FlowX.AI

TECHNICAL PAPER

GOVERNED AUTONOMY, VERIFIED BY AN EVIDENCE LAYER (GAVEL)

Governance That Runs Where the Agent Runs

Why governing autonomous AI agents means enforcing policy and producing audit evidence at runtime, over the agent's own telemetry, rather than asserting compliance on paper.

Bogdan Răduță
Head of Research, FlowX.AI
bogdan.raduta@flowx.ai

FlowX.AI · Applied Research

May 2026 · Version 1.0

ABSTRACT

Enterprises are granting AI agents more autonomy than they can currently govern. In a 2026 survey, 71% of organizations deploying agents had no formal governance framework even as a majority planned to widen agent autonomy, and 80% reported risky agent behaviors such as unauthorized data access[6]. The dominant failure mode in production is not the model hallucinating but the agent exceeding its authority, chaining actions across systems faster than any human review cycle can follow. Regulation is closing in: the EU AI Act reaches full effect in August 2026 with binding requirements for human oversight and multi-year record-keeping[1], and standards bodies have opened tracks for agent identity, action logging, and containment[3]. Yet the tooling on offer splits into two halves that do not meet: governance-risk-and-compliance suites that document policy in prose disconnected from what agents do, and runtime control planes that enforce limits but generate no compliance evidence. We argue that AI agent governance only works when it is *enforced and evidenced at runtime*, in the same place the agent runs and over the same telemetry the agent produces. We present **GAVEL (Governed Autonomy, Verified by an Evidence Layer)**, the governance architecture of the FlowX.AI Platform's Observatory plane. GAVEL compiles organizational and regulatory requirements into machine-checkable policies, evaluates them against live execution telemetry rather than against a questionnaire, harvests compliance evidence automatically from traces, binds each regulatory requirement to the policies and evidence that satisfy it, and records every enforcement action and human review in an immutable audit trail. An EU AI Act audit is then answered not by assembling a binder but by walking evidence back through policy to requirement, every link grounded in what the agent actually did. We detail the architecture, the regulation-to-runtime evidence chain, and an illustrative human-oversight case, and we discuss how governance grounded in observability scales autonomy with risk rather than trading one against the other.

KEYWORDS AI agent governance · runtime policy enforcement · compliance evidence · EU AI Act · audit trail · bounded autonomy · human oversight · observability

1 Introduction

An autonomous agent is, by definition, a system to which an organization has delegated authority: the right to read data, call tools, move money, or decide a case, with limited human involvement per action. Governance is the discipline of bounding that delegated authority, and it is the discipline enterprises are most visibly failing at. A 2026 survey of organizations deploying agents found that 71% had no formal governance framework, even as a majority planned to widen autonomy in the following year, and that 80% had already observed risky agent behaviors, from unauthorized data access to unexpected system interactions[6]. The instructive detail is what goes wrong: the headline industry worry is hallucination, but the dominant failure mode in deployed systems is *privilege excess*, an agent doing something it was technically able to do but never should have, and chaining such actions across systems faster than a human review cycle can observe.

Regulation has noticed. The EU AI Act reaches full effect in August 2026, with binding obligations for high-risk systems including human oversight, transparency, risk management, and record-keeping that must be retained for years[1]. The NIST AI Risk Management Framework[2] and a new standards initiative on agent identity, action logging, and containment[3], alongside ISO/IEC 42001[4] and sector guidance[9], are directionally aligned on the same primitives: accountability, behavioral transparency, and human oversight. The compliance question is no longer hypothetical; it has a date.

The market has responded with two kinds of tool that, tellingly, do not connect. On one side are governance-risk-and-compliance suites: registers, questionnaires, and policy documents that describe how agents *should* behave, maintained in a system that has no visibility into how they actually behave. On the other are runtime control planes: identity, authorization, and guardrail layers that enforce limits in the moment but emit operational logs, not the evidence an auditor or regulator will ask for. The first camp can pass a paper audit while the agents misbehave; the second can stop an agent while being unable to prove, months later, that it was ever governed at all. Between them lies the gap where real agent governance is supposed to happen.

