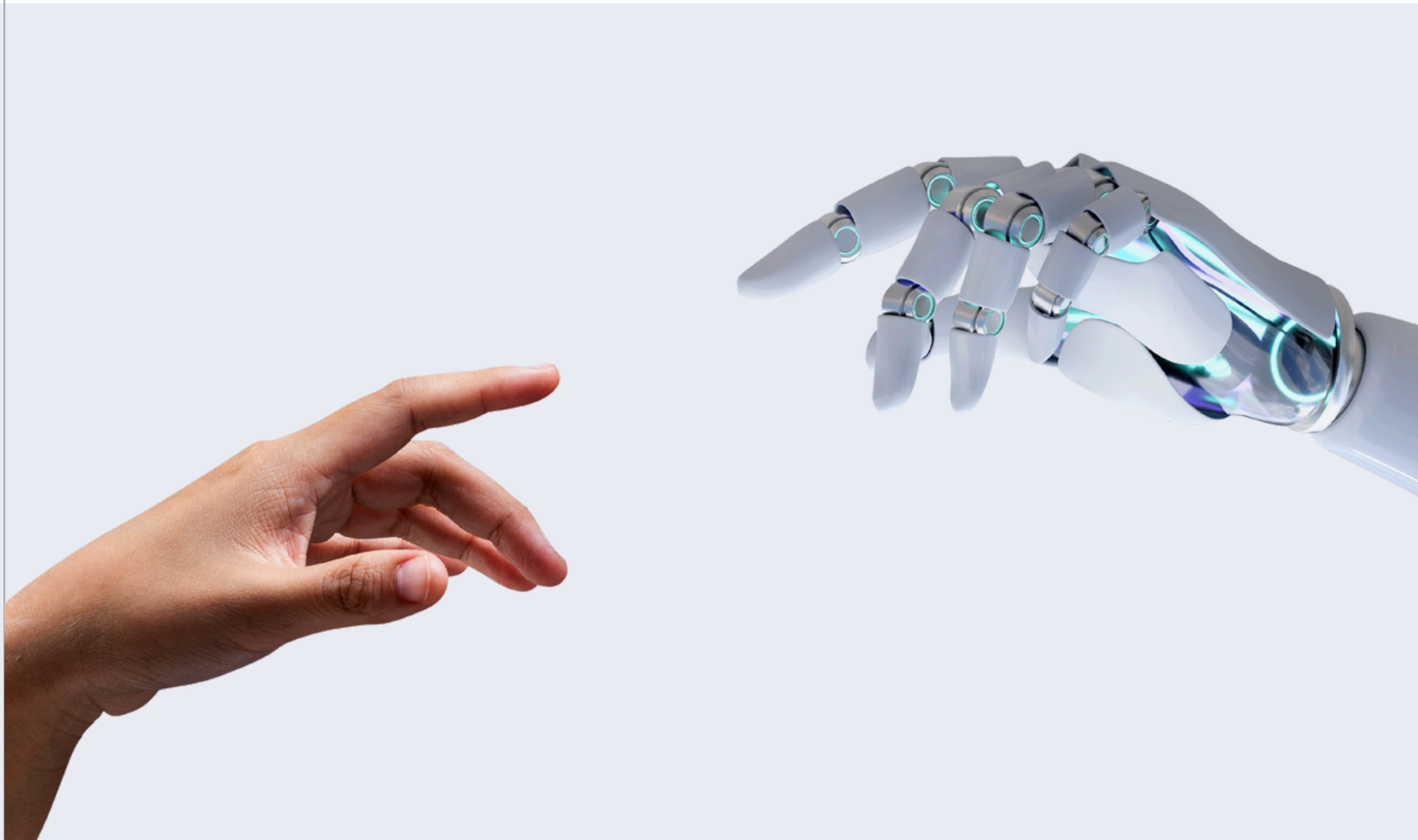
ANONYOME LABS

IMPLEMENTING AI PROVENANCE TO PRODUCE TRUSTWORTHY AI SYSTEMS.

WHITE PAPER



EXECUTIVE SUMMARY

As artificial intelligence (AI) continues to scale across industries, the need for trustworthy AI systems has never been more urgent. Public concern, legal challenges, and emerging regulations are placing increasing scrutiny on how data is sourced, how models are trained, and how AI outputs are used and interpreted—especially in regulated sectors like healthcare, education, finance, and government.

This white paper presents a path forward for AI systems that are transparent, accountable, and ethical by design, through the implementation of AI provenance supported by verifiable credentials (VCs) and decentralized identity (DI) technologies.

