Trustworthy AI with Akka

At Akka, we believe that the next wave of innovation in agentic AI will only be possible if it is built on a foundation of trust.

Trustworthy AI is not a feature you can simply add at the end; it must be an inherent part of the framework's design. Akka is engineered from the ground up to provide the bedrock of trust, but it is a shared responsibility. While our platform offers the essential tools and architectural principles, organizations using Akka must actively configure and implement the controls necessary to build trustworthy AI applications.

This document outlines how Akka enables the core principles of Trustworthy AI, our vision for "continuous compliance", and what you must do to leverage these capabilities effectively.

Akka's Role in Enabling Trustworthy AI

Akka's unique architecture, which includes the actor model and a Zero Trust approach, directly address the key principles of global Trustworthy AI frameworks from ISO, EU, NIST, and the OECD.

Transparency and Explainability

A key tenet of Trustworthy AI is the ability to understand and explain an AI system's behavior. Akka provides a complete audit trail that makes this possible.

- **Immutable Audit Trail**: All agent actions and orchestrations are cryptographically traced and stored in an immutable, append-only ledger. This prevents tampering and ensures an unalterable record of all activity.
- **Traceable Reasoning**: Every action an agent takes—from an LLM call to a tool execution—is logged as a discrete event, complete with the agent's state and the full context of the decision. This creates a provable chain of custody for every AI decision.
- **Standardized Provenance**: Detailed traces are published in a standardized format, ensuring they are interoperable and can be easily integrated into third-party auditing tools for independent verification.

Accountability

Akka's architecture establishes clear lines of responsibility, making it possible to trace every agent action back to its source.

- **Secure Workflow Invocation**: Orchestrations can only be invoked by secure endpoints or other authenticated systems, ensuring that only authorized users or processes can initiate a workflow.
- **Immutable and Signed Artifacts**: Workflow definitions are statically defined, versioned, and cryptographically signed upon deployment. This process prevents unauthorized changes and provides an immutable record of a workflow's history.

- **Robust Logging and Auditing**: Every action, delegation, and authorization decision is logged in a tamper-proof audit log. This capability ensures that administrators can trace any action—even one executed by an agent—back to the human who initiated it.

Privacy and Data Governance

Handling sensitive data ethically and securely is non-negotiable for trustworthy AI. Akka provides multiple layers of protection.

- **End-to-End Encryption**: All Akka network communications, including between external data sources and Akka, are secured with TLS. Furthermore, all agent-to-agent, system-to-agent, and system-to-system communications are end-to-end encrypted under our Zero Trust principles.
- **Fine-Grained Access Control**: Akka enforces fine-grained access control for data retrieval, enabling you to limit an agent's access to the principle of least privilege.
- **PII Redaction**: Akka allows you to use inline evaluations to filter and redact Personally Identifiable Information (PII) before it is passed to an agent or an LLM, adding a critical layer of data privacy.

Safety and Robustness

Akka is built to be resilient, secure, and safe in an unpredictable environment.

- **Zero Trust Architecture**: All services and network communications with Akka operate under Zero Trust principles, as they are cryptographically signed and secured by mutual TLS (mTLS). This prevents message tampering, spoofing, and replay attacks.
- **Isolation and Sandboxing**: Akka's actor-based architecture provides strong isolation. Each agent and service is effectively sandboxed, preventing a compromised component from affecting the integrity of the rest of the system.
- **Proactive Threat Mitigation**: Akka includes built-in mechanisms to defend against AI-specific threats. For example, it provides loop limits and timeouts to prevent infinite loops, and it automatically scans agent packages for known vulnerabilities.
- **Real-time Guardrails**: Orchestrations can load policies from a governance catalog. This enables inline, real-time evaluations that can be executed before and after each step to enforce policies and block potential violations.

Your Responsibility: Building Trustworthy Systems with Akka

Akka provides the framework, but the ultimate responsibility for a trustworthy AI system lies with your organization.

Each organization has its own approach to building trust and managing risk. With AI, you need to structure functions that are implemented continuously through the AI lifecycle, some of which includes establishing a culture of risk management, setting policies for AI, identifying and assessing potential risks and their sources, identifying mitigation plans for each risk, and establishing internal metrics and controls, and ongoing monitoring of implementation.

Your work with AI risk management occurs across every system that is part of AI including the models themselves. With Akka, you build agentic systems that interact with AI models and knowledge systems. With your agentic systems, you must:

- **Enforce Input Validation**: Treat all external data as untrusted. Use Akka's inline evaluations to apply robust input sanitization and filtering mechanisms that defend against prompt injection and jailbreaking attacks.
- **Apply the Principle of Least Privilege**: Limit each agent's permissions to only what is absolutely necessary for its function. Ensure all sub-agents, tools, and resource calls are statically defined and testable.
- **Integrate Human-in-the-Loop**: For any high-risk or sensitive actions, leverage Akka's human-in-the-loop capabilities to require explicit, real-time confirmation from an operator before an agent can proceed.
- **Continuously Monitor**: Use Akka's real-time monitoring and observability tools to track agent behavior and resource consumption. Implement anomaly detection using inline evaluations and endpoint filters to identify unusual activity.
- **Secure Your Environment**: Recognize that the security of your local agent development environment is your responsibility. This must be managed within your corporate security policies.

Certifications and Resources

Akka has robust controls that align with numerous compliance standards, making the systems you build with our framework "compliance-ready." We have undergone rigorous third-party audits and hold certifications such as SOC2 Type II and SOC3. More information, including our penetration testing reports, is available on our trust center.

We have a vision to enable any system built and run within Akka to achieve “continuous compliance”, where it is possible to instantly determine your risk profile and compliance readiness from within each Akka system. We are tracking and monitoring emerging standards which are defining metrics, security requirements, and governance controls that must be provided within every agentic system. This includes ISO/IEC 42001 and the NIST AI Risk Management Framework.