

Article

Mythos-Class AI and Blockchain Systemic Risk: A Comparative Analysis of Bitcoin and Ethereum/L2 Architectures

Robert Campbell

Independent Researcher, Upper Marlboro, MD 20774, USA; rc@medcybersecurity.com

Abstract

Cryptocurrency market infrastructure—public blockchains and cross-chain bridges supporting tens of billions in liquidity—is monitored as a systemic-risk surface by the Financial Stability Board and equivalent bodies, with defensive posture calibrated against human-level adversaries. Anthropic’s April 2026 release of Claude Mythos Preview has prompted institutional response across financial regulation but no blockchain-specific analytical framework. This paper develops one by defining Mythos-class as a vendor-neutral capability profile: a set of frontier autonomous offensive capabilities specified independently of any single model or vendor (defined by five constituent capability primitives). The central analytical claim is friction inversion: the patch primitives, segmentation, vendor-coordinated disclosure, and credential rotation that constrain Mythos-class capability in conventional IT environments are structurally absent on-chain. This makes blockchain exposure positioned differently in kind, not degree, from enterprise IT. The paper instantiates this finding against Bitcoin and Ethereum/L2 architectures through analysis of four major bridge exploits totaling over \$1.74 billion in losses. Vendor-neutral defensive and governance frameworks defined against the capability profile rather than any specific model release are the correct unit of analysis. On this basis the paper offers general recommendations for protocol governance, audit and verification cadence, and regulatory posture, developed as an analytical framework rather than as empirically validated risk estimates.

Keywords: blockchain security; artificial intelligence; cross-chain bridges; systemic risk; smart contracts; Ethereum; Bitcoin; DeFi risk governance; AI governance; responsible disclosure

Academic Editor: Maurizio Talamo

Received: 10 May 2026

Revised: 4 July 2026

Accepted: 7 July 2026

Published: 10 July 2026

Copyright: © 2026 by the author.
Licensee MDPI, Basel, Switzerland.
This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Cryptocurrency market infrastructure—public blockchains and the cross-chain bridges that link them—has produced over \$1.74 billion in bridge-exploit losses since 2021 and remains a focus of financial-stability monitoring for the Financial Stability Board and equivalent bodies [1,2]. The defensive and governance posture of this infrastructure has been calibrated against adversaries operating at human capability levels. On 7 April 2026, Anthropic released Claude Mythos Preview under restricted access through Project Glasswing, citing autonomous offensive cyber capabilities the company described as posing unprecedented risks [3,4]. The UK AI Security Institute’s independent evaluation corrob-

orated a meaningful step up in capability: autonomous discovery and exploitation of vulnerabilities, completion of multi-stage attack chains, and a 73% success rate on expert-level hacking tasks against a baseline in which prior models had completed none of those tasks [5]. The institutional response was immediate. The Bank of England intensified AI risk testing; German banking regulators consulted authorities and cyber experts; the term “Mythos-class” entered active use in the cybersecurity risk literature and policy commentary as a descriptor for the capability profile [6,7]. None of this institutional response has yet translated into blockchain-specific defensive, governance, or regulatory frameworks.

That gap is the subject of this paper. The existing literature on AI in blockchain security has addressed AI-assisted smart contract auditing using large language models [8], automated static analysis of bridge smart contracts [9], machine-learning approaches to anomaly detection in DeFi [10], and the security analysis and attack detection of cross-chain bridge architectures more broadly [11,12]. Recent work in the financial-technology literature has examined AI agents in financial markets and their systemic implications [13], the connectedness between AI-related assets and cryptocurrencies [14], and security-framework construction for adjacent financial infrastructure such as open banking [15]. None of this work treats frontier autonomous offensive AI capability—the capability profile Mythos-class denotes—as a systemic risk to blockchain financial infrastructure at the architectural level.

The bridge attack history of 2021–2022, totaling over \$1.74 billion in losses across four major exploits, was produced by adversaries operating without Mythos-class capability against architectural conditions that will face Mythos-class adversaries on timescales relevant to blockchain governance. The defensive posture appropriate to the latter has not been developed.

This paper makes three contributions. First, it operationalizes “Mythos-class” as a vendor-neutral capability profile defined by five primitives, decoupling the analytical category from any specific model release and engaging the contested boundary between maximalist [16] and distributional [17] framings of the capability. Second, it develops the central architectural claim that the friction mechanisms constraining Mythos-class impact in conventional IT environments are not reduced on-chain but structurally absent, and instantiates this claim through a capability-surface-impact matrix for the Ethereum/L2 programmable surface, with Bitcoin treated through comparative architectural analysis. Third, it grounds the analysis in case-study evidence from four major bridge exploits—Ronin, Wormhole, Nomad, and Poly Network—and develops policy and defensive prescriptions appropriate to the friction-inversion conditions the analysis identifies. More broadly, the contribution is a transferable analytical method: a capability-defined, vendor-neutral way of reasoning about how a step-change in adversarial capability interacts with the defensive friction a target environment does or does not provide. Although instantiated here for blockchain, the friction-inversion lens is intended to generalize to other critical-infrastructure classes whose defensive assumptions were calibrated against human-paced adversaries.

The friction-inversion finding the paper develops (formally stated and defended in Section 2.3)—by friction inversion we mean that the defensive mechanisms which slow a capable adversary in conventional IT environments (patching, segmentation, coordinated disclosure, credential rotation) are not merely reduced but structurally absent on-chain, so that the same capability meets less resistance rather than more—is empirically constrained in three directions. The friction primitives might be re-instantiable on-chain in functionally equivalent forms; the inversion might be load-bearing only at narrow architectural margins; or Mythos-class capability might fail to reach the surfaces identified here on timescales relevant to blockchain governance. Sections 3 through 5 examine each con-

straint and assess the evidence; Section 6 opens with an explicit constraint check that returns to all three.

Three research questions structure the paper: How should Mythos-class capability be defined operationally, independent of any single vendor? How do Bitcoin and Ethereum/L2 architectures differ in exposure to this capability profile? What governance, defensive, and regulatory posture follows? Section 2 develops the conceptual background and operational definition. Section 3 presents the analytical framework. Section 4 instantiates it through comparative architectural analysis and bridge case studies. Section 5 develops the systemic-risk implications. Section 6 outlines defensive prescriptions and a research agenda. Section 7 concludes. A Responsible Disclosure statement addressing dual-use considerations is provided in the Data Availability Statement.

The central durable claim the paper develops—advanced as an analytical, framework-level proposition rather than an empirically validated result—is that **Mythos-class is a capability category, not a single vendor moment**—and that vendor-neutral, capability-defined frameworks are the only unit of analysis that will remain valid on the time-scales relevant to blockchain governance.

2. Conceptual Background

2.1. Mythos-Class AI: An Operational Definition

The term “Mythos-class,” now in active use in the cybersecurity risk literature [17] and policy commentary [6,7], originated as a descriptor for the capability profile publicly disclosed by Anthropic for its Claude Mythos Preview model (released 7 April 2026) [3,16]. For the purposes of this paper, that origin is incidental. A definition that depends on a single vendor’s release would be analytically brittle—it would expire as model lineages evolve, and it would understate the threat by excluding any model exhibiting the same capability profile from a different developer. We therefore define Mythos-class as a capability profile, decoupled from any specific model, vendor, or release. The five primitives below are not introduced de novo; they consolidate capability dimensions already identified in the AI-security evaluation literature. Autonomous vulnerability discovery and multi-step exploit chaining correspond to the offensive-cyber capability categories used in the UK AI Security Institute evaluation methodology [5] and in Anthropic’s own capability reporting [16]; agentic execution with tool use reflects the agentic-evaluation paradigm now standard in frontier-model cyber assessments [5,16]; and the distributional analysis of where these capabilities are concentrated draws on independent replication work [17]. The contribution here is not the individual primitives but their consolidation into a single vendor-neutral classification gate applied to blockchain architecture.

A model is **Mythos-class** if it demonstrably exhibits all five of the following capability primitives:

1. **Autonomous vulnerability discovery at codebase scale.** The capacity to read, reason about, and identify exploitable defects in production-scale codebases (operating systems, browsers, smart contract systems, network protocols) without human-directed task decomposition. Anchor evidence: Anthropic’s reported identification of zero-day vulnerabilities across every major operating system and browser during pre-release evaluation [16]; AISI’s finding that the model autonomously discovered and exploited flaws in controlled testing [5].

2. **Multi-step exploit chaining.** The capacity to compose individual defects into operational exploit chains—privilege escalation sequences, sandbox escapes, write-primitive construction across multiple program states. Anchor evidence: documented multi-vulnerability privilege escalation chains in the Linux kernel, JIT heap sprays escaping browser sandboxes, and an autonomously written remote code execution exploit against FreeBSD [16].

3. Agentic execution with tool use. The capacity to plan, sequence, and carry out multi-stage operations against networked targets when given access to standard tooling, rather than only producing static analysis output. Anchor evidence: AISI's 73% success rate on expert-level hacking tasks and demonstrated completion of multi-stage attack chains estimated to require days of professional human work [5].

4. Sub-day weaponization timeline. The capacity to compress the discovery-to-deployable-exploit window from weeks or months of expert effort to hours of model-driven work. In this paper's usage, *weaponization* refers to exploit construction and testing—the engineering steps from analytical understanding of a vulnerability to a reliable, working exploit artifact ready to be deployed. It does **not** include deployment against live targets or post-deployment propagation, which Figure 1 represents as separate exploit-lifecycle phases. The Anthropic anchor evidence corresponds to this scope: a full exploit-development pipeline (analytical reasoning about the vulnerability, exploit code construction, and validation against target conditions) completed in under 24 h at a cost below USD 2000 [16]. Deployment was not tested in that evaluation.

5. Generality across target classes. The capability profile is not narrow tooling specialized for a single attack class but emerges from general-purpose agentic coding and reasoning capacity, and is therefore directable across heterogeneous target environments [16].

This definition is **forward-compatible and vendor-neutral**: any current or future model meeting all five criteria is Mythos-class regardless of provenance. It is also **conservative**: it excludes models that exhibit one or two of these primitives without the full profile (for example, narrow vulnerability scanners or models capable of analysis but not autonomous exploitation).

The definition admits one important contested boundary, which we note here and address in Section 2.2. Independent evaluation by AISLE [17]—distinct from AISI, the UK AI Security Institute—found that small open-weights models recovered much of the analytical work demonstrated in Anthropic's showcase vulnerabilities, including the FreeBSD exploit, suggesting that the analytical primitive (1) and partial chaining primitive (2) may be more diffuse across the model ecosystem than the Mythos branding implies. The boundary that may genuinely separate Mythos-class capability from broadly available capability appears to lie in the constructive engineering step—translating exploitable analysis into a reusable building block in a delivery chain, particularly under agentic conditions with live tool access. This paper takes the conservative position that the Mythos-class profile, defined by all five primitives operating in concert, is currently demonstrated by a small set of frontier models but should be expected to become approachable from a wider set of model families on the timescales relevant to blockchain security planning.

For the analytical work that follows, **Mythos-class** denotes the threat capability defined above. Where the paper refers to Anthropic's specific Claude Mythos Preview release, we use the proper noun. Where the paper refers to the general capability profile—as a category of adversary that defenders, regulators, and protocol governance must plan against—we use Mythos-class. The capability-surface-impact matrix in Section 3.1 instantiates this definition against the Ethereum/L2 programmable surface; the case studies in Section 4.3 evaluate historical bridge exploits against the same five primitives. Bitcoin's protocol-layer exposure is analyzed separately in Section 4.1 through architectural argument rather than against the same matrix.

Figure 1 illustrates the temporal dimension of the fourth primitive: an order-of-magnitude comparison of typical exploit-lifecycle phase durations from human-paced (months to days) to Mythos-paced (hours to seconds). The Mythos-paced Weaponization bar is anchored directly to Anthropic's reported exploit-development pipeline, which completed in under 24 h [16]. The other Mythos-paced phase durations and all human-

paced phase durations are illustrative bounds drawn from conventional cybersecurity-operations cadence and intended to convey the relative compression—multiple orders of magnitude across phases—rather than precise figures. The collapse is category-defining for blockchain systems whose defensive cadences assume human-paced discovery and weaponization.

