NATO OTAN
Science & Technology Organization (STO)

Securing Military Applications of Generative AI

A Framework for Resilient and Trustworthy Deployment

Jason Samarin (US Navy PEO C4I) , Andrés Vega (M42)

[redacted]

NATO IST-HFM-225 Research Specialists Meeting

M42

Jason Samarin

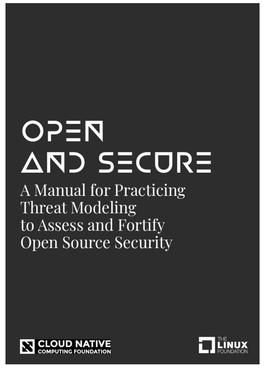
Principal Software Engineer, PEO C4I

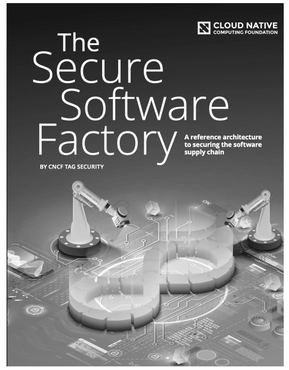
Andrés Vega

M42

Founder and CEO

Publications





NATO IST-HFM-225 Research Specialists Meeting

"The root tenets of command and control are timeless—but they have been lost in the chase for new technologies.

Commanders must exert exacting control over their forces to advance their plans if they are to defeat that future adversary who is multidimensional, well equipped, well trained, willing to fight, and intending to win."

- Rediscover the Art of Command and Control, Vice Admiral Robert F. Willard, USN

Effective C2 in Modern Warfare

- Command grants authority; Control directs action.
- C2 delivers timely, decisive moves.
- Commanders need insight beyond the front line.
- Goal: Real-time enemy picture and rapid response.
- Tech multiplies force through instant data and execution.

Use Cases for AI Enabled Differentiation

- **Real-time Situational Awareness:** Live geospatial intelligence, blue force tracking
- **Secure Data Exchange:** Cross-classification data sharing, coalition interoperability
- **Mission Collaboration:** Secure real-time translation, chat, annotations
- **Intelligent ISR Integration:** Live drone and sensor data

Additional Tactical Capabilities:

- **Advanced Targeting**: AI-driven precision targeting
- **Offline Sync**: Persistent access to mission-critical data
- **Dynamic Planning**: Real-time mission adaptation
- **Biometric Verification**: Rapid identity management
- **Threat Alerting**: Immediate sensor-based threat detection

The Growing Risk of AI in C2

AI relies on sensitive data, expanding breach opportunities:

- Misused AI data can lead to theft, harm, and mission failures
- Breaches in AI systems allow lateral movement, amplifying damage
- Poor isolation in AI systems enables systemic misuse

AI Safety Levels and Risk

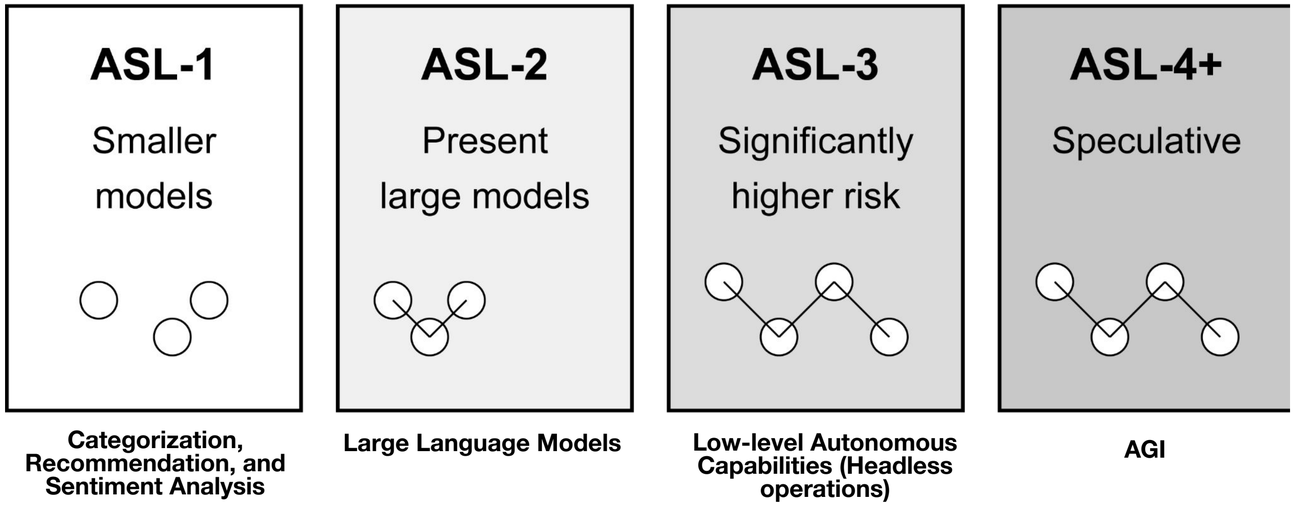
ASL-1: Low safety concerns, negligible impact of errors (e.g., recommendation systems, spell-check).

ASL-2: Moderate safety concerns, some regulatory or ethical considerations (e.g., financial forecasting, marketing tools).

ASL-3: High safety concerns, significant impact on lives or financial stability (e.g., autonomous vehicles, medical diagnostics).

ASL-4: Critical safety concerns, life-and-death or societal impact (e.g., predictive policing, AGI, autonomous weaponry).

ASL-5+: Not yet defined as it is too far from present systems, but will likely involve qualitative escalations in catastrophic misuse potential and autonom



Source: Anthropic's Responsible Scaling Policy, introducing AI Safety Levels (ASL) to manage risks in advanced AI systems.
Read more

When threat modeling AI, the primary consideration is to assume potential compromise or poisoning of both the training data and the data provider.

As defenders, we must create the ability to detect anomalous and malicious data entries, differentiate between them, and implement strategies for recovery.

Expanded Attack Surface

AI systems introduce unique challenges to security, expanding the attack surface in ways traditional systems do not:

1. AI systems evolve rapidly, exposing new vulnerabilities.
2. Dynamic Threats: Susceptible to adversarial attacks, data drift, and model degradation.
3. Black Box Risk: Hard to detect poisoned data or malicious inputs.
4. Operational Weaknesses: Poor security enables prompt injection and model inversion attacks.
5. Mitigation Required: Tailored strategies are essential to counter AI-specific threats.

This combination of evolving threats, opaque models, and operational weaknesses creates a critical security challenge, requiring tailored strategies to mitigate these risks.

Military AI Threat Landscape

Adversarial Threats

- Adversarial Inputs: Malicious signals deceiving ISR and targeting systems
- Data Poisoning: Corrupted sensor data degrading model reliability
- Model Theft: Extracting decision logic for adversary advantage

Operational Risks

- Black Box Risk: Difficulty detecting subtle malicious inputs
- Data Drift: Models failing under real-world battlefield variations
- Prompt Injection: Manipulated inputs distorting intelligence outputs

Command Integrity

- Securing information flows against unauthorized access
- Ensuring AI enhances commander situational awareness
- Preserving human judgment under degraded conditions

Key Questions for an AI Security Review

Data Integrity & Poisoning

- How do you detect tampered data?
- Is input data validated and documented?