> *A policy that lives in a document cannot govern an agent that chains a dozen tool calls in a second. Governance has to run where the agent runs, check what the agent actually did, and leave behind the evidence that it did.*
>
> GAVEL · CORE THESIS

We take the position that AI agent governance is not a document and not a guardrail in isolation, but a property of the system that observes the agent. It must be enforced at runtime, against the agent's actual execution rather than its specification, and it must produce, as a byproduct of that enforcement, the evidence that demonstrates compliance to a third party. We present **GAVEL (Governed Autonomy, Verified by an Evidence Layer)**, the governance architecture of the FlowX.AI Platform's Observatory plane. GAVEL turns governance into an executable layer over agent telemetry: it compiles organizational and regulatory requirements into machine-checkable policies, evaluates those policies against live traces, collects compliance evidence automatically from the same traces, maps every regulatory requirement to the policies and evidence that satisfy it, and records every enforcement action and human review in an immutable audit trail. The result is governance that is simul-

taneously operational and audit-ready, and that lets an organization raise autonomy where the evidence supports it rather than freezing it out of fear.

1.1 Contributions

- A reframing of AI agent governance from documentation to a runtime, evidence-generating property of the observability layer, and an architecture, GAVEL, that operationalizes it.
- A policy engine that compiles organizational and regulatory rules into machine-checkable policies and evaluates them against live execution telemetry rather than against a questionnaire.
- An automated evidence pipeline that harvests compliance evidence from traces, so that an audit is answered by walking evidence back through policy to requirement.
- A regulation-to-runtime mapping that binds requirements (for example, EU AI Act articles) to the policies, evidence, and assessments that satisfy them, with continuous gap analysis.
- A model of risk-based, governed autonomy with human oversight and an immutable audit trail, illustrated on an EU AI Act human-oversight obligation.

2 The Governance Gap

2.1 Two camps that do not meet

Enterprise AI governance tooling has converged on a now-standard set of pillars: bounded autonomy, identity and authorization, auditability, accountability, human oversight, and continuous monitoring[7,8]. The disagreement is not about the pillars but about where they live. Governance-risk-and-compliance platforms implement them as records: a policy is a document, a control is a checklist item, evidence is a screenshot uploaded by an analyst. This satisfies an auditor who reads binders, but it is structurally blind to runtime; nothing in the register changes when an agent quietly exceeds its remit. Runtime control planes implement the same pillars as enforcement: identity is a token, authorization is a privilege check, a guardrail blocks a tool call. This stops bad actions, but it produces operational telemetry, not the durable, requirement-linked evidence a regulator expects, and it has no notion of which regulatory obligation a given block satisfies.

2.2 The evidence gap

The space between the two camps is where agent governance actually fails. A documented policy that is never evaluated against behavior is an assertion, not a control; a runtime block that is never linked to a requirement is an event, not evidence. The EU AI Act, NIST AI RMF, and OWASP guidance converge on a single hard demand that neither camp meets alone: comprehensive, tamper-evident audit logs that connect what the system did to the obligations it was bound by, retained for years[1,2,5]. Closing this gap requires governance that is enforced on live execution *and* that emits requirement-linked evidence as a byproduct. Table 1 frames the contrast that GAVEL is designed to resolve.

Table 1. Where the three approaches place each governance pillar.

Capability	GRC / document suites	Runtime control planes	GAVEL
Policy form	Prose document	Code / config	Machine-checkable rule
Evaluated against	Questionnaire	Live calls	Live execution telemetry
Produces evidence?	Manual uploads	Operational logs only	Auto-collected from traces
Linked to regulation?	Yes, on paper	No	Requirement to policy to evidence
Answers an audit by	Assembling a binder	Cannot, directly	Walking the evidence chain

3 The GAVEL Architecture

GAVEL sits as a governance layer over the Observatory telemetry that already captures every agent run, tool call, retrieval, and decision. Rather than maintaining a parallel record of how agents are supposed to behave, it reads the record of how they did behave and evaluates it against policy. Figure 1 shows the arrangement: a telemetry substrate at the base, three governance subsystems that read from it, a regulatory mapping that organizes their output by obligation, and an accountability spine of role-based access control and an immutable audit trail.