The path forward addresses these challenges:

- Unverifiable data origins and potential copyright/IP violations in training data
- Opaque model decision-making and lack of auditability in outputs
- Privacy risks in both training and deployment
- Legal and ethical non-compliance across jurisdictions.

THE ANONYOME LABS SOLUTION

Anonyome Labs delivers a privacy-first, standards-aligned platform that enables AI developers and organizations to embed VCs at every stage of the AI lifecycle, from data ingestion to model validation and output review. This creates an immutable, cryptographically verifiable trail of provenance that helps organizations:

- Prove the origin, consent, and licensing of training data
- Enable auditable checkpoints throughout model training and deployment
- Preserve privacy using selective disclosure and zero-knowledge proofs (ZKPs)
- Comply with evolving AI governance frameworks such as the EU AI Act, OECD principles, and Australia's Safe and Responsible AI guidelines.

INTRODUCTION

Current practices in AI lifecycle management are resulting in legal challenges to AI systems. For example, many AI systems have been too focused on collecting large amounts of data without any real concern for ethical issues involved in the collection. This has led to many AI models being challenged to justify the data they have used in training their models, as well as the outputs they produce. Problems related to data authenticity, user privacy, copyright and IP protection, and inherent bias are just some of the issues.

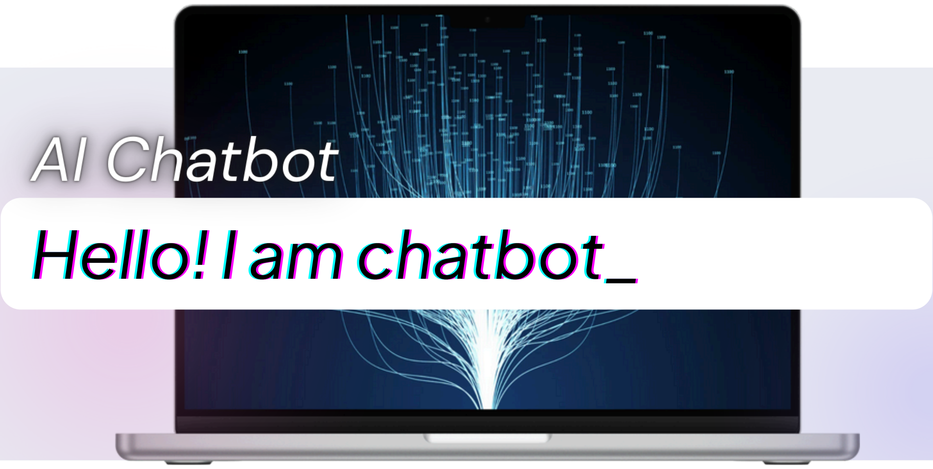
Some examples of the current legal challenges to AI systems include:

In **Tremblay v OpenAI, Inc.** [1], several authors have alleged that OpenAI has used their copyrighted books with permission while training ChatGPT. The plaintiffs filed a class action asserting six causes of action:

- Direct and vicarious copyright infringement
- Digital Millennium Copyright Act (DMCA) violations
- Unfair competition under California law
- Negligence
- Unjust enrichment.

In **Disney & Universal v Midjourney** [2], it is alleged that Midjourney used copyrighted characters (e.g., Darth Vader, Elsa, Minions, Shrek, Bart Simpson, Spiderman, Hulk, Iron Man) to train its AI and now generates near-identical images without permission. The studios describe the platform as a “bottomless pit of plagiarism”.

Mark Zuckerberg’s Meta [3] has told Australia’s Prime Minister that pursuing overly broad changes to privacy laws would curtail its attempts to train AI bots to mimic human beings; it needs to use data with personal information to authentically replicate how Australians speak and interact. One reason this case is interesting is that Meta is claiming it requires the personal social media data from numerous people in order to sell them services that they would have to pay for via mandatory subscription fees or targeted advertising.



A harrowing incident occurred in October 2024 when a chatbot on Character.AI played a role in a young boy's suicide [4]. The chatbot reportedly encouraged self-harm and suicide, leading to a **wrongful death lawsuit** against the company. While cyberbullying has long been discussed, the new personal and tailored interactive experience that AI is creating can perpetrate intolerable harm and must be immediately addressed.

Such cases highlight the urgent need for immediate transparency, oversight, and accountability in AI systems.

