SOLUTION & WHITE PAPER

Verifiable identity for every AI agent

KOBIL Agent Trust gives each agent its own cryptographic identity, checks every tool call against policy before it executes, and writes every decision to a tamper-evident audit trail, without changing your models or your APIs.

8 platform services, one Helm chart

3 SDK packages per language, TypeScript and Python

3 agent frameworks: Anthropic, OpenAI, Google ADK

FIG. 1 An agent identity, rendered as a credential: issuer, agent ID, delegated capabilities and limits, public key, and a verification endpoint.

Executive Summary

An AI agent stops being a chatbot the moment it can call a tool. From then on it can move money, change a record or send something on the organisation's behalf. The credential model most deployments inherited was never designed for that.

01

Identity

Every agent proves who it is with its own certificate and credentials, issued and revocable one agent at a time.

02

Authorization

Every action is checked against policy before it executes, by a component the agent cannot instruct.

03

Accountability

Every decision is logged where the agent cannot edit it: what happened, why, and who approved it.

WHY IT MATTERS

Identity, authorization and accountability are three separate controls, and each has to hold when the agent itself is the thing that has been manipulated. The rest of this paper describes where those controls sit, what they are built from, and what remains outside their scope.

WHAT CHANGES, AT A GLANCE

INHERITED MODEL	WITH AGENT TRUST
A shared service account behind many agents	One certificate per agent, revocable alone
Scopes granted once, at deployment time	A policy decision per call, at execution time
Evidence assembled from logs after the fact	A correlated record written by the control path

01 / EXECUTIVE SUMMARY

The Challenge

Scoping permissions does not close the gap. Scope is a statement about categories; the risk is in the instance: this agent, this argument, this moment.

HOW AN INCIDENT TRAVELS

01
Untrusted input
Documents, pages, retrieved passages

→

02
Agent reasoning
Non-deterministic tool choice

→

03
Privileged action
API, payment, record change

→

04
Business effect
Irreversible, and attributable to you

THE FEARED EVENT

An agent holding standing authority is manipulated through content it did not write, and uses valid credentials to take a real action, with no independent record of who authorised it.

WHY THE USUAL ANSWERS FALL SHORT

Narrower scopes
Still a standing grant. It says which category of action is allowed, never which instance.

Prompt-level guardrails
They live inside the component that has been manipulated, and depend on it cooperating.

Application logs
Written by the caller, in the caller's format, and editable by whatever wrote them.

A human in the loop
Only evidence if the approver sees the resolved call, on a channel the agent cannot reach.

Market Development & Trends

Adoption, autonomy and regulation are advancing together. The combination turns agent identity from an architecture preference into an operational requirement.

62%

experiment with or scale AI agents

McKinsey, State of AI 2025

23%

scale at least one agentic system

McKinsey, State of AI 2025

<10%

scale agents in any single function

McKinsey, State of AI 2025

Obligation, not intention

DORA has applied to in-scope financial entities since 17 January 2025. It asks for demonstrable control over automated processes, not documented good intent.

Agentic risk is now named

OWASP published a Top 10 for Agentic Applications in December 2025, separating identity abuse, inter-agent trust and tool misuse from ordinary application risk.

Human oversight becomes testable

EU AI Act Articles 12 and 14 require automatic logging and effective human oversight. The AI Omnibus deferred them to December 2027; it did not repeal them.

The building blocks already exist

CIBA has been a final OpenID spec since 2021 and DPoP an RFC since 2023; NIST's 2026 paper names both. What is new is applying them per agent, per action.

REGULATORY TIMELINE

SEP 2023	JAN 2025	DEC 2025	DEC 2027
DPoP standardised as RFC 9449	DORA applies to in-scope financial entities	OWASP Top 10 for Agentic Applications published	EU AI Act high-risk obligations apply to Annex III systems

MARKET SIGNAL

Adoption is broad; governed production is rare. The constraint is no longer whether an agent can act, but whether an organisation can prove what it did and stop it when it should not.

The Solution

Agent Trust sits at the point where an agent's output becomes an action. Five steps run on every tool call, and none of them depend on the agent cooperating.

