are worst at pricing before it crystallises.

This is where the mispricing lives. Not in assets where the AI risk is known and priced, but in assets where the risk is real, material, and not yet legible in the frameworks the market is currently using to value them.

What This Means for Capital Allocation

The argument leads to a clear hierarchy by risk-adjusted defensibility. Simple, overcollateralised, battle-tested capital deployment — lending markets with long track records and minimal external dependencies — could remain defensible if and only if yields reprice upward significantly; at 3–4% against 4.5% Treasuries, they do not pass the test today. Complex, multi-protocol compositions — yield aggregators, leveraged strategies, protocols built on protocols — multiply attack surface without commensurate yield compensation. And DeFi governance tokens, at current valuations, are the weakest expression of all: they embed TVL growth assumptions and fee revenue extrapolations that are negatively correlated with the direction AI exploit risk is moving, with no self-correction mechanism when the thesis deteriorates.

The conclusion is not that DeFi dies. It is that different DeFi exposures should carry very different required returns. The market is pricing DeFi capital deployment and DeFi token ownership as similar risk expressions when they are structurally different, and that both are being valued as if AI is a neutral variable when it is a negative one for the security side of the equation. The yield premium required to justify DeFi capital deployment is higher than current rates imply, and the token premium required to justify DeFi governance token ownership is lower than current multiples reflect.

Until DeFi yields reprice upward to compensate for the new attack surface — or until AI defensive tooling demonstrably closes the attacker/defender gap — the rational framework is to demand significantly higher yield premiums for capital deployment, and to treat DeFi governance token multiples with considerable scepticism.

In 2020, Robinson and Konstantopoulos warned that the dark forest was only going to get scarier. They were right.

---

written June 2026