AI Safety Startup - Identity System for AIs

Artem Grigor
UCL

This proposal addresses the urgent need for robust identification and verification systems for AI models, particularly as AI becomes more integrated into daily life and increasingly autonomous. The problem lies in the lack of a standardized method to trace the origins and actions of AI models, which poses significant risks including misinformation, malicious use, and the potential for AI to operate without accountability. Our solution leverages cryptographic techniques to assign unique identities to AI models and generate verifiable proofs for their outputs. By implementing this system, we ensure that every AI action is traceable, authenticated, and transparent, thereby enhancing AI safety and establishing a foundation for responsible AI governance.

1. Motivation

The motivation for this proposal comes from a recent work on IDs for AI systems paper which has shown how a rather trivial yet vital problem has seemingly gatherned little attention recently.

2. Problem overview

As AI systems, particularly large language models (LLMs), become deeply integrated into our daily lives, they are transforming how we work, communicate, and make decisions. AI tools are now as ubiquitous as smartphones, enhancing productivity and enabling tasks that were once unimaginable. However, just as cars, firearms, and pharmaceuticals require stringent identification and regulation due to their potential for harm, so too do AI systems. The absence of a standardised identification mechanism for AI models presents significant risks, both now and in the future.

Currently, AI systems operate without clear, enforceable identities, making it difficult to track the origin and accountability of their outputs. This gap leaves room for misuse—ranging from the spread of deepfakes and misinformation to more malicious activities like automated hacking or the creation of undetectable malware. For instance, the rapid spread of AI-generated fake news during election cycles or the creation of realistic yet false images of public figures has shown how easily AI can be weaponized without traceability. Without the ability to trace these actions back to a specific AI model or agent, it becomes nearly impossible to hold the responsible parties accountable, eroding public trust in AI technologies.

Looking ahead, the challenge intensifies as AI evolves from a mere tool into autonomous agents capable of making independent decisions. Imagine a scenario where one AI impersonates another, potentially leading to fraudulent transactions, unauthorized access to sensitive data, or even automated decision-making that bypasses human oversight. Without robust identification mechanisms, these scenarios could become reality, posing existential risks to both individual privacy and global security. Just as VINs and serial numbers have become indispensable in regulating vehicles and firearms, a similar system is urgently needed for AI—one that is cryptographically secure, immutable, and universally recognized.

3. Your solution

In addressing the urgent need for AI agent identification and accountability, we propose a robust cryptographic system that assigns unique, verifiable identities to machine learning models and links all their outputs to them.

Components:

- **Model Identity via Cryptographic Hashing**: Every AI model is assigned a unique identity using a cryptographic SHA256 hash, which acts as a digital fingerprint. This identity is immutable and ensures that each model instance is uniquely identifiable and deterministic.
- **Tracking AI Outputs with Cryptographic Stamps**: very piece of data generated by the AI is stamped with cryptographic identifiers, which is an unforgeable link to the model identity that produced this output using cryptographic SNARKs. This guarantees that every output is traceable, preventing unauthorised alterations and ensuring transparency and authenticity of the AI's actions.
- **Publically Available Verification System**: Finally, we have a way for anyone to validate that the certificate is actually valid and has not been forged by an adversary, enabling trust in the system.

3. Additional Features

Multi-Layered Identity Structure: The system can support a more granular level of identification beyond just the model itself, to include its fine-tuned versions or specific configurations (e.g., user-defined prompts). This ensures that each output can be traced back not just to a generic model but to the exact instance and context in which it was used.

Solving Scalability: Most times a cryptographic system is applied to ML, it does not work on any SOTA models, such as GPT-4, due to their complexity. However, here, we can scale our system independent of the size of the models if we put some assumption into trusted participation, such as OpenAI to be honest and use tools as TLSNotary to avoid SNARK costs.

4. Pilot experiment or demo

Objective of the Pilot

The primary goal of our pilot was to demonstrate the feasibility of assigning unique digital identities to AI models and verifying the authenticity of their outputs. We aimed to create a practical proof-of-concept that would serve as a foundation for developing a robust framework for AI agent identification and accountability.

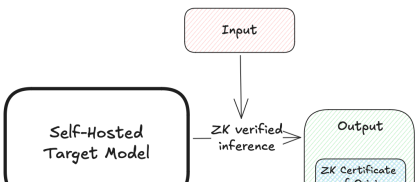
Methodology and Tools Used

We focused on two scenarios: **self-hosted models** and **server-hosted models**.

1. **Self-hosted Models**: The user has access to the model and is running it locally. We used the EZKL library to generate cryptographic proofs and create unique identifiers for AI models, ensuring that each model instance had a distinct, traceable identity.
2. **Server-hosted Models**: For models accessed via a black box API, such as SOTA OpenAI's ChatGPT, we explored the use of the TLS Notary system. This system allows us to prove that specific data was received from a model provider, leveraging the security of TLS communications and Multi-Party Computations.