EVERY TOOL CALL, FIVE STEPS

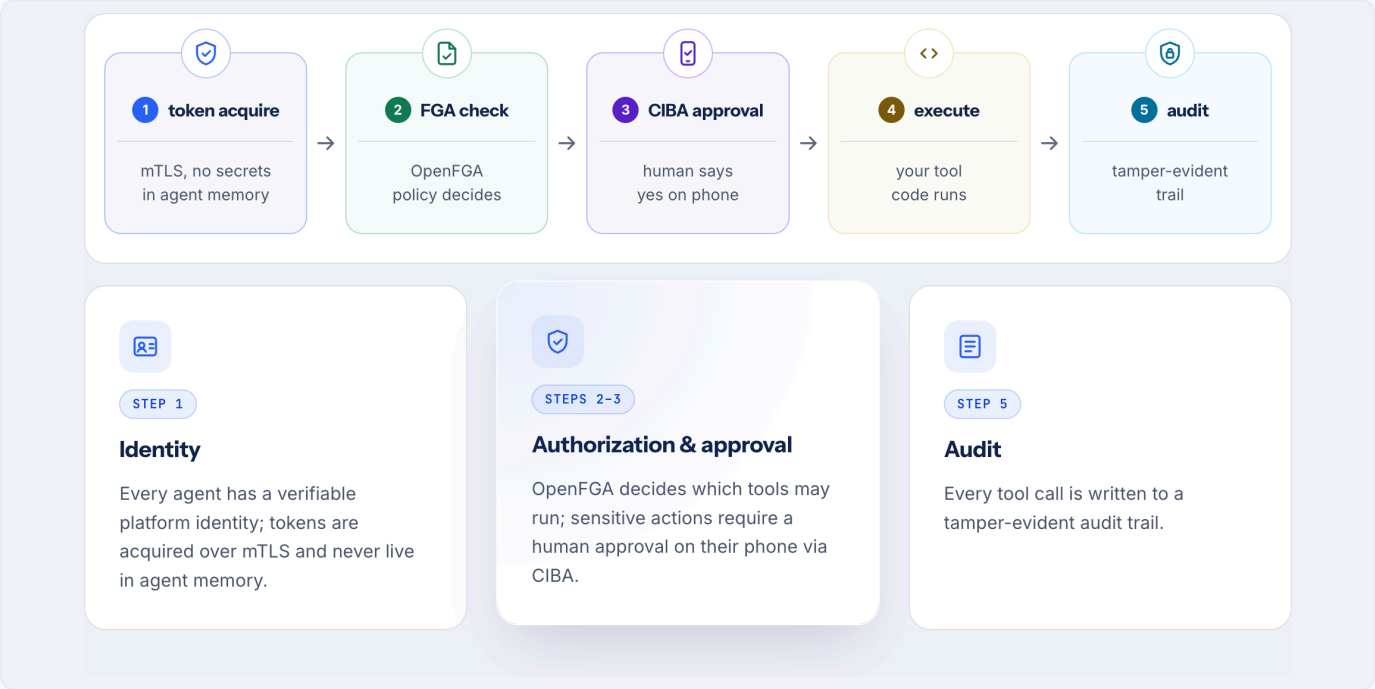

FIG. 2 Register your tools once; every call the model makes runs through the same guarded pipeline.

WHERE THE PLATFORM SITS

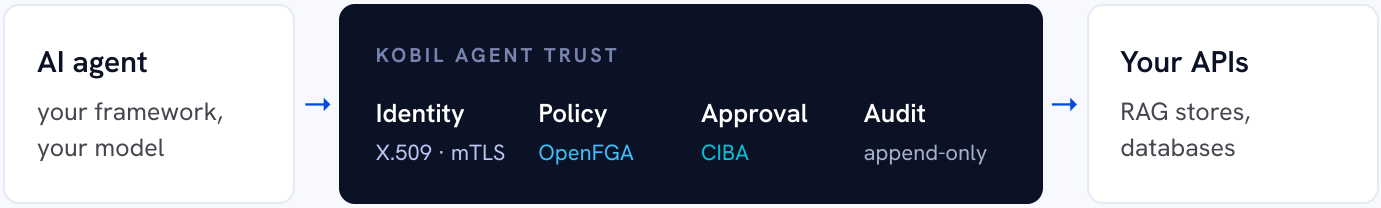

FIG. 3 The agent never holds the credential, never evaluates the policy, and never writes the record.

The credential is issued, not held

Certificates are minted per agent by the control plane and can be revoked one at a time.

The decision sits outside

Policy is evaluated by a service the agent can call but cannot instruct or bypass.

The record is a side effect

Each step appends to one correlated entry, written on the control path itself.

