TDW on Trustworthy AI

Theme Development Workshops

March 6th, 2026
Paris, France (Hybrid event)

The integrity challenge of the AIaaS model

Dr. Georgios Spathoulas
Associate Professor IIK@NTNU
ENFIELD Technical Coordinator

Funded by the European Union

INTRODUCTION

The Rise of AI as a Service

- Artificial Intelligence has revolutionized modern life, enabling systems that learn from data and make autonomous decisions
- However, the **computational demands** and **expertise** required for AI development have created **barriers** for many organizations
- **AIaaS** emerged as the solution, allowing **major cloud providers** like Amazon Web Services, Google Cloud, Microsoft Azure, and OpenAI to offer **sophisticated AI capabilities through simple API calls**
- This democratizes access to cutting-edge AI without requiring massive infrastructure investments or specialized expertise.

4

Major Providers

Leading AIaaS platforms dominating the market

The Black Box Problem

Limited Visibility

Users cannot directly access or inspect the AI models they're using

Trust Gap

Clients must rely on provider reputation rather than technical verification

Probabilistic Outputs

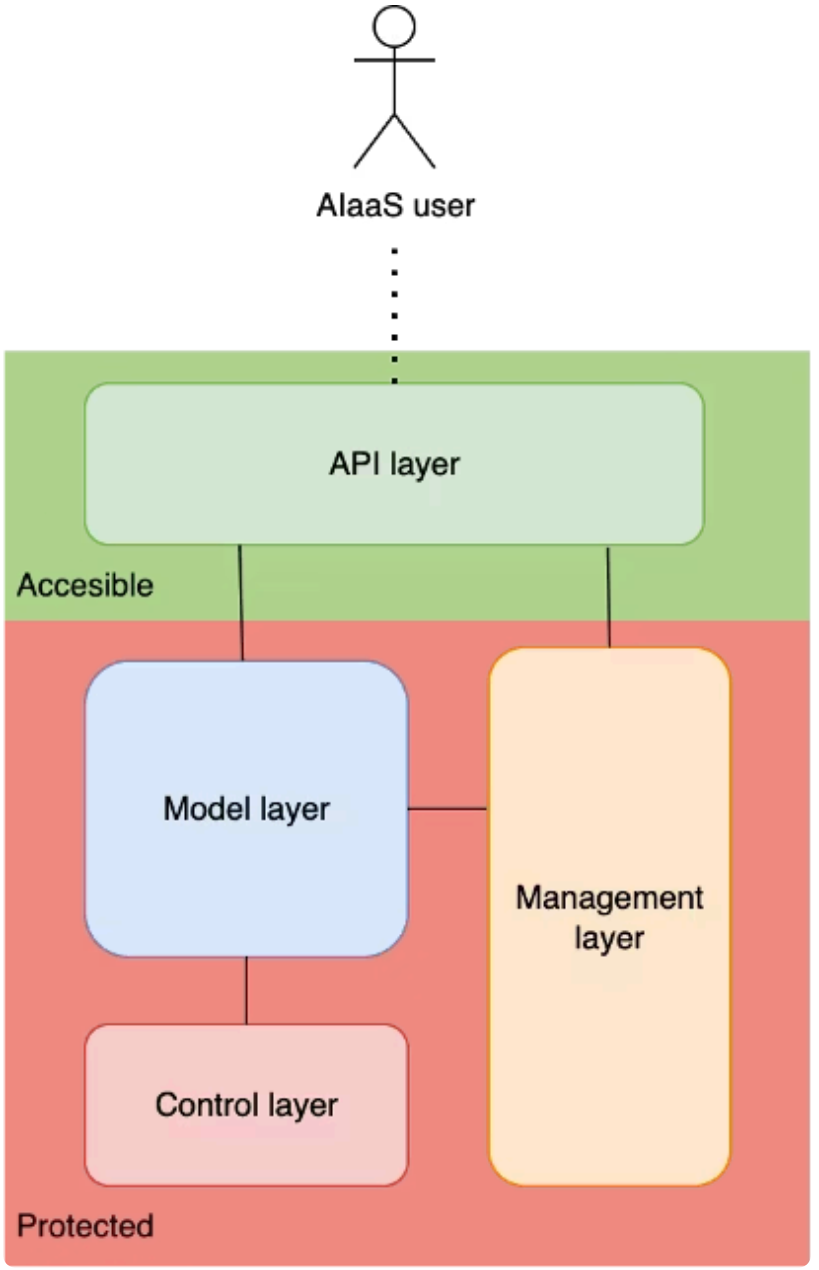
Non-deterministic results make it difficult to detect model substitutions

ARCHITECTURE

AIaaS Reference Architecture

- **Model Layer** Secured trained models, inaccessible to consumers
- **API Layer** Consumer access point, separated from actual models
- **Management Layer** Authentication, authorization, monitoring, and logging
- **Control Layer** Model versioning, updates, and fine-tuning management

This separation creates the **fundamental integrity concern**: users cannot technically verify that their requests are served by the contracted model.



