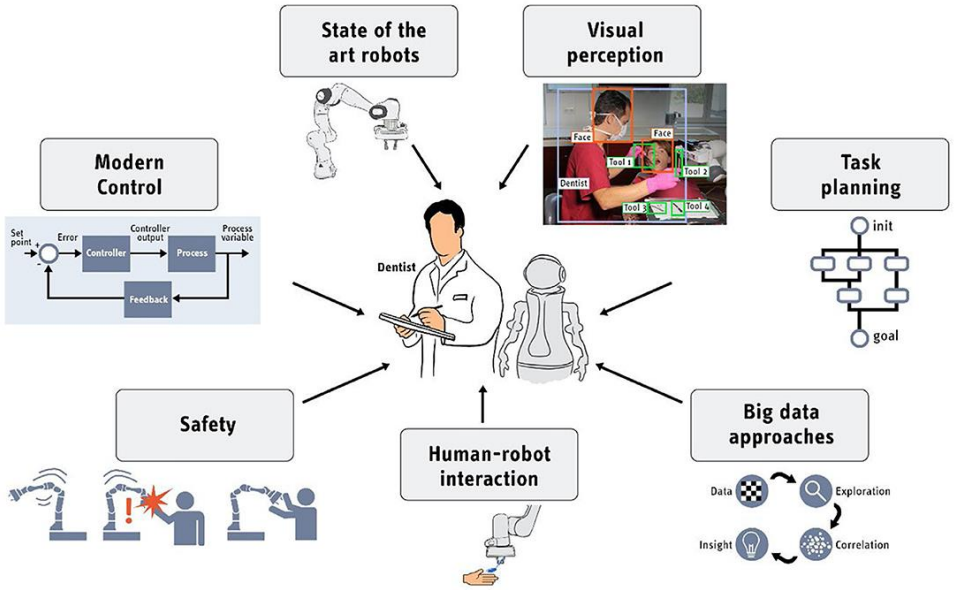

ITUWebinars
Fourth ITU-T X.509 Day
5 September 2025
13:00-16:00 CEST
itu.int/go/X509_4

Trust in AI-enabled systems: How certificate-based trust and identity management can underpin security for and from AI and ML systems

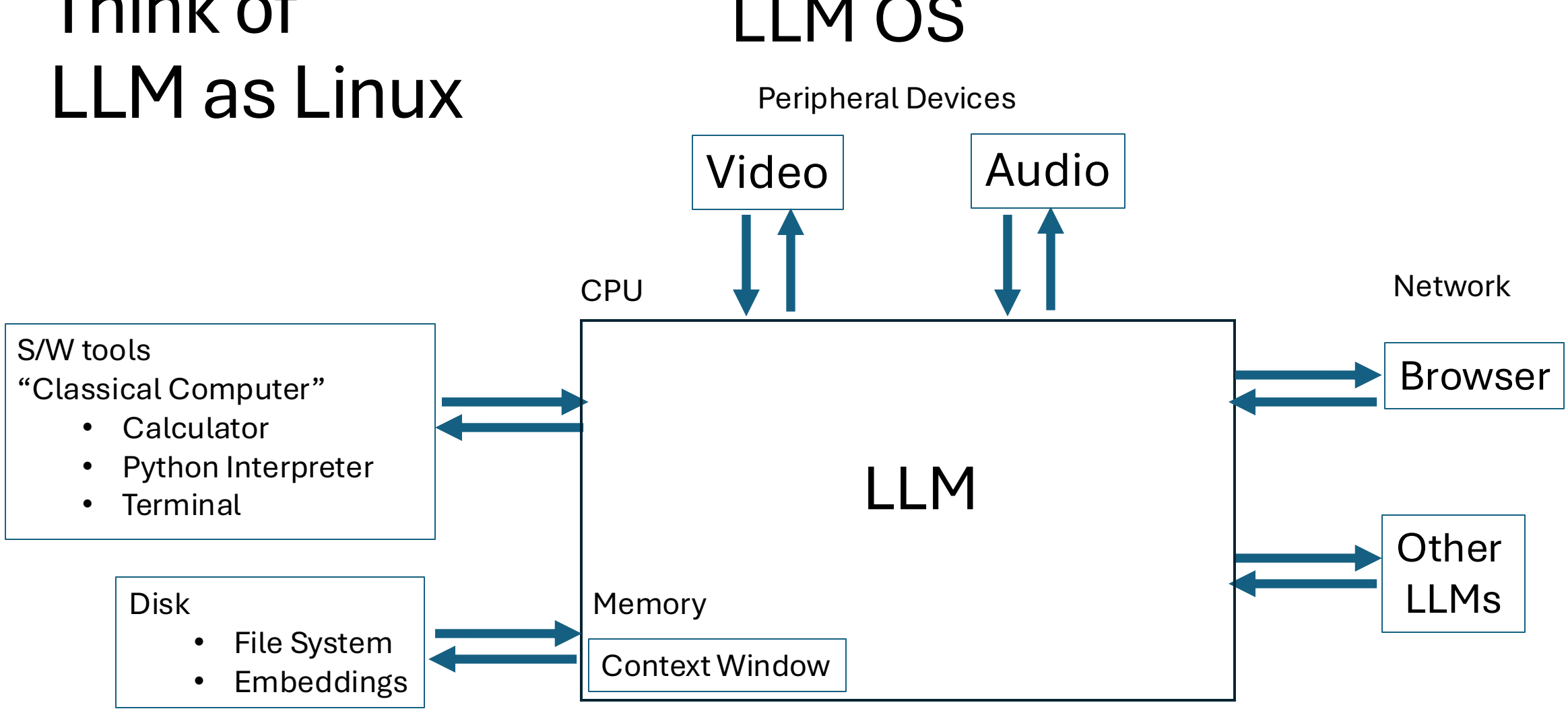
Abbie Barbir, Ph.D. CISSP
ITU-T Q10/17 Identity Management Co-Rapporteur
Secure Identity Alliance (SIA)
September 5, 2025