FIG. 1 · GAVEL: GOVERNANCE AS AN EVIDENCE LAYER OVER AGENT TELEMETRY

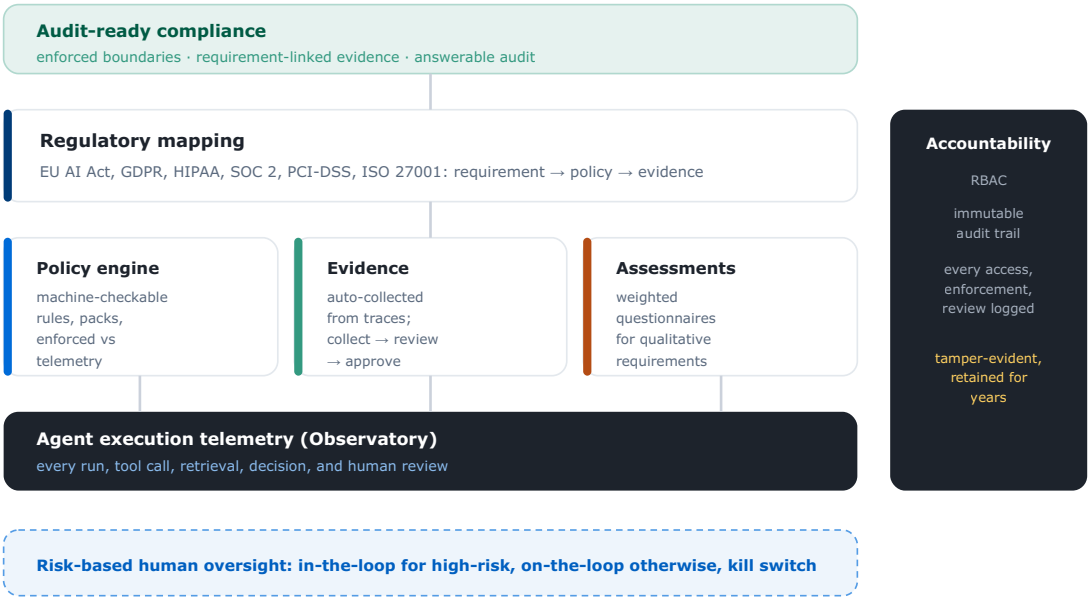

Figure 1. GAVEL reads the same telemetry the agent produces. Three subsystems (policy, evidence, assessment) evaluate behavior and gather proof; a regulatory mapping organizes that proof by obligation so an audit is answerable; an accountability rail of role-based access control and an immutable, tamper-evident audit trail records every access, enforcement action, and human review; and a risk-based oversight band determines when a human must approve rather than merely monitor.

3.1 The policy engine

A GAVEL policy is not prose but a structured, machine-checkable rule set: a metric, an operator, and a threshold, composed through a visual builder, with an enforcement mode of *mandatory* or *advisory* and a category that ties it to a regulatory family. Related policies are grouped into packs, such as an EU AI Act High-Risk Pack, and a pack is assigned to a project as a unit. The defining move is evaluation: a policy is run against the project's actual telemetry and produces a pass or fail result with detailed findings, and an aggregate evaluation reports the project's overall compliance status. Because policies are evaluated against execution, not against a self-report, a policy that is being violated says so, continuously, rather than waiting for an annual review to discover it.

3.2 Evidence collected from execution

Evidence in GAVEL is, wherever possible, a byproduct of running the system rather than a separate clerical task. Automated collection scans telemetry for policy-relevant artifacts (run records, evaluation results, assessment scores), creates evidence linked to the matching policy, and tags it as telemetry-sourced. Each piece of evidence moves through a *collect, review, approve* workflow so that a human signs off where judgment is required, and time-bound evidence carries an expiry so stale proof does not silently count. A gap analysis compares the evidence each policy requires against what has been collected and returns the shortfall, turning compliance from a once-a-year scramble into a continuously visible backlog.

3.3 Assessments for what telemetry cannot see

Not every obligation is a number. "Have you documented the model's intended use?" is a real regulatory question that no metric answers. GAVEL handles these with weighted assessment questionnaires whose questions carry types and weights and whose submission computes a score and, importantly, automatically creates an evidence record of type *assessment* linked to the relevant policy. The qualitative and the quantitative thus land in the same evidence store, addressable by the same audit. Policy captures what compliance looks like, evidence captures whether it is honored, and assessment captures the judgments thresholds cannot express.

3.4 Accountability: RBAC and an immutable audit trail

Governance that cannot say *who* is not governance. GAVEL records, in an append-only audit trail, every access to a sensitive trace, every policy enforcement action, and every human review, attributed to an identity and a role through role-based access control. The audit trail is tamper-evident and retained to satisfy record-keeping obligations that, under the EU AI Act, run to years for high-risk systems[1]. This is the layer that converts "the agent did something wrong" from an argument into a record, and that lets accountability attach to a person, not to an abstraction.