Training Data Security

- How are model-data links secured?
- Can data sources alert to compromise?
- Is sensitivity of data assessed and cataloged?

Model Output & Privacy

- Could your model leak sensitive information?
- Are unnecessary or risky outputs exposed?
- Can attackers extract training data?

Anomaly Detection & Recovery

- How do you trace accuracy declines?
- How are invalid or malformed inputs managed
- Can silent output errors be detected?

Adversarial Resilience

- Is training robust against adversarial inputs?
- How quickly can you recover or revert models?
- Can threats be isolated and mitigated?

Data Lineage & Provenance

- Can data issues be traced to their origin?
- Which data points are vulnerable to manipulation?
- Who contributes to your data, and how might they be exploited?

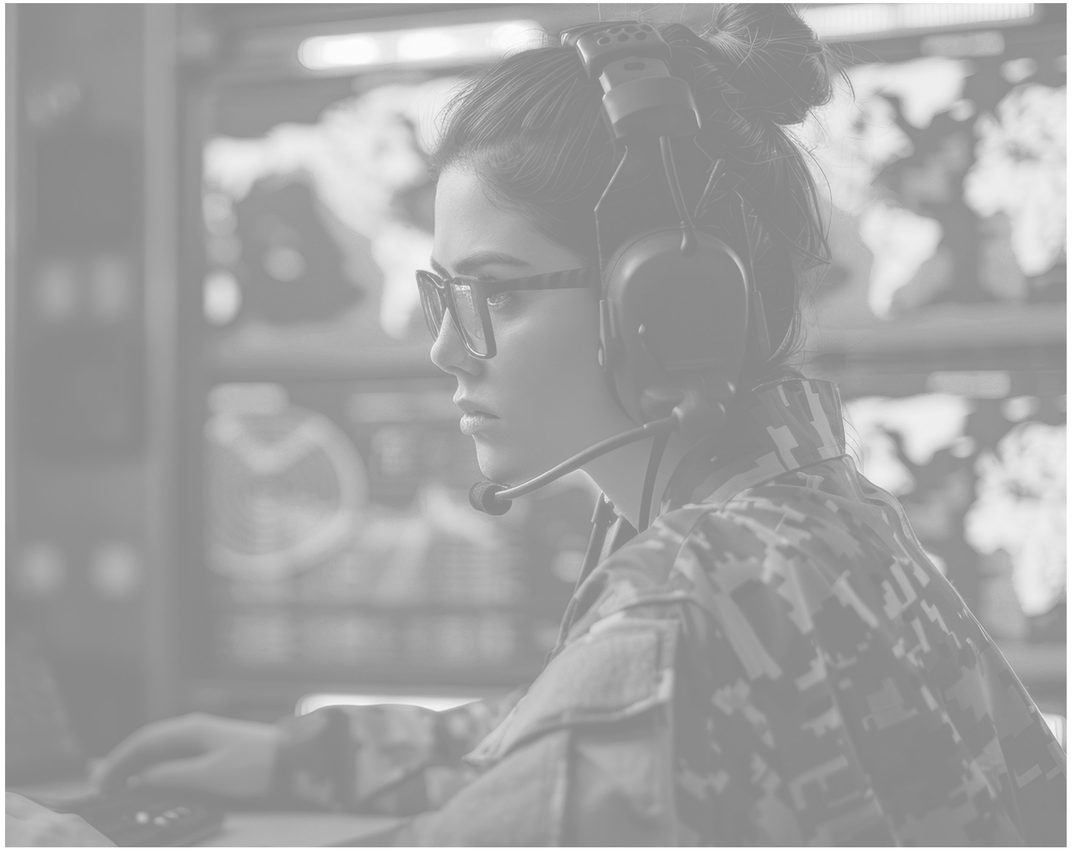
Supply Chain & External Dependencies

Supply Chain Risks:

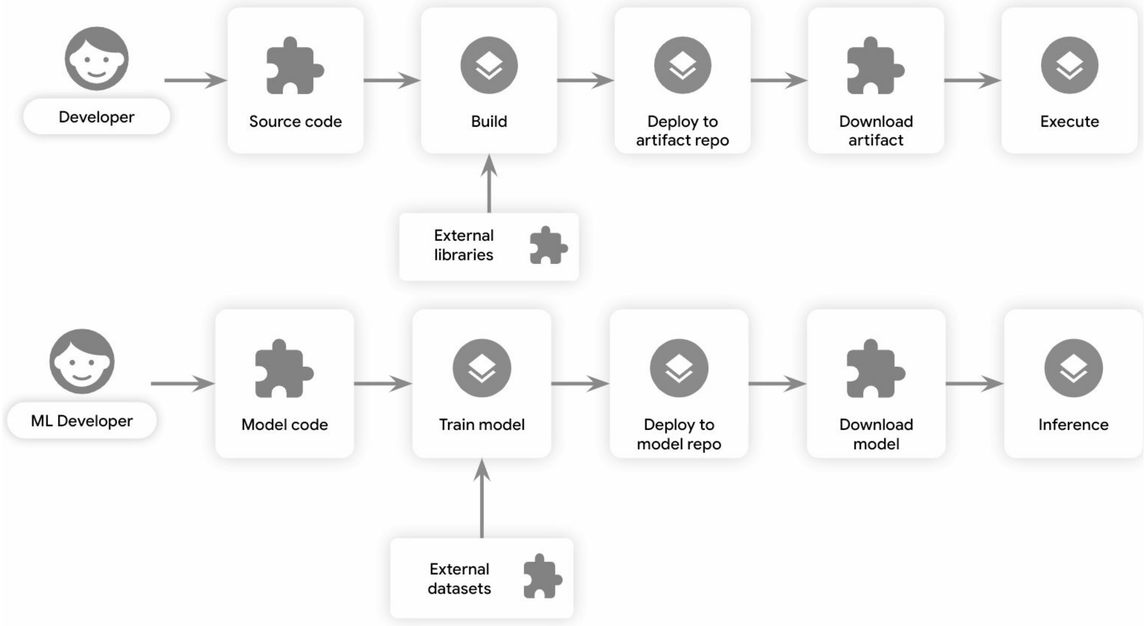
- Military AI relies on classified data and proprietary models.
- Threats from compromised third-party AI/ML providers.

Mitigation Strategies:

- Strict verification of AI supply chain components.
- Implement military-grade authentication & encryption.
- Continuous monitoring of AI dependencies and third-party software.
- Secure AI/ML deployment with controlled access to sensitive algorithms.

