

Global AI Safety Notary: A Decentralised Protocol for International AI Incident Reporting

Augusta Lina¹, Caleb Rudnick², and Dennis Kiptanui³

With Apart Research

Abstract.

International AI governance suffers from a fundamental trust deficit. Nations are reluctant to report AI safety incidents to centralised bodies controlled by potential adversaries, creating dangerous information silos where critical safety lessons go unshared. We present Global AI Safety Notary, a decentralised incident reporting platform that creates immutable, verifiable records of AI safety incidents on the Ethereum blockchain. Built on the Ethereum Attestation Service (EAS), the platform enables researchers, developers, and organisations to report and track AI failures, including bias, safety violations, and data leaks, with full transparency and accountability. By leveraging blockchain technology, our system creates a sovereign-neutral registry where no single government or institution controls the data. All participants, including geopolitical rivals, can verify that records are tamper-proof and that no party has privileged access. We demonstrate a working prototype deployed on Ethereum Sepolia that enables structured incident submission, searchable registry browsing, and independent blockchain verification through EASScan. Global AI Safety Notary provides the technical infrastructure for evidence-based international AI policy, enabling cooperation without requiring trust.

1. Introduction

As AI systems become more capable, the risks they pose increasingly cross national borders. A model trained in one country can cause harm in another; a safety incident at one lab may reveal vulnerabilities shared across the industry. The EU AI Act took effect in August 2024, establishing the first comprehensive regulatory framework for AI systems [5]. However, practical tools for international incident coordination remain scarce. The core problem is trust. When an AI safety institute in one nation discovers a critical vulnerability, sharing that information with a centralised international body controlled by a potential adversary creates risks: the information could be weaponised, access could be asymmetric, or records could be altered. This creates dangerous information silos where critical safety lessons go unshared, increasing global risk while nations remain blind to incidents elsewhere.

This trust deficit has precedent in other safety-critical domains. Aviation safety reporting faced similar challenges before frameworks like the Aviation Safety Reporting System (ASRS) established confidential, non-punitive reporting mechanisms. Healthcare incident reporting evolved through systems like the Patient Safety Reporting System. More recently, researchers have proposed blockchain-based approaches for construction safety [1], cybersecurity incident management [10], and healthcare adverse event tracking [6]. We present Global AI Safety Notary, a decentralised platform that addresses the international AI governance trust deficit through blockchain technology. Our contributions are:

1. A structured schema for AI safety incidents compatible with emerging governance frameworks, capturing incident type, severity, model identifier, and mitigation status.
2. An implementation using the Ethereum Attestation Service (EAS) that creates immutable, cryptographically verifiable incident records without requiring trust in a central authority.
3. A working prototype demonstrating the complete workflow: incident submission via Web3 wallet, on-chain attestation, and independent verification through public blockchain explorers.
4. Analysis of how decentralised infrastructure enables international cooperation by removing the political barrier of centralised control.

2. Background & Related Work

2.1. Blockchain for Safety Incident Reporting

Multiple domains have explored blockchain and distributed ledger technology for safety incident management. Aktas et al. [2] proposed a blockchain framework for aviation safety reporting, arguing that distributed ledgers could address confidentiality concerns while maintaining auditability. Morteza et al. [9] developed a blockchain information management framework for construction safety, demonstrating how permissioned ledgers enable multi-stakeholder coordination. In healthcare, Marbough et al. [6] presented a blockchain-based incident reporting system for patient safety, emphasising the importance of immutability and provenance for medical error tracking. Milne-Ives et al. [8] proposed DeLIVER, a distributed ledger infrastructure for adverse event reporting with privacy-preserving features. For cybersecurity, Olukoya [10] analysed blockchain requirements for digital investigation platforms, and Preuveneers et al. [11] developed a distributed security framework for threat intelligence sharing. These systems share common requirements, specifically immutability, decentralised trust and auditability [3].

2.2. Ethereum Attestation Service

The Ethereum Attestation Service (EAS) is an open-source infrastructure for creating and verifying on-chain attestations. Unlike custom smart contracts, EAS provides a standardised protocol where attesters can make structured claims about any subject using custom schemas. EAS attestations are stored on-chain, queryable via GraphQL, and independently verifiable through EASScan explorers. This infrastructure is deployed on Ethereum mainnet and multiple testnets including Sepolia.

