Received: 06 June 2025, Revised: 14 June 2025, Accepted: 07 August 2025, Online: 20 August 2025

DOI: https://dx.doi.org/10.55708/js0408004

Magnetic AI Explainability: Retrofit Agents for Post-Hoc Transparency in Deployed Machine-Learning Systems

Maikel Leon*

Department of Business Technology, Miami Herbert Business School, University of Miami, Miami, Florida, USA

*Corresponding author. Email: mleon@miami.edu

ABSTRACT: Artificial intelligence already influences credit allocation, medical diagnosis, and staff recruitment, yet most deployed models remain opaque to decision makers, regulators, and the citizens they affect. A new wave of transparency mandates across multiple jurisdictions will soon require organizations to justify automated decisions without disrupting tightly coupled production pipelines that have evolved over the years. We advance a conceptual proposal to address this tension: the magnetic AI agent. This external, attachable software layer learns a faithful surrogate of any target model, delivering audience-tailored explanations on demand. The paper first synthesizes fragmented scholarship on post-hoc explainability, sociotechnical alignment, and model governance, revealing an unmet need for lightweight retrofits that minimize downtime. It then creates a basic framework based on design principles, explaining methods for data collection, ongoing learning processes, and user-friendly explanation tools. A plan for evaluation lists both numerical and descriptive measures, including how closely a model matches reality and how much extra time it takes, as well as the mental effort required and how well policies work, which users can adjust for different fields like credit scoring, medical imaging, and predictive maintenance. Overall, the work contributes a roadmap for upgrading the installed base of black-box systems while aligning with emergent regulatory frameworks and ethical guidelines for trustworthy AI.

KEYWORDS: Magnetic AI, Explainable Artificial Intelligence, Agentic AI, Retrofit Transparency, Design-Science Research, Policy Compliance.

1. Introduction

Artificial Intelligence (AI) systems that once resided in research labs now power high-stakes finance, health care, logistics, national security, and public administration decisions. These models deliver unprecedented speed and predictive accuracy, yet they rarely reveal the internal logic that drives their outputs. This asymmetry between performance and interpretability poses reputational, operational, and legal risks for organizations that rely on opaque algorithms. Recent incidents—such as biased credit approvals, flawed recidivism predictions, and inconsistent medical triage decisions—demonstrate how opacity can erode stakeholder trust and invite regulatory scrutiny [1].

Last century, AI research surged on the back of expert systems, decision trees, and the early "neural nets" revival. Success was measured almost entirely by how precisely these models could predict outcomes, whether diagnosing disease, flagging credit risk, or recognizing handwritten digits. Researchers fine-tuned rule bases or tweaked hidden-layer weights to squeeze out a few extra percentage points of accuracy, and industry adopters celebrated any gains that outperformed human benchmarks. Yet this accuracy-first mindset treated the models as opaque black boxes: engineers rarely asked why a particular rule fired or a neuron activated, and users seldom demanded a justification. As a result, explainability remained an afterthought; the momentum and funding of the era were channeled into sharpening predictive performance, not into opening the "black box" so stakeholders could trust and understand the reasoning inside it.

Across major jurisdictions, regulation is converging on a common requirement that AI systems be explainable: the European Union's AI Act, recent U.S. executive directives, and China's updated generative-AI rules all mandate that high-impact models provide meaningful information about how they reach their outputs. This amounts to an emerging right for everyday users to demand clear, human-readable reasons for automated predictions or decisions, even when those decisions come from complex neural networks. Anticipating audits, fines, and reputational risks, companies are building explanation layers into their products—dashboards that visualize feature contributions, surrogate models that translate deep-learning logic into plain language, and customer portals that show "what-if" scenarios—because meeting this new transparency baseline is becoming less a nice-to-have and more a competitive necessity.

Societal expectations for transparency have accelerated. Policymakers on both sides of the Atlantic have enacted or proposed frameworks that place the burden of justification on automated decision-makers. The European Union's AI Act, the United Kingdom's Algorithmic Transparency Standard, and various U.S. proposals such as the Algo-

rithmic Accountability Act collectively signal a shift from self-regulation to explicit accountability. These initiatives often focus on two intertwined requirements: the ability to generate human-understandable explanations and the capacity to audit models throughout their life cycle. Organizations, therefore, face the dual challenge of upgrading legacy AI assets and operationalizing governance processes at scale.

