Bessemer Venture Partners

A governance framework for defensible, trustworthy AI

A practical framework for chain of custody, data provenance, and the seven actions boards should take before regulators define the standard for them.

Research prepared by AI Overwatch Group, distributed by Bessemer Venture Partners

George DeCesare
Operating Advisor
Bessemer Venture Partners
AI Overwatch Group

AI OVERWATCH GROUP™
White Paper Series | 2026

Data Provenance & Chain of Custody in Artificial Intelligence Systems

A Governance Framework for Defensible, Trustworthy AI

Prepared by
AI Overwatch Group™, LLC
March 2026

About AI Overwatch Group™, LLC

AI Overwatch Group founded by George DeCesare, advises boards, founders, CEOs, and institutional investors on AI governance, cyber risk, and enterprise resilience. We operate at the intersection of technology, regulatory oversight, and fiduciary accountability.

Our mission is simple:

Protect the integrity of intelligence before it becomes institutional authority.

Executive Summary

Artificial Intelligence systems now influence healthcare diagnoses, financial underwriting, securities trading, regulatory compliance, and enterprise decision-making. Yet, in many cases, most organizations cannot prove the origin, transformation history, or lawful authorization of the data powering those systems.

In legal proceedings, evidence must maintain a documented chain of custody to be admissible. Similar to the established practices commonly used to manage chain of custody in the digital data space, AI systems increasingly require the same standard and practices.

Without enforceable data provenance:

- AI models may train on poisoned or unlawful datasets
- Organizations may fail regulatory audits
- Boards may lack defensible oversight
- Litigation risk escalates

- Public trust erodes

This paper outlines a practical governance and technical framework to ensure AI systems maintain evidentiary-grade data integrity. Utilizing frameworks such as the National Institute of Science and Technology (NIST) AI Risk Management Framework, the European Union's Artificial Intelligence Act, or COBIT, an organization can implement the proper controls that will mitigate the risks discussed hereunder. As there is not a one size fits all solution that works for every organization, selecting the appropriate framework for your organization's strategy and needs will yield the best results in mitigating the risks detailed in hereunder.

---

1. The Strategic Imperative

Traditional cybersecurity protects data and systems. AI governance must protect intelligence formation. Traditional governance models were designed for systems that stored and processed information. Security frameworks focused on protecting data from theft, alteration, or unauthorized access. If the database was secure and the software functioned as designed, governance obligations were largely satisfied.

Artificial Intelligence fundamentally changes this paradigm. This transformation process can be described as intelligence formation. AI systems do not simply process data—they create intelligence from data. Through statistical learning, pattern recognition, and probabilistic inference, AI models transform raw information into insights, predictions, and recommendations that increasingly guide enterprise decisions. This significantly impacts traditional risk profiles, shifting the way confidentiality, integrity, and availability were viewed, to accommodate for how AI systems work. The risk profile then looks more like:

Cybersecurity Risk	AI-Era Risk
Data	Data Poisoning
Unauthorized Access	Unauthorized Influence
Malware	Model Manipulation
Log Tampering	Statistical Corruption

AI systems create a distinct governance imperative: when organizations delegate decision-making to machines, both management and Boards become jointly accountable for ensuring that AI-generated outputs are lawful, trustworthy, and defensible.

Management and Boards are now accountable not only for protecting data — but for validating the integrity of the intelligence derived from AI systems. Historically, management was accountable for implementation and operational assurance. Boards were responsible for oversight of:

- financial reporting accuracy
- regulatory compliance
- enterprise risk management
- cybersecurity posture

Artificial Intelligence introduces a new set of governance obligations. Boards must oversee the integrity of machine-generated intelligence that influences strategic or operational decisions. Management must ensure the integrity of AI-derived intelligence as it depends on the day-to-day technical controls that only management can implement. Inherent to AI systems, this now adds additional risks to the list that Management and Boards must pay attention to, such as:

- ethical considerations
- data quality bias
- workforce displacement and gaps
- socioeconomic inequality
- AI system technical failures, errors and vulnerabilities

1.1. Management Responsibilities in the AI Era

In traditional IT systems, Management has developed numerous controls to protect the confidentiality, integrity and availability of these systems and their data. AI systems require the same execution of controls, but at a different level. Management must now ensure the data governance and provenance of the datasets used in the AI system, document origin, verify legal authority for use, conduct integrity validation, and track through lineage systems. They must also oversee the entire model lifecycle from data ingestion, model training, testing, validation, deployment, and on-going monitoring through retirement.

As in traditional IT systems, management is responsible for implementing technical protections of the system and data. This includes controls such as cryptographic dataset hashing, immutable logging, secure model registries, controlled training pipelines, and managing identities and access. This set of controls provide the evidence foundation for chain-of-custody verification.

Lastly, Management must ensure that AI systems comply with applicable laws such as state and federal privacy and consumer protection laws, financial oversight requirements, healthcare regulations, and any emerging AI accountability frameworks. As regulations are rapidly evolving in this space, documentation is key to future proofing compliance. Maintaining records of dataset origins, data transformation history, model training

configurations, and testing results will provide strong documentation and become a key component of the chain-of-custody evidence foundation.

1.2. Board Responsibilities in the AI Era

Boards have similar changes to their scope of responsibilities in AI systems. These new responsibilities emerge because AI systems increasingly function as decision-support or decision-automation mechanisms. When AI outputs affect healthcare outcomes, financial transactions, regulatory disclosures, or safety systems, the board cannot treat those outputs as opaque technical artifacts. They become institutional knowledge assets.

If those assets are corrupted, manipulated, or derived from unlawful datasets, the board's duty of care and duty of loyalty may be implicated.

Therefore, Boards must require organizations to adopt formal AI governance policies that include data provenance standards, AI lifecycle controls, risk classification frameworks, and independent oversight mechanisms.

Not unlike traditional Board oversight responsibilities, AI risks will require Board oversight of operational risk of the AI system, legal and regulatory risk, reputational risk, cybersecurity, and intellectual property, but add to this the risks of ethical bias, discrimination, and workforce and leadership impact. Depending the organization, industry, and rate of AI adoption, the Board's requirements will vary as to the level of controls and procedural governance, therefore, this document does not prescribe specific responsibilities for Board directors. At a minimum, Boards and Management should consider the following protocols and practices:

- trust but verify
- validate acceptable use policies
- designate an accountable AI lead
- review cost analysis
- verify that cybersecurity programs are adapting to AI in the environment
- require audits and traceability of data
- develop and approve a set of AI ethics
- encourage societal adaptation

While Management is ultimately responsible for ensuring chain of custody processes are established for AI systems, the Board has oversight responsibility to ensure that governance structures of AI systems are in place and adequate for their respective organizations, which should include chain-of-custody implications.

2. Defining Data Provenance in AI

Understanding how the origin, transformation, and custody of data determine the reliability and defensibility of AI-driven decisions. Documenting the origin and handling of data is essential to ensuring the integrity of AI models and their outputs. The origin and handling of data determine the reliability and defensibility of AI-driven decisions. Data lineage, how information is sourced, transformed, and passed through systems, establishes the integrity, legality, and trustworthiness of machine-generated insights. Without verifiable data lineage, organizations cannot prove that their AI-generated intelligence meets any of those standards.

Data Provenance is the verifiable documentation of:

- data origin
- ownership and authorization
- transformation history
- training inclusion
- inference usage
- model impact traceability

In AI systems, provenance must span:

1. raw ingestion
2. labeling and annotation
3. data transformation
4. training and fine-tuning
5. embedding generation
6. retrieval-augmented inputs
7. model outputs

Unlike traditional databases, AI models internalize data influence in latent space — making traceability exponentially more complex. Many AI systems operate within what researchers call latent space. Latent space is a mathematical representation where complex relationships between data points are encoded as vectors.

Understanding the distinction between how traditional databases store data and how AI models internalize data is essential for developing appropriate governance, security, and provenance controls.