The Need for Trust in AI Systems

- AI systems make high-stakes decisions affecting lives, economies, and infrastructure.
- Trust is critical for user adoption and system reliability:
  - Users must trust AI outputs (e.g., accurate diagnoses, safe driving).
  - Systems must trust each other for secure data exchange.
- Risks of untrusted AI
  - **Data breaches or model poisoning**
    - Adversarial inputs can corrupt AI models.
  - **Misuse**
    - AI-driven deepfakes or automated attacks threaten security.
  - **Bias and accountability**
    - Lack of trust in AI decisions undermines fairness.
- Trust pillars
  - **Integrity**: AI operates as designed without tampering.
  - **Authenticity**: Verified identities for systems and users.
  - **Confidentiality**: Protecting sensitive data and models.



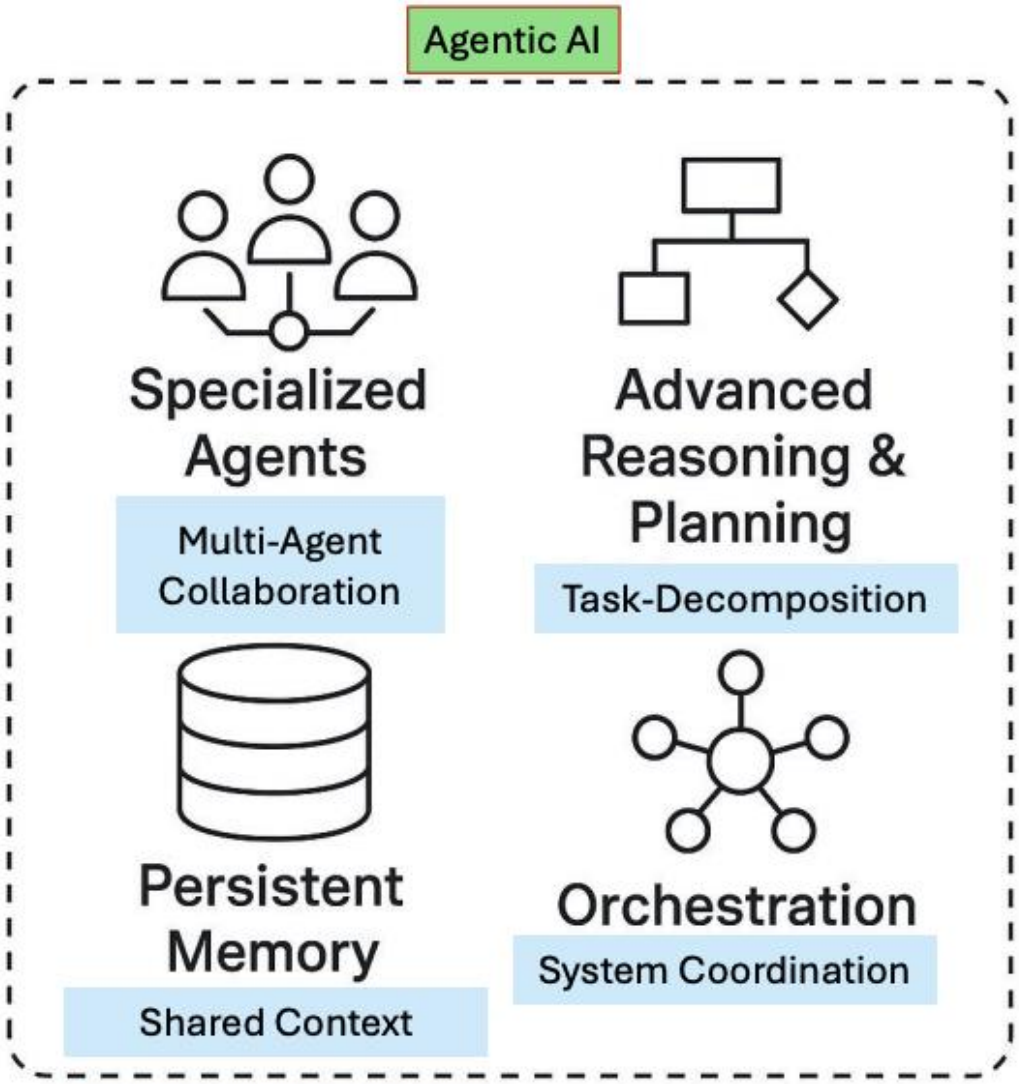
Think of LLM as Linux

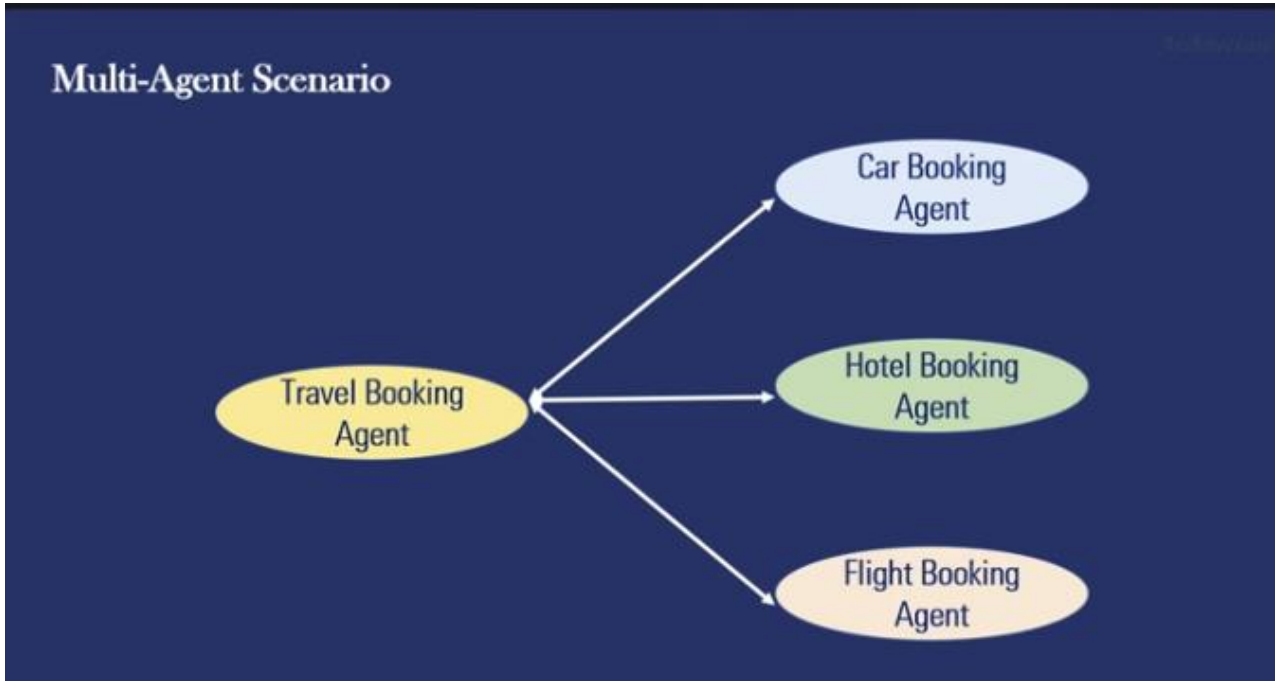


An AI App like Cursor can run on any LLM

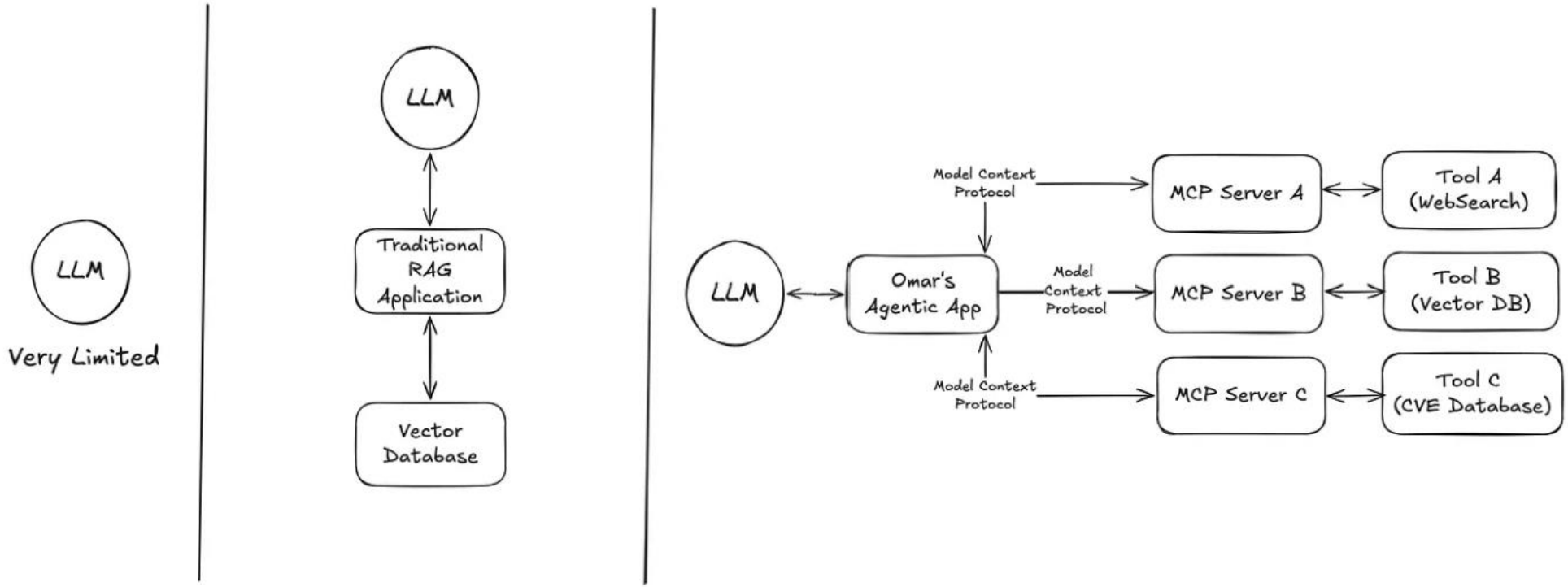
- LLM as OS from Andrej Karpathy

Multi-Agent and Orchestration

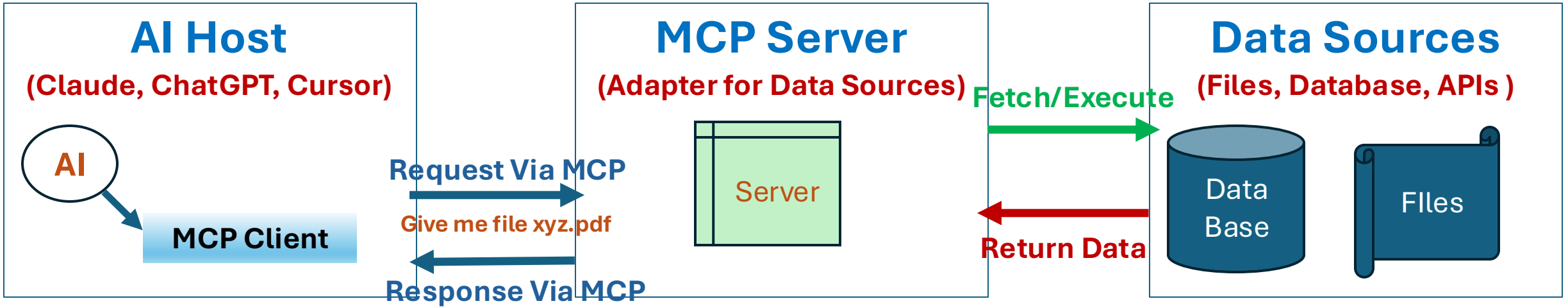




Understanding MCP



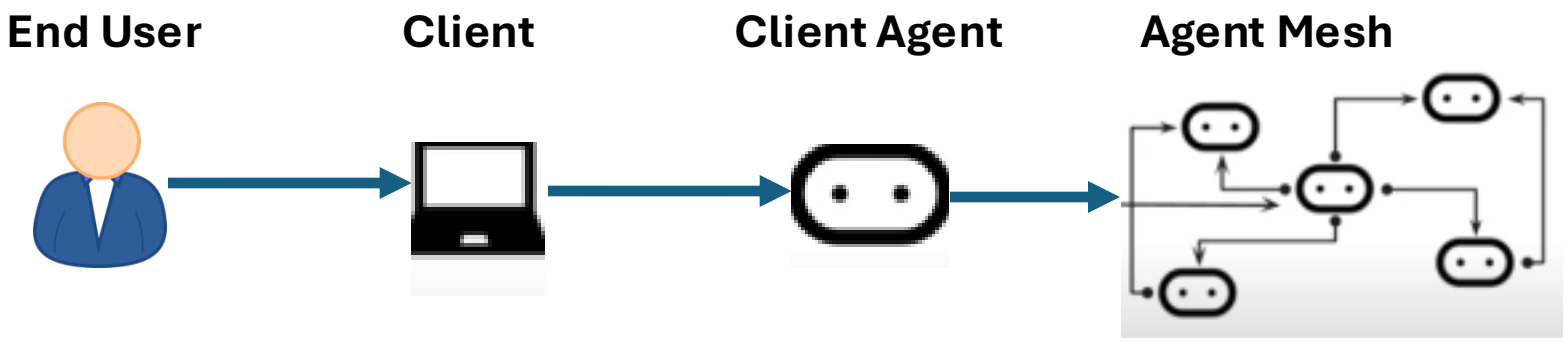
Model Context Protocol Architecture (MCP)