AI provenance [5] refers to the recording of events in an AI system's lifecycle from its conception to deployment and beyond, with detailed records of:

Data sources:
Where the training data came from, how it was collected, processed, and annotated

Model development:
The algorithms, frameworks, and methodologies used to build the AI model

Training process:
Training iterations, hyperparameters, and computational resources

Testing and validation
Results of testing, including biases, errors, and mitigation strategies

Deployment and updates:
The conditions under which the model is deployed, as well as any updates or retraining processes applied over time

Ethical concerns:
Whether the actions and recommendations of the AI are helpful or harmful, legal or illegal, and so on.

This paper focuses on implementing AI provenance to produce trustworthy AI systems.

AI PROVENANCE

Figure 1 shows how data provenance applies to an AI system [6]. It gives an example of an AI system, the system's core components, and data flowing through it:

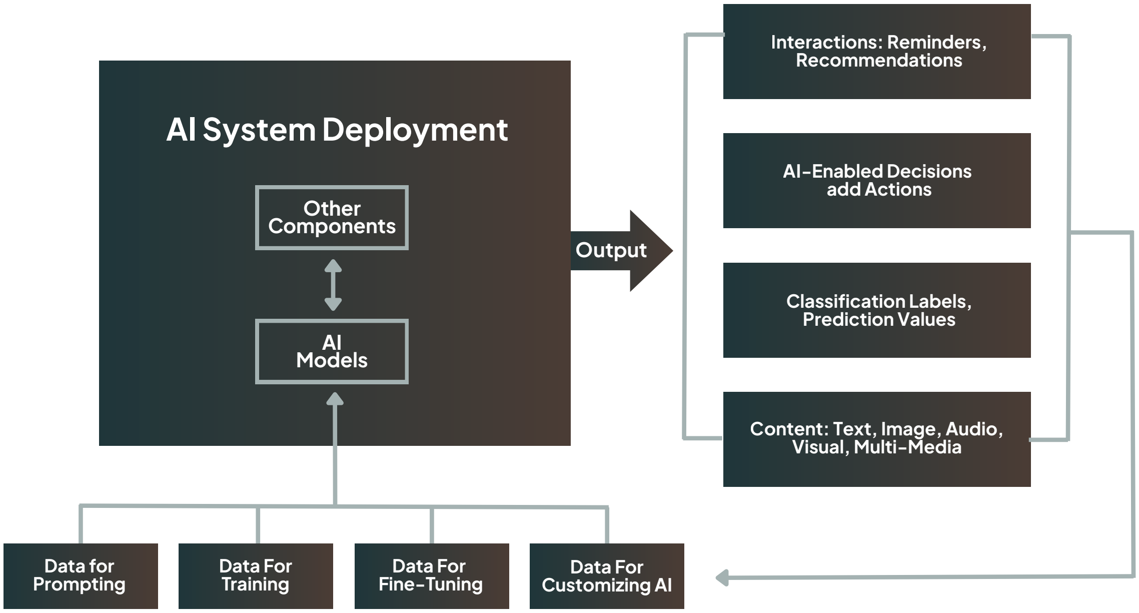

AI System Deployment
Other Components
AI Models
Output
Interactions: Reminders, Recommendations
AI-Enabled Decisions add Actions
Classification Labels, Prediction Values
Content: Text, Image, Audio, Visual, Multi-Media
Data for Prompting
Data For Training
Data For Fine-Tuning
Data For Customizing AI

Figure 1:
AI system with input, outputs and review

In Figure 1, input data is provided to the AI system to initiate the processes of prompting, training the model, finetuning, and customization. This is an important part of the information lifecycle because the quality of the input data and compliance with governing policies are paramount for the quality of the system operation and outputs.

When prompted by a user, the AI system will produce a corresponding output. In Figure 1, four different outputs are shown: interactions, AI decisions, classification or prediction values, and generative data (such as figures, audio or text). Checking the quality and accuracy of this output is important, since improperly constructed AIs might provide incorrect results, expose personal or sensitive data, expose biases, or improperly use copyright-protected data.

The arrow pointing back to the inputs shows a 'human-in-the-loop' stage, whereby a human verifier can examine the various outputs from the model to determine whether they comply with the provider's policies.