The Evolution of Machine Learning Workflows

Traditional Approach

- Single entity controls entire pipeline
- Data acquisition through deployment
- High upfront infrastructure costs
- Limited scalability
- Complete visibility and control

AIaaS Model

- Decomposed workflow across actors
- Provider handles training and deployment
- Pay-as-you-go cost efficiency
- Unlimited scalability
- Black-box user experience

⚠ CHALLENGES

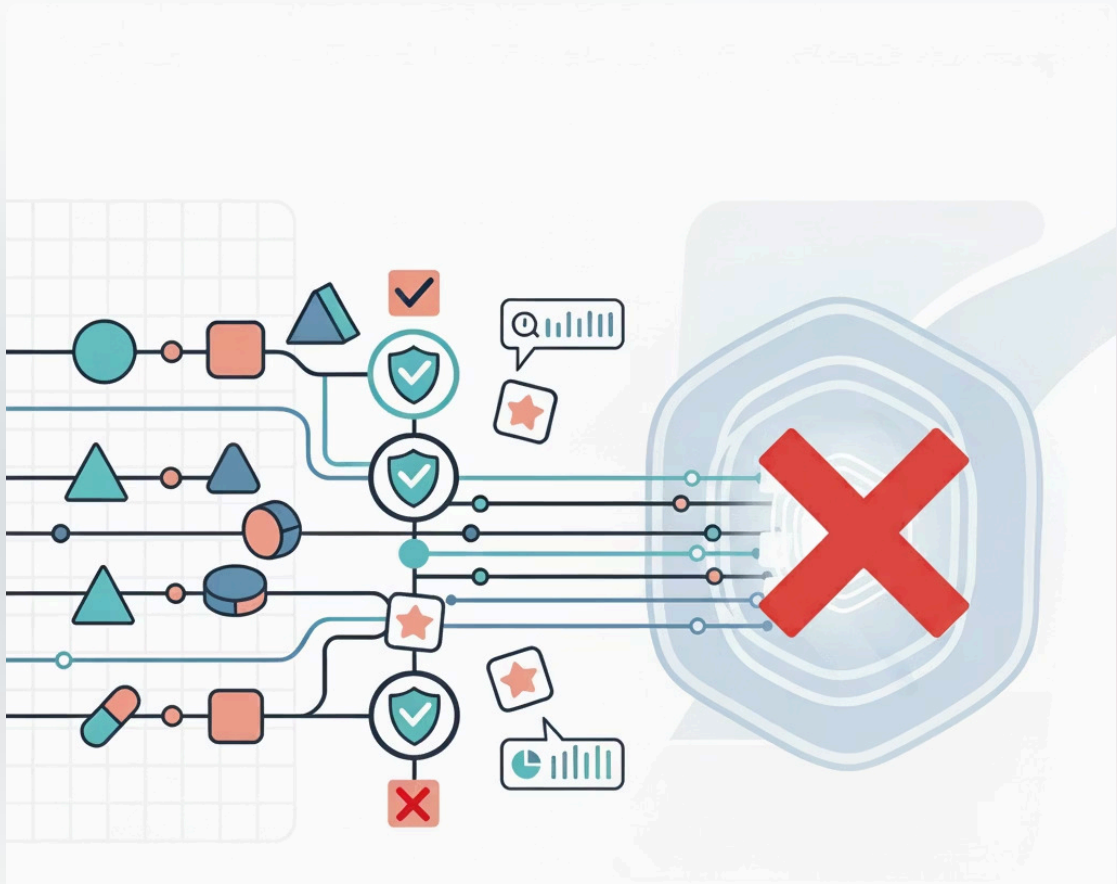
Direct Model Substitution

- Providers may replace agreed-upon models with cheaper alternatives
- The probabilistic nature of AI outputs makes detection difficult
- Operation can cost from $0.03 to $75+ per million tokens depending on the model.