Traditional information systems store data in a structured, retrievable form. Artificial intelligence systems, by contrast, absorb data into mathematical representations that influence model behavior but cannot be easily traced back to individual records.

This difference fundamentally changes how organizations must think about traceability, accountability, and data governance.

In conventional database systems, data is stored in clearly defined structures such as:

- relational tables
- documents
- key-value pairs
- data warehouses

Each record is individually identifiable and retrievable.

AI models, particularly machine learning and deep learning systems, operate very differently.

Rather than storing data explicitly, AI models learn statistical relationships within the data during the training process. This training process transforms datasets into model parameters, typically represented as large collections of numerical weights within neural networks.

During training:

- images, text, or signals are converted into numerical vectors
- the model organizes these vectors in a high-dimensional space
- relationships between vectors capture meaning or similarity

For example, in a language model utilized in the healthcare space:

The vector representing "doctor" may exist near vectors representing:

- hospital
- patient
- diagnosis
- treatment

These relationships emerge from training data, but the original sentences used to create them cannot be directly reconstructed. This abstraction layer is powerful—but it complicates provenance and accountability.

Another key distinction is that AI training is largely irreversible.

Once a model has been trained:

- the original training data cannot easily be extracted
- the influence of individual data points cannot easily be isolated
- removing a specific record may require retraining the entire model

This creates a significant governance challenge. For example:

If a privacy regulation requires deletion of a specific individual's data, organizations may struggle to determine whether that data influenced a trained model. Traditional systems allow record deletion. AI systems may require model retraining to fully remove influence.

Because AI models internalize data rather than storing it directly, traditional provenance methods are insufficient.

Organizations must instead track

- training dataset origin
- where the data used to train the model originated
- data transformation history
- model training lineage
- how the data was cleaned, labeled, and modified
- which datasets contributed to each model version
- model evolution (how models change over time through retraining or fine-tuning)

Without this information, it may be impossible to determine how specific data influenced a model's behavior. This can then lead to issues of attribution difficulty, legal defensibility, investigation bias, and data removal obligations.

3. AI Data Governance: The Missing Layer

Traditional cybersecurity has long relied on the Confidentiality, Integrity, and Availability (CIA) Triad as its foundational model for protecting information assets. The framework has served organizations well because it answers three essential security questions:

- Confidentiality: Can unauthorized individuals access the data?
- Integrity: Has the data been altered without authorization?
- Availability: Is the data accessible when the business needs it?

These remain indispensable security objectives. AI systems cannot be trusted without them. However, AI introduces an entirely different governance challenge. The fundamental question is no longer simply whether data is protected. It is whether data deserves to influence machine reasoning.

An AI model does not merely store information. It learns from it. Every dataset used for training, fine-tuning, retrieval, or reinforcement shapes how the model reasons, makes recommendations, and increasingly, makes autonomous decisions on behalf of the organization.

Perfectly protected data can still produce an unsafe AI system if that data is inappropriate, biased, unlawfully acquired, poorly governed, or used outside its intended purpose.

In other words, organizations can satisfy every traditional cybersecurity control and still build an AI system that regulators, customers, shareholders, or courts determine should never have been trusted. This requires an additional governance framework.

3.1. The AI Data Governance Framework

Before any dataset is permitted to influence an AI system, organizations should classify it across three governance dimensions:

Sensitivity

Sensitivity extends beyond confidentiality. Confidentiality asks whether unauthorized individuals can access information. Sensitivity asks whether the information should be used by an AI system at all, and under what conditions.

Organizations should evaluate questions such as:

- Does the dataset contain regulated personal information?
- Could misuse create harm to individuals or society?
- Does it contain intellectual property or confidential business information?
- Could the information reveal sensitive business strategies or competitive intelligence?
- Would customers reasonably expect this data to be used to train or influence AI?

Not every dataset that is technically accessible should be eligible for AI use. Sensitivity determines the governance obligations associated with that data.

Criticality

Availability ensures systems remain operational. Criticality determines the potential organizational consequences if the AI relies upon incorrect, incomplete, poisoned, or manipulated data. The higher the consequence of an AI-assisted decision, the greater the governance requirements surrounding the data that produced it.

For example:

- Healthcare diagnosis
- Financial reporting
- Credit underwriting
- Fraud detection
- Public safety
- Pharmaceutical research
- National security
- Critical infrastructure

These systems require substantially greater governance than AI used to summarize meetings or generate marketing content because failures directly affect human lives, financial outcomes, or regulatory obligations.

Criticality should drive the level of oversight, validation, monitoring, testing, and executive review required before deployment.

Compliance

Integrity confirms that data has not been altered. Compliance determines whether the organization possesses the legal, contractual, ethical, and regulatory authority to use that data throughout the AI lifecycle.

Organizations should be able to answer:

- Was the data lawfully obtained?
- Do licensing agreements permit AI training or fine-tuning?
- Are there contractual restrictions governing its use?
- Are retention requirements satisfied?
- Does the intended use comply with privacy regulations?
- Have cross-border transfer requirements been addressed?
- Does model training create new regulatory obligations?

Compliance is no longer simply a legal review conducted at project approval. It becomes a continuous governance process that follows data throughout its lifecycle.

3.2. Security Protects Data. Governance Protects Decisions.

Traditional cybersecurity focuses on protecting information assets. AI governance focuses on protecting the decisions that emerge from those assets.

Security asks: *"Can unauthorized parties access or modify this data?"*

Governance asks: *"Should this data influence enterprise intelligence?"*

Those are fundamentally different questions requiring different controls.

Extending the CIA Model

Rather than replacing traditional cybersecurity principles, AI governance builds upon them.

Traditional Security	AI Governance
Confidentiality protects information from unauthorized disclosure.	Sensitivity determines whether information should influence AI systems and under what conditions.
Integrity ensures information remains accurate and unaltered.	Compliance determines whether information may legally, ethically, and contractually be used throughout the AI lifecycle.
Availability ensures information is accessible when needed.	Criticality determines the business consequences if AI makes decisions using that information and establishes the level of governance required.

The CIA Triad remains the foundation of cybersecurity. The AI Data Governance Framework extends that foundation by ensuring organizations govern not only how data is protected, but how it shapes machine reasoning.

3.3. CEO/Board Implications

Boards and senior leaders should require management to demonstrate that every material AI system has classified its underlying data according to both security and governance requirements.

Specifically, management should be prepared to answer:

- Which datasets are sufficiently sensitive to require additional governance before AI use?
- Which AI systems operate in high-criticality environments requiring enhanced oversight?
- Which regulatory, contractual, licensing, or ethical obligations govern every dataset used throughout the model lifecycle?
- How are these classifications maintained as datasets evolve, models are retrained, and AI capabilities expand?

Without this governance discipline, organizations cannot credibly demonstrate that their AI systems are trustworthy, defensible, or suitable for consequential decision-making.

4. Chain of Custody Applied to AI

The concept of chain of custody originates in legal and forensic disciplines, where it ensures that physical evidence presented in court can be trusted as authentic and unaltered. Every transfer of evidence—from collection to storage to analysis—must be documented and verifiable.

In the context of artificial intelligence systems, these same principles must be applied to the data and models that generate machine intelligence. Because AI systems derive insights from complex data pipelines and training processes, maintaining an auditable chain of custody is essential for ensuring that the intelligence produced by these systems is reliable, lawful, and defensible.

AI systems transform datasets into distributed statistical representations, Because of this, the chain of custody must be established before training occurs. Once the training process has internalized the data, the ability to reconstruct lineage becomes far more difficult. This is why provenance tracking must occur at:

- data ingestion
- dataset preparation
- training pipeline execution
- model versioning

Establishing chain of custody in these early stages of model formation will form the evidence trail required to demonstrate the integrity of the intelligence produced by the model.