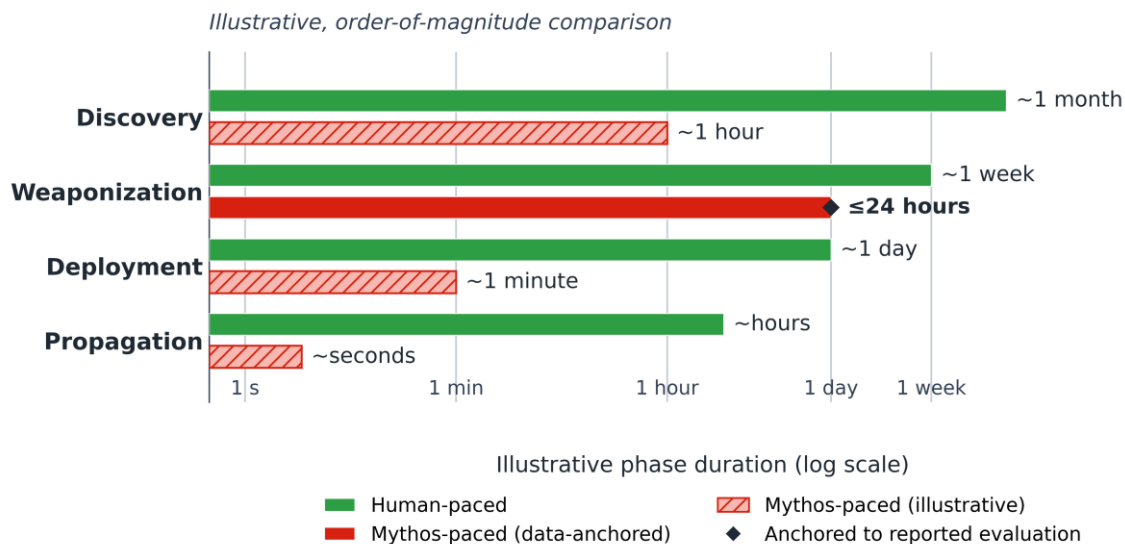


Figure 1. AI-accelerated exploit timeline compression—schematic, order-of-magnitude comparison (not to scale beyond log-axis approximation). Phase durations on a logarithmic scale for human-paced (green) and Mythos-paced (red) attackers across the four exploit-lifecycle phases. The Mythos-paced Weaponization bar is anchored to Anthropic’s reported exploit-development pipeline, which completed in under 24 h [16], and is labeled accordingly (≤ 24 h); other phase durations are bounds drawn from conventional cybersecurity-operations cadence. In the revised figure the single data-anchored bar (Mythos-paced Weaponization) is drawn solid and marked with a diamond, while the illustrative bars are hatched, so the distinction between the one measured value and the order-of-magnitude bounds is visible in the figure itself and not only in this caption.

2.2. The Capability Debate

The operational definition in Section 2.1 is conservative by design, but its analytical force depends on whether the Mythos-class capability profile is genuinely a category boundary or merely a high-water mark on a continuous distribution. The question is contested in the public technical literature, and the answer materially shapes what defenders, regulators, and protocol governance should plan against.

The maximalist framing. Anthropic’s characterization positions Mythos Preview as a category shift: internal evaluations identified zero-day vulnerabilities across every major operating system and web browser, with 99% unpatched at disclosure [16]. The model was deemed too dangerous for public release and distributed instead through Project Glasswing—a closed consortium including Microsoft, Google, Apple, AWS, JPMorgan Chase, and Nvidia [4]. The independent evaluation by the UK AI Security Institute corroborated a meaningful step up: autonomous discovery and exploitation of vulnerabilities, completion of multi-stage attack chains, and a 73% success rate on expert-level hacking tasks against a baseline in which prior models had completed none of those tasks [5]. The Bank of England intensified AI risk testing in response [7]. The maximalist framing rests on three claims: that the capability is qualitatively new, that it is currently held by a

small set of frontier models, and that it justifies treating Mythos-class as a distinct adversary category.

The distributional counter-argument. Independent evaluation by AISLE [17] tested the specific vulnerabilities Anthropic showcased—including the FreeBSD remote code execution exploit and Linux kernel privilege escalation chains—against small, cheap, open-weights models. Eight out of eight models tested detected the FreeBSD exploit, including one with only 3.6 billion active parameters. AISLE’s conclusion is that AI cybersecurity capability is “very jagged”: it does not scale smoothly with model size, and the analytical primitive of reasoning about exploitability is more diffuse across the model ecosystem than the Mythos branding implies. AISLE’s tests were not conducted with full agentic infrastructure, and the team identifies the genuine boundary as the constructive engineering step—translating exploitable analysis into a reusable building block under live agentic conditions. AISLE thus does not refute the maximalist framing; it relocates the boundary from analytical capability (primitive 1) to agentic execution and weaponization (primitives 3–4).

What this paper takes from the debate. The two critiques cut in opposite directions but converge on a single methodological consequence: the threat surface is wider than Anthropic’s framing suggests, but operational impact within that surface depends on the friction the target environment provides. For enterprise IT, the defensive-friction argument is load-bearing. For blockchain systemic risk, the opposite is true. Smart contract systems present almost none of the friction conventional environments provide: code is public, deterministic, replayable, and immutable once deployed; defenders cannot patch in the operational sense; high-value targets are addressable on-chain. The friction critique therefore weakens the case for enterprise concern and strengthens the case for blockchain concern. We develop this argument in Section 2.3.

The AISLE distributional critique cuts decisively in favor of treating Mythos-class as a category defenders must plan for. If the boundary is the agentic-execution and weaponization primitive, capability that today resides in a small set of frontier models will become approachable from a wider set of model families on the timescales relevant to blockchain governance. Vendor-neutral, capability-defined frameworks are therefore the correct unit of analysis.

This paper accordingly takes the operational definition in Section 2.1 as the analytical anchor; does not assume Mythos-class capability is uniquely Anthropic’s; does not assume it is uniformly distributed across the model ecosystem; and treats it as a capability category any sophisticated adversary should be expected to access on planning-relevant timescales.

2.3. Why Blockchain Is Distinctively Exposed

The defensive-friction argument from Section 2.2—that even concentrated Mythos-class capability faces real-world friction against well-defended systems—is the strongest available rebuttal to maximalist threat framing, and the central reason a Mythos-class threat applied to enterprise IT is not a category emergency. Patch cycles, network segmentation, vendor-coordinated disclosure, endpoint detection, and credential rotation produce friction that compresses but does not collapse defender response times.

This paper’s central architectural claim is that **those friction mechanisms are not natively available as universal, operationally immediate primitives at the deployed-contract layer**; where they exist—through upgrade proxies, pause functions, timelocks, governance upgrades, guardian committees, validator-set changes, or bridge emergency controls—they are governance-dependent, unevenly implemented, and often slower than on-chain exploit propagation. The properties that make blockchain systems valuable—public, deterministic, immutable, composable, addressable—are the same properties that strip

out the friction Mythos-class capability must overcome elsewhere. The result is not a marginal increase in risk relative to enterprise IT; it is a structural difference in the defender's position. We develop this through five architectural properties, each inverting a friction mechanism that constrains Mythos-class impact in conventional environments. Methodologically, each property is developed as a paired contrast: a named friction mechanism that is operative in enterprise IT is set against its on-chain status, and the inversion is evidenced either by a documented incident (for example, Wormhole for the absent patch primitive, Nomad for replayability) or by an intrinsic architectural property of public deterministic execution. This paired-contrast structure is what allows the five properties to be read as a single systematically related inversion rather than an unordered list of risks, a point made explicit at the end of the subsection.

Immutable deployed code. Smart contracts are not patchable in the operational-IT sense. Mitigation requires migration (user action, liquidity reallocation), governance-mediated upgrades (multisig coordination, pre-existing upgrade architecture), or hardfork (a coordination cost most protocols cannot bear for a single defect). The Wormhole exploit is illustrative: the vulnerability was patched in source and committed to a public GitHub repository, but the deployed mainnet contracts remained vulnerable for hours, and an attacker reverse-engineered the fix and weaponized it before deployment [18]. The conventional defender confronts a vendor-coordinated disclosure window of hours to days; the on-chain defender has no patch primitive at all.

Public, deterministic, replayable execution. Production smart contract code is public by design. Adversaries can read it, simulate it against forked chain state, and execute unbounded test exploits at zero risk before submitting any on-chain transaction. Every primitive in the operational Mythos-class definition is enabled, not constrained, by blockchain's transparency: autonomous discovery operates on textual code; multi-step chaining operates on simulated state transitions; agentic execution operates on transactions whose effects are fully predictable from public state. In conventional environments, attackers confront opacity at multiple layers—closed binaries, unobservable runtime state, behavioral defenses. On-chain, the cost of discovery against public contracts is reduced to the cost of model inference.

Addressable, quantifiable, high-value targets. A bridge contract holding \$400 M in TVL is not an inferred target; it is a public address with a public balance. The targeting problem that consumes substantial adversary effort against enterprise victims—identifying which systems hold valuable assets, which credentials grant access, and which paths lead from initial foothold to monetizable outcome—is essentially solved before the adversary begins. The four bridge exploits analyzed in Section 4.3 collectively held billions of dollars in TVL preceding their exploits; none required identification work. Adversaries without Mythos-class capability already converge on these targets. The relevant question is what changes when adversaries with capability arrive.

Composability as attack amplifier. DeFi composability turns a single-contract compromise into a multi-protocol incident. A flaw in a widely integrated lending protocol, stablecoin, or oracle propagates through every protocol that depends on it. Flash-loan-enabled exploitation chains permit assembly of large attack-time capital across protocols within a single transaction, making attack profitability a function of combined liquidity reachable in one block. Mythos-class multi-step-chaining capability operates natively in this environment: a profile demonstrated to chain operating-system vulnerabilities through privilege escalation [16] is structurally well-matched to chaining DeFi protocol interactions. The conventional analog—a multi-host lateral movement campaign—typically takes a skilled human operator days to weeks. The DeFi analog executes in a single transaction.

Compressed disclosure-to-exploit windows and replayability. The Nomad bridge case demonstrates a property unique to blockchain: post-discovery dynamics. Once the initial Nomad exploit transaction was on-chain, the exploit was trivially replayable; Mandiant's incident reconstruction documents approximately 300 addresses participating in the cascade, generating 960 transactions over roughly 150 min, with MEV bots automating replication within minutes [19]. No analogous dynamic exists in enterprise IT—a successfully exploited corporate target does not create a public, executable artifact. On-chain, every exploit transaction is a public proof-of-concept. A Mythos-class capability that surfaces a Nomad-class flaw produces not a single attacker but a crowd-loot dynamic measured in minutes.

The friction inversion. The five properties above are not independent risk factors to be summed; they are systematically related. The relationship is generative rather than additive: each property removes a specific defensive primitive (patching, opacity, targeting cost, containment, single-shot exploitation), and the removals compound because they share a common cause—the public, deterministic, immutable, composable, addressable character of on-chain code. This is the same structural logic that systemic-risk analyses of financial infrastructure describe when a single architectural property drives multiple correlated failure channels rather than independent ones [1,2]; treating the five as summable independent factors would understate the exposure precisely because they are not independent. Each inverts a friction mechanism that constrains Mythos-class impact in conventional environments, and together they produce a defender position with no direct analog in enterprise IT. The conventional defender against a Mythos-class adversary has reduced friction but retains patch primitives, segmentation primitives, and credential rotation primitives. The on-chain defender often lacks immediate functional equivalents. The argument is therefore not that blockchain is more vulnerable than enterprise IT in degree but that it is differently positioned in kind. The defensive-friction critique that materially weakens the case for enterprise crisis does not apply on-chain; the friction is not reduced but structurally absent.

This claim has three immediate consequences. First, the Section 3 capability-surface-impact matrix must instantiate Mythos-class primitives against a defender model with no patch primitive. Second, the Section 4 comparative analysis cannot be a simple severity ranking but must analyze how the five properties manifest differently across architectures with different programmability profiles. Third, Sections 5 and 6 implications must be constructed for a defender position lacking the friction mechanisms regulatory and defensive frameworks have implicitly assumed for decades.

Figure 2 visualizes the consequence at the architectural level: the six architectural surfaces introduced above, classified by exposure tier under the friction-inversion conditions just developed. Bridge contracts emerge as the singular Severe-tier exposure surface; the execution layer occupies the High tier; client implementations, wallet/firmware, and governance occupy the Medium tier; consensus alone is bounded at Low. Section 3.1 develops the underlying ratings against the full capability-primitive matrix; Section 4 instantiates them comparatively across Bitcoin and Ethereum/L2.