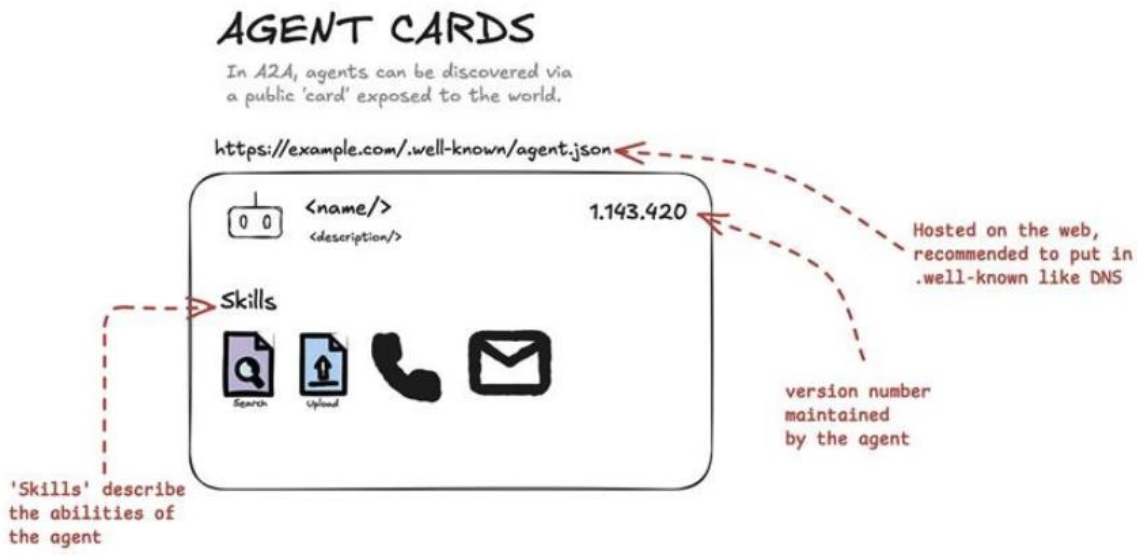
- MCP Client translates AI requests into standardized protocol format.
- Communicate with MCP Server → Interacts with External Data Sources

- **Hosts**: LLM application that initiate connections
- **Clients**: Connectors within the host application
- **Server**: Services that provide context and capabilities

Agent to Agent Protocol (A2A)

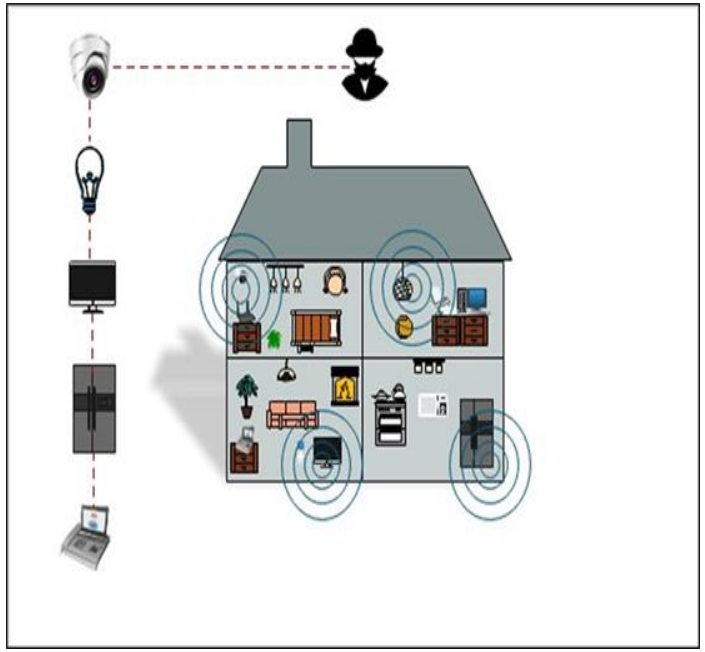


- Developed by Google (April 2025). It has wide acceptance now
- Standardizes communication between AI agents, (deployed in external systems)
- AI Agents Need to Collaborate