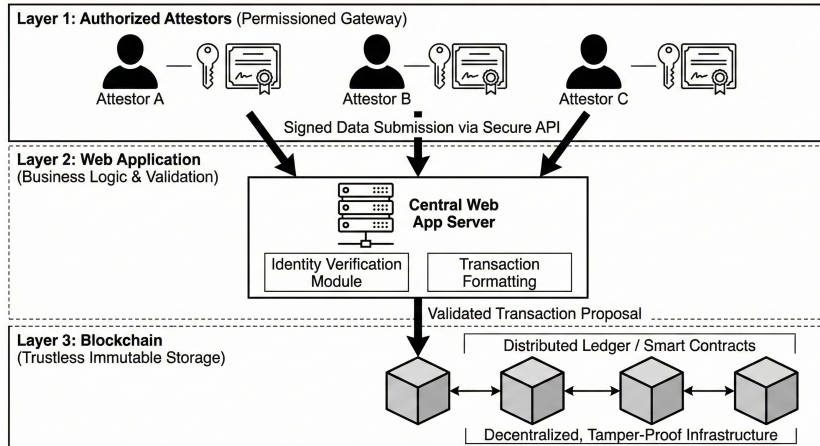


Figure 1: System architecture showing the three-layer design: authorised attestors at the top, the web application in the middle, and the blockchain infrastructure at the bottom.

2.3. International AI Governance Context

Current international AI governance relies on voluntary frameworks and national regulations without robust verification infrastructure. The EU AI Act establishes risk-based requirements but compliance verification remains challenging. The Bletchley Declaration (2023) acknowledged the need for international cooperation on AI safety but provided no technical mechanisms. Proposals for international AI safety institutes, analogous to the International Atomic Energy Agency (IAEA), face the fundamental challenge of verification between parties without full trust. How can nations confirm that incident reports are accurate without privileged access? How can they ensure records are not altered or selectively disclosed? Blockchain technology directly addresses these challenges. A decentralised ledger provides:

- **Sovereign neutrality:** No nation “owns” the registry
- **Immutability:** Records cannot be altered retroactively

- **Transparency:** All participants see the same data
- **Cryptographic verification:** Attestations are signed by known parties
- **No privileged access:** Data is equally available to all

3. Methods

Global AI Safety Notary implements a three-layer architecture (Figure 1):

1. **Presentation Layer:** Next.js web application with Web3 wallet integration
2. **Blockchain Layer:** Ethereum Attestation Service on Sepolia testnet
3. **Query Layer:** GraphQL API for retrieving and filtering attestations

Users connect via MetaMask or compatible Web3 wallets. Incident submissions create on-chain attestations through the EAS smart contract. The registry queries attestations via the EAS GraphQL endpoint and displays them with links to EASScan for independent verification.

4. Implementation

4.1. Technology Stack

The prototype was implemented using Next.js 16 with React 19 as the frontend framework, providing a modern web application architecture with server-side rendering capabilities. For blockchain integration, we deployed on Ethereum Sepolia testnet, enabling cost-free testing while maintaining compatibility with mainnet deployment. Attestation functionality was implemented using the official @ethereum-attestation-service/eas-sdk, which provides a standardised interface for creating and querying on-chain attestations. Web3 wallet connectivity was achieved through a combination of wagmi for React hooks, viem for low-level Ethereum interactions, and RainbowKit for a polished wallet connection interface supporting MetaMask and other popular wallets. The user interface was styled using Tailwind CSS, enabling rapid development of a responsive, accessible design.

4.2. Core Components

The SchemaEncoder from the EAS SDK handles data serialisation. The attestation creation function connects to the EAS contract and submits the encoded incident data. GraphQL queries retrieve attestations filtered by schema, enabling the registry to display all incidents matching our custom schema.