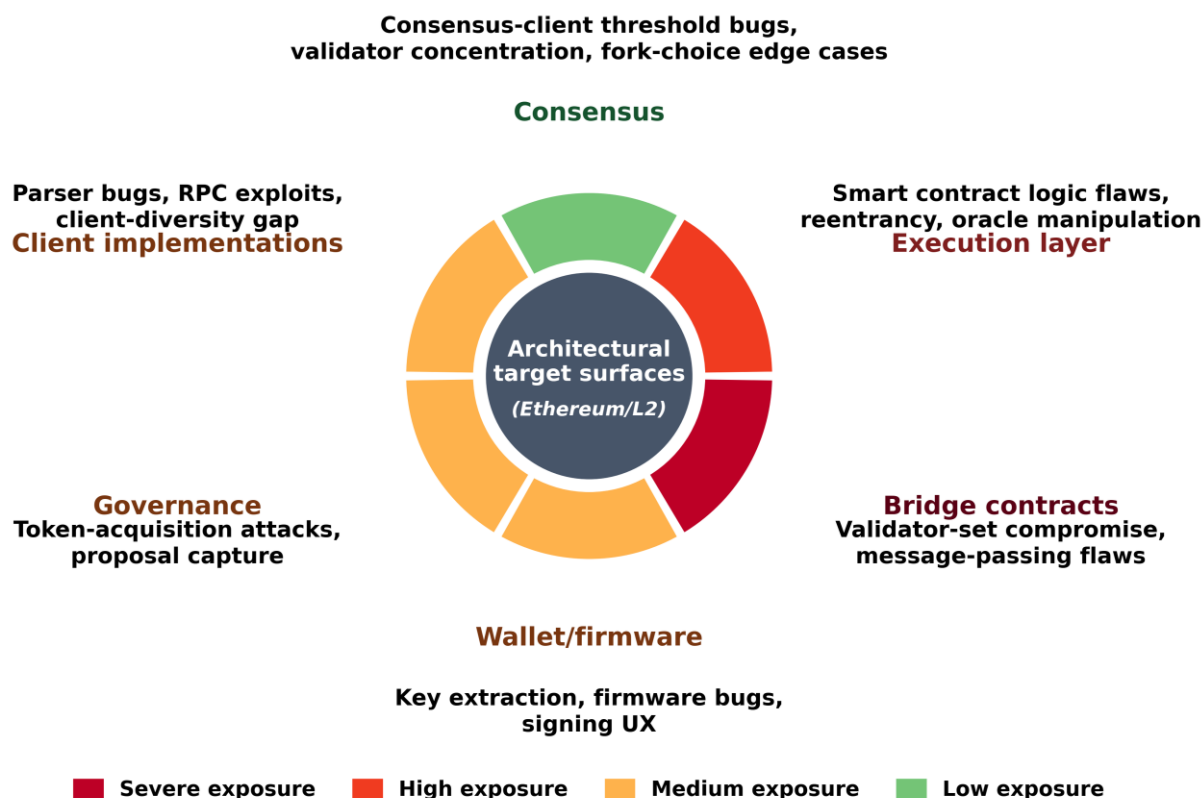


Figure 2. Architectural target surfaces under Mythos-class threat. The six architectural surfaces of the Ethereum/L2 programmable substrate, classified by aggregate exposure tier (Low/Medium/High/Severe). Aggregate tier reflects the modal rating per surface across the capability-surface-impact matrix primitives (Section 3.1), not a quantitative risk score.

2.4. Bitcoin and Ethereum/L2 Architectural Profiles

This subsection establishes the technical baseline against which the comparative analysis in Section 4 is conducted. The treatment is intentionally brief; the architectural details below are well-established in the existing blockchain literature [20–22].

Bitcoin. Bitcoin’s protocol architecture is intentionally minimal. Script, the language used for transaction validation, is non-Turing-complete and provides a small, deliberately constrained operational set, with no looping primitives and no persistent state across transactions [21]. Consensus operates through Nakamoto-style proof-of-work with conservative protocol evolution; recent upgrades (SegWit, Taproot) have extended scripting capability incrementally without altering the underlying minimal-execution philosophy. The Lightning Network, Bitcoin’s primary layer-2, operates through bilateral payment channels with on-chain settlement, introducing complex state-machine logic at the channel layer but not at the protocol layer. Wallet implementations, exchange integrations, and Lightning channel software constitute Bitcoin’s ecosystem layer; the operational distinction between protocol and ecosystem is sharper for Bitcoin than for any subsequent major blockchain.

Ethereum. Ethereum’s architecture inverts Bitcoin’s minimal-execution philosophy. The Ethereum Virtual Machine (EVM) is Turing-complete, supporting arbitrary computation subject to gas-cost bounding [22]. Smart contracts deployed to the EVM are immutable post-deployment unless explicit upgrade mechanisms are implemented at the contract level. Consensus operates through proof-of-stake with validator clients implemented across multiple independent codebases—Geth, Nethermind, Reth, Besu, and Erigon at the execution layer; Prysm, Lighthouse, Teku, Nimbus, and Lodestar at the consensus layer

[23]. Client diversity is a deliberate architectural defense against single-implementation flaws.

The L2 Stack. Ethereum scaling has consolidated around layer-2 rollup architectures, with optimistic rollups (Arbitrum, Optimism, Base) and zero-knowledge rollups (zkSync, StarkNet, Polygon zkEVM, Linea) constituting the dominant scaling pattern. Rollups inherit Ethereum’s L1 security through proof or fraud-proof mechanisms but introduce additional architectural surface—sequencer software, data availability layers, bridge contracts, and proof systems [24]—each in earlier operational maturity than Ethereum L1 itself, and growing faster than the audit capacity available to it.

The architectural divergence between Bitcoin and Ethereum/L2 is the variable Section 4 isolates. The Section 3 capability-surface-impact matrix instantiates Mythos-class capability primitives against this divergence.

To position this contribution against existing work at the intersection of AI and blockchain security, Table 1 compares representative recent approaches along four dimensions: the security task addressed, the analytical unit, whether the work treats frontier autonomous offensive capability as a threat model, and whether it addresses cross-chain bridge or systemic-risk exposure. The comparison shows that prior work concentrates on tool-building for defensive detection (LLM-assisted vulnerability detection, static bridge analysis, DeFi anomaly detection, cross-chain attack-transaction detection) and on cross-chain bridge security as an attack-surface taxonomy, but that none adopts a vendor-neutral frontier-capability threat model or develops the architectural friction-inversion argument this paper contributes.

Table 1. Positioning against representative state-of-the-art work at the AI-blockchain security intersection.

Approach/Work	Security Task	Analytical Unit	Frontier Autonomous-Capability Threat Model?	Cross-Chain/Systemic Scope?
LLM-powered contract vulnerability detection [8]	Defensive detection	Individual contract	No	No
SmartAxe static cross-chain analysis [9]	Defensive detection	Bridge contract	No	Bridge (detection only)
DeFiGuard GNN price-manipulation detection [10]	Defensive detection	DeFi transaction graph	No	DeFi (detection only)
Attack-transaction detection on bridges [12]	Defensive detection	Bridge transaction	No	Bridge (detection only)
SoK: cross-chain bridge security [11]	Attack-surface taxonomy	Bridge architecture	No	Bridge (taxonomy)
This work	Architectural/systemic risk analysis	Capability profile × architecture	Yes (vendor-neutral Mythos-class profile)	Bitcoin vs. Ethereum/L2 + systemic

3. Materials and Methods

This section instantiates the operational definition from Section 2.1 against the architectural properties from Section 2.3. The framework is presented as a capability-surface-impact matrix (Section 3.1), grounded in a defined threat-modeling approach (Section 3.2), and bounded by explicit assumptions (Section 3.3). The matrix is referenced throughout Sections 4 and 5.

3.1. The Capability-Surface-Impact Matrix

Table 2 maps the five Mythos-class capability primitives (rows) against six architectural surfaces (columns) of the Ethereum/L2 programmable substrate. The matrix is scoped to Ethereum/L2 because the friction-inversion thesis developed in Section 2.3 is about programmable on-chain code: the capability primitives interact with codebase complexity, multi-step composability, and bridge connectivity, all of which are properties of the programmable surface. Bitcoin’s protocol-layer exposure is structurally different (small non-Turing-complete Script surface, fifteen-plus years of adversarial pressure, conservative protocol evolution) and is analyzed through architectural argument in Section 4.1 rather than through the same matrix. Each cell of Table 2 records the assessed exposure under three criteria stated in the Scoring Criteria section: codebase accessibility, historical exploit precedent, and demonstrated capability primitives in published Mythos-class evaluations.

Table 2. Capability-surface-impact matrix (Ethereum/L2 programmable surface).

Capability Primitive	Consensus	Execution Layer	Client Implementations	Bridge Contracts	Wallet/Firmware	Governance
1. Autonomous discovery at codebase scale	Low	High	Medium	Severe	Medium	Medium
2. Multi-step exploit chaining	Low	High	Medium	Severe	Low	High
3. Agentic execution with tool use	Low	High	Low	Severe	Medium	Medium
4. Sub-day weaponization	Medium	High	Medium	Severe	Medium	Medium
5. Generality across target classes	Low	High	Medium	High	Medium	Medium

The matrix surfaces three findings the paper will develop. First, the bridge-contract column is the only column with Severe ratings across four of five primitives. Bridges are the architectural surface most fully exposed to the Mythos-class capability profile, and Section 4.3 develops this as the central comparative finding. Second, consensus-layer exposure is predominantly low, with one medium rating for sub-day weaponization: Ethereum has undergone extensive adversarial scrutiny for more than a decade; its post-Merge proof-of-stake consensus layer has operated under adversarial scrutiny since 2022 and presents a small, well-audited surface, and Mythos-class capability does not change that aggregate posture. Third, the execution-layer column rates High across all five primitives, reflecting the composability and codebase-complexity properties developed in Section 2.3, while the governance column shows a single High rating specifically on the multi-step chaining primitive—the surface where compositional attack paths most directly amplify governance exposure.

Scoring Criteria

Cells are rated **Low/Medium/High/Severe** according to three combined criteria:

1. Codebase accessibility. Public source availability, codebase size, and the degree to which the surface is exposed to autonomous code-reading capability. Public Solidity contracts score higher exposure than closed-source wallet firmware.

2. Historical exploit precedent. The volume and severity of exploits historically observed against the surface, weighted toward 2021–2025 incidents to reflect current adversary capability. Bridges, with \$1.74 billion in losses across the four cases analyzed in Section 4.3, score strongly on this criterion—one of the inputs supporting Severe ratings, not by itself sufficient.

3. Demonstrated capability primitives. Whether the specific primitive has been demonstrated against analogous targets in published Mythos-class evaluations [5,16,17]. The autonomous-discovery primitive against operating system codebases, for example, transfers cleanly to autonomous discovery against smart contract codebases of comparable size and complexity.

Cell ratings aggregate the three criteria above into a four-level scale per the rubric in Table 3. The rubric makes the scoring logic explicit and transparent: a third party scoring the same surface should be able to apply the same criteria transparently and arrive at a defensible rating. The Severe rating requires not only that all three criteria are strongly present but also that the surface satisfies an additional gate—high-value targets (typical loss magnitude per incident measured in tens of millions or more) and low-friction exploitability (public, replayable, deterministic execution; no patch primitive).

A criterion is treated as strongly present when there is either direct historical incident precedent on the surface, public code accessibility at production scale, or demonstrated transferability from published Mythos-class evaluations to an analogous target class. Ratings are qualitative and diagnostic, not probabilistic.

Table 3. Scoring rubric for Table 2’s capability-surface-impact ratings.

Rating	Required Condition
Low	No strong alignment across the three criteria (codebase accessibility, historical exploit precedent, demonstrated capability primitives)
Medium	One criterion strongly present
High	Two criteria strongly present
Severe	All three criteria strongly present, plus an additional gate: high-value targets and low-friction exploitability (public, replayable, deterministic execution with no patch primitive at the deployed-contract layer)

3.2. Threat Modeling Approach

The framework adapts established threat-modeling methodology rather than constructing a novel scheme. Two reference frameworks are used in combination.