⚠ CHALLENGES

Training Data Integrity

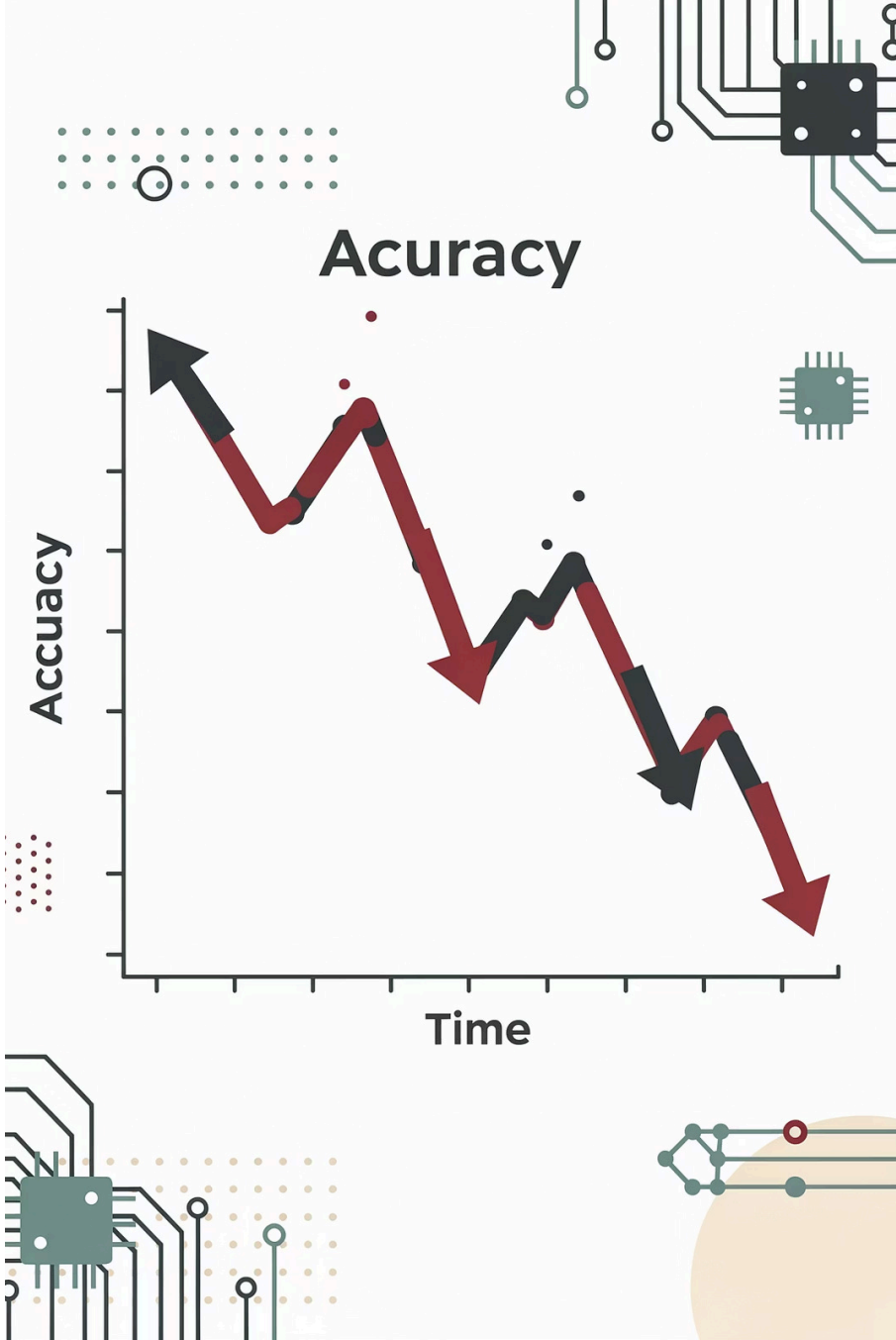
- Model quality depends entirely on training data quality
- Data poisoning attacks or accidental alterations during preprocessing can compromise model performance
- Since training occurs on the provider's side, users cannot verify data integrity.



⚠ CHALLENGES

Training Drift

- Continuous training and periodic fine-tuning create moving targets
- Performance variations may result from legitimate model evolution or problematic drift.
- Without access to the model, users struggle to distinguish between expected adaptation and concerning degradation, making it difficult to maintain service quality standards.



SOLUTIONS

Technical Countermeasures

Model Provenance

Cryptographic tracking of model lifecycle with verifiable identifiers for each version

Model Watermarking

Embedded patterns enabling authentication without affecting performance

Verifiable Execution

Trusted execution environments providing cryptographic proof of correct model usage

Data Validation

Pipeline quality controls detecting poisoning and anomalies in training data

Drift Monitoring

Continuous performance tracking with explainability tools for transparency

Cryptographic Logging

Immutable audit trails ensuring accountability and non-repudiation

Model Provenance: Tracking the Lifecycle

- Model provenance creates an auditable trail from model creation through deployment
- Each version receives a **cryptographically verifiable identifier**—such as a hash or digital signature—that clients can validate.
- When providers update models, they share new identifiers with clients, who can then **verify** the model in use matches the agreed version
- This approach mirrors software supply chain verification but adapts to machine learning artifacts' unique characteristics

Watermarking and Fingerprinting

Watermarking embeds subtle patterns into model parameters or behavior without degrading performance. These patterns remain undetectable during normal use but can be verified through predefined processes to confirm model authenticity.

Benefits: Enables authentication, prevents theft, secures supply chain, and allows unauthorized substitution detection

Fingerprinting takes a complementary approach by observing specific model features and correlating them with known versions. For deterministic models, recording outputs on predefined test inputs creates a unique signature for identification.