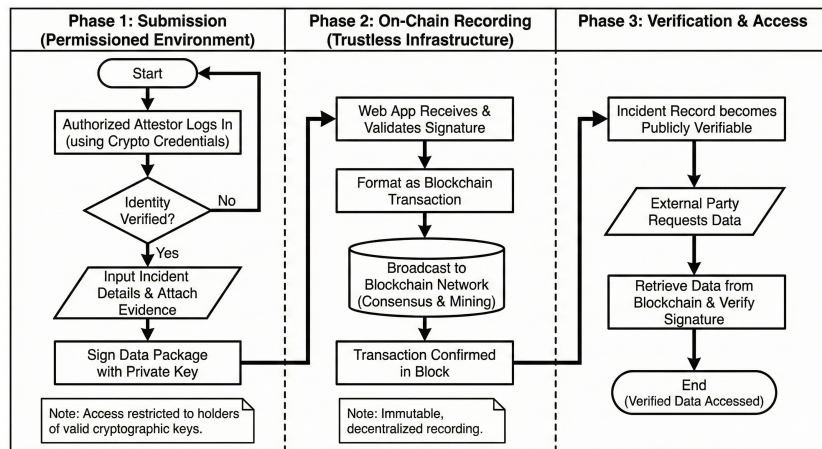


Figure 2: Incident attestation lifecycle showing the three phases: submission in a permissioned environment, on-chain recording via trustless infrastructure, and verification & access.

5. Discussion

5.1. Addressing the International Trust Deficit

Global AI Safety Notary demonstrates how blockchain infrastructure can enable international cooperation without requiring trust in central authorities. When organisations report incidents to the decentralised registry rather than a national agency, they contribute to a shared global knowledge base while retaining sovereignty—they do not cede authority to any single nation. Participants can independently verify that records are tamper-proof through cryptographic proofs, and they have assurance that no party has privileged access to the data. This creates a credible commitment mechanism for international AI safety coordination. Even geopolitical rivals can participate in the same registry because the blockchain’s cryptographic guarantees replace institutional trust.

5.2. Permissioned Attestors: Leveraging the Development Bottleneck

The notary system can theoretically be applied in a context where only authorised organisations are able to submit based on a database of approved organisation hashes. This prevents against false submissions which result from permissionless participation. The authorised contributor approach makes use of the bottleneck of major frontier AI contributors, whereby only a few players are currently contributing to the advancement of frontier AI models and compute resources. In 2024, fewer than twenty laboratories worldwide possessed the compute, talent, and capital to train frontier models [4]. This development bottleneck creates a natural quality control mechanism. Rather than allowing arbitrary wallet addresses to submit incidents, Global AI Safety Notary is designed to support an allowlist of authorised attestors:

Authorised Attestor Categories:

- **AI Safety Institutes:** National bodies such as the UK AI Safety Institute, US AI Safety Institute, and equivalents in other jurisdictions
- **Frontier AI Laboratories:** Organisations developing models above defined capability thresholds (e.g., 10^{25} FLOP training compute)
- **Accredited Auditors:** Third-party evaluation organisations with demonstrated technical competence
- **Academic Research Groups:** University-affiliated AI safety researchers with institutional accountability

This permissioned model provides several advantages. Incidents submitted by authorised parties carry institutional credibility, as false or frivolous reports risk reputational damage to the submitting organisation, creating natural incentives for accuracy. Because attestor addresses map to known organisations, the registry becomes a tool for accountability tracking, monitoring which labs are reporting incidents and which are not. Patterns of non-reporting may themselves be significant signals worthy of regulatory attention. Authorised attestor status can be linked to regulatory compliance, enabling labs required to report incidents under frameworks like the EU AI Act to demonstrate compliance through on-chain attestations from their registered wallet addresses. This approach supports scalable governance: as more organisations develop frontier capabilities, the allowlist can expand, and the blockchain infrastructure scales independently of the number of authorised parties.

Implementation uses EAS’s built-in filtering: the registry interface queries only attestations from addresses on the authorised list, while the underlying blockchain remains permissionless for transparency. Any party can verify that authorised attestors submitted specific records, but only authorised parties can add records to the curated registry view. This approach transforms the concentrated structure of frontier AI development from a governance challenge into an advantage. With few players, comprehensive monitoring becomes tractable. Every major laboratory can be assigned a cryptographic identity, and their reporting behaviour, or absence thereof, becomes part of the permanent, auditable record.

5.3. Alignment with Governance Frameworks

Our incident taxonomy aligns with the EU AI Act’s risk categories and emerging frameworks from AI safety institutes. The structured schema enables aggregation of incidents by type for policy analysis, allowing regulators and researchers to identify systemic patterns across AI systems. Severity-based classification supports prioritisation for regulatory attention, ensuring critical incidents receive appropriate scrutiny. The decentralised nature of the registry facilitates cross-border incident tracking and pattern detection, as all participating nations access the same immutable dataset. Furthermore, the timestamped record of incidents and their resolution enables longitudinal analysis of mitigation effectiveness, providing empirical grounding for evidence-based AI policy.