Technology & Product Overview

Four layers describe the deployment. The platform owns the middle two; the agent framework above it and the estate below it stay as they are.

AGENT LAYER	Anthropic SDK	OpenAI SDK	Google ADK	Your own framework
DEVELOPER SDKS	aitrust meta-package	agent-sdk identity core	agent-audit audit client	TypeScript · Python at parity
CONTROL PLANE	SHIFT IDP	OpenFGA	Management API	Audit service
YOUR ESTATE	REST APIs	RAG stores	Databases	Internal services

FIG. 4 Nothing in the top row or the bottom row is replaced; the two middle rows are what a deployment adds. The control plane row names the four services a developer calls; Istio, PostgreSQL, Redis and Elasticsearch make up the remaining four.

INTEGRATION SCOPE

The platform wraps the tool call. It does not host your model, replace your agent framework or require code changes in the services the agent calls. One package install from KOBIL's registry, and your API gateway pointed at the platform's JWKS endpoint.

Use Cases

The strongest fit is where agent speed meets money, personal data or a regulator.

1 Banking & payments

SCENARIO

An agent prepares a transfer and must not complete it alone once the amount crosses a threshold.

CONTROLS

CIBA approval on a separate device, policy threshold, correlated audit record

2 Insurance claims

SCENARIO

An agent assesses damage from photos, drafts a settlement and touches customer records. KOBIL ships this one as a working reference agent in TypeScript and Python.

CONTROLS

Per-resource authorization, claims-manager approval before settlement, tamper-evident trail

3 Shared tool servers

SCENARIO

Agents reach shared tool servers that have only the caller's word to go on.

CONTROLS

mTLS between agents, scoped credentials injected per call, per-tool policy, replay prevention

4 Regulated audit

SCENARIO

A supervisor asks what one agent did on a named date, and who allowed it.

CONTROLS

Append-only records, SIEM export, per-agent attribution

PILOT PRINCIPLE

Start with one agent and a small set of consequential actions. Gating everything produces reflexive approval, which is worse evidence than no approval at all. Gartner expects more than 40% of agentic AI projects to be cancelled by the end of 2027, inadequate risk controls among the reasons. Scope is the lever you control.

Business Impact

The case is not one security feature. It is the ability to put agents into production at all in an environment where someone must answer for what they did.

Production becomes possible

Agents reach real systems under a control an auditor recognises, instead of stalling in proof-of-concept.

Blast radius is bounded

Per-agent identity means one compromised agent is revoked alone, and what it did is attributable.

Evidence is a by-product

The record is produced by the control path itself, rather than assembled from logs after a question is asked.

One control, many frameworks

The same identity and policy layer covers Anthropic, OpenAI and Google ADK agents without per-framework work.

Pilot measurement framework

Time to governed production

Elapsed time from a chosen use case to an agent running under enforced policy.

Policy coverage

Share of consequential actions with an explicit rule rather than a default allow.

Approval quality

Proportion of gated actions where the approver saw platform-composed detail, and the approval rate.

Evidence completeness

Share of actions with a full correlated record: agent, principal, resource, decision, approver, timestamp.

EVIDENCE STATUS

No verified customer or ROI metrics are published in this paper. The measures above are the objectives a design-partner pilot is intended to establish.

Technical Details

Eight services deployed as a single Helm chart, with the realm, certificate authority, authorization model and mesh routing pre-configured. No per-customer build.

SHIFT IDP

Keycloak 26. Tokens, login, CIBA and DPoP on a pre-configured realm, with OAuth 2.1 and PKCE.

OpenFGA

Relationship-based authorization: may agent Z reach document Y for user X?

Management API

The agent identity control plane. Ten API groups; token exchange touches all three backends.

Audit service

One ingestion path for SDK events over REST and platform events over Kafka, indexed in Elasticsearch.

Istio

Service mesh: mTLS, traffic management and path-based routing through a VirtualService.

PostgreSQL · Redis · Elasticsearch

One database with three schemas, a config cache with DPoP nonce rotation and anomaly counters, and five audit indices.