The **MITRE ATT&CK** structure [25] provides the kill-chain decomposition (reconnaissance, initial access, execution, persistence, exfiltration). Where MITRE ATT&CK assumes enterprise IT targets, we substitute blockchain-specific equivalents: target identification becomes on-chain TVL ranking; initial access becomes contract function call; execution becomes transaction submission; persistence is generally absent on-chain; exfiltration becomes asset transfer to attacker-controlled addresses. The substitution is straightforward because the kill-chain structure is target-agnostic; what changes is the friction at each stage, as developed in Section 2.3.

The **STRIDE** taxonomy [26] (Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of privilege) provides the vulnerability-class decomposition applied to each architectural surface. The four bridge case studies in Section 4.3 distribute across STRIDE categories as follows: Ronin (Spoofing—forged validator signatures via key compromise); Wormhole (Spoofing—fake sysvar account bypassing signature verification); Nomad (Tampering—manipulated Merkle root state allowing forged messages); Poly Network (Elevation of privilege—privileged contract abused to swap Keeper authority). This distribution is itself a finding: bridge exploits cluster on Spoofing

and Elevation-of-privilege classes, both of which are well within the demonstrated Mythos-class capability profile.

3.3. Assumptions, Boundaries, and Epistemic Position

The analysis rests on five stated assumptions.

First, the operational definition of Mythos-class capability is derived from publicly available evidence: the Anthropic Mythos Preview system card and announcement, the UK AI Security Institute's April 2026 evaluation, AISLE's independent technical analysis, and reporting from Fortune and Gilbert + Tobin. No proprietary data, classified evaluation, or private vendor disclosure informs the capability definition. The conservative posture—requiring all five primitives for Mythos-class classification—is intended to remain valid as the public evidence base expands.

Second, the analysis is **architectural and systemic**, not protocol-audit. We do not claim to identify novel vulnerabilities in any specific protocol; we analyze how the Mythos-class capability profile interacts with architectural classes. Specific protocol exposure within an architectural class will vary by implementation quality, audit coverage, and governance maturity.

Third, the bridge case studies in Section 4.3 are **historical and post-disclosure**. All vulnerabilities analyzed are publicly disclosed, post-patch, and have been subjects of multiple prior post-mortem analyses. The contribution is the comparative analysis against the operational Mythos-class profile, not novel disclosure.

Fourth, the analysis assumes Mythos-class capability is **approachable from multiple model families** on the timescales relevant to blockchain governance, consistent with the AISLE distributional finding [17]. Conclusions remain valid under both the maximalist (single-vendor) and distributional (multi-vendor) framings of Section 2.2; recommendations are constructed to remain valid across the full range.

Fifth, the analysis treats blockchain systems as the defender position. We do not address scenarios in which Mythos-class capability is deployed defensively (e.g., autonomous formal verification, continuous audit) except in Section 6, where defensive applications are framed as future work rather than load-bearing claims.

Source quality and evidence hierarchy. The reference base aggregates four classes of sources with distinct evidentiary status, and load-bearing claims are anchored accordingly. *Primary technical sources*—foundational protocol papers, peer-reviewed publications, and threat-modeling standards (the Ethereum white and yellow papers, MITRE ATT&CK, Shostack's threat-modeling reference, and prior peer-reviewed work in the Journal of the British Blockchain Association)—anchor the architectural argument and the threat-modeling vocabulary used throughout. *Regulatory and institutional sources*—the UK AI Security Institute, the Financial Stability Board, the UK Treasury Select Committee, and Anthropic's first-party technical disclosures—anchor capability and systemic-risk claims. *Official postmortems and formal incident-response reports*—Sky Mavis on Ronin and Mandiant on Nomad—establish first-party incident facts for two of the four bridge cases. *Independent technical analysis and commentary*—security-firm postmortems (Halborn, CertiK, SlowMist), blockchain forensics (Elliptic), analytics dashboards (L2BEAT), exchange security writeups (Kraken, Coinbase), independent capability research (AISLE), news reporting (Fortune), and legal commentary (Gilbert + Tobin)—corroborate post-disclosure incident reconstruction in Section 4.3, where multiple independent technical writeups exist for each major bridge exploit. Commentary is used to corroborate rather than to establish factual claims; where a single claim depends on commentary alone, the text marks this explicitly.

Limitations. The analysis does not estimate loss probability, does not test live protocols, and does not claim that any specific bridge currently contains an exploitable vulnerability; it evaluates architectural exposure under a capability-defined threat model.

The framework's intended use is comparative and diagnostic. It is not a quantitative risk model and does not produce loss estimates. It produces an ordered understanding of which architectural surfaces demand the earliest defensive and governance attention under a Mythos-class threat assumption.

Scientific contribution and external validity. The contribution of this paper is an analytical framework, not an experimentally validated risk model. It does not report experimentation, and it does not claim replication or external validity in the empirical sense. Its ratings—the Low/Medium/High/Severe values in Table 2 and the counterfactual uplift ratings developed in Section 4.3—are structured qualitative judgments derived from the three stated scoring criteria (Scoring Criteria section) and the explicit uplift criteria (Section 4.3), not measurements. The framework's validity claim is therefore internal and methodological: given the stated criteria and the public evidence base, the ratings are reproducible and defensible, and a third party applying the same criteria to the same surfaces should arrive at the same ordering. Establishing external validity—for example by constructing blockchain-specific capability benchmarks, replaying historical exploits under Mythos-class conditions in a controlled cross-chain simulation environment, or red-team evaluation against instrumented testnets—is deliberately framed as future work in Section 6.5. We make this positioning explicit here so that the qualitative ratings throughout Sections 4 and 5 are read as analytical instruments for prioritization rather than as probabilistic or empirical estimates. Section 4 instantiates the framework against Bitcoin and Ethereum/L2 architectures; Section 5 develops the systemic-risk implications; Section 6 maps the framework cells to defensive priorities.

4. Results: Comparative Analysis of Bitcoin and Ethereum/L2

This section instantiates the framework from Section 3 against the two architectural classes selected in Section 2.4. The analysis is intentionally asymmetric: Bitcoin (Section 4.1) and Ethereum/L2 (Section 4.2) receive comparative-architecture treatment, while bridges (Section 4.3) receive case-study treatment because the bridge attack history provides the empirical foundation on which the rest of the paper's claims rest. Section 4.4 synthesizes the comparative finding.

4.1. Bitcoin: Bounded Exposure

Bitcoin's protocol-layer attack surface is small by design. Script is intentionally non-Turing-complete; consensus has undergone over fifteen years of adversarial pressure; protocol evolution is conservative and slow. The Section 3.1 matrix is scoped to the Ethereum/L2 programmable surface and is therefore not used to rate Bitcoin protocol-layer exposure; the architectural argument here stands on its own. None of the five Mythos-class capability primitives engage strongly against this surface: autonomous discovery against well-audited, small-surface protocol code yields marginal returns; multi-step chaining has limited substrate to chain through; agentic execution finds limited high-value endpoints at the protocol layer; sub-day weaponization is constrained by the absence of rapidly deployable on-chain logic; and the generality of the capability profile cannot compensate for the absence of programmable surface to direct it against.

Bitcoin's exposure concentrates in the **ecosystem layer**: wallet implementations (software and hardware), Lightning Network channel logic, exchange integrations, and adjacent infrastructure such as mining pool software. These surfaces present higher exposure than the protocol—wallet codebases are larger, Lightning channel logic involves complex state machines, and exchange integrations are operationally rather than cryptographically

defended. Historical loss precedent reflects this: the largest Bitcoin-denominated incidents (Mt. Gox, Bitfinex, multiple subsequent exchange breaches) have been operational and custodial rather than protocol-level.

Two ecosystem-layer surfaces warrant closer treatment under a Mythos-class assumption, because they concentrate exactly the codebase-complexity and state-machine properties the capability profile is best matched to. The first is **Lightning Network channel state machines**. A Lightning channel is a bilateral, off-chain state machine whose security depends on each party holding the correct latest commitment transaction and on the timely, correct use of penalty transactions and justice mechanisms when a counterparty broadcasts a revoked state. This is precisely the kind of multi-step, state-dependent logic against which autonomous discovery (primitive 1) and multi-step chaining (primitive 2) have leverage: the attack surface includes revoked-state broadcast, watchtower failure or eviction, channel-jamming and grieving vectors, and race conditions around time-lock expiry. Unlike the Bitcoin base layer, channel software is comparatively young, implemented across multiple independent codebases (for example LND, Core Lightning, and Eclair), and evolving quickly, which widens the codebase an autonomous-discovery capability can read and shortens the adversarial-maturity record any single implementation has accumulated. The second is **hardware wallet firmware**. Hardware wallets defend high-value keys through a combination of secure-element isolation, firmware signature verification, and a constrained input path, but their firmware is nonetheless a bounded, analyzable codebase, and the historical record already includes supply-chain, fault-injection, and firmware-parsing vulnerabilities discovered by human researchers. A Mythos-class capability directed at published firmware images, PSBT and descriptor parsers, and update mechanisms raises the floor rate at which such implementation flaws are surfaced, even where the secure element itself resists direct extraction. In both cases the strategic conclusion is unchanged—these are ecosystem-layer, not protocol-layer, exposures, and their mitigations remain largely operational (signed reproducible builds, watchtower redundancy, firmware attestation, and coordinated disclosure) rather than requiring changes to the Bitcoin protocol. What the Mythos-class assumption changes is the cadence at which these surfaces must be audited and the assumption that discovery of such flaws remains human-paced.

A Mythos-class capability profile applied to Bitcoin therefore raises **floor risk** modestly without changing strategic profile. Wallet firmware and Lightning channel software become more attractive targets under autonomous-discovery capability; exchange integrations remain primarily operationally exposed. The categorical risk shift the paper develops for Ethereum/L2 in Section 4.2 does not apply to Bitcoin. This finding is itself analytically useful: it isolates **programmability**, not chain identity, as the variable that interacts with Mythos-class capability.

4.2. Ethereum and the L2 Stack: Expanding Exposure

Ethereum's exposure profile inverts Bitcoin's. The Turing-complete EVM presents a substantial execution-layer surface. Client diversity (Geth, Nethermind, Reth, Besu, Erigon) reduces consensus-layer risk through implementation heterogeneity but multiplies the codebase Mythos-class autonomous discovery can read. The deployed contract ecosystem is large, public, and growing. Against the Section 3.1 matrix, Ethereum's execution-layer column rates High across all five capability primitives.

The L2 stack adds a second-order expansion. Optimistic and zero-knowledge rollups, data availability layers, sequencer logic, and L1↔L2 messaging contracts have each grown faster than the defensive tooling around them. Sequencer software is in early operational maturity relative to L1 client software; data availability proof systems involve novel cryptographic constructions whose adversarial pressure is still accumulating; the

bridge contracts that connect L1 and L2 inherit the architectural exposure documented in Section 4.3. The L2 ecosystem’s growth rate—measured in deployed TVL, sequencer count, and rollup framework diversity—has outpaced the audit and formal verification capacity available to it.

A Mythos-class capability profile applied to Ethereum/L2 therefore acts as a **multiplier on existing programmability risk** at a moment when programmability surface is expanding faster than defensive capability. The Section 2.3 friction-inversion argument applies fully here: deployed contracts are immutable, public, and composable; high-value targets are addressable; replayability dynamics apply. The categorical risk shift is real.

Two governance-adjacent surfaces deserve specific mention. Governance contracts—token-weighted voting systems controlling protocol parameters—present a target class where multi-step chaining capability has unusual leverage, because successful exploitation of a governance vector can change protocol parameters rather than extract a one-time loss. The Section 3.1 matrix rates governance High on the chaining primitive specifically. Validator client diversity, while genuinely protective at the consensus layer, also expands the codebase autonomous-discovery capability can read; a discovered flaw in a single client implementation reaching threshold network share is a consensus risk despite client diversity.

4.3. Bridges as the Mythos-Class Prime Target

Bridges are the single architectural surface where the Section 3.1 matrix rates Severe across four of five Mythos-class capability primitives. They are also the architectural class with the largest historical loss precedent: the four exploits compiled in Tables 4 and 5 below total over \$1.74 billion in losses. The convergence of architectural exposure and historical precedent makes bridges the empirical foundation for the paper’s central claims. Table 4 records the factual incident profile of each exploit; Table 5 applies the Mythos-class framework to assess capability uplift. The split separates what happened from how the paper’s analytical apparatus interprets it.