3.5 Risk-based, governed autonomy

The point of governance is not to minimize autonomy but to make it safe to grant. GAVEL supports a risk-based posture in which low-risk actions run autonomously under basic logging while high-impact, irreversible, or regulated actions require a human in the loop, with everything in between monitored on the loop and an escalation path and kill switch always available. Crucially, the oversight requirement is itself expressed as policy and checked against telemetry: a rule such as "every adverse decision must carry a recorded human review" is enforced and evidenced in exactly the same way as any other control. Autonomy can then expand where the evidence shows it is warranted, rather than being capped uniformly by the riskiest case.

4 From Regulation to Runtime

The capability that ties the architecture together is the chain from an external obligation to a runtime check and back to an answerable audit. GAVEL ships pre-built regulatory frameworks (EU AI Act, GDPR, HIPAA, SOC 2, PCI-DSS, ISO 27001) whose requirements are structured rows, each with an official code, the risk levels it applies to, the evidence types that satisfy it, and pre-mapped suggested policies and assessment questions. A compliance mapping links each requirement to a project and to the specific policies, evidence, and assessments that discharge it, and an auto-mapping step proposes those links by matching categories and content. Figure 2 traces a single obligation through the chain.

FIG. 2 · THE REGULATION-TO-RUNTIME EVIDENCE CHAIN

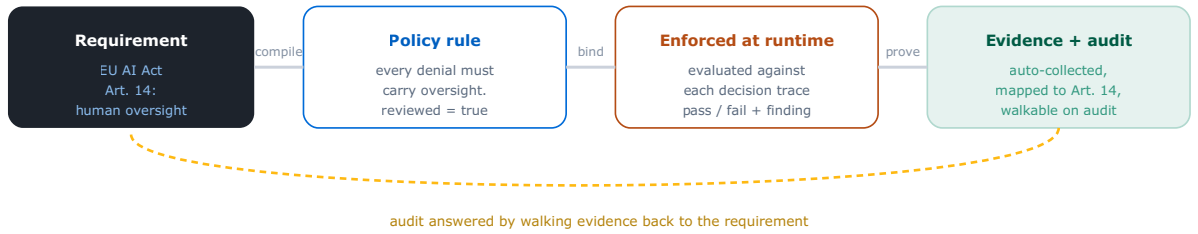

Figure 2. A regulatory obligation is compiled into a machine-checkable policy, bound to the agent's runtime so it is evaluated against each relevant decision trace, and proven by evidence that is auto-collected and mapped back to the obligation. The audit is then answered in reverse: from a requirement, to the evidence that discharges it, to the actual traces underneath, with no manual binder in between.

Two products fall out of the chain almost for free. A compliance report renders per-requirement status, an overall percentage, and the specific gaps (requirements not started or non-compliant), while an organization overview presents a color-coded project-by-framework matrix. Both are live views over the same evidence, so "where do we stand on the EU AI Act?" is a query, not a project. And because the evidence is requirement-linked, the same data answers an auditor's narrower question, "show me human oversight on adverse decisions in the last quarter," by returning the mapped evidence and, beneath it, the traces themselves.

5 Illustrative Application

Consider Meridian Insurance, an illustrative regulated insurer running an agent that adjudicates claims, including the authority to deny them. Under the EU AI Act, an automated adverse decision of this kind is high-risk and falls under Article 14, which requires meaningful human oversight. The figures and configuration here are illustrative and chosen to expose the mechanism.

Meridian assigns the EU AI Act High-Risk Pack to the claims project. Within it, a mandatory policy expresses Article 14 as a checkable rule: *every automated denial must carry an oversight.reviewed = true field and a recorded reviewer identity.* Nothing about this is a document; it is a rule evaluated against the telemetry of every denial the agent produces. A denial that a reviewer signed off on passes, and its trace becomes automated evidence linked to Article 14. A denial emitted without the oversight field fails the policy: the finding surfaces in the project's compliance status, the gap analysis records a shortfall against Article 14, and, because the policy is mandatory, the action can be held for review rather than released. The human-oversight obligation is, in other words, enforced at the moment of the decision and not reconstructed afterward.