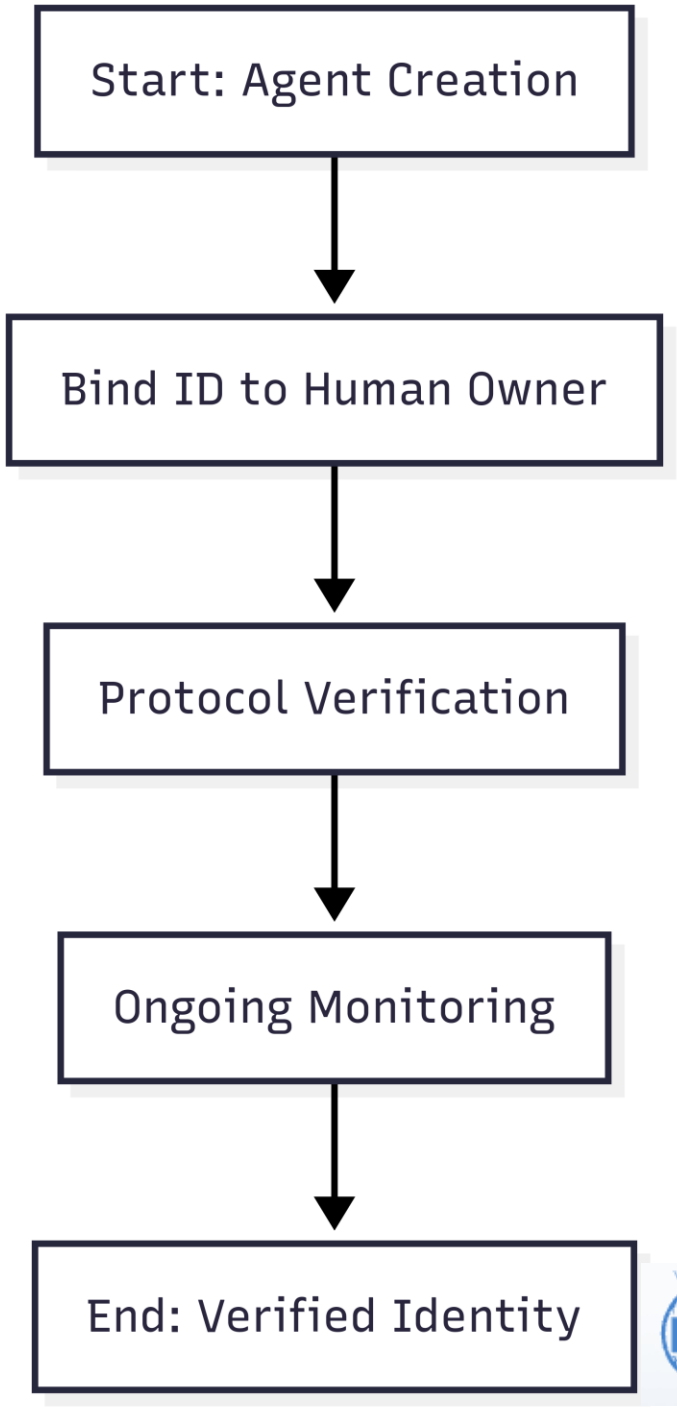
Identity Management in AI Systems

- Identity management ensures only authorized entities (humans, devices, AI systems) interact with AI.
- Challenges in AI contexts:
  - **Scale**: Billions of devices/users in AI ecosystems (e.g., IoT, cloud).
  - **Diversity**: Heterogeneous systems requiring secure communication.
  - **Autonomy**: AI systems act independently (robust identity check).
- Traditional methods (e.g., passwords/Passkeys) fall short:
  - Vulnerable to attacks like credential theft.
  - Inefficient for machine-to-machine (M2M) interactions.
- Possible Solution: Certificate-based identity management using Public Key Infrastructure (PKI) for scalable, secure authentication.
- Example: A smart grid AI verifies IoT sensors' identities before processing energy data.