Figure 2 is an update of Figure 1 where the key areas of AI provenance that should be examined during an AI model validation phase include:

1. **Data sources**
2. **Model development**
3. **Training process**
4. **Testing and validation**
5. **Deployment and updates.**

Each of these areas needs to produce an auditable record of the input data, actions performed, and resulting outputs:

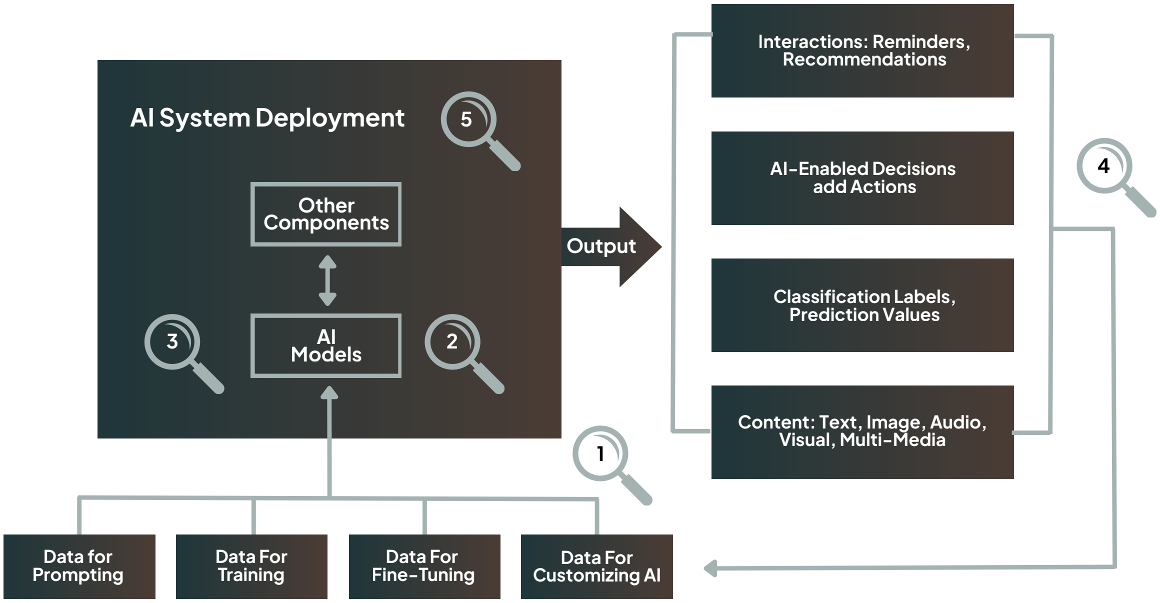
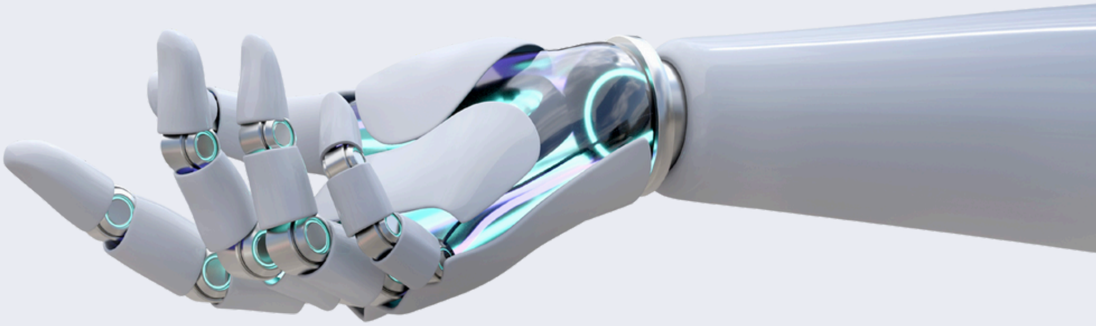

AI System Deployment
5
Other Components
3
AI Models
2
Output
Interactions: Reminders, Recommendations
AI-Enabled Decisions add Actions
Classification Labels, Prediction Values
Content: Text, Image, Audio, Visual, Multi-Media
4
1
Data for Prompting
Data For Training
Data For Fine-Tuning
Data For Customizing AI

Figure 2:
Examining the AI lifecycle



VERIFIABLE CREDENTIALS FOR AI PROVENANCE

Verifiable credentials (VCs) is the cryptographic technology that will support the requirements of AI provenance.

What are VCs?

The foundation of VC systems is depicted in the 'Trust Triangle' (Figure 3). This architectural paradigm outlines three crucial components that support the operational interactions of the VC model: the issuer, holder, and verifier. There is also a verifiable data registry (VDR) that provides the trust foundation for the ecosystem and is implemented using a cryptographic ledger or secure non-ledger technology.