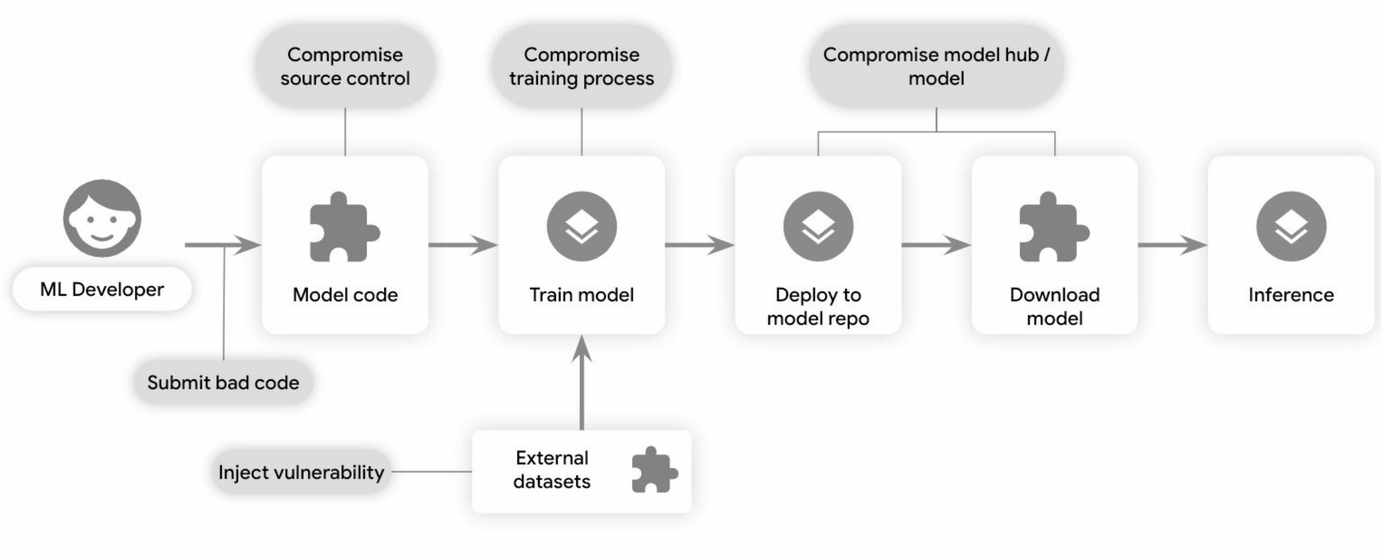


Similarities between software development and ML model development



Source: Increasing transparency in AI security - Google Security Blog

Attack vectors on ML through the lens of the ML supply chain



Who published the **model**? Are they trustworthy? Did they use safe practices?

For open source **models**, what was the **training** code?

What **datasets** went into **training** that **model**?

Could the **model** have been replaced by a tampered version following publication? Could this have occurred during **training** time?

And what about… Information Aggregation Risks in Military AI?

Hypothetical scenario:

Each service member's device telemetry; location pings, login times, and training-schedule metadata, is fed into a transformer model.

Potential adversary insights from the aggregate:

- Strategic or operational intent
- Force posture shifts
- Command priorities or upcoming missions

Seemingly unclassified metadata, once combined, can expose sensitive patterns and create a new intelligence vulnerability.

Question: Should unclassified data be reclassified at a higher level when aggregated and processed by AI systems?

Microsoft AI Research Data Leak (2023)

What Happened:

Microsoft exposed **38TB** of sensitive internal data due to a misconfigured **Shared Access Signature (SAS) token** in Azure Blob Storage. The leaked data included private keys, passwords, and backup workstation images.

Key Takeaways:

- **Static Credentials:** The SAS token, which was long-lived and overly permissive, led to unauthorized access risks.
- **Lack of Key Access Control:** Exposed credentials enabled potential misuse of sensitive data encryption keys.
- **Poor Isolation:** Misconfigured permissions allowed access to unrelated sensitive workloads.

Applicable lessons:

- **Encrypting Data at Rest:** Ensure all sensitive data is encrypted with robust algorithms to prevent exposure, even in case of misconfigurations.
- **Key Access Control:** Implement strict controls to determine which workloads or users can access keys, reducing the blast radius of breaches.
- **Dynamic, Short-Lived Credentials:** Replace static tokens with time-bound credentials to minimize the risk of unauthorized access.
- **Continuous Monitoring:** Regularly scan for and address misconfigurations in storage and key management systems.

Compensating Controls

1. **Credential Exposure**	More workloads mean more opportunities for stolen credentials. **Control:** Use short-lived, dynamic credentials and limit access to critical systems.
2. **Autonomous Workloads**	Privileged AI systems can act unpredictably or escalate risks without oversight. **Control:** Enforce strong identity controls and limit actions based on attested behavior.
3. **Data Integrity**	AI depends on sensitive data that can be poisoned or inadvertently exposed. Secure data pipelines and implement training data provenance tracking.

Autonomous agents access unprecedented amounts of data and systems, making robust safeguards essential to prevent unimaginable harm.

- Key protection secures data from unauthorized access.
- Hard isolation prevents lateral movement and secures workloads.
- Credential management reduces persistent threats.
- Zero trust ensures secure, efficient credential distribution.

The Copilot Oversharing Problem

Key Issues

- AI accessed data across intended boundaries
- Traditional permissions failed to contain exposure
- Data aggregation created unexpected connections

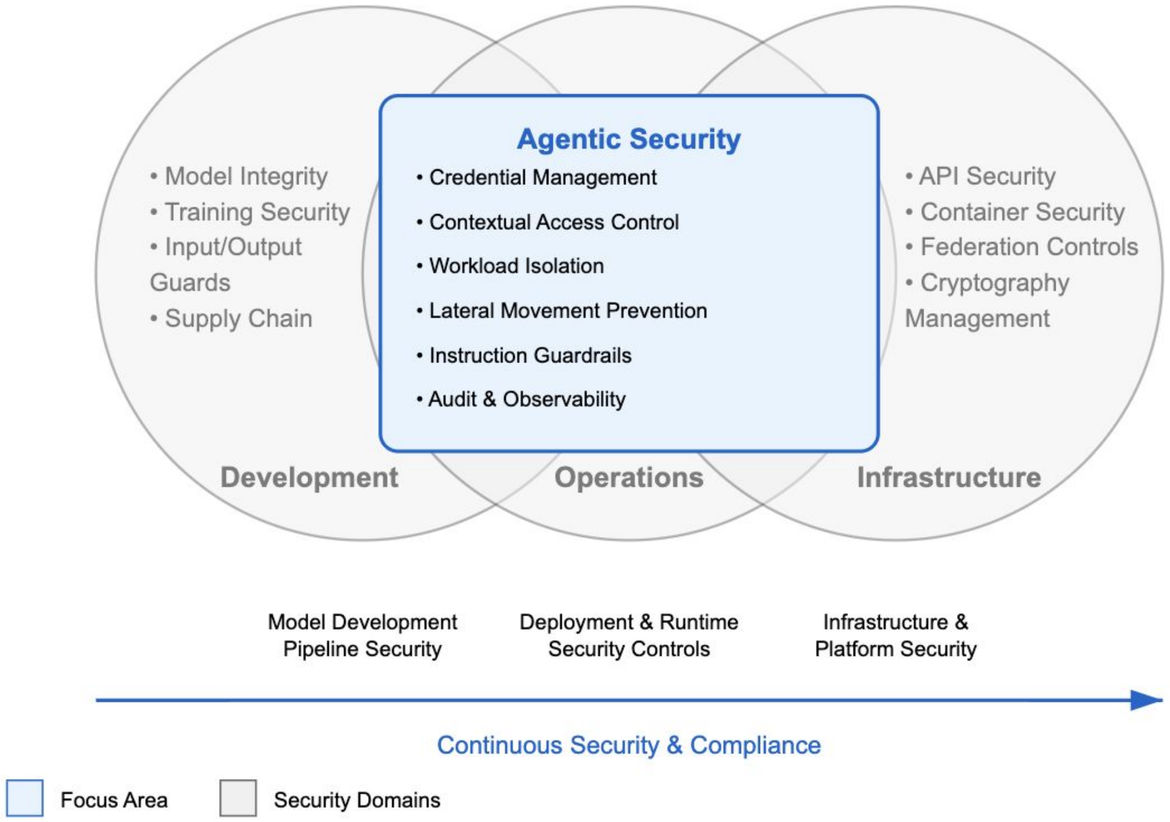
Business Impact

- Customer trust erosion
- Regulatory scrutiny
- Required architecture changes

Lessons for Security

- Need for contextually aware access controls
- Granular data segmentation based on context
- Context-sensitive monitoring and detection

Closing Security Gaps Across the AI/ML Lifecycle



- Security must span the full AI lifecycle.
- Each phase requires tailored safeguards.
- Agentic security bridges boundaries.
- Continuous compliance is crucial.
- Development, operations, and infrastructure must integrate seamlessly.