Our work is available in two Github repositories:

- **Self-hosted Models**: ai-id-hackathon (video demo)
- **Server-hosted Models**: tlsn-openai

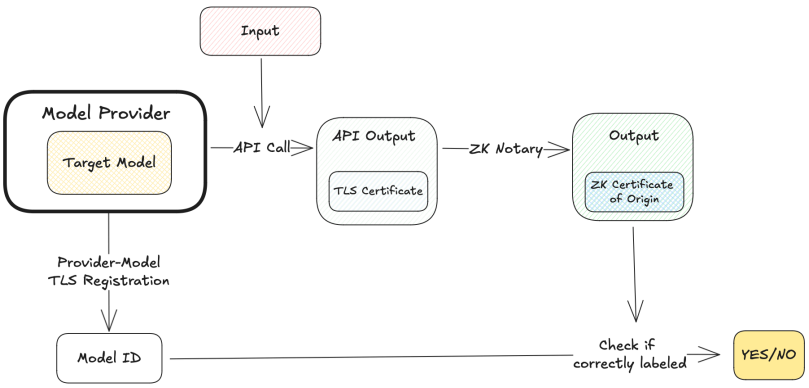


Process for Self-Hosted Models

1. **Model ID Generation**: We began by writing a program to generate identities for models passed in the ONNX format. We did that by generating a unique SHA256 hash as the model's digital fingerprint, which was saved alongside the model for future verification.
2. **Proof Generation**: We then wrote a program to generate a cryptographic proof that ties specific outputs to the model's identity, ensuring that the output was produced by the identified model instance. This proof was saved in a *proof.json* file, and an origin certificate was created to document the model's ID and the time of proof generation.
3. **Verification**: Finally, we developed a script to verify the proof against the model's ID. The script confirms the integrity of the process by ensuring that the proof and model ID match, outputting a success message if the verification is successful.

Process for Server-Hosted Models

For server-hosted models, such as those accessed via APIs (e.g., OpenAI's ChatGPT), we used the TLS Notary system. This system allows us to create a verifiable chain of trust that proves data was received from the API, tied to the model provider. While we successfully demonstrated this approach, we encountered challenges in creating a user-friendly web interface and could not complete the full implementation within the hackathon timeframe. Overall, this method has proved to be scalable and able to create unique identities for even very large models.



Result

Our pilot successfully demonstrated that it is feasible to assign unique identities to AI, even large black box models and verify their outputs using cryptographic methods. The system we developed is scalable and can be integrated into larger frameworks for AI governance and accountability.

5. Process

Timeframe	What will you do?
Next 3 months	**Prototype Expansion:** Enhance prototype features and robustness, develop use cases for Web2 (AI agents acting via APIs). Identify and engage potential partners in AI safety. Collect feedback from early users to refine the prototype.
2025	**Technical Development**: Integrate zkTLS for secure interactions with third-party AI providers like ChatGPT. Develop frameworks for AI agent authentication in Web3 and adapt for Web2. **Open-Source & Partnerships:** Launch an open-source version

	of the authentication framework with tools for developers. Partner with major AI providers to integrate identity systems and scale infrastructure.
2026	**Combatting Unmarked Content**: Advocate for mandatory AI identities and develop content marking standards. Create an AI safety dashboard for tracking and verifying LLM outputs. Implement behavioural authentication to detect unmarked AI content.
2027	**Expansion & Global Impact**:Develop deepfake mitigation solutions using AI identification frameworks. Educate the public on AI content verification and the risks of unmarked AI outputs. Work towards global standards for AI identification, verification, and content marking.

6. Impact on AI safety & key risks

Proactive Risk Identification:
The system enables the tracking of AI models, allowing us to identify and respond to emerging safety concerns in real-time. If a particular model is repeatedly linked to harmful behavior or misuse, our system can flag it for further investigation and potential restriction, thereby preventing further harm.

Combating Misinformation:
By associating every AI-generated output with a verifiable origin certificate, our solution makes it possible to distinguish between genuine and manipulated content. This is crucial in preventing the spread of misinformation, particularly in sensitive areas like news, politics, and public health, where the authenticity of information is paramount.

Preventing AI Manipulation:
Our system prevents AI models from masking their true capabilities or actions by hiding under other models. Every action taken by an AI model must be authenticated, whether performed by a human operator or autonomously by the AI. This ensures that no model can operate in the shadows, unnoticed or unaccounted for.

Controlled AI Usage:
Our approach allows for the controlled use of AI without imposing overly strict restrictions that could stifle innovation. By ensuring that all AI actions are traceable and verifiable, we can maintain a balance between freedom of use and safety, enabling responsible AI development and deployment.

7. Sources

Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021). Fighting Deepfakes: Media and Internet Giants' Converging and Diverging Strategies Against Hi-Tech Misinformation. Media and Communication. https://doi.org/10.17645/MAC.V9I1.3494.

Blauth, T., Gstrein, O., & Zwitter, A. (2022). Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI. IEEE Access, 10, 77110-77122. https://doi.org/10.1109/access.2022.3191790.

Akin, C. (2023). Understanding strategic deception and deceptive alignment. Apollo Research. Retrieved from https://www.apolloresearch.ai/blog/understanding-strategic-deception-and-deceptive-alignment.

Groth, J. (2016). On the Size of Pairing-based Non-interactive Arguments. Cryptology ePrint Archive, Paper 2016/260. Retrieved from https://eprint.iacr.org/2016/260.