When Meridian's regulator later asks the standard question, "demonstrate human oversight of adverse decisions," the answer is not a binder assembled under deadline. The compliance report shows Article 14 as compliant, backed by a population of evidence rows; each row links to a specific denial trace and the reviewer who approved it; and the immutable audit trail shows, for each, who accessed the trace and when the review occurred. A documented policy could only have asserted that oversight existed; GAVEL shows it, decision by decision. The same evidence, incidentally, feeds the platform's compliance-audit ROI view, where the automated-versus-manual evidence ratio translates the governance work into auditor-hours saved, so the investment in governance appears on the financial dashboard rather than only on the risk register.

The agent retains real autonomy: it adjudicates the great majority of claims without a human touching them. What Article 14 governs, and what GAVEL enforces and evidences, is the narrow, high-impact class of actions where oversight is required. Autonomy and accountability are not traded against each other; the evidence is what lets both rise together.

6 Discussion

Treating governance as a runtime, evidence-generating layer changes the procurement question an enterprise should ask. Not “do you have a policy for that?” (a binder can answer yes) but “evaluate the policy against last week’s traces, show me the evidence, and walk it back to the regulation.” That question is answerable only by a system that lives where the agents run. It also reframes the relationship between governance and autonomy. The reflex under uncertainty is to cap autonomy uniformly; but if oversight obligations are themselves enforced and evidenced, an organization can grant more autonomy precisely where the evidence shows controls are holding, and reserve human review for the cases that warrant it. Governance becomes the enabler of autonomy rather than its brake.

GAVEL is also the connective layer beneath the rest of the platform’s assurance story. Node-level self-adaptation, evidence-graded return on investment, and hallucination containment each generate telemetry; GAVEL is what binds that telemetry to obligations and renders it auditable, and what monetizes it as a compliance-audit return rather than a cost. As regulation moves from principles to dated obligations[1,3,9], the organizations that fare best will be those for whom compliance is a query over what their agents actually did, not a document about what they were supposed to do.

7 Limitations

- **Governance is only as good as the telemetry.** GAVEL evaluates policy against what Observatory captures; an action taken outside the instrumented platform is invisible to it. Coverage of the agent's real action surface is a precondition, not a guarantee.
- **A policy still has to be written correctly.** Compiling a regulatory obligation into a machine-checkable rule is an act of interpretation; a poorly specified rule can pass while the spirit of the obligation is violated. Auto-mapping suggests links but does not absolve a human of validating them.
- **Evidence can be over-trusted.** Automated, telemetry-sourced evidence is strong, but a mapping that marks a requirement compliant on thin or stale evidence is a false comfort; expiry and gap analysis mitigate this but depend on correct configuration.
- **Assurance of the governance layer itself.** Mature frameworks call for proving the governance layer has not been tampered with, ideally via hardware-backed attestation[7]. Tamper-evident logging is a step toward this, not the whole of it.
- **Regulation is a moving target.** Pre-built frameworks must track amendments and new guidance; a framework that drifts out of date silently weakens every mapping built on it.

8 Conclusion

The enterprise is delegating real authority to autonomous agents faster than its governance can follow, and the failures that result are less about models inventing facts than about agents exceeding the authority they were given. Documents cannot govern such systems, and runtime guards alone cannot prove they were governed. The resolution is to make governance a property of the layer that watches the agent: compile obligations into machine-checkable policies, enforce them against live execution, harvest the evidence as a byproduct, bind it to the regulations it satisfies, and keep an immutable account of who did what. GAVEL is that layer. It does not ask an organization to choose between autonomy and accountability; it makes the evidence that lets both grow at once, and turns the looming compliance deadline from a binder to assemble into a question the system can already answer.

References

[1] European Parliament and Council. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union, 2024. (Art. 12 record-keeping; Art. 14 human oversight.)

[2] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* NIST AI 100-1, 2023.

[3] National Institute of Standards and Technology, NCCoE. *AI Agent Standards Initiative: Concept Paper on Agent Identity, Action Logging, and Containment.* NIST, 2026.

[4] International Organization for Standardization. *ISO/IEC 42001:2023, Information technology, Artificial intelligence, Management system.* ISO, 2023.

[5] OWASP Foundation. *OWASP Top 10 for Large Language Model Applications.* OWASP, 2025.