In order to establish a defensible chain of custody process, there are six foundational principles that should be established during the early stages. These principles are:

- identifiable origin
- documented possession
- tamper evidence
- integrity verification
- access control logs
- preservation

In AI systems, this translates to:

Legal Standard	AI Equivalent
Evidence seal	Cryptographic dataset hashing
Custodian log	Immutable pipeline logs
Forensic validation	Reproducible training pipelines
Admissibility	Regulatory defensibility

4.1. Identifiable Origin

The first principle of chain of custody is clearly identifying where the data originated. In AI systems, datasets may be collected from a wide range of sources, including:

- internal enterprise data repositories
- third-party vendors
- public datasets
- licensed commercial data providers
- web-scraped content
- synthetic data generation systems

Without clear documentation of origin, it becomes difficult to determine whether the data was legally obtained, if it complies with privacy regulations, if any proprietary or copyrighted information is in use, and if any malicious data has been injected.

Establishing identifiable origin requires organizations to maintain detailed metadata describing:

- the source organization or system
- date and method of acquisition

- licensing or contractual rights governing the data
- collection methodology
- applicable regulatory classifications

In practice, this often involves implementing dataset registries or metadata catalogs that track the provenance of each dataset used in model training or inference.

In summary, identifiable origin is particularly important for AI because once data is incorporated into a model's training process, its influence becomes embedded in the model's parameters and may no longer be directly observable. Establishing provenance at the point of ingestion is therefore critical.

4.2. Documented Possession

The second principle requires that every transfer or handling of the data be recorded.

In legal chain-of-custody procedures, evidence logs document who had possession of an item, when it was transferred, and under what conditions. This ensures that there are no unexplained gaps in custody that could allow evidence to be tampered with.

For AI systems, documented possession means tracking every stage of the data lifecycle, including:

- data ingestion into the organization
- transfer between data storage systems
- movement into training pipelines
- access by data scientists or engineers
- transformation processes such as labeling or feature engineering
- integration into model training runs

Each stage should record:

- the identity of the system or user performing the action
- the timestamp of the event
- the operation performed
- the version of the dataset affected

This information is typically captured through immutable audit logs and data lineage systems.

Documented possession is an important practice, as it allows organizations to reconstruct the complete history of how data moved through their AI infrastructure. If an anomaly or failure occurs, investigators can trace the source of the issue by reviewing the custody record.

4.3. Tamper Evidence

Another critical principle is ensuring that any unauthorized alteration of the data is detectable.

In physical evidence handling, tamper-evident seals are used to indicate whether evidence containers have been opened or modified. For AI systems, tamper evidence is typically achieved through cryptographic techniques and secure logging mechanisms.

Examples include:

- cryptographic hashing of datasets upon ingestion
- digital signatures applied to dataset versions
- immutable storage architectures
- write-once logging systems
- blockchain or distributed ledger audit records

When a dataset is hashed, a unique fingerprint of the data is created. If the dataset is altered in any way—even by a single byte—the hash value changes. By comparing current hashes with recorded hashes, allows organizations to detect unauthorized modifications.

Tamper evidence is particularly important in defending against data poisoning attacks, in which adversaries intentionally introduce malicious data into training datasets in order to manipulate model behavior. Not only are tamper-evident controls a key element of a chain of custody practice, without tamper-evident controls, data poisoning attacks may go unnoticed.

4.4. Integrity Verification

Integrity verification refers to the ability to confirm that data and models remain authentic and unaltered throughout their lifecycle. While tamper evidence reveals whether modifications occurred, integrity verification actively confirms that the data remains consistent with its original state.

This typically involves processes such as:

- periodic hash validation checks
- dataset version comparison
- secure checksum verification
- validation of digital signatures
- reproducibility testing for model training pipelines

Integrity verification is also essential for ensuring model reproducibility. If an organization cannot reproduce a model using the documented training datasets and configuration parameters, it may indicate that:

- undocumented data sources were used
- datasets were modified after ingestion

- training processes were altered without proper logging

Therefore, while integrity validation plays a vital role in chain of custody controls, it is also an essential practice in reproducibility of the model.

4.5. Access Control Logs

Access control logs support several important objectives in AI governance.

First, they establish accountability by linking every data interaction to a specific identity, whether human or machine.

Second, they provide transparency into data handling, enabling organizations to reconstruct how datasets moved through training pipelines and development environments.

Third, they support security monitoring and incident investigation by revealing unusual or unauthorized access patterns.

Finally, they provide evidence necessary to demonstrate compliance with regulatory and governance standards. Without reliable access logs, it becomes difficult to prove that datasets were handled appropriately or that models were trained using authorized data.

4.6. Preservation

The final principle is the preservation of the evidence itself. In legal proceedings, physical evidence must be preserved in conditions that prevent degradation or alteration.

In AI systems, preservation refers to maintaining historical records of:

- datasets used in model training
- dataset versions over time
- model configurations and parameters
- training logs and hyper-parameters
- evaluation results and validation tests

Preservation is essential for several reasons.

First, it allows organizations to demonstrate regulatory compliance by proving what data was used at the time a model was deployed.

Second, it enables retrospective analysis if a model later produces unexpected or problematic outcomes.

Third, preservation supports forensic investigations into incidents such as model bias, regulatory violations, or adversarial manipulation.

Many organizations implement preservation through model registries and dataset archives, which store versioned artifacts of both data and models.1313If AI influences healthcare, credit decisions, or public markets, evidentiary standards will apply — explicitly or implicitly.

5. Risk Domains

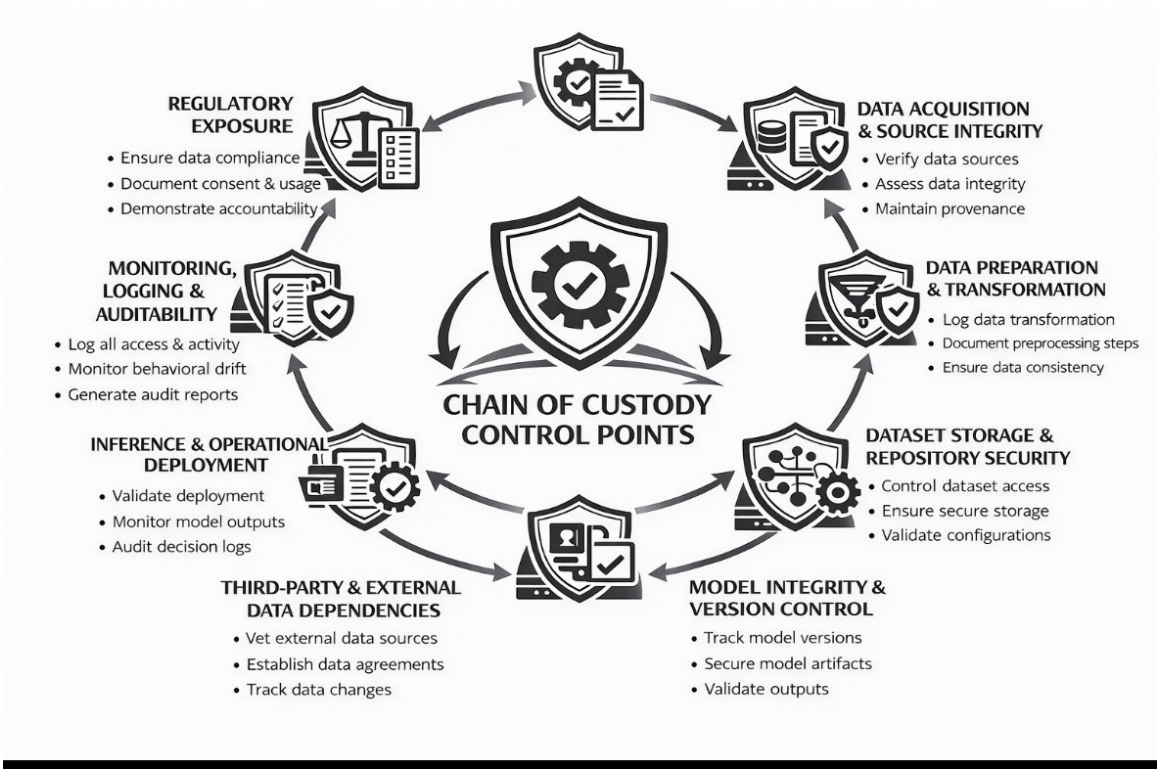

Image: Chain of Custody Risk Domain Controls Workflow

In traditional legal and forensic contexts, chain of custody ensures that physical evidence remains authentic, unaltered, and traceable from the moment it is collected until it is presented in court. Every stage of handling is scrutinized because the integrity of the evidence determines whether it can be trusted.