Benefits: Non-invasive verification, version identification, and compatibility with existing models

Trusted Execution Environments

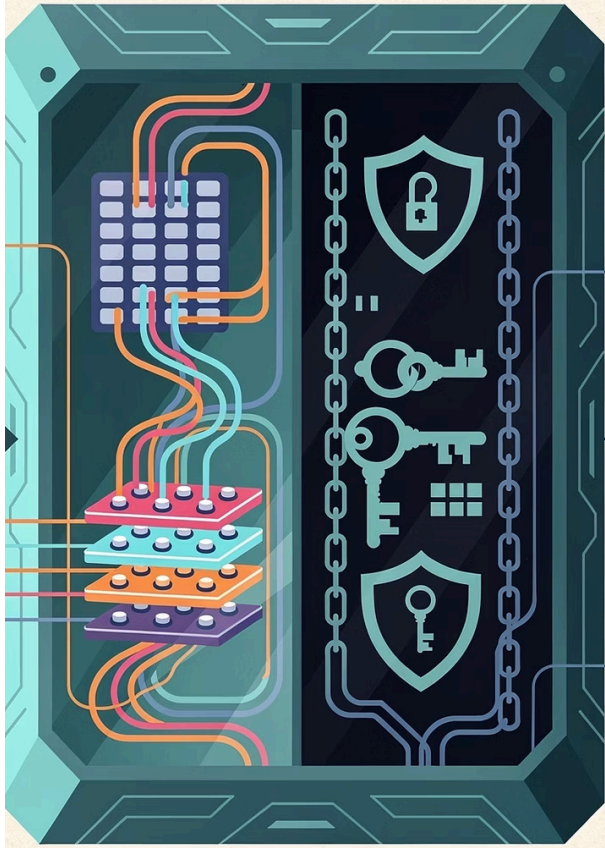
Intel SGX

Hardware-isolated secure enclaves

ARM TrustZone

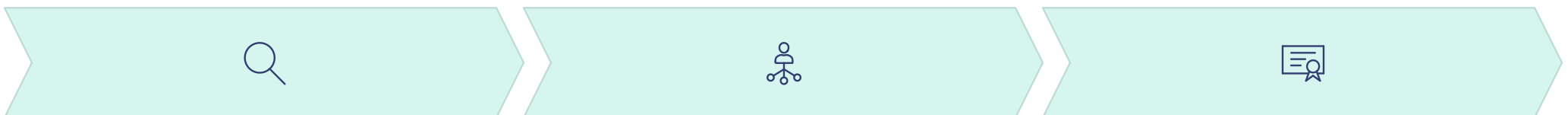
Processor-level security zones

- Trusted Execution Environments (TEEs) enable processing within hardware-protected enclaves that generate cryptographic proofs of correct execution
- For AIaaS, this means providers can prove inference occurred with the specified model without tampering.
- While TEEs add computational overhead, they provide crucial assurance for regulated industries



Data Pipeline Quality Validation

Training data integrity directly impacts model reliability. Countermeasures focus on continuous observation and validation:



Anomaly Detection

Statistical analysis identifies outlier data points that may indicate poisoning attempts

Source Redundancy

Cross-validation across independent pipelines before training integration

Certified Defenses

Formal guarantees about resilience to bounded poisoned samples

Drift Monitoring and Explainability

- **Model drift** poses unique challenges in continuous training scenarios
- Without a static reference model, integrity verification shifts to **quality preservation**
- **Continuous monitoring** tracks performance metrics, compares incoming data distributions with historical baselines
- **Explainability** tools complement this by revealing sudden changes in feature importance that may indicate drift before performance degrades

Cryptographic Logging and Auditing

When real-time prevention fails, post-hoc accountability becomes essential. Cryptographic logging systems provide secure, non-repudiable records of all model operations:

Event Recording

Every model update, inference request, and response is logged with digital signatures

Immutable Storage

Blockchain-backed ledgers or zero-knowledge proof systems prevent tampering

Violation Detection

Integrity breaches trigger alerts and initiate investigation protocols

Event Reconstruction

Complete audit trails enable high-confidence forensic analysis

GOVERNANCE

Organizational Countermeasures

Technical protections alone cannot ensure AIaaS integrity

Organizational measures create governance structures, contractual frameworks, and cultural practices that complement technical safeguards:

- Transparency Policies
- Independent Auditing
- Contractual Liability
- Awareness Training

Transparency as Policy

Mandating **transparency** transforms the provider-client relationship from blind trust to informed partnership

Providers must disclose all model updates, retraining events, and version changes, including expected performance differences

This information flow enables clients to validate new versions, prepare for changes, or refuse updates that don't meet their requirements

- **Version Identifiers** Unique tags for each model release
- **Change Documentation** Detailed descriptions of modifications
- **Performance Metrics** Comparative benchmarks across versions
- **Advance Notice** Sufficient time for client preparation