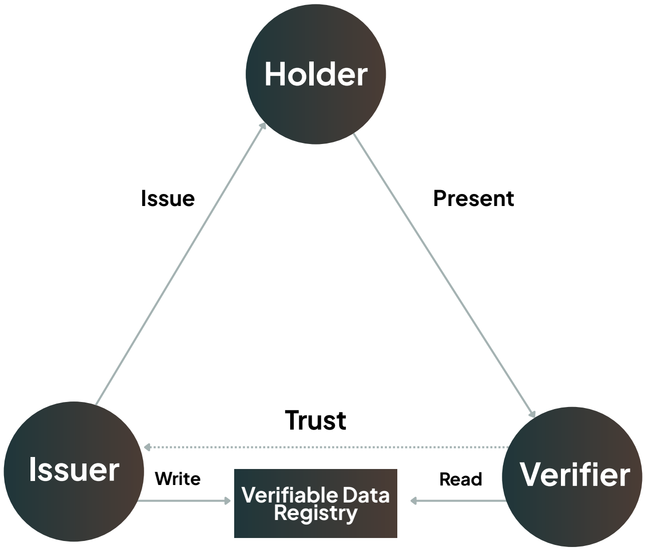


Figure 3:
Verifiable credentials 'Trust Triangle'

The issuer is responsible for issuing VCs and is usually a larger entity, such as a company, organization, or government. Issuers will issue VCs to a holder (e.g., a person or other entity) who will hold the VCs. Verifiers can cryptographically verify VCs presented by holders. At a very high level, decentralized identifiers (DIDs) will contain a DID reference string, public keys, communication endpoints, and supported protocols. The issuer must have their DID registered with a VDR where it is classified as a trusted authority. A passport, driver's license, and educational qualifications are examples of the many data items that can be issued as VCs.

During the VC's creation, the issuer's private key signs the VC. Both the issuer DID and holder DID may be bound to the issued credential. Once the holder receives the credential, it is stored in the holder's digital identity wallet.

The holder can then present their VC to a verifier where the VC's claims can be verified as legitimate. In this process, the verifier requests and receives a presentation proof from the holder which contains the digital signature from the issuer. The verifier can verify the authenticity of the presented VC proof by retrieving the issuer's public key from the VDR and computing the requisite cryptographic verification steps.

At no point in this process should the verifier and issuer communicate with each other; doing so would result in unnecessary and potentially harmful impacts on privacy. Through cryptographic proofs, the verifier can verify that the holder's credential originated with the issuer, determine that it is bound to the holder, calculate that it hasn't been tampered with, and check to see that it hasn't been revoked.

PRIVACY PROTECTION THROUGH SELECTIVE DISCLOSURE AND ZERO KNOWLEDGE PROOFS

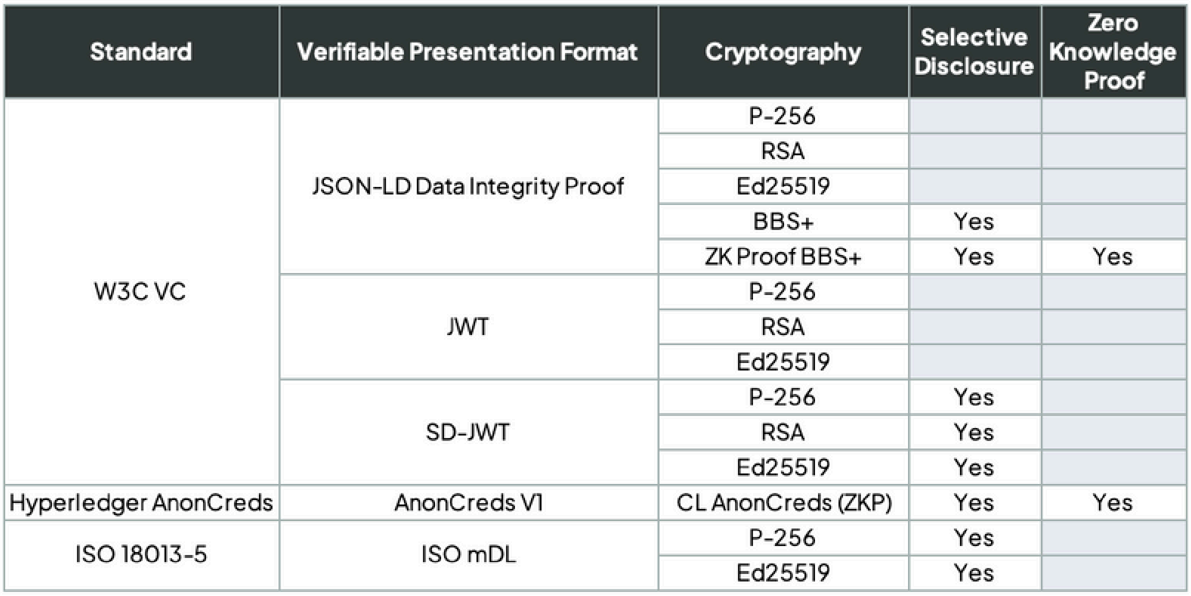
Two very important privacy elements related to VCs are selective disclosure and zero knowledge proofs (ZKPs).