Artificial intelligence systems require a similar framework. However, the complexity of modern AI data pipelines introduces numerous points where the integrity of the system can be compromised. These points of vulnerability are often referred to as risk domains.

Risk domains represent the specific categories of exposure where the chain of custody for AI data or models may be broken, compromised, or rendered unverifiable. Identifying these domains is critical because they define where governance controls must be applied to preserve the integrity of the intelligence produced by AI systems. Without clearly defined risk domains, organizations cannot systematically identify where custody controls are required, nor can they ensure that the intelligence derived from AI systems remains trustworthy.

5.1. Data Acquisition and Source Integrity

The first and most fundamental risk domain involves the origin and legitimacy of the data entering the AI system. If the initial dataset cannot be verified, the entire downstream intelligence formation process becomes questionable.

Data acquisition represents the point of origin for AI intelligence formation. At this stage, organizations determine which datasets will shape how models perceive patterns, relationships, and correlations. If these datasets contain errors, bias, malicious manipulation, or unlawful content, those issues may become embedded in the resulting model.

The primary objective of this risk domain is to ensure that the organization can clearly answer a fundamental governance question, "Where did this data come from?"

Unlike traditional systems, where data can be inspected and verified at the time of use, AI models internalize the statistical influence of training data during the training process. Once this occurs, the original data may no longer be directly observable within the model. For this reason, the integrity of data sources must be validated before the data is introduced into training pipelines. If the chain of custody is compromised at the acquisition stage, every downstream stage of the AI lifecycle—including data preparation, model training, and inference—may inherit those flaws.

Along with point of data origin, data source integrity becomes a key factor in establishing proper data sources. Source integrity validates that the organization has the legal right to use the data.

This is particularly important in AI systems, where large datasets may be aggregated from multiple sources, including publicly available or scraped information.

Source integrity controls also play a critical role in protecting AI systems from data poisoning attacks. In a data poisoning attack, adversaries deliberately introduce malicious or misleading data into training datasets with the goal of influencing the behavior of the model.

Examples include:

- injecting fraudulent transactions into financial training datasets
- inserting manipulated images into medical imaging datasets
- embedding adversarial examples into open-source datasets

Because machine learning algorithms learn patterns from training data, even a small number of malicious samples can alter model behavior in subtle but impactful ways. Organizations must therefore implement validation processes that screen incoming datasets for anomalies, malicious content, or unexpected statistical patterns before allowing the data to enter training pipelines.

Risks associated with improper data authorization include:

- copyright infringement
- violation of data licensing agreements
- misuse of personal or protected data
- violation of regional privacy regulations

As regulatory frameworks governing AI continue to evolve, organizations will increasingly be expected to demonstrate that the data used to train their models was obtained and used in accordance with applicable laws. Failure to do so may result in regulatory enforcement actions or legal challenges to the legitimacy of the AI system itself. If organizations cannot prove where their training data came from or how it was collected, they may face legal exposure and lose the ability to defend the integrity of their AI systems.

For chain of custody, this domain requires the retention of metadata that includes:

- the original data source (organization, platform, or repository)
- the method of collection (survey, sensors, web scraping, licensing, etc.)
- date of acquisition
- licensing or contractual permissions governing use
- regulatory classification of the data (personal, confidential, public, etc.)
- geographic or jurisdictional origin of the data

5.2. Data Preparation and Transformation

Once data has been acquired and its source integrity established, it typically undergoes a series of preparation and transformation processes before it can be used for training or inference. These processes are necessary because raw data is rarely suitable for direct use in machine learning systems.

However, this stage introduces significant chain-of-custody risk because data is actively modified, interpreted, and reshaped by both automated systems and human operators. If these transformations are not carefully documented and controlled, the lineage of the data may become unclear, making it difficult to verify how the final training dataset was constructed.

In the context of AI governance, the data preparation phase represents the point where raw evidence becomes analytical input. Maintaining a verifiable chain of custody through this transformation process is essential for ensuring that the intelligence generated by the model can be trusted.

Unlike the acquisition stage—where data is collected but not yet altered, the preparation and transformation stage involves intentional modification of the dataset. These modifications may include cleaning, normalization, filtering, labeling, and restructuring of data elements. Each transformation step potentially alters the characteristics of the dataset in ways that influence model behavior. If the transformation process is not properly controlled, it may introduce errors, bias, or manipulation that ultimately shape the model's outputs.

Because AI models internalize patterns from the training data, these modifications can have lasting consequences. Once the model has been trained, the impact of a transformation step may no longer be easily identifiable. This is why chain-of-custody governance must ensure that every transformation applied to a dataset is documented and traceable.

Data preparation typically involves several types of transformation operations. Each introduces specific custody risks that must be managed.

Data Cleaning

Data cleaning removes incomplete, corrupted, or inconsistent records from the dataset.

Typical cleaning actions include:

- removing duplicate records
- correcting formatting inconsistencies
- filling missing values
- filtering irrelevant data points

While these steps improve dataset quality, they also alter the original dataset. Without proper documentation, it may become impossible to reconstruct what the original dataset looked like or how it changed.

Chain-of-custody controls should therefore track:

- which records were removed
- which values were modified
- which rules governed the cleaning process

Data Labeling and Annotation

Many machine learning models require labeled datasets in order to learn patterns effectively. Labeling may involve assigning categories, classifications, or contextual annotations to raw data.

Examples include:

- labeling images with object categories
- annotating medical images with diagnostic findings
- tagging text data with sentiment or topic labels
- identifying fraudulent vs legitimate transactions

This process often involves human judgment, which introduces the possibility of error or bias.

To maintain chain of custody, organizations should record:

- who performed the annotation
- which labeling guidelines were used
- when the labeling occurred
- which version of the dataset was labeled

Feature Engineering

Feature engineering transforms raw data into variables optimized for use by machine learning algorithms.

Examples include:

- converting timestamps into time-of-day features
- aggregating transactional history into behavioral indicators
- transforming text into vector embeddings
- scaling or normalizing numerical values

These transformations often involve complex mathematical operations that significantly reshape the dataset. If feature engineering processes are not documented, organizations may be unable to explain how certain model behaviors emerged.

Maintaining chain of custody requires recording:

- the algorithms used to transform features
- parameter settings used in the transformation
- the dataset version on which the transformation was applied

Data Filtering and Sampling

Large datasets may be filtered or sampled to create smaller training sets.

For example, engineers may:

- remove outliers
- sample a subset of records
- balance datasets across categories
- exclude certain demographic attributes

While these actions are often necessary for model performance, they can introduce unintended bias or alter statistical relationships within the data.

Chain-of-custody documentation should record:

- why records were excluded
- how sampling was performed
- what criteria were used for filtering

Before data can be used for training, it typically undergoes multiple transformations such as cleaning, normalization, feature engineering, and labeling. These transformations introduce additional risks because data may be altered in ways that affect model behavior.

Chain-of-custody controls here require:

- transformation logs
- dataset versioning
- annotation audit trails
- validation procedures

The data preparation and transformation domain plays a critical role in AI accountability because it defines how raw data is converted into training inputs that ultimately shape model behavior.