Table 4. Bridge exploit case studies—factual incident record.

Dimension	Ronin (March 2022)	Wormhole (February 2022)	Nomad (August 2022)	Poly Network (August 2021)
Loss	~\$625 M	~\$320–326 M	~\$190 M	~\$611 M
Primary vector	Validator key compromise (5-of-9 multisig forged)	Fake sysvar account bypassed signature verification	Init bug (committed-Root = 0x00): messages always verified	Privilege-architecture flaw + selector collision swapped Keeper key
Root-cause class	Operational + governance hygiene	On-chain code flaw (Solana/Rust BPF)	On-chain code flaw (Solidity initialization)	On-chain code flaw (Solidity privilege architecture)
Discovery/dwell-time fact	Permission residue ~4 mo: Axie DAO validator delegation (November 2021) unrevoked at exploit [27]	Hours: public source commit → mainnet exploit [18]	~6 weeks: proxy upgrade (June 2022) → exploit (August 2022) [19]	Pre-exploit dwell not publicly characterized [28]
Replayability/crowd-loot	Not replayable (key-dependent)	Not trivially replayable	Trivially replayable—~300 addresses, 960 tx, ~150 min	Not replayable post-exploit
Resolution	Sky Mavis covered losses; validator set expanded	Jump Crypto replaced 120 k wETH; \$140 M counter-recovered Feb 2023 [29]	~\$33–37 M returned by white-hats; majority unrecovered	All funds returned by attacker

Key supporting sources	[27,30,31]	[18,32]	[19,33,34]	[28]
------------------------	------------	---------	------------	------

Note on loss figures in Table 4: All values are reported in commonly cited exploit-date USD, computed at asset prices prevailing at the time of the malicious transactions. Source-to-source variation (e.g., the Ronin loss is variously reported as ~\$540 M and ~\$625 M across the cited postmortems and forensics writeups) reflects rounding conventions and the inclusion or exclusion of partially recovered or pre-exploit-frozen tranches; post-event recovery dynamics are reported separately in the Resolution row.

Table 4 summarizes the six dimensions most load-bearing for the paper’s analytical claim. Additional incident detail—discovery method, vulnerability dwell time pre-exploit, time-to-detection, and attribution—is reported in the bridge-by-bridge narrative in this section rather than in the table, both to keep the table compact for journal column layout and to preserve the connective tissue between facts and analytical interpretation.

Table 5 applies the Mythos-class framework (Section 3) to the same four cases. It rates the code-reading capability the historical attack required, identifies which Mythos-class capability primitives map most directly to the vulnerability class, and assigns an uplift assessment for what Mythos-class capability would change about each attack’s discovery cost and timeline. The uplift ratings are counterfactual judgments—what Mythos-class capability would have changed about each historical exploit—and not empirical measurements; no Mythos-class actor was involved in the four incidents, all of which predate the April 2026 disclosure. The four uplift levels are assigned against explicit criteria, stated here so that a third party can apply them reproducibly. **Severe** uplift requires that the capability primitives in the operational definition (Section 2.1) map directly onto the vulnerability class and that their application would have collapsed time-to-discovery from the weeks-to-months of the historical timeline to hours-to-days. **High** uplift requires that the primitives accelerate the attack but do not transform its class—the discovery bottleneck is eased, not removed. **Medium** uplift denotes marginal acceleration, typically where the decisive step lies partly outside the capability profile (for example, social-engineering-led access). **Low** uplift denotes an attack whose decisive step lies outside the capability profile altogether. The same four-level scale and gating logic used for the surface-exposure ratings (Table 3) governs the uplift ratings; the distinction is that Table 3 rates static architectural exposure whereas the uplift ratings rate counterfactual change to a specific historical attack.

Table 5. Bridge exploit case studies—Mythos-class uplift assessment.

Dimension	Ronin (March 2022)	Wormhole (February 2022)	Nomad (August 2022)	Poly Network (August 2021)
Code-reading capability required	Low (social-engineering-led)	High (Solana sysvar/Rust BPF)	Medium–High (proxy upgrade + Solidity storage defaults)	Very High (multi-contract privilege graph + selector preimage search)
Mythos primitives most relevant	Agentic execution (recon, social engineering)	Autonomous discovery + multi-step chaining	Autonomous discovery + agentic execution (crowd-loot amplification)	Autonomous discovery + multi-step chaining (multi-contract reasoning)
Mythos-class uplift assessment	Medium—improves social-engineering yield but not the discriminator	Severe—sub-day discovery from public diff is paradigmatic	Severe—collapses the 6-week discovery window	Severe—multi-contract reasoning is core demonstrated capability

Uplift ratings are counterfactual analytical judgments, not empirical measurements: the four incidents predate the April 2026 Mythos Preview disclosure and no Mythos-class actor was involved. Scoring criteria: Severe = capability primitives in the operational definition (Section 2.1) map directly and would have collapsed time-to-discovery from weeks/months to hours-to-days; High = capability primitives accelerate but do not transform the attack class; Medium = capability primitives provide marginal uplift; Low = attack primarily relies on vectors outside the capability profile.

Three of the four bridge exploits this paper analyzes—Wormhole, Nomad, and Poly Network, accounting for roughly \$1.12 billion in losses—were on-chain code flaws discoverable through code review of public smart contract source [18,28,32,34]. None required social engineering, key compromise, or operational compromise. Each was a single-actor discovery; in two cases (Wormhole, Poly Network) by an unknown individual reading complex multi-contract logic in publicly available code.

This is precisely the capability class the operational Mythos-class definition (Section 2.1) describes. AISI reports the capability profile completing parts of multi-stage attack chains that would take human professionals days of work [5]; the bridge exploit history shows what happens when that level of capability is applied to public, deterministic, replayable code with high-value targets. The Wormhole case is particularly diagnostic: the vulnerability was visible in a public GitHub commit before the patch was deployed to mainnet, and an attacker found and weaponized it within hours [18]. A Mythos-class capability monitoring public bridge repositories is the limit case of that exact dynamic, generalized.

The Nomad case introduces a second-order risk the paper calls out explicitly: post-discovery dynamics. Once the initial Nomad transaction was on-chain, the exploit was trivially replayable by anyone capable of editing transaction calldata. Mandiant's reconstruction documents nearly 300 addresses participating, generating 960 transactions over roughly 150 min [19]; Coinbase's earlier post-incident analysis corroborates the general pattern of trivially replayable transaction calldata and rapid crowd participation [33]. MEV bots automated replication within minutes. A Mythos-class capability that surfaces a similarly structured vulnerability—and either publishes it or executes it observably—does not produce a single attacker; it produces a crowd-loot dynamic measured in minutes. The discovery-to-cascade pipeline collapses from the conventional weeks-to-months timeline to a window the defender cannot meaningfully respond within. The pre-discovery side of this dynamic is also notable: the vulnerable initialization state in the Nomad Replica contract persisted from the proxy upgrade in June 2022 until the August 2022 exploit, creating an approximately six-week exposure window [19].

The Ronin case is included as the architectural counterexample. Ronin's loss was operational (spear-phishing of a Sky Mavis engineer with a malicious PDF) and governance-related (a never-revoked Axie DAO validator delegation from November 2021) [27,30,31]. Mythos-class capability would offer marginal uplift on social-engineering execution but is not the primary discriminator of this attack class. The Ronin counterexample is analytically important: not all bridge risk is Mythos-class risk, and the defensive prescriptions for code-flaw bridges and operational/governance bridges are structurally different. Section 6 develops both prescriptions separately.

The AISLE distributional finding (Section 2.2) strengthens rather than weakens this analysis. If the analytical capability that surfaces Wormhole-, Nomad-, and Poly-class flaws is approachable from a wider set of model families than the maximalist Mythos framing implies, the threat actor pool against bridges expands accordingly [17]. The bridge architectural surface offers no friction against this expansion; the Section 2.3 friction-inversion conditions apply in full.

4.4. Comparative Synthesis

Figure 3 summarizes the comparative risk profile across the four highest-load architectural domains, drawn from the capability-surface-impact analysis in Section 3.1 and the architectural review in Sections 4.1–4.3. The “Client implementation surface” row rates the codebase exposed to autonomous-discovery review; client diversity is a mitigation that reduces single-implementation blast radius but is not the rated quantity. The “Native/bridge-mediated cross-chain surface” row corresponds to Table 2’s Bridge contracts column; the label distinguishes Bitcoin’s indirect cross-chain exposure—via wrapped-BTC (wBTC) representations on Ethereum—from Ethereum/L2’s broader cross-chain surface, while the underlying ratings derive from Table 2. Ethereum/L2 ratings are synthesized from Table 2’s capability-surface-impact matrix, scoped to the programmable surface—the Severe rating on the cross-chain row reflects Table 2’s Bridge contracts column (Severe in four of five capability primitives). Bitcoin ratings are derived from the architectural argument in Section 4.1: small non-Turing-complete Script surface, fifteen-plus years of adversarial pressure on consensus, conservative protocol evolution, and the absence of native cross-chain bridges other than wBTC bridging exposure on the Ethereum side.

	Bitcoin	Ethereum/L2
Consensus	Low	Low
Execution	Low	High
Client implementation surface	Medium	Medium
Native / bridge-mediated cross-chain surface	Low	Severe

■ Low
 ■ Medium
 ■ High
 ■ Severe

Figure 3. Architecture-aligned risk map: Mythos-class adversary exposure for Bitcoin (protocol surface) and Ethereum/L2 (programmable surface) across four risk domains—consensus, execution, client implementation surface, and cross-chain bridging. Ratings use the same four-level scale as Table 2 (Low/Medium/High/Severe). Bitcoin’s bounded execution and limited cross-chain surface produce concentrated low-risk exposure; Ethereum/L2’s expanding programmable surface produces High exposure on execution and Severe exposure on cross-chain bridging.

The architectural conclusion is that **programmability and bridge connectivity, not chain identity, determine Mythos-class exposure**. Bitcoin’s exposure under Mythos-class threat is bounded and concentrated in ecosystem layers, where conventional operational defenses retain force. Ethereum/L2 exposure is structurally expanding, and the bridges connecting the L1↔L2 stack and cross-chain ecosystem represent the single architectural surface most fully exposed to the Mythos-class capability profile. The paper’s policy and defensive recommendations (Sections 5 and 6) are constructed against this comparative finding: bridge architecture is the highest-priority defensive and governance target under Mythos-class threat, and the L2 expansion that has driven the most growth in Ethereum exposure has also driven the largest growth in the bridge attack surface.

5. Discussion: Systemic Risk Implications

The architectural findings in Section 4 establish that bridges and the broader programmable-contract surface are the architectural classes most fully exposed to Mythos-class capability. This section first states the systemic-risk synthesis method (Section 5.1), then develops the systemic implications of that exposure across three dimensions: liquidity cascade and composability dynamics (Section 5.2), governance and validator-level risk (Section 5.3), and the policy and regulatory posture appropriate to the threat (Section 5.4). The treatment is the systemic-implication side of Section 2.3's architectural argument: each subsection identifies a systemic-risk pathway that exists because on-chain systems lack the friction conventional environments provide, and that propagates through composability rather than being contained by it.

5.1. Method for the Systemic-Risk Synthesis

The systemic-risk implications developed in this section are synthesized through an explicit three-step method, stated here so that the pathways in Sections 5.2 and 5.3 are read as the output of a defined procedure rather than as narrative extrapolation. The method takes as input the architectural exposure ratings established in Section 4 (the capability-surface-impact matrix, Table 2, and the bridge uplift assessment, Table 5) and produces as output a small set of named systemic-risk pathways, each traced from a single architectural exposure to a system-level consequence. It is an analytical synthesis method, consistent with the framework-level positioning stated in Section 3.3; it does not estimate probabilities or loss magnitudes.

Step 1: propagation channel identification. For each architectural surface rated High or Severe in Section 4, the analysis identifies the mechanism by which a compromise of that surface propagates beyond the directly affected contract. Three propagation channels are recognized, corresponding to the composability, governance, and consensus properties developed in Section 2.3: value-dependency propagation (a compromised component is a collateral, oracle, or liquidity input to other protocols), authority propagation (a compromise changes protocol parameters or permissions rather than extracting a one-time loss), and replication propagation (a compromise creates a public, executable artifact that others reproduce). This channel taxonomy is drawn from the documented dynamics of the historical incidents analyzed in Section 4.3 and from the established systemic-risk literature on financial infrastructure, in which contagion is understood to travel through correlated exposures rather than through independent shocks [1,2].