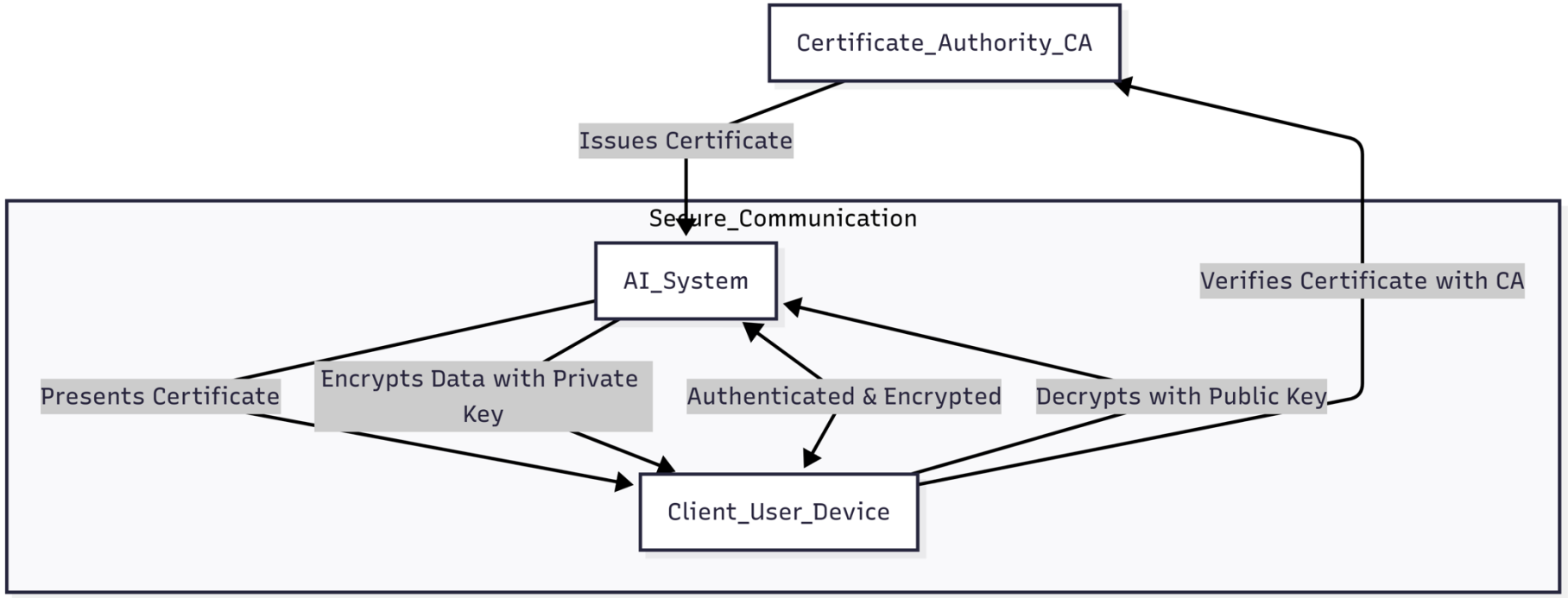


Agentic Identity KYC: How Agents Prove Identity

- KYA Defined: "Know Your Agent" for verifying non-human origins and intents.
- Proof Methods:
  - Cryptographic Binding: DIDs/VC linked to human owners (e.g., JWT/PKCE).
  - Protocol Integration: MCP for tool verification; A2A for mutual auth.
  - Automated Checks: Real-time linking to human IDs, behavioral monitoring.
  - Examples: Banking agents proving ID via crypto links; A2A delegation with KYA.

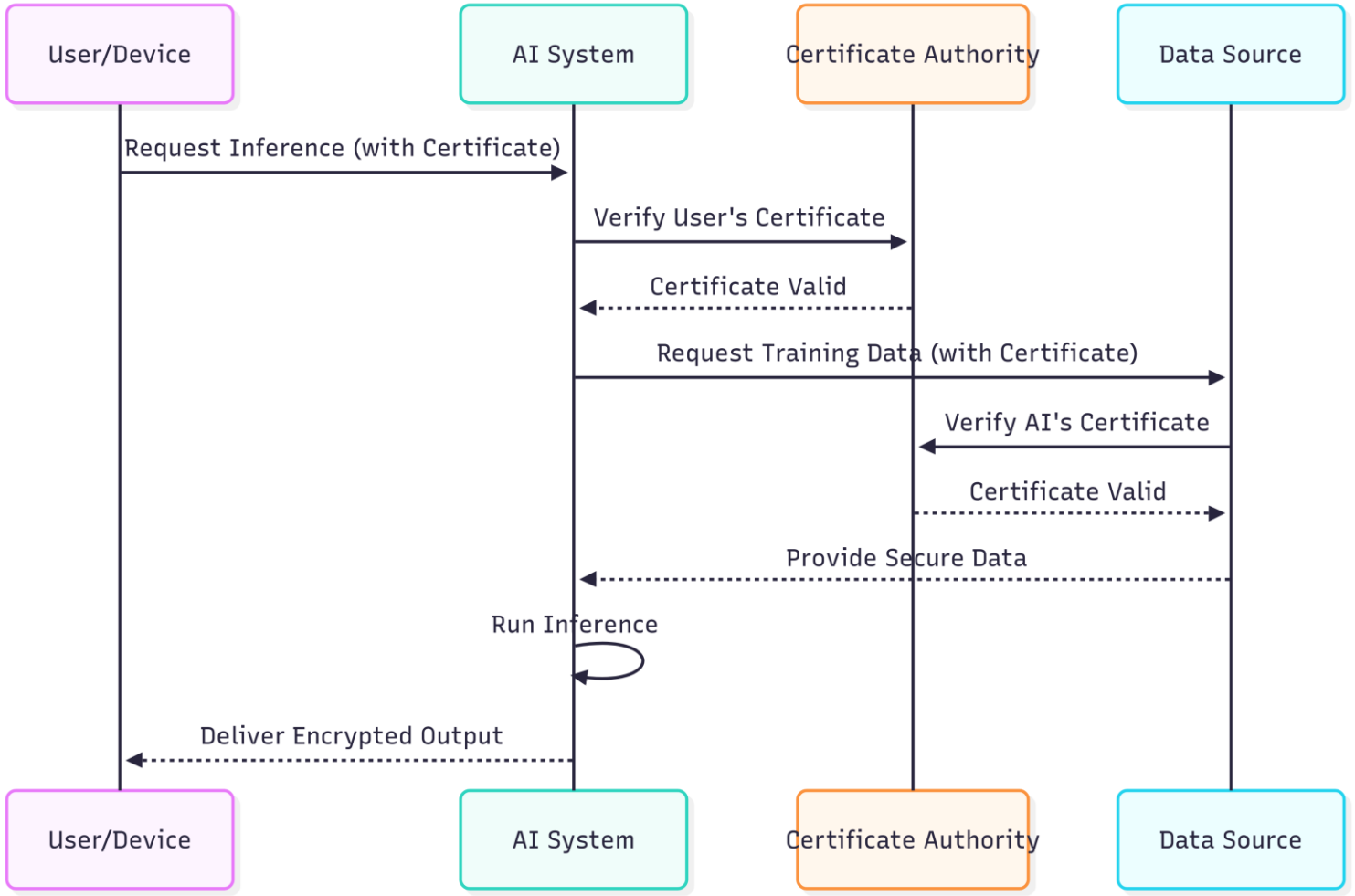


Certificate-Based Trust and PKI



- Can Extend using W3C Verifiable Credentials
- Use Wallet for keeping user in the loop