5.3. Dataset Storage and Repository Security

Once datasets have been acquired and prepared, they are typically stored within internal repositories where they can be accessed by data scientists, automated pipelines, and model training systems. These repositories may reside in cloud storage environments, internal data lakes, machine learning platforms, or specialized dataset registries.

This stage represents a critical risk domain in the chain of custody for AI systems because stored datasets become the authoritative source from which models learn. If the integrity of these repositories is compromised, the organization may unknowingly train models on altered or unauthorized data. Maintaining strong controls over dataset storage is therefore essential to preserving the integrity of the AI intelligence formation process.

Unlike traditional application data, training datasets often remain stored for extended periods of time and may be reused across multiple model training cycles.

During this time, datasets may be accessed by numerous systems and individuals, including:

- data scientists conducting experiments
- machine learning engineers running training jobs
- automated feature engineering pipelines
- quality assurance systems validating data
- governance teams conducting audits

Each of these access points introduces the possibility of modification, corruption, or unauthorized duplication. If custody controls are weak at the storage stage, organizations may lose confidence that the dataset used to train a model is the same dataset that was originally approved for training. This creates a governance challenge because the integrity of AI outputs ultimately depends on the integrity of the stored training data.

Several types of risks emerge when datasets are stored in repositories that lack proper custody controls.

Unauthorized Dataset Modification

If individuals with excessive privileges can modify stored datasets, they may unintentionally or intentionally alter training data.

Examples include:

- editing labeled outcomes
- inserting new training samples
- modifying feature values
- replacing subsets of the dataset

Even small changes can significantly influence how machine learning models interpret patterns. Without version tracking and integrity controls, these changes may go undetected.

Dataset Substitution

Another significant risk involves the replacement of an approved dataset with a different dataset that has not undergone proper validation.

This may occur when:

- training pipelines reference the wrong dataset version
- automated processes overwrite existing files
- experimental datasets are mistakenly used in production training runs

Dataset substitution can lead to models being trained on data that lacks verified provenance, potentially breaking the chain of custody.

Insider Threats

Insiders with legitimate access to AI repositories may introduce malicious or biased data in order to manipulate model behavior.

Examples include:

- altering labeled examples to skew model predictions
- introducing fabricated records into training data
- modifying sensitive attributes to influence bias outcomes

Because these individuals may have legitimate credentials, their actions may be difficult to detect without strong auditing mechanisms.

Data Leakage and Exfiltration

Training datasets often contain sensitive or proprietary information, including:

- personal data
- financial records

- intellectual property
- proprietary operational data

If these datasets are copied or exported from the repository without authorization, the organization may face both security and regulatory risks.

Furthermore, leaked datasets may later appear in external environments where they can influence other AI systems in uncontrolled ways. This creates risks related to unauthorized access or modification.

To maintain custody integrity at the storage stage, organizations must implement a combination of technical and procedural controls that ensure datasets remain authentic and unchanged. These controls include strong access control mechanisms, data version controls, cryptographic integrity verification, immutable storage architecture, and comprehensive access logging.

Strong Access Control Mechanisms

Access to dataset repositories should be restricted using role-based or attribute-based access control policies.

This ensures that only authorized individuals or systems can:

- view datasets
- modify data records
- upload new data
- initiate training processes

Access privileges should follow the principle of least privilege, meaning users are granted only the permissions necessary to perform their roles. Privileged access should also be periodically reviewed to prevent unnecessary exposure.

Dataset Version Control

To preserve the chain of custody, organizations should maintain versioned records of datasets. Every time a dataset is modified or updated, a new version should be created while preserving prior versions for reference.

Version control enables organizations to:

- reconstruct which dataset version was used for model training
- compare dataset changes over time
- investigate the source of unexpected model behavior

Without versioning, the historical lineage of a dataset may be lost.

Cryptographic Integrity Verification

Organizations often apply cryptographic hashing to datasets when they are stored in repositories. Hashing generates a unique fingerprint for the dataset. If the dataset is altered even slightly, the hash value changes.

By periodically validating dataset hashes, organizations can confirm that the stored dataset has not been modified since it was approved for use. This technique is widely used in digital forensics and software supply chain security.

Immutable Storage Architecture

Some organizations implement storage frameworks designed to prevent modification after datasets are approved.

Examples include:

- write-once storage systems
- append-only data structures
- immutable object storage configurations

These mechanisms help ensure that once a dataset is finalized and approved for training, it cannot be altered without detection.

Comprehensive Access Logging

Every interaction with dataset repositories should be recorded in access control logs.

Logs should capture:

- user or system identity
- timestamp of access
- dataset accessed
- action performed (read, write, delete, copy)

These logs allow investigators to reconstruct the chain of custody if anomalies arise in model behavior. They also provide evidence for regulatory audits and compliance reviews.

5.4. Training Pipeline Integrity

The training pipeline is the stage of the AI lifecycle where datasets are transformed into model parameters that encode learned patterns and relationships. During this process, algorithms analyze training data, adjust internal weights, and ultimately produce a trained model capable of generating predictions or insights. Within the chain-of-custody framework, the training pipeline represents the point at which data becomes intelligence.

Because of this, it is one of the most critical risk domains in the governance of AI systems. If the integrity of the training pipeline is compromised, the resulting model may internalize manipulated or unauthorized data without any visible indication that the training process was altered.

Once a model has been trained, the influence of individual data points is typically distributed across thousands or millions of parameters. As a result, it can be extremely difficult to detect or reverse the effects of compromised training processes after the fact.

For this reason, maintaining strict custody controls within the training pipeline is essential to ensure that the intelligence generated by AI systems remains trustworthy and defensible.The training pipeline is the process through which datasets are loaded into machine learning frameworks and used to generate model parameters.

This domain is particularly sensitive because the actual formation of AI intelligence occurs during training.

Several types of risks can arise within the training pipeline domain. These include dataset substitution during training, unauthorized training runs, adversarial manipulation of training data, and manipulation of training parameters.

Dataset Substitution During Training

One of the most significant risks is the possibility that a model is trained on a dataset different from the dataset that was originally approved.

This can occur when:

- training scripts reference the wrong dataset path

- pipelines automatically retrieve updated datasets without validation
- temporary datasets used for experimentation are mistakenly used in production training

Because training processes often involve automated workflows, these substitutions may occur without immediate detection. If a model is trained on an unauthorized dataset, the resulting intelligence may be based on data that lacks verified provenance.

Unauthorized Training Runs

In many AI environments, engineers and researchers have the ability to initiate training jobs directly.

If governance controls are weak, individuals may run training experiments that:

- use unapproved datasets
- modify training parameters
- introduce experimental architectures

While experimentation is essential for AI development, the results of these experiments must be carefully managed to prevent experimental models from entering production environments. Without oversight, unauthorized training runs may create models whose lineage cannot be verified.

Adversarial Manipulation of Training Data

Training pipelines are also vulnerable to data poisoning attacks, in which malicious data is intentionally inserted into the training dataset.

These attacks may occur when:

- attackers compromise dataset repositories
- malicious records are injected into data ingestion pipelines
- external datasets contain adversarial samples

The goal of such attacks is often to manipulate model behavior in subtle ways, such as causing the model to misclassify specific inputs or exhibit hidden biases. Because the effects of poisoning are embedded within the trained model, identifying the source of manipulation can be extremely challenging without detailed training records.

Manipulation of Training Parameters

Another risk arises from changes to training configurations, including:

- learning rates
- optimization algorithms
- batch sizes
- training duration
- regularization techniques

These parameters influence how the model learns from data and can significantly alter model performance and behavior.

If parameter changes are not properly documented, organizations may be unable to reproduce training outcomes or understand why a model behaves differently from previous versions.

Maintaining custody over training parameters is therefore an essential part of preserving model lineage.

Among all risk domains in the AI lifecycle, training pipeline integrity may be the most critical because it represents the moment when data becomes embedded within the model.

If custody breaks before training, it may still be possible to correct the issue.

If custody breaks during training, the resulting model may permanently encode corrupted or manipulated data patterns.

