Proving AI Compliance: Deterministic Evidence for Autonomous AI

Who this is for

This paper is written for people accountable when an AI system gets compliance wrong. Different readers will find different parts more useful than others:

Data protection and legal readers will get most from Sections 6 and 7, on what Obquan checks against and the evidence it produces.

Security and architecture readers will want Sections 5, 8 and 9, on how it works, how it deploys, and what it deliberately does not cover.

Risk and accountability readers will want Sections 2, 3 and 10, on what happens after an AI Agent has system permissions, why the usual tools miss it, and why it matters now.

1. Summary

Firms are giving AI agents real credentials and real access. Those agents now read customer data and act on decisions at machine speed, but the controls already in place cannot tell whether any given action was lawful. The evidence a regulator expects, showing that AI activity was monitored against specific obligations, is not being produced.

Obquan checks every AI interaction against named regulatory rules at the moment it happens and returns a decision: allow, warn, or deny. It does this deterministically with a fixed rule engine rather than another AI model, so the same interaction always produces the same result and every result maps to a specific regulatory article. The output is a record a compliance team can defend in front of the FCA or the ICO.

This paper explains what happens after an AI agent logs in, why the usual tools miss it, how Obquan works, and what Obquan is and is not built to do.

2. Current Systems

When an AI agent runs inside a regulated firm it uses the same credentials as any other piece of software. It holds an API key or an OAuth token, it authenticates against the systems it needs, and it is treated as authorised because it is. From that point it can read

customer records and act on or draft the decisions it was scoped to handle. None of that looks any different to the firm's controls from ordinary authorised traffic.

The controls most firms rely on here are firewalls, identity and access management (IAM), and SIEM logging. They were designed to establish whether an identity should be let in and they do that well. This is not the question that matters once the agent is already inside and working. What a regulator wants to know is whether the interaction the agent had was lawful; the access layer has no view of that.

An example makes it concrete. A customer service agent authenticates with a valid token and reaches the systems it is allowed to reach. On the way through it might process special category health data with no recorded lawful basis, or produce output that strays into regulated financial advice, or keep acting on a customer whose erasure request has not been closed. The access layer sees none of this. The token checked out and the request went through normally, so nothing is flagged at the moment it happens. It usually comes to light later when someone is auditing the system or investigating a breach.

Access control confirms the agent was allowed in but it doesn’t show whether the interaction that followed was compliant.

3. Current Limitations

Firms are not ignoring this, but they are mostly trying to cover it with tools they already run, tools that were built for other work.

The most common response, a periodic audit, has become far less appropriate for addressing the problem. Whereas a quarterly review looks at a sample of interactions after they have happened, an agent can act thousands of times a day, so a sample taken three months later says very little about what is running now and cannot catch anything early enough to prevent it.

General observability platforms are the next thing firms reach for. Datadog and tools like it are good at what they were built for, which is latency, uptime, error rates and the general health of a system. They retain the content of an interaction whilst letting compliance issues pass through, simply because they were never built to evaluate them.

The newest approach is to put one AI model in charge of watching another, and this is the one that fails most clearly here. The problem is that a probabilistic system checking another probabilistic system does not give the same answer twice. Run the same interaction through it again and the verdict can move. If the output is only an internal suggestion, it might be tolerable, but it stops being tolerable the moment the output has to serve as evidence because a DPO cannot hand a regulator a judgement that was roughly confident at the time and cannot be reproduced months later.

Compliance evidence has to hold up when it’s checked subsequent times. Run the same interaction through the same rules and you should get the same result mapped to the same regulatory article. A DPO can take that to a regulator and defend it. A verdict that comes out differently on a re-run cannot do that job however confident it looked at the time. This is the reason that Obquan is deterministic rather than model-based.

4. Principles

Five principles underlie the design that explain most of the decisions that follow.

Every AI interaction is checked against a named regulatory obligation and not just a general notion of good behaviour.

Checks are deterministic and reproducible - the same input yields the same result every time.

The firm's own declarations are the source of truth for the compliance facts around an interaction.

The engine records its own verdict independent of the application it is watching, so the evidence reflects what Obquan decided rather than what the application reported.

The output is built to be filed with a regulator and not just read internally.

5. How it works

Integration is a single call. Each time an AI agent is about to handle an interaction, the application sends it to Obquan through the SDK to the /evaluate endpoint. Obquan runs every applicable rule and returns one of three outcomes:

Allow: means nothing was triggered and the interaction proceeds.

Warn: means something needs attention but is not a hard stop.

Deny: means a rule was breached and the interaction should not go ahead.

The overall decision is the most severe outcome across all the rules that ran, so a single deny among many allows makes the whole call a deny. Rules that do not apply to a given request are skipped rather than guessed at.