Despite rapid advances in post-hoc interpretability techniques, most production environments cannot easily accommodate invasive code changes, extensive retraining cycles, or computational overhead that might jeopardize service-level agreements. Enterprise Machine Learning pipelines typically integrate proprietary libraries, tightly coupled microservices, and third-party APIs that preclude direct intervention. A non-disruptive alternative is to attach an explanatory agent to the outside of an existing pipeline, much like a magnetic device that snaps onto the surface of a machine without changing its internal workings. We label this solution the magnetic AI agent. The magnetic analogy underscores three salient properties: passive attachment, minimal friction, and continuous real-time learning [2].

While the concept of attaching post-hoc interpretability layers has precedent in techniques such as shadow models, knowledge distillation, and wrapper-based surrogates, the magnetic AI agent diverges in critical ways. Unlike shadow models that mimic predictions for evaluation purposes or distillation methods that compress complex models into simpler ones, the magnetic agent is designed to operate continuously alongside the original model without approximation or replacement [3]. Its emphasis is not only on interpretability but also on modular deployment, governance integration, and lifecycle adaptability in real-world production systems. The magnetic metaphor is not a rhetorical flourish—it reflects an architectural philosophy: to enable passive but intelligent observability without disrupting the core model's functioning or retraining requirements.

The remainder of the paper deepens the conceptual foundation, formalizes the design space, and proposes an actionable evaluation pathway for magnetic AI. While empirical results are not presented here, this absence is by design: the work is intended as a conceptual proposal that lays the groundwork for future implementation and experimentation. Its primary aim is to contribute a structured framework, design rationale, and deployment blueprint that researchers and practitioners can build upon. First, Section 2 surveys the multidisciplinary literature on explainable AI and model-agnostic wrappers, identifying persistent gaps that motivate a new approach. Section 3 introduces the conceptual framework that positions the retrofit agent within sociotechnological constraints and elaborates design principles, reference architecture, and governance interfaces. Section 4 describes a design-science research strategy and methodological considerations for constructing and refining the artifact. Section 5 details an evaluation blueprint that organizations can replicate or adapt in their domains. Section 6 discusses operational, ethical, and societal implications, mapping the proposal onto current regulatory trends. Section 7 concludes by summarizing contributions, delineating limitations, and articulating a future research agenda that includes full-scale prototypes, multimodal extensions, and integration with next-generation foundation models.

2. Related Work

Research on explainability spans multiple disciplines, each supplying partial answers to how automated systems should justify their outputs. Algorithmic contributions range from ante-hoc transparent models to post-hoc attribution methods such as LIME, SHAP, and integrated gradients to compression techniques that create interpretable surrogates. Human-computer interaction studies examine the cognitive load of different explanation formats, user mental-model accuracy, and the conditions under which explanations raise or erode calibrated trust. Work in organizational behavior documents how power dynamics, siloed incentives, and technical debt shape whether explanations are acted upon or ignored. Legal scholarship and policy analyses frame transparency as a right, exploring liability, due-process entitlements, and the evolving notion of algorithmic accountability [4].

This review weaves the strands together, pinpointing where they fall short and how they complement one another. Algorithmic methods often optimize fidelity or sparsity but rarely address maintenance overhead once a model is in production. HCI experiments illuminate user comprehension in laboratory settings, yet evidence remains sparse on sustained behavior change in real workflows. Organizational case studies highlight governance bottlenecks but seldom tie them to concrete design artifacts. Legal work identifies transparency duties but leaves practitioners with little guidance on technical implementation. Magnetic AI draws on the strengths of each field while addressing their gaps: a passive attachment strategy respects intellectual-property boundaries emphasized in law, continuous fidelity auditing answers organizational concerns about drift and technical debt, and explanation pluralism accommodates the heterogeneous user needs documented in HCI research [5].