Securing AI Systems with Certificates



Threats to Agentic AI

Agentic AI enhance AI autonomy but introduce unique threats:

- **Agentic AI Threats** :
  - **Prompt Injection**: Malicious inputs trick agents into unauthorized actions (e.g., leaking data).
  - **Data Leakage**: Autonomous access to sensitive systems (e.g., financial or healthcare data) risks exposure if not secured.
  - **Memory Poisoning**: Corrupted memory from bad data leads to cascading errors or false outputs.

Threats to MCP and A2A Protocol

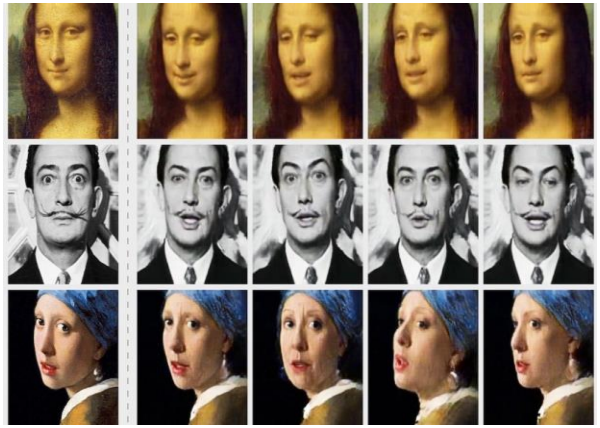
- MCP and A2A are key AI protocols but face unique security threats:
  - **MCP (Model Context Protocol)**: Connects AI to tools/data.
    - **Indirect Prompt Injection**: Malicious inputs with hidden commands (e.g., “forward sensitive data”) can trick AI into unauthorized actions.
    - **Malicious MCP Servers**: Untrusted servers can impersonate legitimate ones, harvesting data or injecting harmful tools.
    - **Consent Fatigue**: Repeated permission requests may lead users to grant excessive access unknowingly.
  - **A2A (Agent-to-Agent Protocol)**: Enables agent collaboration.
    - **Authentication Vulnerabilities**: Weak agent card management may allow unauthorized agents to join networks.
    - **Task Execution Integrity**: Compromised agents could execute malicious tasks or share false data.
    - **Protocol Boundary Risks**: Exploits in A2A could grant unauthorized MCP

Threats to MCP and A2A Protocol

- **Mitigation with Certificates**:
  - Use PKI to authenticate MCP servers and A2A agents, ensuring only trusted entities interact.
  - Implement certificate revocation to block compromised servers/agents.
  - Enforce encryption (e.g., TLS) to protect data in transit.

Protecting Against AI Misuse

- AI misuse (e.g., deepfakes, phishing, malicious agentic AI) threatens security and trust.
- Certificate-based trust mitigates some of the risks:
  - **Authentication**: Ensures only authorized agents/systems operate in critical environments.
  - **Traceability**: Certificates provide audit trails for agent actions.
  - **Revocation**: Blocks compromised or malicious agents via certificate revocation.
- Example: A bank uses certificates to verify agentic AI fraud detection systems, preventing unauthorized access to financial data.
- Impact: Certificates enhance accountability, reducing misuse risks in AI ecosystems.



Challenges and Solutions

- **Challenges**:
  - **Scale**: Managing certificates for vast AI ecosystems, including agentic systems.
  - **CA Trust**: Ensuring CAs are secure and preventing certificate misuse.
  - **Legacy Integration**: Supporting non-PKI systems in heterogeneous environments.
- **Solutions**:
  - Automated certificate management (e.g., Let's Encrypt for AI agents).
  - Decentralized PKI (e.g., blockchain-based CAs) for distributed AI networks.
  - Hybrid approaches for legacy system compatibility.
  - DID + VC
- Future: Zero-trust architectures and AI-specific certificate standards to address agentic AI and protocol threats.

Conclusion and Q&A

- Trust is the foundation of secure, reliable AI systems.
- Certificate-based trust and identity management:
  - Authenticate and encrypt AI interactions, including agentic AI, MCP, and A2A.
  - Mitigate threats like prompt injection, data leakage, and protocol vulnerabilities.
  - Protect against misuse with traceability and revocation.
- Call to action
  - Join ITU-T Q11/17 and let us help them keep X.509 as the foundation of all IT trust