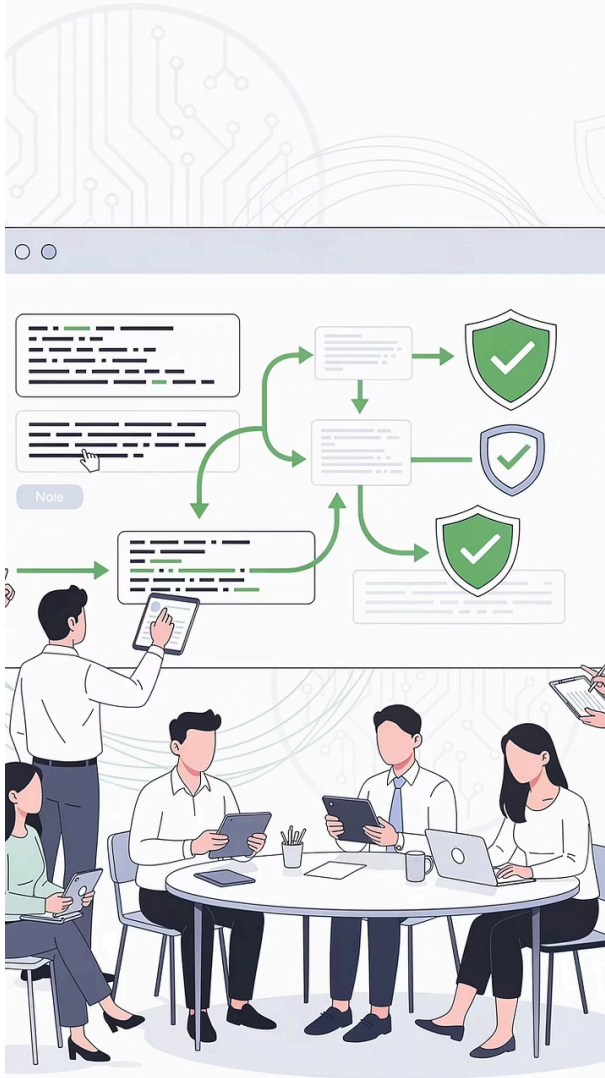
Independent Auditing

Third-party audits provide independent oversight that voluntary transparency cannot guarantee

External auditors examine fairness, robustness, data protection practices, and versioning integrity against industry standards

- Audit Planning
- Technical Review
- Report Generation
- Certification
- Continuous Monitoring

Industry-wide certification schemes create recognizable seals of integrity, enabling clients to **compare providers** on **trustworthiness** alongside **performance** and **cost** considerations



Contractual Liability Frameworks

Service-level agreements (SLAs) can enforce integrity obligations through contractual mechanisms. Well-designed SLAs specify model version guarantees, notification requirements, audit rights, and liability allocation for integrity violations.

Version Guarantees

Contractual commitment to specific model versions with defined update protocols

Notification Requirements

Mandatory advance disclosure of changes with sufficient preparation time

Audit Rights

Client authority to request independent verification of service integrity

Liability Clauses

Financial responsibility for damages resulting from integrity violations

Building an Integrity-Aware Culture

- Training programs must educate personnel about AIaaS integrity importance.
- On the provider side, training ensures service delivery according to agreed terms
- On the client side, it enables continuous monitoring and assessment of received services
- This cultural shift treats AIaaS like any outsourced service requiring appropriate scrutiny.