Maintaining strict custody controls in this domain ensures that the intelligence generated by the model can be traced back to verified data sources and properly governed development processes. In this way, training pipeline integrity forms the core of a defensible chain of custody for artificial intelligence systems.

Because model training is computationally complex, subtle manipulations may go unnoticed but still influence the resulting model.

Chain-of-custody protections require:

- controlled training environments
- reproducible training pipelines
- signed training configurations
- logging of dataset versions used for training

5.5. Model Integrity and Version Control

Once a model has been trained, the output of the training process becomes a model artifact —a file or collection of parameters that encode the statistical relationships learned from

training data. These artifacts may be stored in model registries, deployed into operational systems, or used as the foundation for further fine-tuning and model development.

Within the chain-of-custody framework, the trained model represents the final product of the intelligence formation process. It is therefore essential to maintain custody over the model itself to ensure that the intelligence it produces remains consistent with the datasets and training procedures that originally generated it.

If the integrity of the model artifact is compromised, the organization may unknowingly deploy or rely upon a model that differs from the one that was validated, approved, or documented. In such cases, the chain of custody between the model's training process and its operational behavior becomes broken. At this stage, the focus shifts from data custody to model custody.

Several types of custody risks may arise after a model has been trained. These risks include unauthorized model modification, model replacement, uncontrolled fine tuning, and model drift and undocumented evolution.

Unauthorized Model Modification

One of the most significant risks is the possibility that a trained model is altered after validation.

Examples include:

- modification of model weights or parameters
- replacement of the model artifact with an altered version
- modification of inference configuration files
- injection of malicious backdoors into model logic

Such modifications may be subtle and difficult to detect through routine operational monitoring. Even small changes to model parameters can significantly affect predictions or decision outcomes. Without integrity verification mechanisms, organizations may be unaware that the deployed model differs from the version originally approved for use.

Model Replacement

Another risk involves the replacement of a validated model with a different model artifact.

This may occur when:

- multiple model versions exist in development environments
- experimental models are mistakenly deployed to production
- automated pipelines reference incorrect model artifacts
- unauthorized actors deploy alternative models

If the deployed model differs from the validated version, the organization may lose the ability to explain how decisions are being made or whether the model satisfies regulatory requirements.

Maintaining chain of custody therefore requires strict controls over which model versions are authorized for deployment.

Uncontrolled Fine-Tuning

Many AI systems continue to evolve after initial training through processes such as fine-tuning, reinforcement learning, or incremental retraining.

While these processes can improve model performance, they also introduce custody risks.

Fine-tuning may involve:

- additional datasets
- modified training objectives
- updated hyperparameters
- changes in model architecture

If these updates occur without proper documentation and governance, the lineage of the model becomes difficult to reconstruct. Organizations may eventually lose the ability to determine which data influenced the current model and how its behavior evolved over time.

Model Drift and Undocumented Evolution

In operational environments, models may gradually change due to retraining cycles or continuous learning mechanisms. This can lead to model drift, where the model's behavior diverges from its original training characteristics.

If version control systems are not maintained, organizations may struggle to determine:

- when the model changed
- what triggered the change
- which datasets were involved in the update

Maintaining custody over model evolution is therefore essential for preserving accountability.

Version control provides the primary mechanism for maintaining chain of custody over AI models. Just as software development teams use version control systems to track changes in code, AI organizations must track changes in models and their associated training artifacts.

Model version control typically includes maintaining records of:

- model architecture and configuration
- training dataset versions
- training parameters and hyperparameters

- evaluation results and validation metrics
- timestamps of training runs
- identities of individuals or systems responsible for the training

Each time a model is retrained or updated, a new version should be created while preserving previous versions for reference. This ensures that organizations can reconstruct the historical evolution of the model.

5.6. Third-Party and External Data Dependencies

Modern AI systems rarely operate solely on internally generated data. Instead, they often rely on a complex ecosystem of external datasets, vendor-provided data services, open-source corpora, and API-based information feeds. These external inputs may be used to train models, enrich internal datasets, or supply real-time data during model inference.

While external data sources can significantly enhance the capabilities of AI systems, they also introduce substantial risks to the chain of custody. Unlike internally controlled datasets, third-party data originates outside the organization's direct governance framework. As a result, the organization must rely on the integrity, security practices, and collection methods of external providers.

From a chain-of-custody perspective, these dependencies extend the AI data lifecycle beyond the organization's boundaries, creating additional custody links that must be validated and monitored.

Many AI systems rely on third-party data sources such as:

- commercial datasets
- APIs providing real-time information
- external training corpora
- open-source datasets

These dependencies introduce risks outside the organization's direct control.

Risks include:

- vendor data contamination
- undocumented changes in third-party datasets
- licensing violations
- compromised external data pipelines

Chain-of-custody governance must therefore extend beyond internal systems to include vendor risk management and contractual data provenance requirements.

The central challenge with third-party data is that organizations often have limited visibility into how the data was collected, curated, and maintained.

This lack of visibility raises several critical questions:

- Was the data obtained lawfully?
- Was the dataset altered or manipulated before distribution?
- Does the dataset contain hidden biases or adversarial content?
- Has the dataset changed since it was originally acquired?

If the organization cannot answer these questions, the chain of custody for the AI system may be incomplete.

Because AI models internalize patterns from training data, any integrity issues present in third-party datasets may become embedded within the model's behavior.

External data dependencies introduce several risks that can disrupt the chain of custody for AI systems. These risks include unknown data provenance, data contamination, dataset drift and undocumented updates, and dependency risk.

Unknown Data Provenance

In some cases, organizations may not know how third-party data was originally collected.

For example:

- web-scraped datasets may contain copyrighted or proprietary content
- personal data may have been collected without appropriate consent
- historical datasets may lack documentation describing collection methods

If the provenance of the dataset cannot be verified, the organization may face legal or regulatory challenges.

Data Contamination

External datasets may contain corrupted or malicious records introduced by adversarial actors.

For example:

- manipulated images inserted into public training datasets

- fabricated financial transactions inserted into fraud datasets
- misleading text data inserted into language model training corpora

Because these datasets often originate from distributed sources, detecting contamination may be difficult.

Dataset Drift or Undocumented Updates

Third-party datasets may change over time as providers update their data collections.

If organizations rely on external datasets that are periodically refreshed, they must track which version of the dataset was used during model training.

Failure to track dataset versions may lead to situations where the model's training data cannot be reconstructed.

Dependency Risk

Organizations that rely heavily on external data providers may become vulnerable to disruptions if those providers experience operational issues, security incidents, or data quality failures.

In such cases, the reliability of the AI system may be directly affected by external events outside the organization's control.

To address these risks, organizations must extend their chain of custody controls to include third-party data providers. This often requires implementing vendor governance practices such as vendor due diligence, contractual provenance requirements, dataset validation and screening, and external dataset version tracking.

Vendor Due Diligence

Before adopting external datasets, organizations should evaluate the provider's data governance practices, including:

- data collection methodology
- provenance documentation
- data security controls
- update procedures and versioning

Contractual Provenance Requirements

Organizations should establish contractual obligations requiring third-party providers to maintain documented data lineage and integrity controls.

Contracts may specify:

- data licensing rights

- obligations to disclose dataset updates
- requirements for data security and integrity verification
- liability for inaccurate or unlawful data

Dataset Validation and Screening

External datasets should undergo validation procedures before being incorporated into training pipelines.

This may include:

- statistical anomaly detection
- integrity checks using cryptographic hashes
- verification of dataset completeness
- screening for adversarial samples

External Dataset Version Tracking

Organizations should record:

- the specific version of the dataset used for training
- the date of acquisition
- any updates or modifications provided by the vendor

Maintaining this information ensures that the lineage of the training data can be reconstructed. Ultimately, third-party data dependencies should be viewed as part of the broader AI intelligence supply chain.

Just as organizations now monitor software supply chains to prevent vulnerabilities, they must also monitor the flow of external data that shapes AI behavior.