Every rule reads one of two things:

Metadata Rules

The first “family” reads metadata. These are the compliance facts the firm declares about an interaction: the stated lawful basis for processing, a consent timestamp, whether the customer was told they are dealing with AI, whether human oversight is declared available, and whether there is an open erasure or objection request that should pause processing. The engine records these declared values, reproducibly, and preserves them as evidence. It does not independently verify them. A declared value such as human oversight is an attestation the engine captures, and reconciling a declaration against what actually happened inside the firm's own systems is on the roadmap. Because these rules read declared values rather than free text, the same interaction checked again over the stored record returns the same answer.

Content Rules

The second “family” reads the message itself: the user input, and the model output where output evaluation is in use (this is rolling out in stages rather than fully available yet). These rules catch common formats of personal data appearing in a request, such as email addresses, phone numbers, identifiers, card numbers and health terms, with coverage of UK-specific formats being expanded. They also flag a maintained set of prohibited phrases across hate speech, unauthorised financial advice, unqualified medical claims, and promotional phrasing that would breach FCA rules on fair and clear communication. The phrase lists are seeded and expanding rather than exhaustive. Where personal data is genuinely needed, the firm can declare a lawful basis or have the data masked before it reaches the model rather than being blocked outright.

In a real interaction both “families” run in the same call. The metadata rules confirm the interaction was set up lawfully and the content rules confirm nothing prohibited is in the message. The engine takes the most severe result, returns the decision, records which rules ran, what each returned, and the regulatory article behind each one.

Severity is set per deployment rather than fixed in each rule. The same rule can deny in a strict configuration and only warn in a lighter one, depending on which frameworks a given customer has switched on. The same engine serves both a commercial bank and a legal firm without any code changes, just a different choice of frameworks.