An Emerging AI Security Taxonomy

Privacy Controls for Sensitive Data	Secure sensitive data access during interactions to prevent exposure or misuse.	**Runtime Security and Safeguards**	Detect and mitigate runtime threats, including unauthorized access, adversarial attacks, and data leakage.
Centralized Governance	Use audit trails to manage training, deployment, updates, and compliance.	**Workload Isolation**	Isolate workloads to prevent lateral movement and limit breach impacts.
Proactive Red Teaming	Simulate attacks to uncover vulnerabilities and improve defenses.	**Data Transformation**	De-identify and tokenize sensitive data to reduce risk while maintaining usability.
Data Vaults and Key Management	Secure sensitive data in vaults with cryptographic key management and auditing.	**Least Privilege Credentialing**	Use short-lived, attested credentials to minimize access and exposure.
Instructional Guardrails	Enforce boundaries for AI behavior to ensure safe and predictable execution.	**Provenance and Lineage Auditing**	Track the origin, transformation, and use of data, training, and actions to ensure integrity, transparency, and accountability.
Model Drift Detection	Monitor for changes in model behavior or accuracy due to data drift or evolving conditions.	**Incident Response and Recovery**	Create processes to detect, respond to, and recover from breaches or failures.
Regulatory Auditing and Liability Management	Ensure compliance with AI standards through documented procedures, oversight, and risk management.	**Content Provenance Verification**	Validate the authenticity and origin of generated content using cryptographic proofs to ensure trust in AI outputs.

An Emerging AI Security Taxonomy

Model Security	Runtime Security	Data Security	Supply Chain	Governance
• Training Security	• Workload Isolation	• Privacy Controls	• Artifact Security	• Compliance
• Model Integrity	• Access Control	• Encryption	• Dependencies	• Audit & Logging
• Evaluation	• Monitoring	• Data Lineage	• Deployment	• Risk Management

Cross-Cutting Concerns:

- Identity and Access Management
- Monitoring and Observability
- Compliance Requirements
- Incident Response

Compartmentalization of Risk

- Isolate workloads to block lateral movement and credential theft.
- Secure processes and kernels to prevent container escapes.
- Segment workloads to reduce blast radius in case of breaches.

Agentic AI workloads require dynamic credentials due to constant interactions with new systems and data, demanding quick-expiring, context-aware access.

CSO

AWS environments compromised through exposed .env files

News
22 Aug 2024 • 7 mins

AWS Lambda | Data and Information Security | Data Breach

Attackers collected Amazon Web Services keys and access tokens to various cloud services from environment variables insecurely stored in tens of thousands of web applications.

Credit: xalien / Shutterstock

"Tens of thousands of applications shared environment variables, giving unintended access across contexts"

Impact

- One leaked secret accessed multiple environments
- No boundaries between applications
- No understanding of how credentials were used → Static secrets can't understand context

→ **AI amplifies the danger of secrets without context**

Applying Established Cybersecurity Principles

The convergence of AI, post-quantum cryptography, and cloud modernization offers a pivotal opportunity to transform your organization's security practices:

- Replace static keys with dynamic credentials and robustly protected secrets.
- Mitigate risks like lateral movement and container escapes to safeguard modern environments.
- Establish consistent practices for securing systems and managing credentials, enabling faster and more confident decision-making.
- Monitor cryptographic activity and credential usage across your environment for enhanced control and rapid response.

Seize the opportunity: Align your organization with the future by addressing key management gaps, reducing risks, and enhancing resilience in an evolving threat landscape.

Aligning AI Security with Recommended Practices

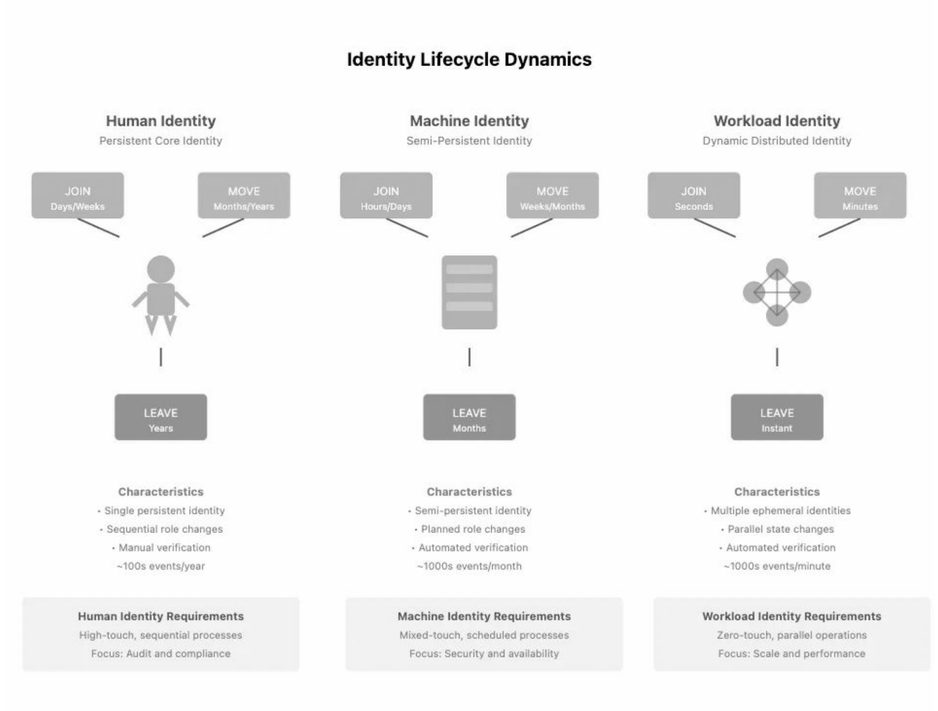
- Applying Least Privilege to AI processing and model execution
- Compartmentalization of Actors through Secure Hardware and Enclaves
- Cryptographic Protections (Post-Quantum and Homomorphic)

Yesterday's Data Exchange Solutions Are Inadequate for Today's Reality

- Physical separation, hardware-based guards do not guarantee Zero Trust
- Mission needs are dynamic—traditional solutions can't adapt quickly to new classification compartments
- Growing quantum threat: Future quantum computers can break classical cryptography, risking data confidentiality
- Legacy solutions hamper agility and scaling Gen AI environments

Adapting Trust to GenAI Systems

- **Human Identity:**
  - Lifespan: Years
  - Manual verification, sequential changes
  - Events: Dozens/year
- **Machine Identity:**
  - Lifespan: Months
  - Manual provisioning and renewal cycles (e.g., CSRs and approvals)
  - Events: ~1,000s/month
- **Workload Identity:**
  - Lifespan: Seconds
  - Automated issuance/verification, real-time attestation
  - Events: ~1,000s/minute



Ephemeral by design – modern workloads may exist for only seconds, so they require dynamic, on-demand credentialing.