From Trust to Verification

1

Current State

Blind trust in provider reputation and contracts

2

Technical Layer

Cryptographic verification of model identity and execution

3

Organizational Layer

Governance frameworks ensuring accountability

4

Future State

Trust in systems and institutions, not just providers

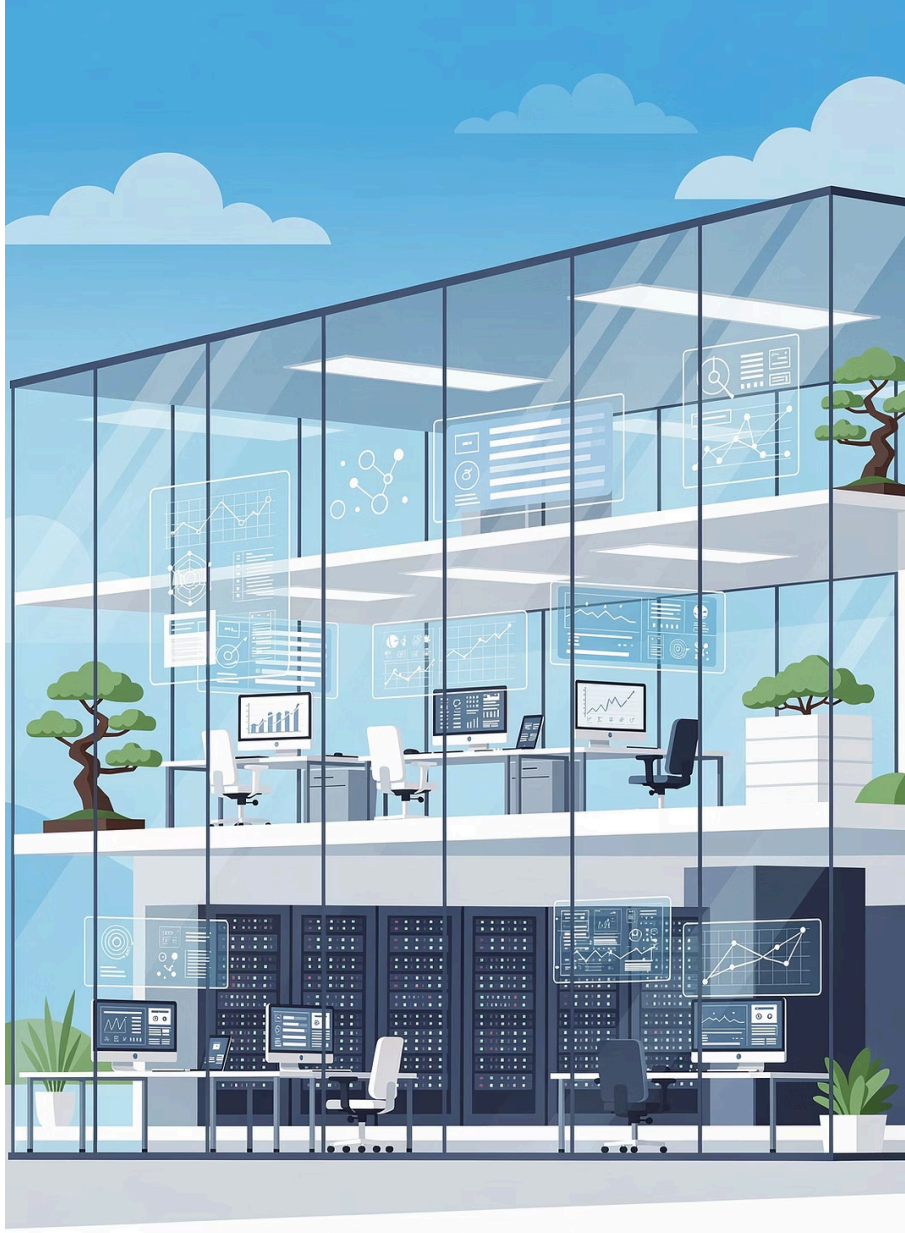
INDUSTRY

Real-World Use Cases

Leading AIaaS providers have begun implementing integrity measures, recognizing that trust and transparency provide competitive advantages in the rapidly evolving AI marketplace.

The following cases demonstrate practical applications of integrity countermeasures.

- Microsoft Azure Confidential AI
- Google's Model Cards Initiative
- OpenAI's Trust Portal
- IBM and Casper Labs: Blockchain for AI Transparency