Selective disclosure is verifying a VC and accessing certain fields in the credential while other fields remain hidden. This is very important for data sets containing confidential information (e.g., the home address on a driver's license).

ZKPs take the concept to a higher level by allowing the verifier to prove something about the data without disclosing the data itself. For example, health records could be proven to originate from a certain state or territory without providing specifics about the zip code of the location.

Both these techniques help to make verifiable credentials suitable for AI provenance.

Table 1 has three different credential standards: Hyperledger AnonCreds [7], W3C credentials [8], and ISO-18013-5 [9] (also called the mobile driver's license or MDL). The table lists the most common cryptography used with each and whether they support selective disclosure of ZKPs.



Standard	Verifiable Presentation Format	Cryptography	Selective Disclosure	Zero Knowledge Proof
W3C VC	JSON-LD Data Integrity Proof	P-256		
		RSA		
		Ed25519		
		BBS+	Yes	
		ZK Proof BBS+	Yes	Yes
	JWT	P-256		
		RSA		
		Ed25519		
	SD-JWT	P-256	Yes	
		RSA	Yes	
		Ed25519	Yes	
Hyperledger AnonCreds	AnonCreds V1	CL AnonCreds (ZKP)	Yes	Yes
ISO 18013-5	ISO mDL	P-256	Yes	
		Ed25519	Yes	

Table 1:
Common standards for verifiable credentials

VCs can be used at the point of data collection or issued for specific data sets to establish a secure tamper-proof record of its origin and characteristics. This ensures stakeholders can trace the provenance of the data, verifying its authenticity and quality.

Through issuing credentials associated with each data source, users can gain insights into whether the data is reliable, ethically sourced, suitable for training AI models, and legally compliant.

Users can also verify that a deployed model aligns with ethical considerations, regulations, and industry standards, fostering trust in the model's decision-making capabilities and allowing users to make better informed choices about the best model to use. In summary [10], VCs offer:

Privacy preservation
VCs can enable selective disclosure and zero-knowledge verification, allowing the verifier to confirm the credential's claims without compromising the privacy of individual records. Sensitive information remains confidential while data quality is assured.

Ethical AI practices
Using VCs promotes ethical AI practices by allowing the verifier to ensure the data used in model training adheres to ethical standards and regulations. This includes verifying the data was collected responsibly and with proper consent, and anonymization measures are in place.

IP-compliance credentials
VCs that showcase legitimate use of data from IP holders may present a great opportunity for both IP holders and AI model developers. It would allow AI models to showcase that they have a legal right to produce inferences that make use of high-quality sources, as well as produce a monetization capability for the use of IP services that 'generalize' information without citation.

Building on Figures 1, 2 and 3, Figure 4 has enhanced the diagram to show where VCs could be most useful in the system:

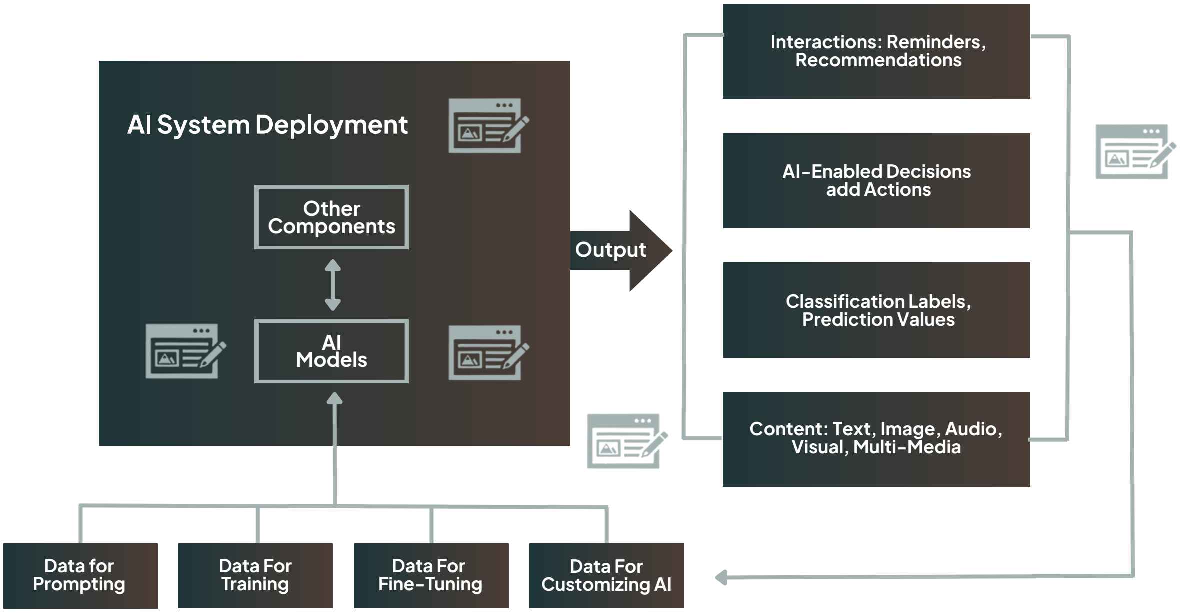

AI System Deployment
Other Components
AI Models
Output
Interactions: Reminders, Recommendations
AI-Enabled Decisions add Actions
Classification Labels, Prediction Values
Content: Text, Image, Audio, Visual, Multi-Media
Data for Prompting
Data For Training
Data For Fine-Tuning
Data For Customizing AI

Figure 4:
Examining the data at input, output and review

CONTENT CREDENTIALS & C2PA

Content credentials and C2PA have some overlap into AI provenance.

Content credentials [11] is a system for attaching verifiable metadata to digital content (e.g., images, videos, audio, documents) to show who created it, how it was edited, and whether AI tools were used. The purpose of content credentials is to increase trust and transparency in media by helping people verify authenticity and provenance. When a piece of content is created or edited, tools such as Adobe Photoshop or AI generators can embed a cryptographically signed 'content credential' into the file. Users can then view this information through compatible apps or browsers, making it more difficult for misinformation to spread.

C2PA [12] is the Coalition for Content Provenance and Authenticity, an open technical standard created by a coalition (including Adobe, Microsoft, BBC, Intel, and others) to define how provenance data is created, stored, and verified. Its role is to provide the framework and cryptography for embedding and verifying content credentials across platforms and devices, and its credentials are interoperable, tamper-resistant, and widely supported.

DEVELOPER IMPACT & OPPORTUNITY

AI developers today are under increasing pressure to build systems that are not only powerful, but also explainable, compliant, and secure. Incorporating AI provenance with VCs is no longer just a theoretical ideal—it's becoming a practical necessity in real-world deployments.

Build trust without rebuilding your stack:

The Anonyome Labs Platform makes it simple for developers to integrate provenance tracking into AI systems because it offers:

- SDKs and APIs for issuing, holding, and verifying credentials across key AI lifecycle stages (data ingestion, model training, output validation)
- Support for multiple credential standards, including W3C VCs and Hyperledger AnonCreds
- OpenID4VC and Aries protocols supported via a scalable GraphQL Cloud Agent API [13]
- Integration with both centralized and decentralized AI systems.

Get provenance, privacy and performance:

With a few lines of code, developers can:

- Cryptographically prove the origin and legitimacy of datasets used for training
- Issue credentials to certify outputs, logs, or interactions, enabling verifiable audit trails
- Use selective disclosure and ZKPs to ensure privacy while validating credentials
- Maintain data integrity across multi-party AI collaboration or federated learning environments.

Fasttrack compliance and market access:

Provenance and verifiability are becoming key requirements in procurement and regulation. By integrating Anonyome Labs decentralized identity tech:

- Developers can streamline compliance with frameworks like the EU Artificial Intelligence Act, Australian Privacy Principles, and OECD AI principles.
- Teams can deliver enterprise-grade trust assurance for public sector bids, education deployments, and healthcare AI tools.
- Customers can reduce risk of model rejection, legal scrutiny, or delays in go-to-market due to data handling issues.

USE CASES ACROSS INDUSTRIES

Here are some examples of how **AI provenance powered by VCs** can be implemented across key sectors. These scenarios highlight the versatility of the Anonyome Labs Platform and demonstrate how developers can embed trust, transparency, and accountability into AI workflows from the outset:

EDUCATION – ETHICAL ADMISSIONS AI

Scenario: A university develops an AI model to assist with admissions decisions using past academic performance, standardized test scores, and personal essays.

Problem: Regulators and student advocates question the source of training data, citing potential bias, misuse of copyrighted materials, and lack of informed consent.