When authenticating an agentic workload, attest the complete execution stack, not just the root-of-trust. Verify the host hardware, firmware, OS/hypervisor, container image, and model runtime before allowing execution

Provenance-Based Trust

Hardware roots of trust (TPM, SGX, enclaves)

Ambient contextuality

Configuration, deployment context, environment state

Context from existing authorization systems and registries

Software Integrity Trust

Code hash, runtime attestation, process integrity and dependencies

When Workloads Act Dynamically, Authorization Must Be Equally Dynamic

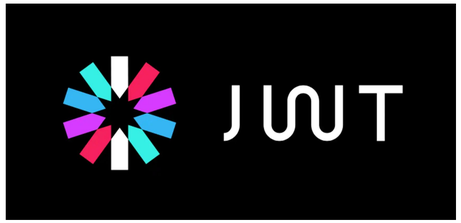
- **Dynamic Policies for Delegated Tasks**
  Evaluate every action based on who requested it, why it's being done, and what is involved.
- **Granular Authorization at Scale**
  Enforce policies for jobs, workloads, and systems acting on behalf of users or themselves.
- **Real-Time Enforcement**
  Tie automated credential issuance to dynamic authorization for secure, auditable operations.

As Trust Scales, Automated High Velocity PKI Makes It Possible to Keep Authorization Promises

- **Automated Identifier Assignment**
  Dynamically assign unique workload-specific identifiers to establish trust foundations.
- **Instant Credential Issuance**
  Issue short-lived credentials tied to workload-specific identifiers in seconds.
- **Policy-Driven Trust Verification**
  Continuously enforce trust through dynamic, policy-based verification.
- **Continuous Monitoring and Rotation**
  Real-time monitoring and automated credential rotation ensure ongoing security.

Standardized way to delegate limited access to resources without sharing full credentials.

X.509



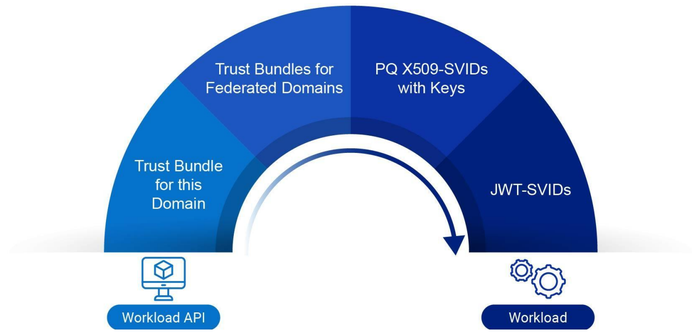
spiffe

Identifier and Trust Bundle

spiffe://nato.int/example/genai-transformer

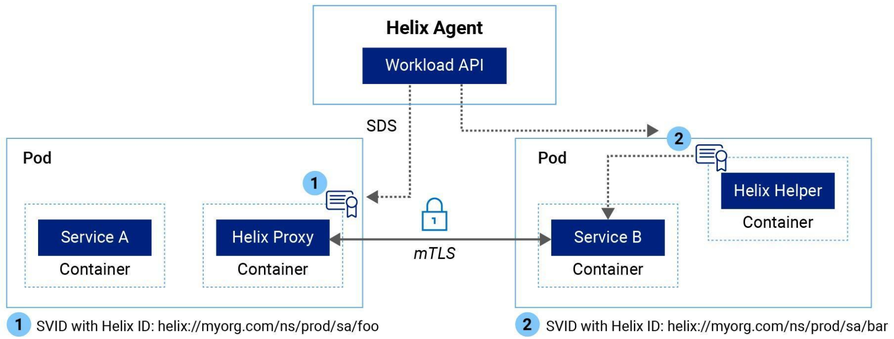
Key pair with certificate chain

Root certificates to trust

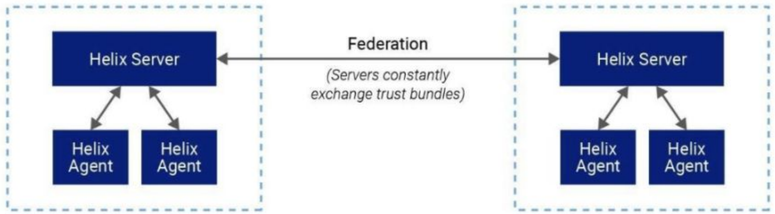


Integration into Existing Systems

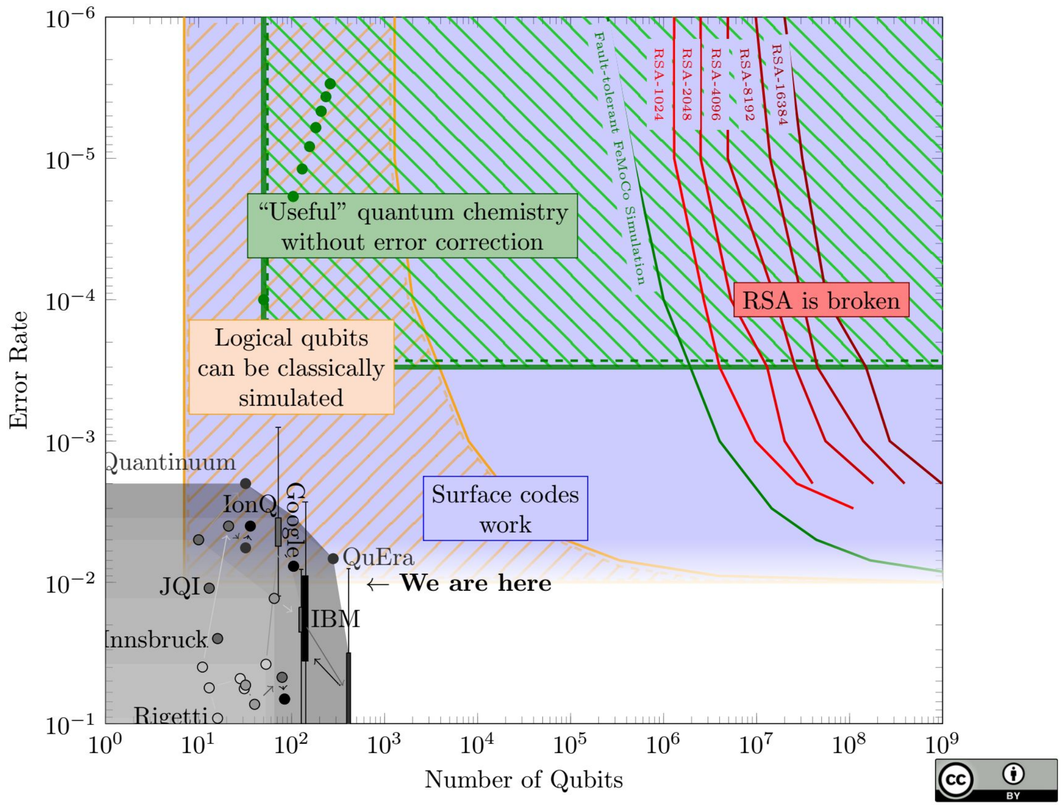
- Micro-service and RPC frameworks
- Smart "sidecar" proxies
- "Off the shelf" systems