Step 2: cascade tracing. Each identified channel is then traced forward to the point at which it produces a system-level rather than protocol-level consequence, and the trace is terminated at the first point where a friction mechanism could arrest it. Because the Section 2.3 friction-inversion argument establishes that the on-chain friction mechanisms are governance-dependent and often slower than exploit propagation, the traces are also annotated with whether any arresting mechanism (pause function, timelock, circuit breaker) is operationally available on the relevant timescale. This step is what distinguishes the two cascade pathways visualized in Figure 4: Pathway A (the liquidity cascade of Section 5.2) is a value-dependency trace, and Pathway B (the validator cascade of Section 5.3) is an authority-and-consensus trace that is explicitly marked as hypothesized because the historical case base does not yet establish its terminal consequence.

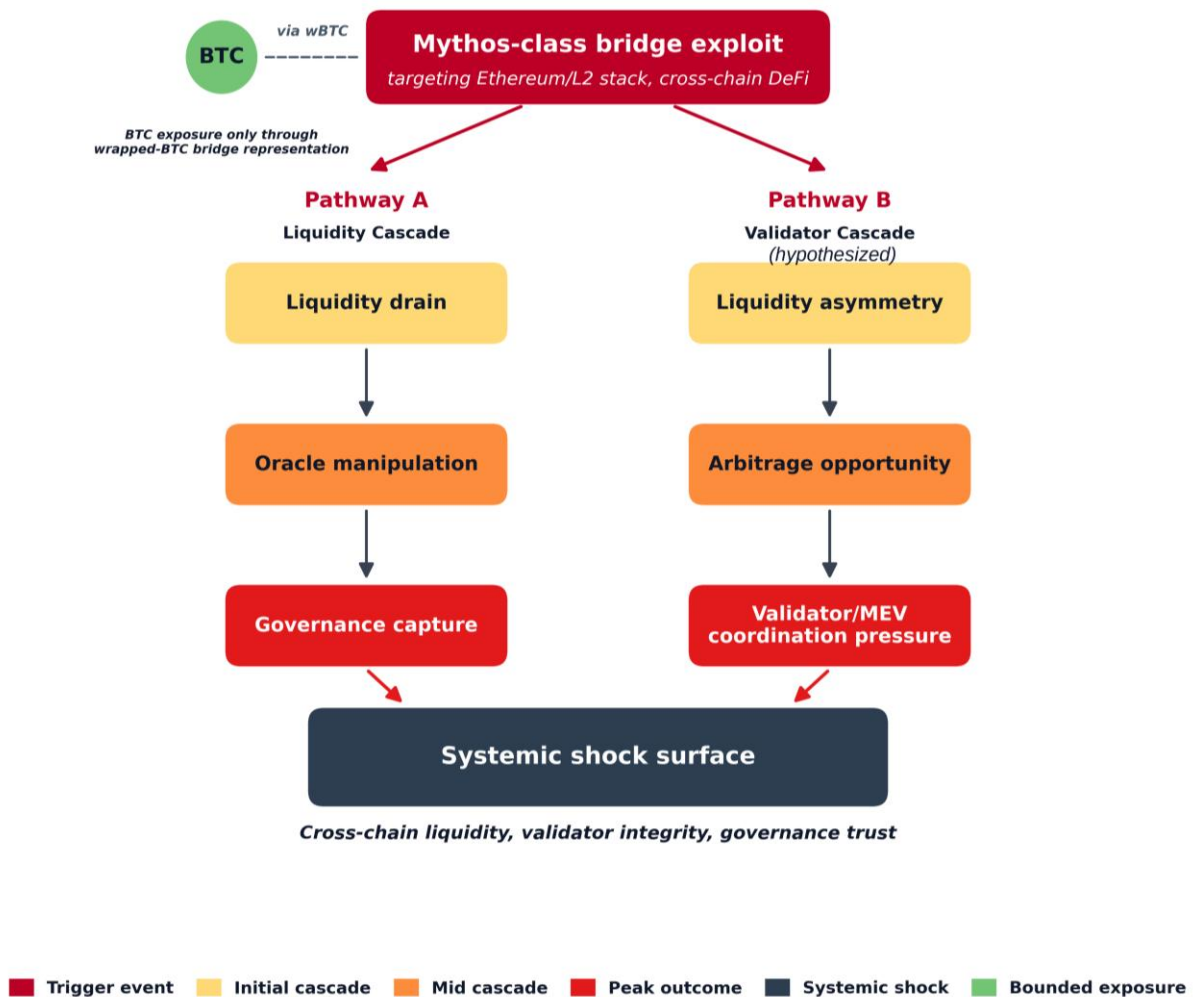


Figure 4. Cross-chain systemic risk pathways under Mythos-class threat. A Mythos-class bridge exploit (targeting Ethereum/L2 stack and cross-chain DeFi) bifurcates into two cascade pathways. Pathway A (Liquidity Cascade): liquidity drain → oracle manipulation → governance capture. Pathway B (Validator Cascade, hypothesized): liquidity asymmetry → arbitrage opportunity → validator/MEV coordination pressure. Both pathways converge on a systemic shock surface spanning cross-chain liquidity, validator integrity, and governance trust. Bitcoin appears peripherally: its exposure is only through the wrapped-BTC (wBTC) bridge representation, as labeled on the figure; the Bitcoin chain itself is not directly exploited. Stage color intensity escalates with cascade severity (light amber → orange → deep red); the trigger event and peak outcomes share the deep-red exposure tier.

Step 3: evidential grading. Each pathway is graded by the strength of its evidential basis, so that demonstrated dynamics are not conflated with plausible but unrealized ones. A pathway is treated as demonstrated where a historical incident exhibits the propagation channel directly (for example, the Terra/Luna contagion for value-dependency propagation, and the Nomad crowd-loot cascade for replication propagation), and as hypothesized where the channel is architecturally supported but has no realized system-level instance under the relevant conditions (for example, Mythos-class-driven validator or MEV-coordination reorg pressure). This grading is carried into the text through the language used to state each pathway, and it is the reason the paper marks the validator cascade as hypothesized while treating the liquidity cascade as demonstrated. The grading is deliberately conservative: where the evidence supports only the existence of a channel and not its terminal consequence, the weaker claim is the one made.

This three-step method is a specialization of the general kill-chain-and-vulnerability-class approach defined in Section 3.2, applied at the system level rather than the single-surface level: propagation-channel identification corresponds to extending the MITRE ATT&CK kill-chain past the exfiltration stage to downstream protocols, and evidential grading applies the same source-quality hierarchy stated in Section 3.3, under which demonstrated dynamics anchored in first-party incident reconstruction outrank commentary-supported extrapolation. The method inherits the framework's limitations accordingly: it produces an ordered, defensible account of which systemic pathways warrant the earliest attention, not a quantitative estimate of their likelihood or cost.

5.2. Liquidity Cascades and Composability Failures

Figure 4 visualizes the two cascade pathways that the remainder of this section and Section 5.3 develop. A bridge exploit triggers two parallel cascades: a liquidity cascade developed in this subsection (Pathway A: liquidity drain → oracle manipulation → governance capture), and a hypothesized validator cascade developed in Section 5.3 (Pathway B: liquidity asymmetry → arbitrage opportunity → validator/MEV coordination pressure). Both converge on a systemic shock surface affecting cross-chain liquidity, validator integrity, and governance trust. Bitcoin's exposure to these pathways is bounded to wBTC bridging only; the analysis below is therefore structurally specific to the Ethereum/L2 stack and the cross-chain DeFi infrastructure built on top of it.

DeFi composability propagates losses in ways that have no direct analog in conventional financial-system stress. Single-protocol failures cascade through every protocol that depends on the failed component as collateral, oracle input, or liquidity source. The 2022 Terra/Luna collapse demonstrated this dynamic without adversarial pressure, with stablecoin depeg propagating through interconnected lending markets and contributing to the failures of Three Arrows Capital, Celsius, and Voyager within weeks [1]. The contagion was structural rather than the result of coordinated attack.

A Mythos-class threat profile applied to this environment introduces adversarial pressure to a contagion mechanism that already operates without it. Three pathways are immediately concerning. **Autonomous discovery against composable lending and stablecoin protocols** identifies single-point-of-failure dependencies—oracle manipulation surfaces, collateral revaluation logic, liquidation thresholds—that propagate widely once exploited. **Multi-step chaining capability** matches the structural demands of flash-loan-enabled exploitation; the canonical multi-protocol DeFi exploit is structurally identical to the multi-host privilege escalation chains the Mythos-class profile has been demonstrated against [16]. **Agentic execution** under MEV-aware conditions introduces post-exploit cascade dynamics analogous to the Nomad crowd-loot finding (Section 4.3): an exploit exposing a composability flaw becomes immediately replayable by automated infrastructure within the same block window.

The defender position against this risk class is constrained by Section 2.3 friction-inversion conditions. There is no patch primitive for deployed composability surfaces; mitigation requires either contract migration (destabilizing under stress) or pre-existing pause functionality (multisig responsiveness measured in the same minutes the cascade operates within). Conventional financial-system circuit breakers have no functional analog at the protocol layer.

5.3. Governance, Validator, and Client-Diversity Risk

Two protocol-adjacent surfaces present systemic risk pathways distinct from the direct-exploitation pathways in Section 5.2.

Governance contracts controlling protocol parameters (interest rate curves, collateral factors, treasury allocations, upgrade authority) are a target class where successful exploi-

tation produces persistent rather than one-time loss. The Section 3.1 matrix rates governance High specifically on the multi-step chaining primitive, because governance attacks typically require composing token acquisition, voting power accumulation, proposal submission, and execution timing into a single coordinated sequence. Mythos-class capability is structurally well-matched to this composition. Historical precedent already exists at lower capability levels: the Beanstalk governance exploit (April 2022, ~\$182 M) used a flash-loan-funded governance proposal to drain the protocol within a single block [35]. Under Mythos-class threat, governance-contract exposure should be treated as a category equivalent to bridge exposure for systemic-risk purposes.

Validator client diversity at the consensus layer is a defense against single-implementation flaws reaching network-threshold share, but it is also an expansion of the code-base autonomous-discovery capability can read. A discovered flaw in a single consensus client implementation that reaches threshold network share is a consensus-layer risk despite client diversity providing baseline protection. Ethereum's current client distribution offers meaningful protection against this scenario; emerging L1 and L2 ecosystems with concentrated client implementations do not. The systemic-risk pathway is **monoculture under Mythos-class discovery pressure**: a capability profile that can autonomously read multiple client implementations and identify the one whose flaw reaches network-threshold consequence is a different threat from human-paced vulnerability research that has historically constrained this risk. Pathway B in Figure 4 is framed as validator/MEV coordination pressure rather than concrete reorg outcome because the four bridge case studies in Section 4.3 do not establish reorgs as a documented consequence; MEV-driven reorg risk is a known theoretical concern in the literature but its empirical realization under Mythos-class capability is open.

A third surface intersects with both concerns. **Multisig signer compromise**—whether through automated social-engineering generation, key-handling infrastructure compromise, or governance-process subversion—operates at the intersection of operational security and protocol governance. Mythos-class agentic-execution capability applied to social-engineering generation (across all signers in a multisig threshold simultaneously) presents a risk profile distinct from Ronin-style single-signer compromise. The defensive prescriptions in Section 6 treat this surface separately because its mitigation primitives (hardware-isolated key management, behavioral analytics, distributed signing thresholds) are operational rather than architectural.

5.4. Policy and Regulatory Implications

The institutional response to the April 2026 Mythos Preview disclosure was substantial and immediate: the Bank of England intensified AI risk testing [7]. None of this institutional response has yet translated into blockchain-specific regulatory frameworks adapted to the Mythos-class threat profile. The architectural findings in Section 4 and the systemic-risk pathways in Sections 5.2 and 5.3 have direct implications for what such frameworks should contain.

Three policy directions follow from the analysis.