Trust boundaries

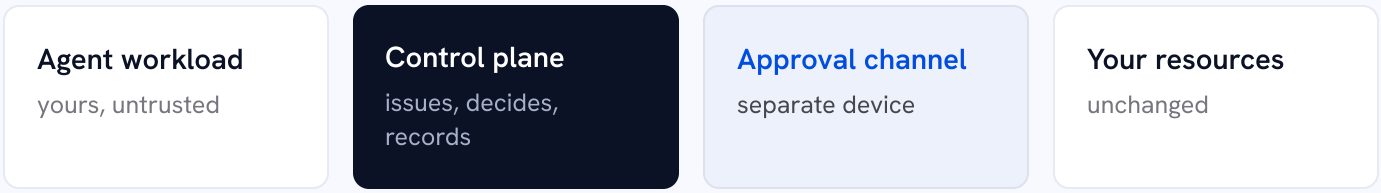

FIG. 5 The agent crosses each boundary with a credential it did not mint and cannot widen.

HONEST SCOPE

The platform authorises the tool call; it does not sandbox what the tool then executes, inspect prompts, or vet the packages your agent loads. It addresses three entries of the OWASP Agentic Top 10, not all ten: ASI02 tool misuse, ASI03 identity and privilege abuse, and ASI07 insecure inter-agent communication.

Maturity & Compliance

Credible adoption needs a clear line between what ships today, what is target state, and what remains the customer's responsibility.

AREA	TODAY	TARGET STATE
Control plane	8 services, single Helm chart	Same, with air-gapped bundle
SDK coverage	3 packages per language, TypeScript and Python at parity	Additional runtimes on demand
Deployment	KOBIL-hosted or self-managed Kubernetes	Dedicated regional tenancy
Approval coverage	CIBA on policy-selected actions	Richer approver routing and delegation
Evidence	Correlated records, SIEM export	Packaged regulatory evidence sets

Regulatory mapping

DORA

Applies since Jan 2025

Enforced control points over automated processes, with supervisory evidence.

NIS2

Transposition incomplete

Strong approver authentication and per-agent access control. Obligations differ by Member State.

EU AI Act, Art. 12 & 14

High-risk from 2 Dec 2027

Automatic per-action logging; oversight the system cannot bypass.

OWASP Agentic Top 10

Published Dec 2025

ASI02, ASI03 and ASI07; the remaining seven named and attributed elsewhere.

This document is not legal advice and is not a certification. Controls may support compliance; the outcome depends on architecture, configuration, process and evidence.

Key Takeaways

1 **The control must sit outside the agent**

A manipulated agent will not honour an instruction to ask permission. Identity, policy and evidence therefore live where the agent cannot reach them.

2 **Approval is a mechanism, not a slogan**

CIBA is a decoupled standard with a banking pedigree. The approver reads what the platform composed from the resolved call, on a device the agent cannot see.

3 **Evidence has to survive the session**

A correlated, append-only record answers the question a supervisor actually asks: what did this agent do, and who allowed it.

4 **Scope honestly, then pilot**

Three of the OWASP ten, not all ten. One agent, a handful of consequential actions, and a measurable objective.

Call to Action

From one prioritised use case to defensible evidence. Design-partner discussions: kobil.com/contact-us

01	02	03	04
Architecture review	**Policy design**	**One agent integrated**	**Evidence & scale**

09 / SOURCES

Sources & Notes

Internal product facts combined with public market, security and regulatory references. External sources accessed 06 Aug 2026.

01	McKinsey: The State of AI 2025	mckinsey.com/.../our-insights/the-state-of-ai
02	Gartner: Agentic AI project cancellations, Jun 2025	gartner.com/en/newsroom/press-releases
03	OWASP: Top 10 for Agentic Applications, 9 Dec 2025	genai.owasp.org/.../agentic-applications-for-2026/
04	Regulation (EU) 2022/2554: DORA	eur-lex.europa.eu/eli/reg/2022/2554/oj
05	Regulation (EU) 2024/1689: AI Act, Art. 12 & 14 (as amended)	eur-lex.europa.eu/eli/reg/2024/1689/oj
06	Directive (EU) 2022/2555: NIS2	eur-lex.europa.eu/eli/dir/2022/2555/oj
07	OpenID Foundation: CIBA Core 1.0 (final, 2021)	openid.net/specs/...-backchannel-authentication-core-1_0.html
08	IETF RFC 9449: OAuth 2.0 DPoP, Sep 2023	rfc-editor.org/rfc/rfc9449.html
09	NIST NCCoE: Software and AI Agent Identity and Authorization, Feb 2026	nccoe.nist.gov
10	KOBIL Agent Trust: platform documentation	Internal source, 06 Aug 2026

DOCUMENT STATUS

Solution and technical white paper based on the KOBIL Agent Trust platform as of 06 Aug 2026. No product certification, legal advice or commitment to future functionality.