- Federation: Exchange a short lived key for another short lived key



Quantum cryptography threats are years away

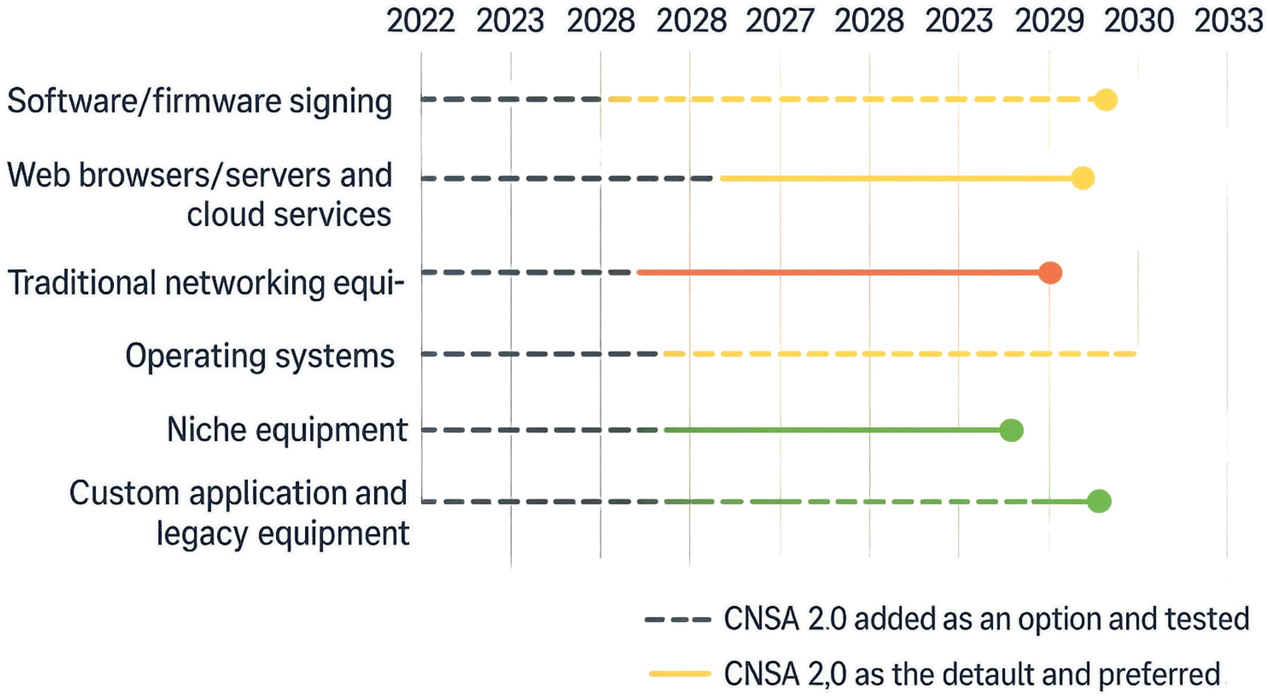


- We may be decades away from the risk of classical cryptography being broken using Quantum Computers.
- Current estimates of the number of physical qubits needed for one logical qubit may exceed 1000 which contributes to this reality.
- IBM recently stated they believe this will happen by the end of the next decade.
- With that said we may experience a 'black swan' event resulting in this happening much sooner.

Phases of Cryptographic Algorithm Transition

- **Algorithm Selection and Development**
  - The creation, standardization, and publication of cryptographic algorithms.
- **Protocol Standardization**
  - Establishing and deprecating standards for protocols that utilize cryptographic algorithms.
- **Implementation**
  - The activities related to incorporating standardized algorithms into software products.
- **Deployment and Usage**
  - This category involves the adoption and utilization of the algorithms and protocols by end-users.

CNSA 2.0 Timeline



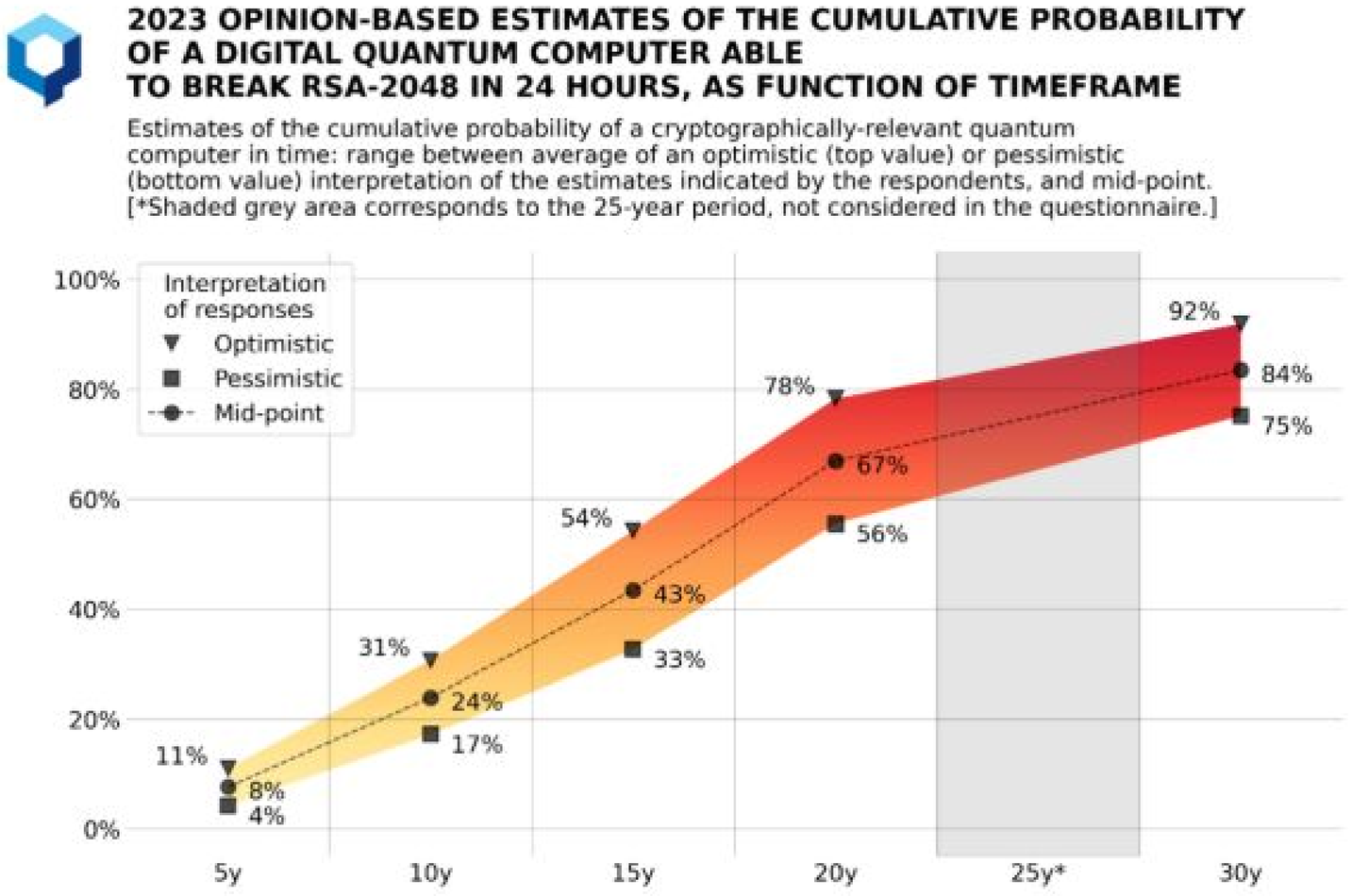


Fig. 1. Global Risk Institute expert estimates for cryptographically relevant quantum computers