Key takeaways that inform the design are as follows:

- Algorithmic insight: incremental surrogates balance fidelity with latency, enabling explanations at line speed without altering the primary model. They learn from a sliding window of recent requests, refresh continuously without full retraining, and respect the intellectual-property boundaries of closed models, making them suitable for third-party APIs and in-house stacks.
- HCI insight: multiple discourse formats—ranked feature tables, layered saliency maps, natural-language counterfactual narratives, and compliance-ready audit summaries—are necessary because data scientists, end users, and regulators each privilege different cues. Adaptive rendering lets the same evidence flow into analyst dashboards, tooltips for consumers, or machine-readable JSON for supervisory authorities.
- Organizational insight: modular deployment decouples the four layers—interception, surrogate learning, explanation rendering, and fidelity auditing—so firms can adopt only the components they lack. This bolt-on architecture avoids rewriting brittle legacy code,

shortens change-management cycles, and reduces the blast radius of defects to a single microservice rather than the full model pipeline.

- Legal insight: persistent audit logs, role-based explanation access, and optional differential-privacy noise satisfy both transparency duties and data-protection rules. The same artifacts can populate internal risk registers, respond to freedom-of-information requests, or demonstrate compliance during external audits, aligning technical controls with emerging statutes such as the EU AI Act and national consumer-protection guidelines [6].

By fusing these lessons, magnetic AI offers a coherent blueprint that advances beyond silo-specific approaches toward an integrated, production-ready solution for trustworthy machine learning.

2.1. Post-Hoc Explainable AI

Early work on interpretability concentrated on "glass-box" algorithms—decision trees, linear or logistic regressions, and simple rule lists—whose parameters and splits can be read like prose. As deep learning's opaque layers dominated predictive accuracy, researchers shifted toward post-hoc techniques that wrap explanations around otherwise black-box models [7].

The most influential of these are LIME and SHAP. Both build local surrogate models that mimic the original model's behavior near a single instance, then report feature attributions: LIME perturbs inputs and fits a sparse linear model, whereas SHAP samples coalitions of features to compute Shapley values that satisfy additivity and consistency. Their appeal lies in domain-agnostic deployment—data scientists can drop in a few lines of code and hand users a ranked list of "which variables mattered most"—yet the price is high computational overhead, sensitivity to sampling noise, and explanations that change when the same point is probed twice [8].

Beyond LIME and SHAP, gradient-based saliency maps track the partial derivatives of a convolutional network to highlight the pixels that nudge an image score upward or downward; attention visualizations in transformer models color the tokens that capture a language model's gaze; counterfactual methods search the input space for the most minor tweak that flips the prediction, offering an actionable "what would need to change?"; and prototype- or example-based explanations surface representative cases that anchor abstract probability scores in concrete, human-readable examples. Each broadens the explanatory toolbox, yet each inherits its drawbacks: saliency maps blur under adversarial noise, attention plots do not always align with causal importance, counterfactuals become infeasible in high-dimensional data, and prototype selection can reinforce majority-class bias [9].

Across the board, explanation strength often comes at the cost of latency, stability, or hardware resources. Empirical studies still debate whether richer explanations meaningfully boost user trust or downstream decision quality, highlighting an unsolved interpretability-accuracy-usability triangle.

2.2. Wrapper and Surrogate Paradigms

Building a simpler model that imitates a complex one is hardly new. In the 1980s, credit bureaus built "shadow" logistic regressions to track the decisions of proprietary loan scoring engines, and in the 1990s, speech-recognition teams used teacher–student pairs to shrink large hidden-Markov networks so they could run on low-power chips. These ideas matured into what is now called knowledge distillation, where an extensive teacher network produces soft targets—probability distributions rather than hard labels—that guide a smaller student network. The result is a faster, lighter model that often matches the teacher's top-line accuracy but may blur fine-grained decision boundaries, especially in rare or ambiguous cases.

Modern workflows try to close that gap by performing distillation continuously. An online student receives a stream of teacher outputs and updates its weights on the fly, or it joins a replay buffer that mixes new observations with old exemplars to resist catastrophic forgetting. Continual-learning variants add regularizers that anchor key teacher activations so the student does not drift when the data distribution shifts. Yet experiments on non-stationary benchmarks show that even these advanced students struggle with concept drift and are highly sensitive to mislabeled or adversarially perturbed examples [10].