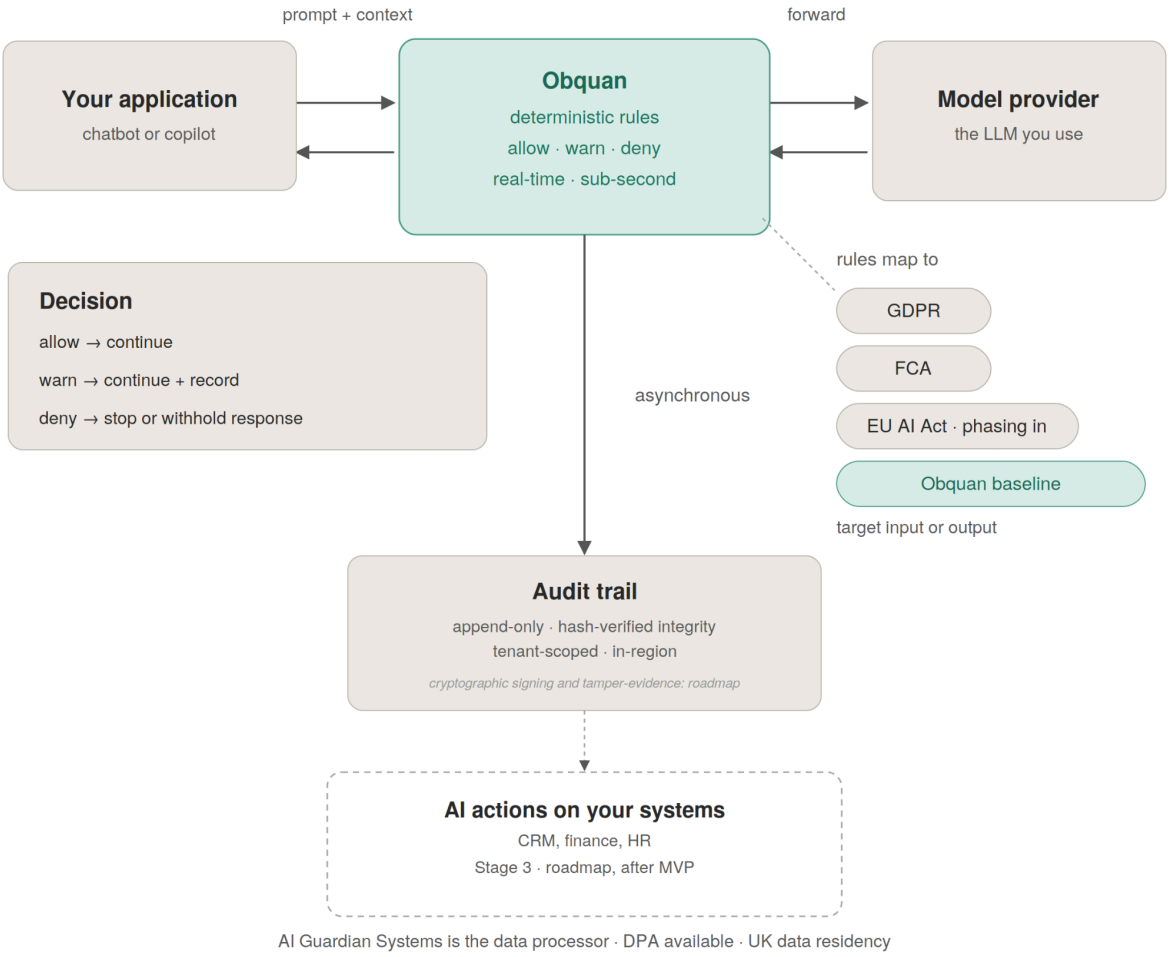
A worked example

Take the customer service agent from Section 2, the one that reaches special category health data with no lawful basis recorded. This is what the engine does with it.

The application sends the interaction to /evaluate. Two rules apply. The content rule scanning the message finds health information in the text and flags it as personal data of a special category. The metadata rule checking lawful basis finds none declared for processing it, which fails the requirement under GDPR Article 6. Both results are collected, the most severe wins, and the call returns deny. The application withholds the response.

The record written at that moment holds the rules that fired, the articles behind them (Article 6 for the missing basis, Article 9 for the special-category health data), the deny outcome, the timestamp, and the agent and tenant it belongs to. No part of the verdict was produced by a model.

Run the same interaction again months later and the engine returns the same deny against the same articles. A DPO can produce the record on demand and it reads the same every time.



Audit trail, framework mapping and Stage 3 covered in Sections 6, 7 and 9

6. What it maps to

The value of a check is only as good as the obligation it maps to. Obquan ties its live rules to specific articles across the frameworks a UK-regulated firm answers to.

GDPR

Live checks cover:

- Lawful basis for processing (Article 6)
- Validity of consent and purpose limitation (Article 5)
- Mandatory notification of AI usage (Articles 13 and 14)
- Erasure and objection request handling (Articles 17 and 21)
- Oversight of automated decision-making with significant effect (Articles 22A to 22D, the UK regime under the Data (Use and Access) Act 2025; Article 22 remains the reference for an EU audience)
- Presence of personal data (Article 5)
- Presence of special-category data such as health information (Article 9)

The accountability principle under Article 5(2) is satisfied by the immutable record generated for every evaluated interaction.

FCA Handbook

Live checks cover:

- Consumer Duty: acting to deliver good outcomes for retail customers (Principle 12 and PRIN 2A), the retail anchor and the largest current FCA theme
- Fair treatment of customers and fair, clear and non-misleading communications (Principles 6 and 7), which cover non-retail business (for retail business these are superseded by the Consumer Duty)
- Availability of human intervention
- Explanation requirements for significant automated decisions
- Audit trail metadata required under Systems and Controls (SYSC) obligations

EU AI Act

The EU AI Act is mapped within the engine. Several live checks map directly to its requirements, specifically regarding transparency of AI interactions and human oversight mechanisms for high-risk systems. Obligations specific to high-risk systems under Annex III carry a staged enforcement timeline (target: December 2027); broader EU AI Act coverage is aligned to that roadmap rather than presented as fully live today.

Roadmap Frameworks

Additional frameworks, including ISO 27001, ISO 42001, and DORA, are scheduled for future release and are explicitly not active in current production deployments.

7. The evidence model

The record Obquan produces is designed to stand on its own in front of a regulator, and two features of it are live today.

Independent Verdict Engine: The engine writes its own verdict. The audit record reflects what Obquan decided, not what the calling application reported. If the application echoes back a different result, the engine's independent record stands. This allows a DPO to present the audit trail as objective evidence rather than a self-attestation.

Append-Only Immutable Log: Evaluated interactions are written once and cannot be mutated. Each entry binds together the rule evaluated, the regulatory article mapped, the outcome (Allow, Warn, or Deny), a precise timestamp, and the associated agent and tenant IDs.

Strategic Direction & Roadmap

Today the full interaction, its input, output and metadata, is stored in the audit trail, because the firm needs that evidence to demonstrate ongoing oversight. Obquan is being built to keep the evidence separate from the underlying content. In that future model, raw prompts and model outputs would remain within the customer's environment, and Obquan would retain only the verdict record and a cryptographic hash of the evaluated payload. Once that ships, a breach of Obquan would expose only verdict records and hashes rather than customer content, and any later tampering with the customer's stored content would break the hash.

Two capabilities are explicitly designated on the roadmap rather than live in production:

- Retaining strictly the verdict record instead of the full payload.
- Cryptographic signing with trusted timestamping of the audit trail.