First, **AI-aware blockchain security standards** should be developed as a category distinct from existing AI risk frameworks (which are oriented toward enterprise IT and consumer harm) and existing crypto-asset regulatory frameworks (which are oriented toward investor protection and anti-money-laundering). The friction-inversion argument from Section 2.3 implies that frameworks designed for either parent category will systematically underestimate blockchain-specific exposure. The relevant precedent is sector-specific cybersecurity standards in critical infrastructure (NERC CIP for the electric grid [36], NIST SP 800-82 for industrial control systems [37]); blockchain warrants treatment in this lineage.

Second, **bridge audit cadence and disclosure norms** require adaptation to compressed weaponization windows. Pre-deployment audit alone is insufficient when the Wormhole case demonstrates that public commits to source repositories can be reverse-engineered into mainnet exploits within hours. Disclosure regimes designed around vendor-coordinated patch cycles do not translate to immutable on-chain code. Specific recommendations include: continuous formal verification rather than discrete audit milestones; private-repository fix protocols with synchronized mainnet deployment; and structured pause-mechanism requirements as a condition of cross-chain bridge operation above defined TVL thresholds.

Third, **systemic-risk monitoring** should extend to the cross-chain bridge graph and high-TVL composability surfaces with an explicit Mythos-class threat assumption. Existing crypto-asset monitoring (Chainalysis, Elliptic, internal exchange systems) is oriented toward post-incident attribution and anti-money-laundering. The threat profile developed in this paper requires pre-incident posture: capability to identify exposure concentration, monitor public-repository patch lifecycles for high-TVL contracts, and intervene before crowd-loot dynamics begin. The Financial Stability Board's existing work on crypto-asset systemic risk [1,2] is the appropriate institutional anchor for this development.

The argument is not that regulation should constrain Mythos-class capability development—that question is being addressed by existing AI-governance initiatives at AISI and equivalent bodies. The argument is that **blockchain-specific defensive and governance frameworks should assume Mythos-class capability is a fixed feature of the threat environment and design accordingly**. The cost of frameworks that do not make this assumption is, on the Section 4.3 evidence base, measured in billion-dollar bridge losses and minute-scale crowd-loot cascades. The institutional response has begun; the blockchain-specific operationalization has not.

6. Defensive Frameworks and Mitigation

Constraint check. Section 1 stated three conditions under which the friction-inversion finding could be narrowed or invalidated. *Re-instantiability*: the on-chain analogs that exist—pause functions, timelocks, multisig emergency controls—are governance-dependent and operate on coordination timescales slower than the minute-scale cascade dynamics demonstrated by Nomad (Section 4.3); they introduce conditional, governance-mediated friction rather than restoring functional equivalents to the absent enterprise-IT primitives. *Marginal scope*: the bridge-contract column in Table 2 rates Severe across four of five primitives, the execution-layer column rates High across all five, and the systemic-risk pathways in Sections 5.2 and 5.3 implicate the entire programmable surface and its cross-chain bridge layer—not narrow architectural margins. The inversion is load-bearing across the load-bearing surfaces, not at the edges. *Capability reach*: the three on-chain code-flaw bridge exploits in Section 4.3—Wormhole, Nomad, Poly Network, totaling \$1.12 billion in losses—were achieved by adversaries operating without Mythos-class capability against the same architectural surfaces Mythos-class adversaries will face. The analytical capability has already reached the relevant surfaces in the historical record; the question is when it arrives at the same surfaces under operational conditions, not whether. Each constraint qualifies the friction-inversion finding without invalidating it; together they sharpen rather than narrow it.

The findings in Sections 4 and 5 establish that bridge architecture and the broader programmable-contract surface are the highest-priority defensive targets under Mythos-class threat, and that the friction-inversion conditions developed in Section 2.3 leave conventional defensive prescriptions inadequate. This section identifies three classes of defensive response (Sections 6.1–6.3), notes the AI-quantum convergence as adjacent future

work (Section 6.4), and outlines a research agenda the architectural findings in this paper open but do not close (Section 6.5).

6.1. AI-Aligned Defensive Modernization

The most direct response to Mythos-class offensive capability is symmetric defensive capability. Three defensive primitives are immediately applicable to the blockchain context.

Continuous formal verification at the bridge layer. Discrete pre-deployment audit is insufficient when the Wormhole case demonstrates that public source commits can be reverse-engineered into mainnet exploits within hours (Section 4.3) [18]. Continuous formal verification—automated proof maintenance against contract invariants, run continuously rather than at audit milestones—closes the source-to-deployment gap that Mythos-class capability exploits most directly. Recent advances in SMT-solver-driven verification of Solidity and Rust smart contracts (Certora, Halmos, Kontrol) provide the technical substrate; the gap is operational integration with deployment pipelines and bridge governance processes, not capability.

Defensive AI agents under explicit dual-use acknowledgment. The same capability profile that surfaces Wormhole-class flaws offensively can surface them defensively, and the Project Glasswing distribution model is the explicit institutional acknowledgment of this dynamic [3]. Blockchain-specific equivalents—autonomous defensive agents monitoring bridge contract state, public repository commits, and on-chain transaction patterns for anomalous activity—are technically straightforward but governance-complex. The dual-use bind is structural: a defensive agent capable of identifying a high-TVL bridge flaw is, by construction, capable of exploiting it. The dual-use considerations stated in the Responsible Disclosure back-matter apply here directly.

Bridge-class architectural hardening. The four cases in Tables 4 and 5 distribute across STRIDE categories that suggest specific architectural mitigations: spoofing-class flaws (Ronin, Wormhole) point to validator-set decentralization and signature-verification simplification; tampering-class flaws (Nomad) point to initialization-invariant testing and proxy-upgrade discipline; elevation-of-privilege flaws (Poly Network) point to least-privilege contract architecture and explicit cross-contract permission graphs. None of these prescriptions is novel as security engineering; what is novel is the requirement that they be applied with the assumption that adversary discovery time has collapsed from months to days.

6.2. Operational and Governance Hardening

The Ronin counterexample in Section 4.3 demonstrated that not all bridge risk is code-flaw risk. Operational and governance pathways require defensive prescriptions distinct from those in Section 6.1.

Hardware-isolated key management for validator and multisig signers reduces exposure to social-engineering and infrastructure-compromise pathways of the kind Lazarus Group exploited against Ronin [30]. Mythos-class agentic capability applied to social-engineering generation across all signers in a multisig threshold simultaneously is a categorical escalation; hardware isolation, distributed signing, and behavioral analytics for multisig-signer activity together constitute the relevant defensive surface.

Governance hygiene—exemplified by the never-revoked Axie DAO whitelist that converted a four-key compromise into a five-key compromise [27]—is a process category that has consistently produced the highest-severity bridge incidents. Defensive prescriptions are well-established (least-privilege defaults, expiring delegations, periodic permission review) but systematically underapplied. Mythos-class threat does not change what is required; it changes the cost of continued non-compliance.

Pause and circuit-breaker primitives at the protocol layer provide the closest functional analog to the response-time friction Mythos-class capability eliminates elsewhere. Multisig-controlled pause functionality on bridges and high-TVL composability surfaces, with response-time targets aligned to the minute-scale cascade dynamics demonstrated by Nomad [33], is an architectural requirement under Mythos-class threat assumption. This is one of the few defensive primitives that genuinely restores friction the Section 2.3 inversion has stripped.

6.3. Audit Cadence and Disclosure Adaptation

The Section 5.4 policy recommendations imply specific operational changes for the audit and disclosure ecosystem. Three are most consequential.

Audit cadence should shift from milestone-driven (deployment, major upgrade) to capability-driven (any change to deployed code, monitored continuously). Private-repository fix protocols with synchronized mainnet deployment address the Wormhole pattern directly: the source-patched-but-not-deployed window is precisely the high-risk interval Mythos-class capability exploits. Coordinated-disclosure norms adapted from conventional vulnerability research—mandatory pre-disclosure pause, time-bounded windows for migration, structured communication with downstream protocols—provide the institutional substrate this adaptation requires.

6.4. AI-Quantum Convergence as Adjacent Future Work

The defensive prescriptions in Sections 6.1–6.3 are constructed against a Mythos-class threat profile assumed to be a fixed feature of the near-term threat environment. A complete defensive architecture must also account for the convergence of AI-driven offensive capability with cryptographic erosion under quantum computing advance—particularly affecting the elliptic-curve signatures that secure validator authority, multisig governance, and wallet infrastructure across the entire blockchain ecosystem [38,39]. The convergence operates through two pathways: AI-driven optimization of cryptanalytic approaches against pre-PQC primitives, and AI-accelerated discovery of implementation flaws in PQC migration code itself. Both pathways reinforce the bridge-centric defensive priority developed in this paper, because cross-chain bridges concentrate cryptographic dependencies precisely where AI-driven offensive capability has the greatest leverage. Detailed treatment of the convergence is beyond this paper's scope; it is identified here as adjacent future work and the appropriate subject of separate analysis.

6.5. Research Agenda

The architectural findings open three research directions the present paper does not close. **Open benchmark suites for blockchain-specific Mythos-class evaluation** would convert the operational definition in Section 2.1 into an empirically testable instrument; existing AI-cybersecurity benchmarks are oriented toward operating-system and network-protocol targets, not Solidity, Rust BPF, or Move. **Cross-chain simulation environments** capable of replaying historical exploits against alternative architectural assumptions, and forward-modeling Mythos-class scenarios against current bridge graphs, are the empirical infrastructure Section 5.2 systemic-risk analysis requires to move from architectural reasoning to quantitative estimation. **Defensive-AI governance frameworks** addressing the dual-use bind in Section 6.1—including how the institutional asymmetry between Project Glasswing-style closed consortia and open defensive infrastructure should be governed—are a research direction this paper opens but cannot resolve, warranting treatment by the AI-governance and blockchain-governance research communities working in concert.

7. Conclusions

Mythos-class is a capability category, not a single vendor moment. Vendor-neutral, capability-defined frameworks are the correct unit of analysis; frameworks built around any specific model release will be obsolete before enforcement. The AISLE distributional finding indicates that the analytical primitive of the capability profile is already approachable from a wider set of model families than the maximalist Mythos framing implies, and that the agentic-execution boundary that genuinely separates frontier from broadly available capability will be approached on timescales relevant to blockchain governance—potentially measured in months rather than years under the distributional-capability assumption.

This paper has argued, as an analytical-framework contribution rather than an empirically validated one, that frontier autonomous offensive AI capability—operationalized through the five-primitive Mythos-class definition in Section 2.1—is a categorical rather than incremental threat to blockchain systems, and that the categorical character emerges not from the capability itself but from blockchain’s architectural inability to provide the friction that constrains the same capability elsewhere. The defensive-friction critique that materially weakens the case for enterprise-IT crisis under Mythos-class threat does not weaken the case for blockchain concern; it strengthens it. This is the friction inversion the paper has developed, and it ties the paper’s elements together: the vendor-neutral capability profile (Section 2.1), the comparative architectural analysis (Sections 4.1 and 4.2), the bridge case-study evidence (Section 4.3), and the systemic-risk pathways (Section 5) all hinge on a single architectural fact—that the friction primitives which weaken the enterprise-IT case under Mythos-class threat are absent at the deployed-contract layer. The category boundary is the absence of those primitives, not the capability of the adversary itself.

The comparative analysis isolates programmability and bridge connectivity, not chain identity, as the variables that determine Mythos-class exposure. Bitcoin’s protocol-layer exposure is bounded; ecosystem-layer exposure increases marginally without changing strategic profile. Ethereum and the L2 stack present an expanding programmability surface; bridges connecting the L1↔L2 stack and the cross-chain ecosystem represent the single architectural surface most fully exposed to the Mythos-class capability profile. The bridge attack history demonstrates that three of the four bridge exploits this paper analyzes—Wormhole, Nomad, Poly Network—were on-chain code flaws discoverable through code review of public smart contract source. Each was a single-actor discovery of exactly the analytical capability class the operational Mythos-class definition describes. The historical record does not require Mythos-class capability to explain; it does indicate what becomes possible when that capability is generally available against the same target class.

The systemic-risk pathways and defensive prescriptions developed in Sections 5 and 6 together describe a posture that does not yet exist at the institutional or protocol level, and that blockchain-specific regulatory and governance frameworks should be constructed in a lineage closer to NERC CIP [36] and NIST SP 800-82 [37] sectoral cybersecurity frameworks than to existing crypto-asset regulation.