[6] Forrester Research. *The State of AI Agent Governance, 2026.* (Survey: 71% without a formal agent-governance framework; 80% reporting risky agent behaviors.) 2026.

[7] *AAGATE: A NIST AI RMF-Aligned Governance Platform for Agentic AI.* arXiv:2510.25863, 2025.

[8] *Identity Management for Agentic AI: Authorization, Authentication, and Security for an AI Agent World.* arXiv:2510.25819, 2025.

[9] Infocomm Media Development Authority (Singapore). *Model AI Governance Framework for Agentic AI.* IMDA / AI Verify Foundation, 2025.

[10] *Systematization of Knowledge: Security and Safety in the Model Context Protocol Ecosystem.* arXiv:2512.08290, 2025.

[11] European Parliament and Council. *Regulation (EU) 2016/679 (General Data Protection Regulation).* Official Journal of the European Union, 2016.

Appendix A Pre-Built Regulatory Frameworks

GAVEL seeds six frameworks on first run; custom org-specific frameworks can be added. Each ships structured requirements with codes, applicable risk levels, and pre-mapped policies and assessment questions.

Framework	Domain	Example requirement	Typical evidence
EU AI Act	Risk-based AI regulation	Art. 14 human oversight; Art. 12 record-keeping	Reviewed-decision traces, audit logs
GDPR	Data privacy	Lawful basis, data minimization	PII-redaction telemetry, DPIAs
HIPAA	Healthcare data	Access controls on PHI	RBAC logs, access audit trail
SOC 2	Service controls	Security, availability, confidentiality	Policy evaluations, monitoring
PCI-DSS	Payment card data	Restrict cardholder-data access	Authorization logs
ISO 27001	Information security	ISMS controls	Assessments, evidence rows

Appendix B Worked Chain: EU AI Act Article 14

This appendix walks a single obligation from regulation to an answerable audit, the steps of Section 4 made concrete.

1. Requirement. The framework row for EU AI Act *Art. 14* (human oversight) declares its risk level (high), the evidence types that satisfy it (records of human review on relevant decisions), and suggested policies and questions.

2. Policy. A mandatory policy in the EU AI Act High-Risk Pack compiles the obligation into a checkable rule: *for every run whose decision is a denial, the trace must contain oversight.reviewed = true and a reviewer identity.* The pack is assigned to the claims project.

3. Enforcement. The policy is evaluated against the project's telemetry. Each denial trace is checked; compliant traces pass, non-compliant ones produce a finding, and because the policy is mandatory the action can be held for review.

4. Evidence. Automated collection turns each compliant, reviewed denial into a telemetry-sourced evidence row linked to the Art. 14 policy; the review event and trace access are written to the immutable audit trail.

5. Mapping and audit. The compliance mapping marks Art. 14 compliant for the project, backed by the evidence population. An audit query walks Art. 14 to its evidence rows to the underlying traces and reviewer identities, with no manual assembly. Gap analysis flags any period whose coverage is thin *before* the auditor does.

Appendix C Governance Pillars Mapped to GAVEL

The field converges on a standard set of governance pillars. The matrix shows where each is realized in GAVEL, and how it is evidenced rather than merely asserted.

Pillar	GAVEL realization	Evidenced by
Bounded autonomy	Risk-based oversight; mandatory policies on high-risk actions	Policy evaluations vs telemetry
Identity & authorization	RBAC over traces and governance actions	Access entries in audit trail
Auditability	Immutable, tamper-evident audit trail	The trail itself, retained for years
Accountability	Every action attributed to identity + role	Attributed audit entries
Human oversight	In-the-loop / on-the-loop, kill switch, review fields	Reviewed-decision evidence
Continuous oversight	Scheduled policy evaluation, gap analysis, drift	Live compliance status
Regulatory alignment	Requirement to policy to evidence mapping	Compliance reports + matrix

FlowX.AI

FlowX.AI builds and orchestrates enterprise-grade AI agents for regulated industries, with a focus on banking, financial services, and insurance. The FlowX.AI Platform pairs LangGraph-native orchestration with the Observatory plane, whose governance engine enforces organizational and regulatory policy against live agent telemetry and produces the evidence that makes autonomous workflows auditable, accountable, and compliant by construction.

FlowX.AI Technical Paper · GAVEL v1.0 · Bogdan Răduță, Head of Research · bogdan.raduta@flowx.ai