Synthetic Recursion Risk

Synthetic recursion risk occurs when AI-generated data is reused to train future AI system, creating a feedback loop that degrades integrity, amplifies bias, and breaks traceability. This occurs when AI-generated data quietly enters training datasets and is mistaken for real-worlddata, causing the model to learn from its own outputs rather than from reality. Ultimately, the model becomes less accurate, less creative, and more brittle, sometimes referred to as model collapse.

Organizations should adopt controls that can distinguish synthetic vs. human-generated data, such as:

- provenance tagging
- data lineage tagging
- synthetic data segmentation
- model input controls

- continuous validation

Without such controls, synthetic data effectively contaminates the training data pipeline,undermining model integrity and creating regulatory and legal exposure.

By extending chain-of-custody controls across external data sources, organizations can ensure that the intelligence produced by their AI systems remains traceable, lawful, and resistant to manipulation.

5.7. Inference and Operational Deployment

After an AI model has been trained, validated, and stored, it is typically deployed into an operational environment where it performs inference—the process of applying learned patterns to new inputs to generate outputs such as predictions, classifications, or recommendations.

Operational deployment represents the stage where AI systems begin influencing real-world outcomes. These may include medical diagnoses, financial decisions, fraud detection alerts, supply-chain recommendations, or customer interactions.

From a chain-of-custody perspective, this domain introduces a new set of risks because the model is now interacting with live data inputs, external systems, and operational workflows. Even if custody has been preserved through data acquisition, training, and model storage, the integrity of the AI system can still be compromised during deployment.

Maintaining custody at this stage ensures that the intelligence produced by the AI system remains consistent with the validated model and the authorized operational environment.

Chain of custody principles must ensure that the deployed model remains the same model that was approved and that operational inputs are monitored for manipulation.

Several types of risks commonly arise during operational deployment, including:

- deployment of incorrect model versions
- runtime environment manipulation
- adversarial inputs and prompt manipulation
- data drift and environmental changes

- undocumented model changes in production

Deployment of Incorrect Model Versions

One of the most common operational risks occurs when the model deployed into production differs from the model that was validated and approved.

This may occur when:

- automated pipelines deploy experimental models instead of production models
- incorrect model version identifiers are referenced
- configuration errors point systems to outdated models
- deployment processes bypass governance approval workflows

If an incorrect model version is deployed, the organization loses the ability to validate, audit, orstand behind the decisions that system produces.

Maintaining custody requires ensuring that the deployed model artifact is cryptographically identical to the approved version stored in the model registry.

Runtime Environment Manipulation

The environment in which a model operates can influence how it behaves. These systems typically include:

- application servers
- containerized runtime environments
- cloud infrastructure
- orchestration frameworks
- monitoring systems

If these environments are compromised or improperly configured, malicious actors may attempt to:

- intercept data flowing into the model
- modify inference code
- alter configuration parameters
- redirect outputs to unauthorized destinations

These manipulations may not change the model itself but can still affect how the system operates.

Protecting the runtime environment is therefore an essential component of preserving custody during inference.

Adversarial Inputs and Prompt Manipulation

Another major risk in operational AI systems is the manipulation of model inputs.

Adversarial actors may craft inputs designed to exploit weaknesses in the model's learned patterns. Examples include:

- adversarial images designed to fool computer vision systems
- manipulated financial transactions intended to evade fraud detection
- malicious prompts designed to influence language model behavior

In systems involving generative AI, attackers may attempt prompt injection attacks, where crafted instructions embedded within input data cause the model to produce unintended outputs.

These attacks do not alter the model itself but instead manipulate how the model interprets incoming data.

From a chain of custody perspective, this represents a break in the integrity of the model's inference process.

Data Drift and Environmental Changes

Operational environments evolve over time. The statistical properties of input data may change as user behavior, market conditions, or environmental factors shift.

This phenomenon, often referred to as data drift, can gradually degrade model performance.

If input data distributions diverge significantly from the datasets used during training, the model may produce unreliable or biased outputs. For example, a fraud detection model trained onhistorical transaction patterns, for instance, may fail to recognize new attack methods thatemerged after its training data was collected.

Monitoring data drift is therefore necessary to ensure that operational inputs remain within the scope of the model's intended use.

Undocumented Model Updates in Production

In some environments, models may be updated or fine-tuned after deployment. While continuous learning can improve performance, it also introduces custody risks if updates are not carefully controlled.

Undocumented updates may occur when:

- developers modify models directly in production environments
- automated retraining systems deploy updated models without governance approval
- configuration files are changed without logging

These changes can alter model behavior without leaving a traceable record.

To preserve custody, any modification to a deployed model must be documented, versioned, and subject to approval workflows. Organizations must implement controls that ensure operational AI systems remain consistent with their approved configuration.

Key custody controls include secure deployment pipelines, runtime monitoring and logging, input validation and screening, and model behavior monitoring,

Secure Deployment Pipelines

Deployment pipelines should verify that only approved models can be deployed to production environments.

This may include:

- cryptographic verification of model artifacts
- automated validation of model version identifiers
- approval workflows for deployment actions
- separation of development and production environments

These controls help ensure that experimental or unauthorized models cannot be introduced into operational systems.

Runtime Monitoring and Logging

Operational AI systems should generate logs describing model activity and system interactions.

Logs should capture information such as:

- model version used for each inference request
- timestamps of model execution
- identities of systems invoking the model
- characteristics of input data

These records allow organizations to reconstruct the inference process if anomalies occur.

Input Validation and Screening

Organizations should implement mechanisms that monitor and validate incoming data used during inference.

These mechanisms may include:

- anomaly detection systems
- adversarial input screening
- rate limiting and access controls
- input normalization procedures

Such controls help prevent malicious inputs from manipulating model outputs.

Model Behavior Monitoring

Continuous monitoring of model outputs can help detect unusual patterns that may indicate operational problems.

Examples include:

- sudden changes in prediction distributions
- unexpected classification patterns
- spikes in model error rates

Monitoring systems can alert operators when the model's behavior diverges from expected patterns.

The inference domain represents the final link in the chain of custody for AI systems.

While earlier stages ensure the integrity of data and model development, operational controls ensure that the intelligence produced by the model remains trustworthy during real-world use. If custody is lost at this stage, even a well-trained and well-governed model may produce unreliable results.

By maintaining rigorous controls over inference and operational deployment, organizations can ensure that AI-generated intelligence remains traceable, verifiable, and consistent with the system that was originally approved for use.

5.8. Monitoring, Logging, and Auditability

The final risk domain involves the visibility and traceability of AI system operations. Even when organizations implement strong controls over data acquisition, training pipelines, model versioning, and deployment environments, the integrity of an AI system ultimately depends on the organization's ability to observe, record, and verify what actually occurs within the system over time.

Monitoring, logging, and auditability form the observability layer of AI chain-of-custody governance. These mechanisms provide the evidentiary record that allows organizations to

reconstruct the lifecycle of AI intelligence—from the origin of the training data to the outputs generated in operational environments.

Without comprehensive monitoring and logging, organizations may have policies and procedures in place but lack the ability to prove that those procedures were followed. In effect, the chain of custody becomes theoretical rather than verifiable.

For this reason, monitoring and auditability are essential for ensuring that AI systems remain transparent, accountable, and defensible.

Monitoring

Monitoring refers to the continuous observation of system behavior to detect anomalies, performance issues, or security incidents.

In the context of AI systems, monitoring should occur across several layers of the lifecycle. These layers include data monitoring, training activity monitoring, and model behavior monitoring.

Data Monitoring

Organizations must monitor how datasets are accessed, modified, and used within training pipelines.

Monitoring systems may detect:

- unusual patterns of dataset access
- unexpected dataset modifications
- unauthorized data transfers
- statistical anomalies in training datasets

These signals may indicate potential custody breaches such as insider manipulation or data poisoning.