The bridge attack history this paper analyzes was produced by adversaries operating at human capability levels against the same architectural conditions that will face Mythos-class adversaries. Over \$1.74 billion in losses across four exploits is the cost of the defensive posture the ecosystem held when capability was scarce. The defensive posture appropriate to capability that is no longer scarce has not yet been built. Building it is the work this paper is intended to support.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. The bridge-exploit case material analyzed is openly available through the public disclosure sources cited in references [18,19,27–34], covering the Ronin, Wormhole, Nomad, and Poly Network incidents. This paper analyzes publicly disclosed historical bridge exploits and a publicly disclosed AI capability profile; no novel vulnerabilities are revealed. The dual-use considerations associated with developing defensive frameworks against frontier offensive AI capability are addressed in Section 6.1; all defensive prescriptions in Section 6 are constructed to be informative to defenders without providing operational uplift to attackers.

Acknowledgments: During the preparation of this manuscript, the author used Claude Opus 4.7 (Anthropic, San Francisco, CA, USA) for editorial review, structural framing assistance, and reference integration. The author has reviewed and edited all output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

Abbreviation	Meaning
AISI	Artificial Intelligence Security Institute (UK)
BPF	Berkeley Packet Filter
DeFi	Decentralized Finance
EVM	Ethereum Virtual Machine
JIT	Just-In-Time (compilation)
L1	Layer 1 (blockchain)
L2	Layer 2 (blockchain)
MEV	Maximum Extractable Value
MITRE	MITRE Adversarial Tactics, Techniques, and Common Knowledge
ATT&CK	
NERC CIP	North American Electric Reliability Corporation Critical Infrastructure Protection
NIST	National Institute of Standards and Technology
PQC	Post-Quantum Cryptography
TVL	Total Value Locked
wBTC	Wrapped Bitcoin (Bitcoin tokenized as an ERC-20 on Ethereum)
wETH	Wrapped Ether (Ethereum’s native ETH tokenized as an ERC-20)

References

1. Financial Stability Board. The Financial Stability Risks of Decentralised Finance. 16 February 2023. Available online: <https://www.fsb.org/2023/02/the-financial-stability-risks-of-decentralised-finance/> (accessed on 3 May 2026).
2. Financial Stability Board and International Monetary Fund. G20 Crypto-Asset Policy Implementation Roadmap: Status Report. October 2024. Available online: <https://www.fsb.org/2024/10/g20-crypto-asset-policy-implementation-roadmap-status-report/> (accessed on 3 May 2026).
3. Anthropic. Project Glasswing: Securing Critical Software for the AI Era. 7 April 2026. Available online: <https://www.anthropic.com/project/glasswing> (accessed on 3 May 2026).
4. Kahn, J. Anthropic ‘Mythos’ AI model representing ‘step change’ in capabilities, leaked documents reveal. *Fortune*, 26 March 2026. Available online: <https://fortune.com/2026/03/26/anthropic-says-testing-mythos-powerful-new-ai-model> (accessed on 3 May 2026).
5. UK AI Security Institute. Our evaluation of Claude Mythos Preview’s cyber capabilities. *AISI Blog*, 13 April 2026. Available online: <https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities> (accessed on 3 May 2026).
6. Gilbert + Tobin. How Mythos-Class AI is Changing Cyber Security Risk. April 2026. Available online: <https://www.gtlaw.com.au/insights/how-mythos-class-ai-is-changing-the-cyber-security-risk> (accessed on 3 May 2026).

7. UK Parliament Treasury Committee. Bank of England and FCA Commit to Action on AI Following Warnings from MPs. 16 April 2026. Available online: <https://committees.parliament.uk/committee/158/treasury-committee/news/213162/bank-of-england-and-fca-commit-to-action-on-ai-following-warnings-from-mps/> (accessed on 3 May 2026).
8. Hu, S.; Huang, T.; İlhan, F.; Tekin, S.F.; Liu, L. Large Language Model-Powered Smart Contract Vulnerability Detection: New Perspectives. *arXiv* **2023**, arXiv:2310.01152. <https://doi.org/10.48550/arXiv.2310.01152>.
9. Liao, Z.; Nan, Y.; Liang, H.; Hao, S.; Zhai, J.; Wu, J.; Zheng, Z. SmartAxe: Detecting Cross-Chain Vulnerabilities in Bridge Smart Contracts via Fine-Grained Static Analysis. *Proc. ACM Softw. Eng.* **2024**, *1*, 249–270. <https://doi.org/10.1145/3643738>.
10. Wang, D.; Wu, B.; Yuan, X.; Wu, L.; Zhou, Y.; Cui, H. DeFiGuard: A Price Manipulation Detection Service in DeFi using Graph Neural Networks. *arXiv* **2024**, arXiv:2406.11157. <https://doi.org/10.48550/arXiv.2406.11157>.
11. Zhang, M.; Zhang, X.; Barbee, J.; Zhang, Y.; Lin, Z. SoK: Security of Cross-chain Bridges: Attack Surfaces, Defenses, and Open Problems. In *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '24)*; ACM: New York, NY, USA, 2024. <https://doi.org/10.1145/3678890.3678894>.
12. Wu, J.; Lin, K.; Lin, D.; Zhang, B.; Wang, J.; Zheng, Z. Safeguarding Blockchain Ecosystem: Understanding and Detecting Attack Transactions on Cross-chain Bridges. *arXiv* **2024**, arXiv:2410.14493. <https://doi.org/10.48550/arXiv.2410.14493>.
13. Gong, H. AI agents in financial markets: Architecture, applications, and systemic implications. *FinTech* **2026**, *5*, 34. <https://doi.org/10.3390/fintech5020034>.
14. Hamori, S. Artificial intelligence and financial market connectedness: Evidence from AI-related equities, cryptocurrencies, and global assets. *FinTech* **2026**, *5*, 40. <https://doi.org/10.3390/fintech5020040>.
15. Wilson, C.; Tam, C. Security challenges in open banking: A systematic review and conceptualisation of a tri-dimensional security framework. *FinTech* **2026**, *5*, 38. <https://doi.org/10.3390/fintech5020038>.
16. Anthropic Red Team. Assessing Claude Mythos Preview's Cybersecurity Capabilities. 7 April 2026. Available online: <https://red.anthropic.com/2026/mythos-preview/> (accessed on 3 May 2026).
17. Fort, S. AI Cybersecurity After Mythos: The Jagged Frontier. *AISLE*, 7 April 2026. Available online: <https://aisle.com/blog/ai-cybersecurity-after-mythos-the-jagged-frontier> (accessed on 3 May 2026).
18. Halborn. Explained: The Wormhole hack (February 2022). *Halborn Blog*, February 2022. Available online: <https://www.halborn.com/blog/post/explained-the-wormhole-hack-february-2022> (accessed on 3 May 2026).
19. Eitzman, R.; Dobson, J. Decentralized robbery: Dissecting the Nomad Bridge hack and following the money. *Mandiant/Google Cloud Threat Intelligence*, 29 November 2022. Available online: <https://cloud.google.com/blog/topics/threat-intelligence/dissecting-nomad-bridge-hack> (accessed on 3 May 2026).
20. Buterin, V. Ethereum: A next-generation smart contract and decentralised application platform. *Ethereum White Paper*, 2014. Available online: <https://ethereum.org/en/whitepaper/> (accessed on 3 May 2026).
21. Antonopoulos, A.M. *Mastering Bitcoin: Programming the Open Blockchain*, 2nd ed.; O'Reilly Media: Sebastopol, CA, USA, 2017.
22. Wood, G. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Yellow Paper*, 2014. Available online: <https://ethereum.github.io/yellowpaper/paper.pdf> (accessed on 3 May 2026).
23. Ethereum Foundation. Nodes and Clients: Client Diversity. 2024. Available online: <https://ethereum.org/developers/docs/nodes-and-clients/client-diversity> (accessed on 3 May 2026).
24. L2BEAT. Layer 2 Ecosystem Statistics. 2026. Available online: <https://l2beat.com> (accessed on 3 May 2026).
25. Strom, B.E.; Applebaum, A.; Miller, D.P.; Nickels, K.C.; Pennington, A.G.; Thomas, C.B. *MITRE ATT&CK: Design and Philosophy*; MITRE Corporation Technical Report MP180360R1; MITRE Corporation: McLean, VA, USA, 2020. Available online: https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf (accessed on 3 May 2026).
26. Shostack, A. *Threat Modeling: Designing for Security*; Wiley: Indianapolis, IN, USA, 2014.
27. SlowMist Security Team. Ronin exploit, largest crypto hack to date. *SlowMist Medium*, 27 April 2022. Available online: <https://slowmist.medium.com/ronin-exploit-largest-crypto-hack-to-date-8b7c581e38fd> (accessed on 3 May 2026).
28. Kraken Security Labs. Abusing smart contracts to steal \$600 million: How the Poly Network hack actually happened. *Kraken Blog*, 22 September 2021. Available online: <https://blog.kraken.com/product/security/abusing-smart-contracts-to-steal-600-million-how-the-poly-network-hack-actually-happened> (accessed on 3 May 2026).
29. Nelson, D. Oasis Exploits Its Own Wallet Software to Seize Crypto Stolen in Wormhole Hack. *CoinDesk*, 24 February 2023. Available online: <https://www.coindesk.com/business/2023/02/24/oasis-exploits-its-own-wallet-software-to-seize-crypto-stolen-in-wormhole-hack> (accessed on 3 May 2026).
30. Sky Mavis. Back to building: Ronin security breach postmortem. *Ronin Blog*, 27 April 2022. Available online: <https://ronin-chain.com/blog/posts/back-to-building-ronin-security-breach-6513cc78a5edc1001b03c364> (accessed on 3 May 2026).

31. Elliptic. North Korea's Lazarus Group identified as exploiters behind \$540 million Ronin bridge heist. *Elliptic Blog*, 14 April 2022. Available online: <https://www.elliptic.co/blog/540-million-stolen-from-the-ronin-defi-bridge> (accessed on 3 May 2026).
32. CertiK. Wormhole bridge exploit incident analysis. *CertiK Blog*, 1 August 2022. Available online: <https://www.certik.com/resources/blog/wormhole-bridge-exploit-incident-analysis> (accessed on 3 May 2026).
33. Kacherginsky, P.; Wilder, H. Nomad Bridge incident analysis. *Coinbase Blog*, 9 August 2022. Available online: <https://medium.com/the-coinbase-blog/nomad-bridge-incident-analysis-899b425b0f34> (accessed on 6 July 2026).
34. CertiK. Nomad bridge exploit incident analysis. *CertiK Blog*, 1 August 2022. Available online: <https://www.certik.com/blog/nomad-bridge-exploit-incident-analysis> (accessed on 3 May 2026).
35. Halborn. Explained: The Beanstalk Hack (April 2022). *Halborn Blog*, April 2022. Available online: <https://www.halborn.com/blog/post/explained-the-beanstalk-hack-april-2022> (accessed on 3 May 2026).
36. North American Electric Reliability Corporation. *Critical Infrastructure Protection (CIP) Reliability Standards*; NERC: Atlanta, GA, USA, 2026. Available online: <https://www.nerc.com/standards/reliability-standards/cip> (accessed on 6 July 2026).
37. Stouffer, K.; Pease, M.; Tang, C.Y.; Zimmerman, T.; Pillitteri, V.; Lightman, S.; Hahn, A.; Saravia, S.; Sherule, A.; Thompson, M. *Guide to Operational Technology (OT) Security*; NIST Special Publication 800-82, Revision 3; NIST: Gaithersburg, MD, USA, 2023. <https://doi.org/10.6028/NIST.SP.800-82r3>.
38. Campbell, R.E., Sr. Evaluation of post-quantum distributed ledger cryptography. *J. Br. Blockchain Assoc.* **2019**, *2*, 1–8. [https://doi.org/10.31585/jbba-2-1-\(4\)2019](https://doi.org/10.31585/jbba-2-1-(4)2019).
39. Campbell, R., Sr. Hybrid post-quantum signatures for Bitcoin and Ethereum: A protocol-level integration strategy. *J. Br. Blockchain Assoc.* **2026**, *9*, 1–13. [https://doi.org/10.31585/jbba-9-1-\(2\)2026](https://doi.org/10.31585/jbba-9-1-(2)2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.