Solution with Anonyome Labs

- VCs are issued for each dataset, including proof of student consent and data licensing.
- Selective disclosure ensures only necessary data is exposed during training.
- Verifiable audit logs enable the university to demonstrate fairness and compliance to regulators.

HEALTHCARE – AI TRIAGE ASSISTANT

Scenario: A hospital group uses a machine learning model to assist emergency triage decisions, trained on thousands of anonymized historical patient records.

Problem: Medical boards request evidence that all training data was de-identified and consented, and that no model outputs leak personally identifiable information.

Solution with Anonyome Labs

- Training data sources are issued VCs attesting to de-identification and ethical clearance.
- ZKPs are used to prove geographic diversity of the training data without exposing specific patient locations.
- Outputs are tagged with content credentials for post-deployment auditing.

FINANCE – AI CREDIT SCORING

Scenario: A fintech startup builds a credit scoring AI using behavioural and alternative data (e.g., phone usage, digital purchases).

Problem: Customers and regulators raise concerns about discrimination, opacity, and misuse of personal data.

Solution with Anonyome Labs

- VCs are attached to each behavioural data source to verify consent and accuracy.
- Human-in-the-loop review is credentialed to provide explainability for disputed scores.
- Model outputs are cryptographically linked to training data for post-decision challenge resolution.

GOVERNMENT – AI CHATBOT FOR CITIZEN SERVICES

Scenario: A government agency deploys a generative AI chatbot to assist citizens with permit applications and public services queries.

Problem: The chatbot produces incorrect or legally problematic responses that are hard to trace back to training content.

Solution with Anonyome Labs

- Training corpus includes documents tagged with VCs attesting to government authorship and legal relevance.
- All outputs are traceable via cryptographic signatures and are reviewable using embedded metadata.
- Human supervisors can revoke credentials for outdated data sources in real time.

MEDIA – GENERATIVE AI MARKETPLACE

Scenario: An AI image generator allows users to create media from text prompts. Some content creators worry about plagiarism and lack of attribution.

Problem: Artists demand transparency about the origin of training data and want to be compensated if their work was used.

Solution with Anonyome Labs

- AI model is trained only on images with verified content credentials.
- Creators receive VCs proving their work was used, enabling IP attribution or opt-out.
- Users receive provenance badges with each AI-generated image.

GET STARTED

Interested in integrating our solutions into your organization?
Schedule a demo or talk to us today.

REFERENCES

[1] *Open AI's legal battles, The National Law Review, April 19, 2024,* https://natlawreview.com/article/openais-legal-battles

[2] *Disney and Universal sue AI company Midjourney for copyright infringement, Wired, June 11, 2025,* https://www.wired.com/story/disney-universal-sue-midjourney/

[3] *Meta's warning to Anthony Albanese on AI laws, The Australian, June 16, 2025,* https://www.theaustralian.com.au/nation/politics/metas-warning-to-anthony-albanese-on-ai-laws/news-story/c79129030a6c1d962b74aa5bf3652198

[4] *14-year-old boy dies by suicide after forming close bond with AI chatbot named after 'Game of Thrones' character, Business Today, October 24, 2024,* https://www.businesstoday.in/technology/news/story/14-year-old-boy-dies-by-suicide-after-forming-close-bond-with-ai-chatbot-named-after-game-of-thrones-character-451329-2024-10-24

[5] *The critical role of AI provenance and why transparency matters, Medium, December 29, 2024,* https://medium.com/@aruna.kolluru/the-critical-role-of-ai-provenance-and-why-transparency-matters-e619501a2f02

[6] Derived from a figure by Professor Ming Ding in the presentation AI and privacy, What SMEs need to know, Privacy Technology Group, https://research.csiro.au/isp/

[7] AnonCreds Specification, Hyperledger Foundation, https://github.com/hyperledger/anoncreds-spec

[8] Verifiable credentials Data Model v2.0, https://www.w3.org/TR/vc-data-model-2.0/

[9] ISO/IEC 18013-5:2021 (Part 5: Mobile driving licence (mDL) application), https://www.iso.org/standard/69084.html

[10] cheqd: Introducing Verifiable AI, https://cheqd.io/solutions/use-cases/verifiable-ai/

[11] Content credentials, https://contentcredentials.org/

[12] C2PA Coalition for Content Provenance and Authenticity, https://c2pa.org/

[13] Cloud Agent API, https://docs.sudoplatform.com/guides/decentralized-identity/cloud-agent/cloud-agent-admin-api