Training Activity Monitoring

Training pipelines generate numerous operational signals that should be monitored in real time.

Examples include:

- initiation of training jobs
- datasets used during training
- changes to training parameters
- duration and resource usage of training runs

Monitoring training activity helps organizations detect unauthorized training runs or unexpected pipeline modifications.

Model Behavior Monitoring

Once models are deployed, organizations must monitor how the model behaves in operational environments.

Indicators of concern may include:

- sudden changes in prediction distributions
- increased model error rates
- abnormal confidence scores
- deviations from expected output patterns

These indicators may suggest that the model is experiencing drift, manipulation, or other operational issues.

Logging

While monitoring focuses on detecting anomalies in real time, logging provides the permanent record of system activity that supports forensic investigation and governance oversight.

Logs capture detailed records of system events and interactions. In the context of AI systems, several types of logs are particularly important. These logs include, data access logs, training logs, model deployment logs, and inference logs.

Data Access Logs

These logs record who accessed datasets and what actions were performed.

Typical log entries include:

- user or system identity
- dataset identifier
- action performed (read, write, delete)
- timestamp of access
- originating system or network location

These records allow organizations to track the handling of training data throughout the AI lifecycle.

Training Logs

Training logs document how models were created.

These logs may include:

- dataset versions used during training
- preprocessing steps applied

- training parameters and hyperparameters
- model architecture information
- start and completion times for training runs

Training logs form a critical component of the chain of custody because they connect the final model artifact to the datasets that produced it.

Model Deployment Logs

Deployment logs track when and how models enter operational environments.

These records may include:

- model version deployed
- deployment environment
- deployment approval status
- system or individual initiating deployment
- timestamp of deployment

Maintaining these logs ensures that organizations can verify that the model used in production corresponds to the validated version stored in the model registry.

Inference Logs

Inference logs capture the operational behavior of deployed models.

Typical inference logs record:

- model version used for each request
- timestamp of inference events
- identity of the calling system or application
- characteristics of input data
- model outputs generated

These records allow organizations to reconstruct how specific AI decisions were produced.

Effective chain of custody governance requires maintaining auditable records across the entire AI lifecycle.

These records support:

- regulatory compliance
- internal audits
- forensic investigations
- model performance analysis

Audibility

Auditability refers to the ability of internal or external reviewers to evaluate how the AI system operates and verify that governance policies have been followed.

Audits may examine records to determine:

- how training datasets were obtained
- how models were developed and validated
- whether appropriate approval processes were followed
- how the system behaves in production environments

Without reliable logs and monitoring records, auditors may be unable to confirm that the organization has maintained proper custody over its AI systems. This can create significant regulatory and legal exposure.

Monitoring, logging, and auditability represent the final layer of protection in the AI chain-of-custody framework. While earlier controls protect the integrity of data and model development processes, monitoring ensures that those controls remain effective over time.

Together, these mechanisms provide the evidence needed to demonstrate that the intelligence produced by an AI system is traceable, trustworthy, and consistent with the organization's governance policies. In this way, observability becomes the mechanism through which the chain of custody for AI systems is continuously validated.

5.9 Risk Domain 2 - Regulatory Exposure

Regulators increasingly expect organizations to demonstrate that AI systems operate under well-defined governance frameworks. Identifying risk domains helps organizations show that they have systematically assessed where custody risks may occur and have implemented controls to mitigate those risks. Specifically, regulators in the United States such as the Federal Trade Commission, Securities and Exchange Commission, and Health and Human Services are focused on the core idea that an organization needs to prove where data came from, how it is handled, and that it wasn't altered in a way that creates risk or harm.

This structured approach supports regulatory expectations related to:

- algorithmic accountability
- data protection and privacy compliance
- transparency in automated decision-making
- auditability of AI systems

From a chain of custody standpoint, AI systems should have the ability to prove end-to-end data origin, data integrity, transformation history, access history, model lineage, and output traceability.

Fundamentally, the regulators viewpoint is that AI risk is no longer just about model accuracy—it's about whether an organization can prove the integrity and lineage of the data and decisions. Without a verifiable chain of custody, AI becomes a source of regulatory and legal exposure rather than competitive advantage.

Organizations must answer:

Can we prove lawful data acquisition and authorized usage?

Most cannot.

5. Conclusion

Chain of custody in AI systems is not merely a technical control—it is a board and executive management responsibility because it directly underpins accountability for decisions that carry financial, legal, and reputational consequences. As AI systems increasingly influence outcomes such as clinical recommendations, financial approvals, fraud detection, and customer interactions, the inability to demonstrate data provenance, integrity, and decision traceability exposes the organization to regulatory scrutiny and liability. Boards are ultimately responsible for overseeing risk, governance, and fiduciary duty; without a verifiable chain of custody, they lack the assurance that AI-driven decisions are reliable, auditable, and compliant. Likewise, executive management must operationalize these controls across data pipelines, models, and access layers to ensure that every input, transformation, and output can be traced and defended. In this context, chain of custody becomes a core element of enterprise risk management—on par with financial controls—requiring active oversight, clear accountability, and continuous validation at the highest levels of the organization.

The integrity of AI systems is no longer defined solely by model performance, but by the trustworthiness of the data and processes that underpin them. Chain of custody has emerged as the foundational control that enables organizations to demonstrate provenance, integrity, and accountability across the AI lifecycle. Without it, even the most advanced models become opaque, unexplainable, and ultimately untrustworthy. As regulatory

expectations accelerate and AI-driven decisions increasingly impact financial, clinical, and societal outcomes, organizations must treat chain of custody not as a technical feature, but as a strategic imperative. Those who invest in verifiable data lineage and governance will not only reduce risk—they will define the standard for trusted AI.

At its core, chain of custody is about accountability. When AI systems influence outcomes that affect people, finances, and safety, organizations must be able to answer a simple question: Can we prove how this decision was made? If the answer is no, the risk is not theoretical—it is immediate and material. Breakdowns in data lineage, integrity, or access control create blind spots that can lead to bias, error, and regulatory exposure. Establishing a robust chain of custody closes these gaps, transforming AI from a source of uncertainty into a system of record that can be trusted, audited, and defended.

From an architectural perspective, chain of custody represents the evolution of security from protecting systems to protecting data and decisions. It requires the integration of identity, access control, logging, and integrity verification across distributed AI pipelines. This includes not only human actors, but also non-human identities—models, agents, APIs, and automated processes. The challenge is not simply capturing data lineage, but ensuring it is cryptographically verifiable, tamper-evident, and continuously monitored. Organizations that operationalize these capabilities will be able to detect anomalies, prevent data contamination, and maintain trust in model outputs at scale. In this sense, chain of custody becomes the backbone of resilient AI systems.

Regulators are rapidly converging on a central expectation: organizations must be able to demonstrate how AI systems are built, trained, and operated with verifiable integrity. Chain of custody provides the mechanism to meet this expectation, enabling traceability from data origin through model output. In its absence, organizations face heightened exposure to regulatory enforcement, litigation, and reputational harm. As frameworks such as the EU AI Act, NIST AI RMF, and sector-specific regulations mature, the ability to produce auditable evidence of data provenance, access controls, and transformation history will become a baseline requirement. Chain of custody is no longer optional—it is the control plane for compliance in the age of AI.

We are entering a new era where trust is no longer inferred—it must be proven. In AI systems, this proof is established through an unbroken chain of custody that connects data, models, and decisions with verifiable integrity. As synthetic data, autonomous agents, and machine-to-machine interactions proliferate, the boundaries of traditional governance dissolve, making provenance and traceability the new anchors of trust. The organizations that succeed will be those that treat data as an asset with identity, lineage, and accountability. In doing so, they will not only safeguard against risk, but unlock the full potential of AI as a trusted partner in decision-making.

Acknowledgments

The author would like to thank Rochelle Benning for her thoughtful review, editorial guidance, and contributions that helped strengthen the clarity and quality of this paper.