The Threat Window

Mosca's Theorem

$$X + Y > Z$$

Where:

X = the amount of time your data must remain secure (data shelf life)

Y = the time needed to migrate to a quantum-safe system (migration time)

Z = the estimated time until a quantum computer can break existing cryptography (Q-day)

Stronger Algorithms from Diverse Sets of Problems

Constant-time implementations

Reduced reliance on RNGs

Deterministic outputs to prevent nonce reuse issues

Secure parameterization

Making Keys and Signatures Post Quantum

- **Enhanced Implementation Error Resistance:** Dilithium3 and Kyber utilize structured lattice-based cryptography, reducing vulnerability to side-channel attacks and implementation flaws compared to traditional cryptosystems.
- **Increased Mathematical Complexity for SNDL Resilience**: Their reliance on hard lattice problems, such as Learning With Errors (LWE) and Module-LWE, makes them resistant to subexponential and quantum attacks, ensuring security against future SNDL (Shor's Non-Deterministic Logarithm) exploits.

FIPS 203

Federal Information Processing Standards Publication

Module-Lattice-Based Key-Encapsulation Mechanism Standard

Category: Computer Security Subcategory: Cryptography

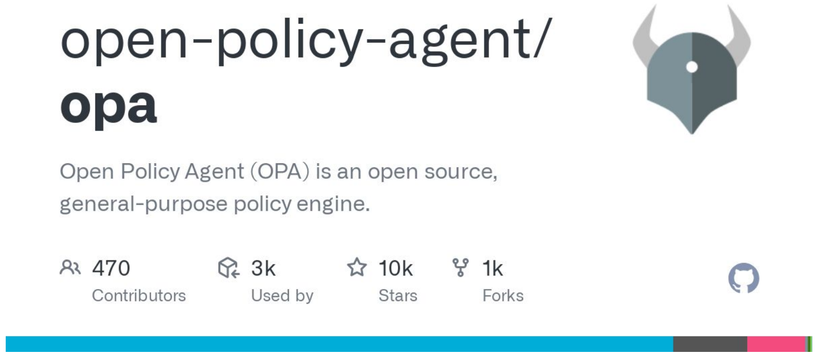
FIPS 204

Federal Information Processing Standards Publication

Module-Lattice-Based Digital Signature Standard

Category: Computer Security Subcategory: Cryptography

Authorization Policy Enforcement

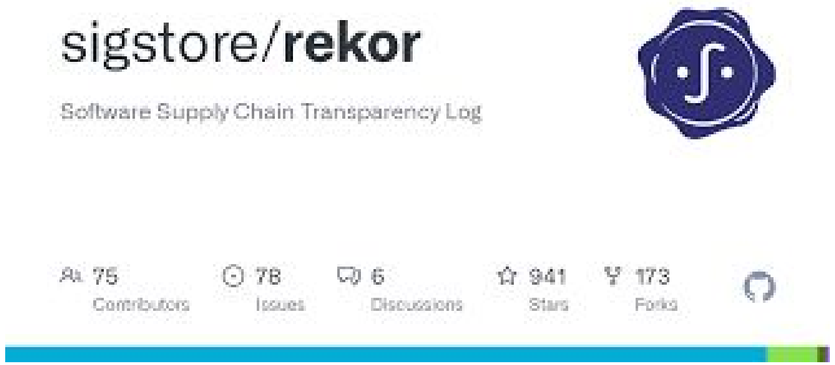


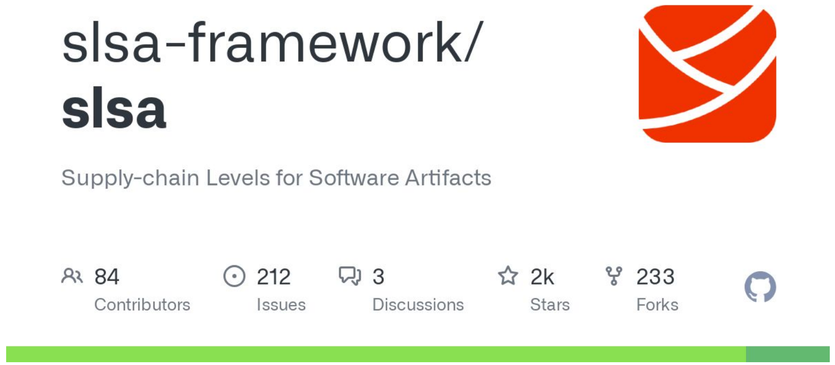
Processing Jobs where Data is Created

- Move data is hard - whether it's video, text, or imagery.
- Moving data is insecure, slow, costly. Even at low cost it can overwhelm your network

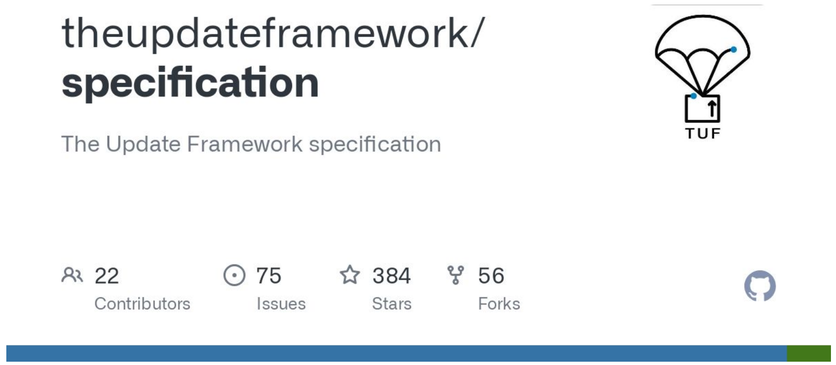
- Moving raw data is risky - takes time to process, slow remediation, often contains PII