8. Deployment

Content rules require only the message payload and work as soon as the SDK is installed. Metadata rules require the firm to provide surrounding operational facts; the SDK is designed to capture these at their natural scope rather than requiring them to be passed on every API call.

Integration uses three distinct metadata levels:

Project Level: Set once for the entire deployment (e.g., whether human oversight mechanisms are active).

Session Level: Configured per user session for stable parameters (e.g., a consent timestamp retrieved from a consent management platform).

Request Level: Passed dynamically per call for changing context (e.g., the specific lawful basis or processing purpose of a given transaction).

The integrator sets each field where it belongs and the SDK assembles them at request time. Some audit fields, such as the model version and a hash of the input, are filled in automatically.

Integration Effort

Implementation complexity depends entirely on existing compliance infrastructure:

Light Integration (Days): A firm that already logs consent, displays AI usage disclosures, and enforces human review can integrate rapidly.

Larger Build: A firm lacking these mechanisms must construct the underlying controls. Because current regulations require these controls regardless of vendor selection, this work is foundational. Obquan provides the structured operational framework to enforce and verify them on every interaction rather than annually on paper.

Four diagnostic questions determine the integration tier:

- Do you obtain explicit consent before users interact with your AI?
- Do you disclose to users that they are interacting with an AI system?
- Is there a named internal owner for AI risk and governance?
- Is a human review step integrated anywhere in the autonomous workflow?

Mostly “yes” points to a light integration. Mostly “no” points to a larger build and usually to a firm that has more regulatory exposure than it has realised.

Adoption path

A firm does not have to switch on blocking from day one, and most should not. The engine supports a staged route in.

The first stage is observe. The engine evaluates every interaction and records the verdict, but nothing is blocked. Rules are set to record rather than stop, so the firm sees what

Obquan would flag against real traffic with no effect on the running service. This gives a true reading of where exposure sits, on live data, before anything about the service changes.

The second stage is enforcement, brought in rule by rule. Once the firm trusts the verdicts on a given rule, it moves that rule from record to block. Because severity is set per deployment, this happens one rule or one framework at a time rather than all at once. A firm might enforce the personal data and financial advice rules first, where a wrong action is clearly worth stopping, and hold others in observe while it tunes them.

This gives a firm a way in that carries no operational risk at the start. It watches first, builds confidence in the engine against its own traffic, and turns on enforcement where it is ready. The interaction flow does not change when it does. The same verdict that was recorded now also stops the request.

9. What Obquan does not do

Obquan evaluates AI interaction behaviour; it does not monitor network traffic, authentication events, credential usage, or infrastructure access patterns. It does not replace a firewall, an IAM platform, or a SIEM, nor is it a general IT security tool.

This boundary is explicit. Evaluating whether an individual AI interaction is lawful is a dedicated job, and Obquan is engineered solely to execute that job and produce defensible evidence. Being clear about what it does not do matters: a platform that claims to cover all security domains provides compliance teams with no verifiable boundary where its evidence holds. Obquan's evidence holds strictly at the interaction layer.

10. Why now

The shift from suggestive AI to autonomous AI creates immediate compliance risk. When models only summarised text or generated drafts for human review, simple logging was sufficient because a human served as the circuit breaker. Autonomous agents now execute transactions, modify databases, and interact with customers directly at machine speed. Post-hoc audits occur too late to prevent or contain a breach.

Regulatory enforcement is already active across core frameworks:

GDPR: Obligations covering lawful basis, transparency, and data subject rights are enforceable today, carrying penalties up to £17.5M or 4% of global turnover under UK GDPR.

FCA Handbook: Requirements for ongoing risk monitoring, Consumer Duty compliance, and fair customer treatment are active, backed by unlimited fining authority (Calculated based on financial harm, benefit, & turnover).

EU AI Act: Staged implementation is progressing, with transparency mandates and high-risk system obligations taking effect according to the statutory timeline. High-risk obligations, the tier that applies to the systems Obquan's customers run, carry penalties up to €15M or 3% of global turnover; the €35M or 7% ceiling applies only to outright prohibited practices.

An enterprise deploying autonomous AI agents today operates inside the enforcement scope of GDPR and FCA regulations regardless of future EU AI Act milestones.

11. What you can prove

With Obquan deployed, a compliance team can demonstrate the following for any AI interaction:

Obligations Checked: The exact regulatory rules evaluated at runtime.

Verdict Rendered: The deterministic decision (Allow, Warn, or Deny) generated by the engine.

Temporal & Legal Traceability: The precise timestamp and the underlying statutory article mapped to the check.

The system can reproduce any historic verdict on demand. It changes the firm's position from claiming its AI is monitored to being able to prove it.