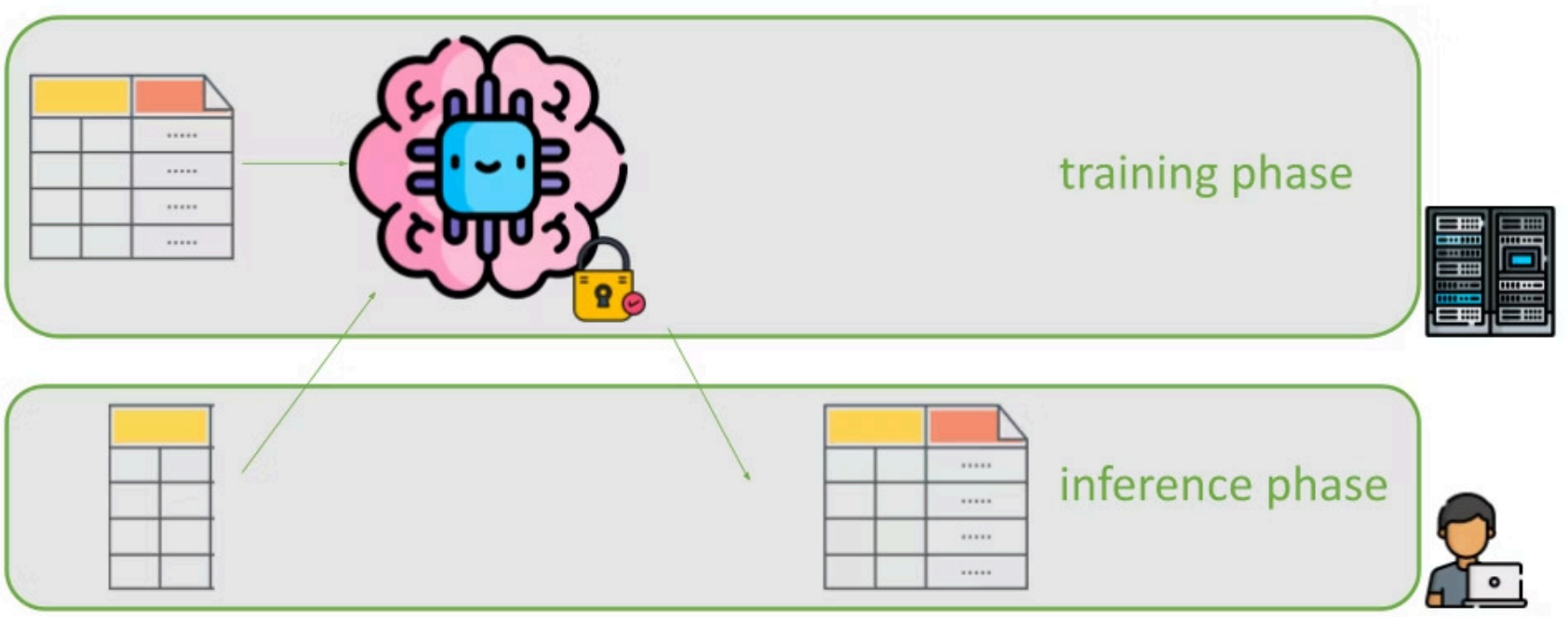
Verifiability through ZKPs in AIaaS

Cryptographic techniques like Zero-Knowledge Proofs can revolutionize auditing and trust in AI systems without exposing sensitive data.

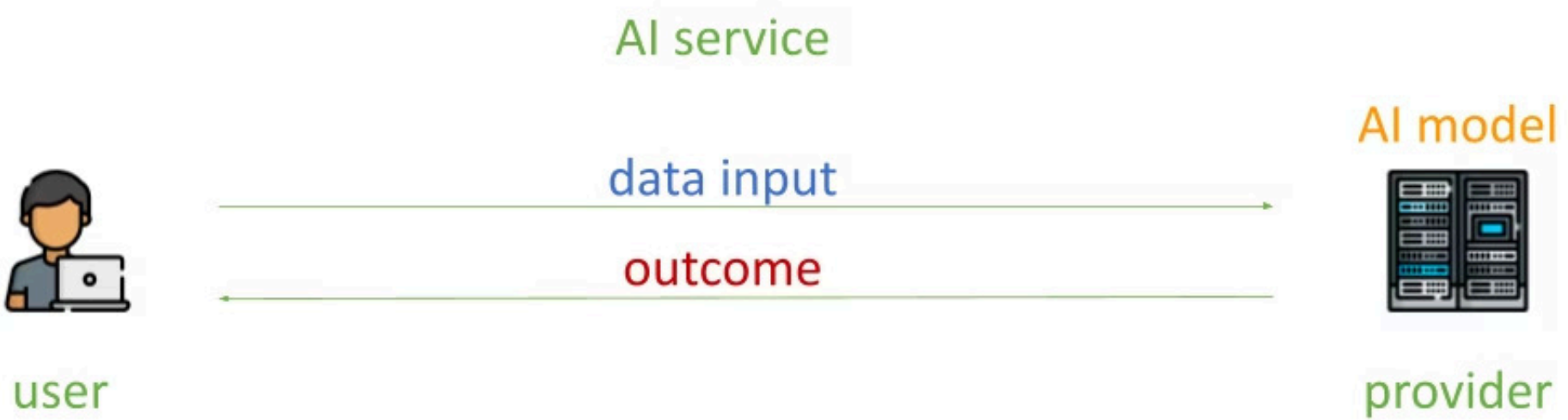
ENFIELD

Traditional AI Workflow

Modern AI Workflow



Verifiability requirement



user requires to be able to verify that outcome has been produced by feeding data input to AI model

Zero Knowledge Proofs

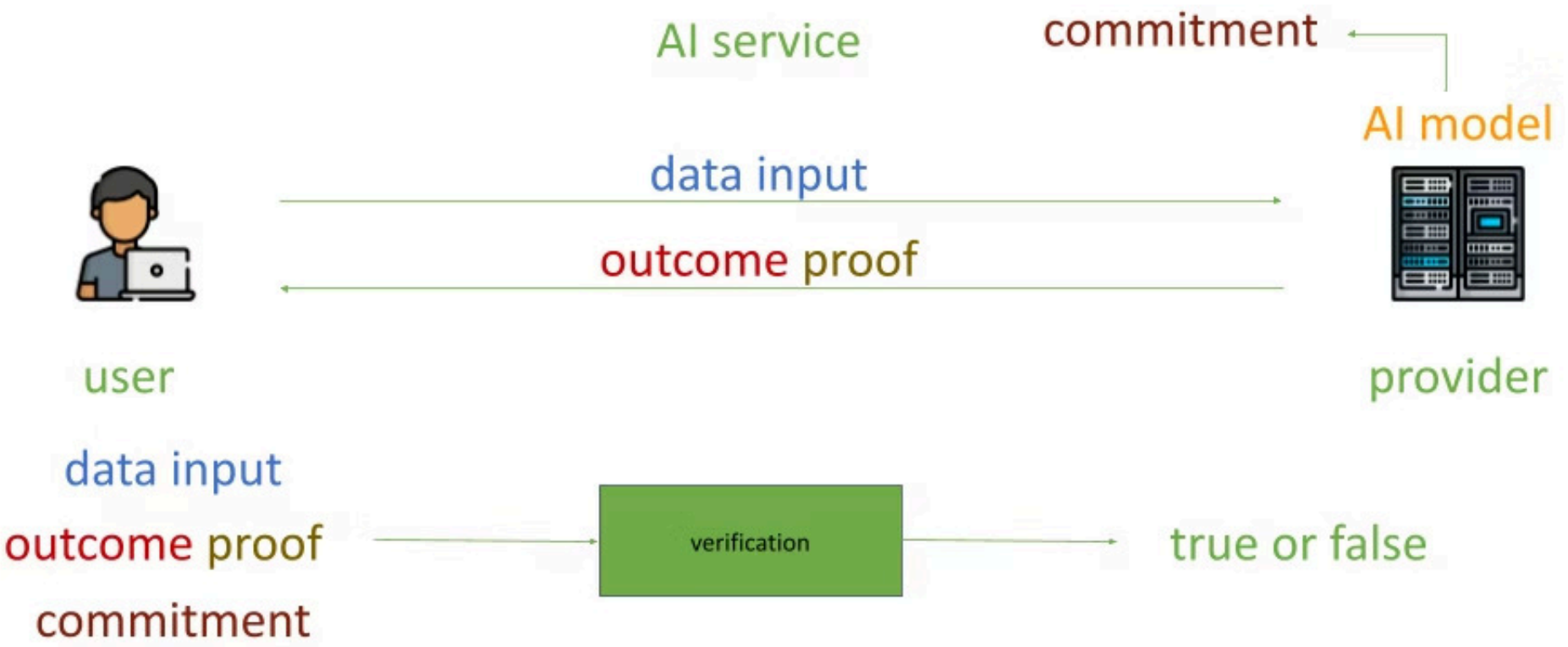
A **zero-knowledge (ZK) proof** is a cryptographic protocol