5.4. Limitations

Current limitations of the system should be acknowledged. The prototype is deployed on Ethereum Sepolia testnet; production use would require mainnet deployment with associated transaction costs and additional security considerations. While attester addresses provide cryptographic identity, they do not inherently verify real-world identity, meaning anonymous reporting remains possible without additional identity verification mechanisms. Large evidence files such as screenshots and logs require off-chain storage solutions, as the current implementation stores only content hashes rather than full evidence. The community verification workflow creates new attestations referencing original reports but lacks formal multi-signature governance for disputed incidents. Finally, schema evolution presents challenges: changes to the attestation schema require new registration on EAS, and migrating historical data to updated schemas is non-trivial.

6. Conclusion

We presented Global AI Safety Notary, a decentralised incident reporting platform that addresses the international trust deficit in AI governance. By leveraging the Ethereum Attestation Service, our system creates immutable, verifiable records of AI safety incidents without requiring trust in central authorities. Our working prototype demonstrates the complete workflow from incident submission through on-chain attestation to independent blockchain verification. The structured schema captures essential incident information while remaining compatible with emerging governance frameworks. As AI systems become more capable and their risks more global, the infrastructure for international cooperation becomes critical. Current proposals for AI safety institutes and international agreements assume technical verification mechanisms that largely do not exist. Global AI Safety Notary contributes a practical building block: a sovereign-neutral, tamper-proof registry where any nation or organisation can report incidents with confidence that records cannot be altered, access cannot be restricted, and verification is independent. Decentralisation removes the political barrier. When the infrastructure is trustless, cooperation becomes possible even between parties who cannot trust each other. This is the technical foundation that international AI governance requires.

References

- [1] S. Ahmadiheykhsarmast, S. Aminbakhsh, R. Sonmez, and F. Uysal, “A transformative solution for construction safety: Blockchain-based system for accident information management,” *Journal of Industrial Information Integration*, vol. 35, p. 100491, 2023.
- [2] E. Aktas, S. Demir, and T. Paksoy, “The Use of Blockchain in Aviation Safety Reporting Systems: A Framework Proposal,” *The International Journal of Aerospace Psychology*, vol. 32, pp. 283–306, 2022.
- [3] E. Diallo, R. Abdallah, M. Dib, and O. Dib, “Decentralized Incident Reporting: Mobilizing Urban Communities with Blockchain,” *Smart Cities*, vol. 7, no. 4, p. 90, 2024.
- [4] Epoch AI, “Trends in Machine Learning Hardware,” Epoch AI, 2024. [Online]. Available: <https://epochai.org/>
- [5] European Commission, “Artificial Intelligence Act,” *Official Journal of the European Union*, 2024.

- [6] D. Marbough, M. Simsekler, K. Salah, R. Jayaraman, and S. Ellahham, “Blockchain-based Incident Reporting System for Patient Safety and Quality in Healthcare,” in *Trust Models for Next-Generation Blockchain Ecosystems*, 2021.
- [7] S. McGregor, “Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database,” in *AAAI Conference on Artificial Intelligence*, 2021.
- [8] M. Milne-Ives, C. Lam, N. Rehman, R. Sharif, and E. Meinert, “Distributed Ledger Infrastructure to Verify Adverse Event Reporting (DeLIVER): Proposal for a Proof-of-Concept Study,” *JMIR Research Protocols*, vol. 10, 2021.
- [9] A. Morteza, M. Ilbeigi, and J. Schwed, “A Blockchain Information Management Framework for Construction Safety,” in *Computing in Civil Engineering 2021*, 2022.
- [10] O. Olukoya, “Distilling blockchain requirements for digital investigation platforms,” *Journal of Information Security and Applications*, vol. 62, p. 102969, 2021.
- [11] D. Preuveneers, W. Joosen, J. Bernabé, and A. Gómez-Skarmeta, “Distributed Security Framework for Reliable Threat Intelligence Sharing,” *Security and Communication Networks*, vol. 2020, 2020.