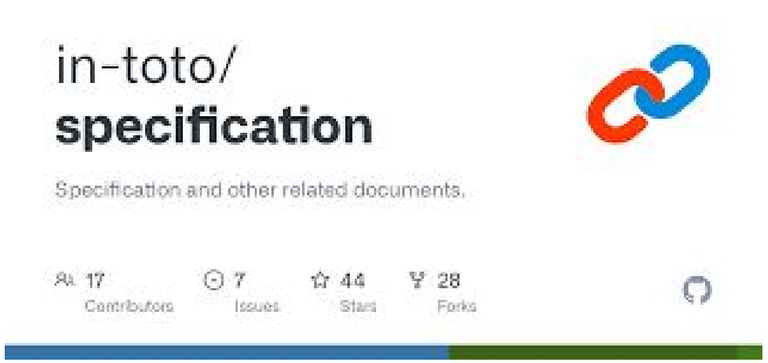
Distributed Ledger and Assurance Levels



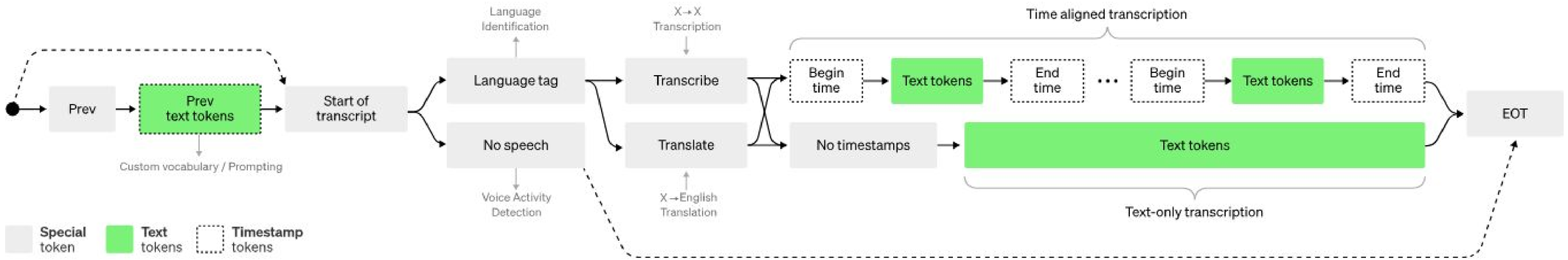


Secure Over-the-Air Updates

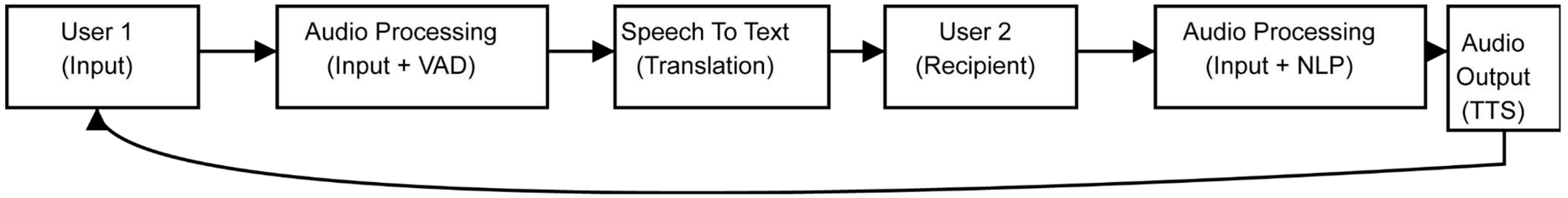




Sample App: Real Time Translator



Sample App: Real Time Translator



Key Components:

Audio Processing

- Captures and normalizes microphone input (16 kHz, mono)
- Uses WebRTC VAD to detect and segment speech
- Streams detected speech for translation

Speech Recognition & Translation

- Whisper-based model (680,000 hours of multilingual and multitask supervised data) for bilingual processing
- Real-time transcription and translation
- Supports bidirectional JA↔EN translation

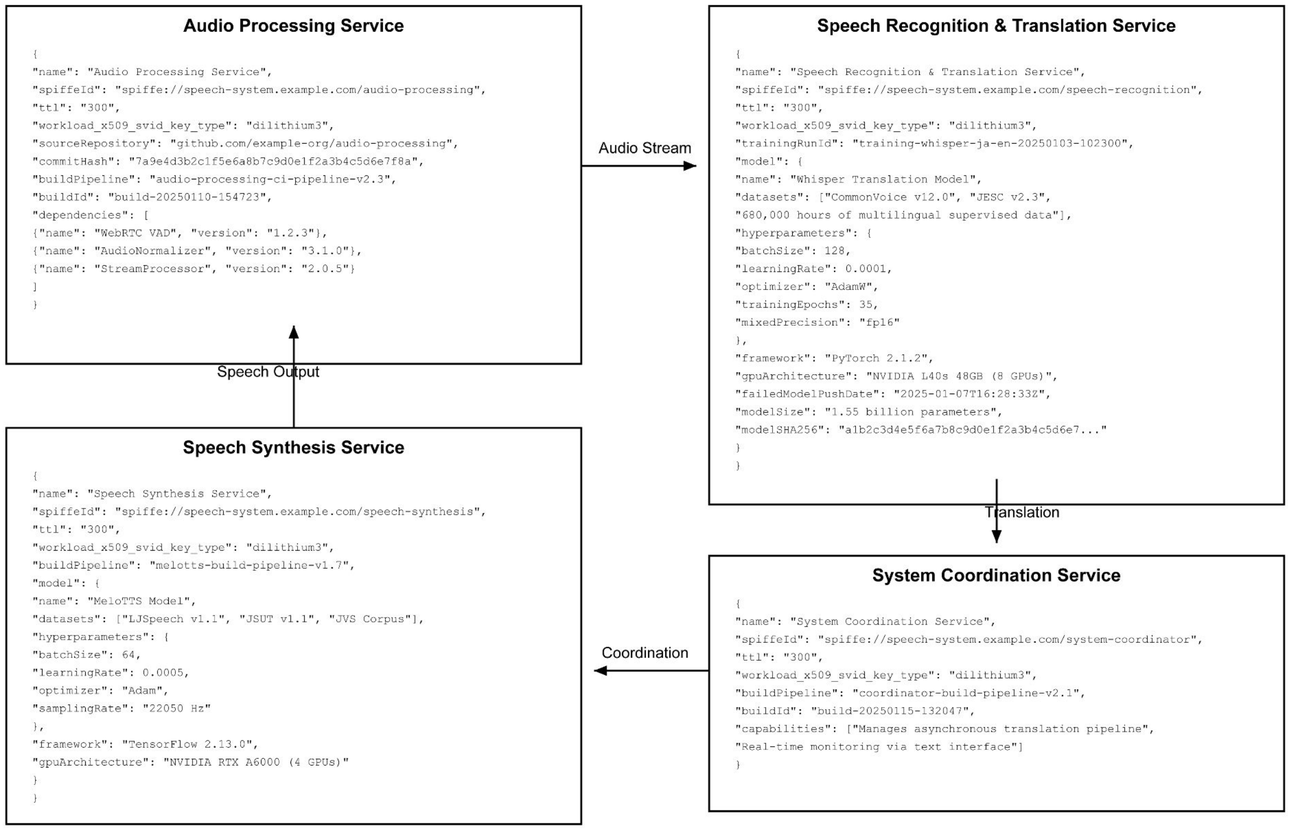
Speech Synthesis

- MeloTTS =for natural speech generation
- Low-latency audio synthesis
- Continuous streaming to user headsets

System Coordination

- Manages asynchronous translation pipeline
- Real-time monitoring via text interface
- Optimized for conversation flow

Anatomy of Attestations in a Transparency Log



AI-driven capabilities restore the defender's edge through relentless vigilance and rapid adaptation

NATO IST-HFM-225 Research Specialists Meeting