One party, the prover, can prove to another party, the verifier, that a given statement is true, without revealing any additional information beyond the fact that the statement is true

ZK is used to create proofs of computational integrity for a set of given computations where:

- The proof is significantly easier to verify than it is to perform the computation itself (succinctness)
- Hiding parts of said computation whilst preserving computational correctness is feasible (zero-knowledge)

Verifiability with Zero Knowledge



Challenges

Parameter distortion during the process of quantifying ML data

- ML models use floating-point numbers
- ZK circuits require the use of fixed-point numbers

High computational requirements for ZK proofs of large-scale models

- Popular zero-knowledge schemes only support small-scale and relatively low computational ML scenarios

CONCLUSIONS

The Path Forward for AIaaS Integrity

Technical Solutions

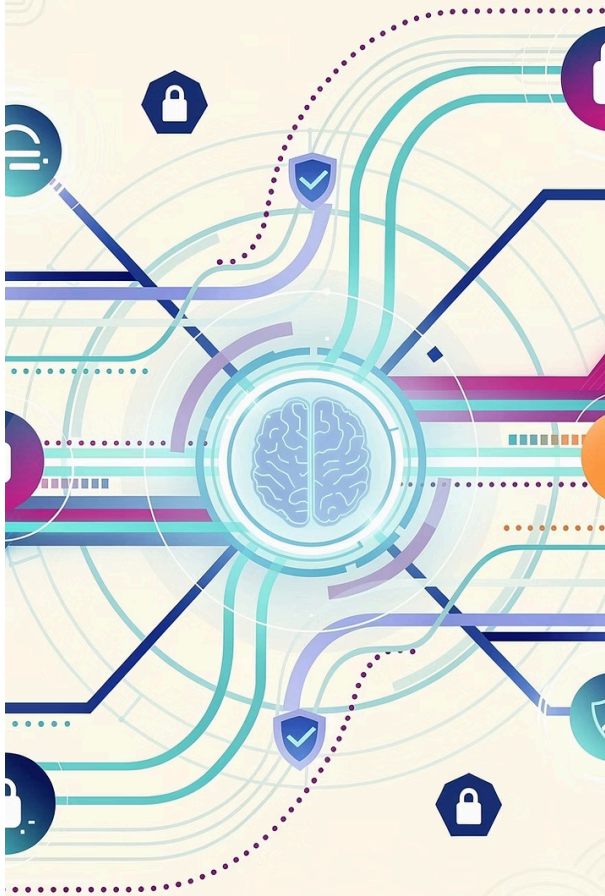
Provenance tracking, watermarking, TEEs, and cryptographic logging provide verification mechanisms

Organizational Frameworks

Transparency policies, auditing, contracts, and training enforce accountability

Industry Leadership

Major providers demonstrate commitment through implemented integrity measures



The Future of Trustworthy AI Services

Integrity is not optional

It is fundamental to AIaaS adoption in critical domains

Solutions exist today

Technical and organizational countermeasures are proven and deployable

Industry momentum is building

Major providers recognize integrity as a competitive differentiator

Collaboration is essential

All stakeholders must work together toward a trusted ecosystem

The future of AI depends on our ability to ensure that users truly get what they are promised

TDW on Trustworthy AI

Theme Development Workshops

March 6th, 2026
Paris, France (Hybrid event)

Funded by the European Union