A parallel line of work forgoes access to internal weights altogether. Instead, engineers wrap the black-box service with a data interceptor that logs inputs and outputs, then train a surrogate, often a decision tree or gradient-boosted ensemble, purely from those pairs. This wrapper strategy sidesteps intellectual-property barriers and can be swapped before any commercial API. Still, it introduces fresh privacy challenges: synthetic or cached query data must be stored outside the original security perimeter, and reconstruction attacks can expose sensitive attributes if the wrapper is breached [11].

Taken together, today's surrogate models fall into two camps. Static snapshots captured once during development grow stale as the real world evolves, while dynamic surrogates that retrain or distill online demand constant monitoring, a computation budget, and careful privacy safeguards. Neither camp fully resolves the tension between efficiency, fidelity, and maintainability in production environments that change by the hour.

2.3. Regulatory and Business Context

Across regions, lawmakers and standard-setters are locking into a shared vocabulary—transparency, accountability, fairness, and meaningful human oversight—and turning it into binding or quasi-binding rules. In Europe, the AI Act labels credit scoring, hiring, medical diagnosis, and other "high-risk" applications. It forces them to generate understandable explanations, document data provenance, and pass third-party conformity assessments before entering the market.

In the United States, the Federal Trade Commission, Consumer Financial Protection Bureau, Department of Justice, and other agencies have warned that undisclosed bias, dark-pattern interfaces, or the sale of inscrutable models

can trigger enforcement actions under existing consumer-protection and civil-rights statutes. At the same time, the White House blueprint for an AI Bill of Rights and the NIST AI Risk-Management Framework give regulators a benchmark for what "reasonable" governance should look like. China's updated Interim Measures on generative AI require providers to watermark outputs, publish model cards, and supply "interpretive" summaries on demand; Canada's forthcoming AI and Data Act mandates impact assessments and real-time monitoring; Brazil and India are drafting parallel bills; and the G7's Hiroshima Process is pressing multinationals to align with these norms wherever they operate.

Industry bodies reinforce the trend: the Partnership on AI, the OECD, the ISO/IEC 42001 management-system standard, and voluntary procurement checklists now ask vendors to show audit logs, bias tests, and plain-language explanations as a condition of sale. Non-compliance can mean multimillion-euro fines, exclusion from public-sector tenders, investor divestment, and reputational damage that stalls digital-transformation roadmaps. Yet most enterprises run on entrenched code bases, brittle data pipelines, and overlapping legacy models; ripping and replacing them is rarely feasible. This clash between external pressure and internal technical debt drives demand for retrofit solutions—lightweight layers that bolt onto existing systems, capture inputs and outputs, monitor drift, and surface user-friendly explanations—so firms can satisfy new governance obligations without rebuilding their entire machine-learning stack [12].

2.4. Gap Analysis

Table 1 contrasts prevailing approaches against operational requirements and spotlights the unresolved disconnect between research prototypes and production realities. While the literature offers algorithmic sophistication, it rarely addresses day-two concerns such as deployment pipelines, monitoring infrastructure, and heterogeneous stakeholder needs. The magnetic AI proposal aims to bridge this gap by integrating passive attachment, continuous fidelity auditing, and human-centered explanation delivery into a unified artifact.

There seems to be a clear trade-off pattern: methods that are easiest to bolt onto any model (LIME, SHAP, Anchors) suffer from high inference latency or heavy sampling, while techniques that are fast enough for production (knowledge-distilled surrogates, ante-hoc interpretable models) often under-fit or drift from the source model without constant retraining. Vision-specific tools like Grad-CAM are efficient but narrow in scope, and counterfactual or prototype-based approaches provide the most human-friendly "what-if" stories yet demand large compute budgets and carefully curated instance libraries [13].

In short, no single technique simultaneously delivers low latency, high fidelity, and broad stakeholder usability. This operational gap motivates a hybrid solution, such as the proposed magnetic AI artifact, that couples passive attachment for real-time capture with continuous fidelity auditing and layered explanation modes tuned to different audiences.

Table 1: Operational gap between explainability techniques and production requirements

Approach	Strengths	Limitations
LIME / SHAP	Model-agnostic; easy to add	High latency in production; explanations local
Knowledge distillation	Compact, fast surrogates	Needs labelled outputs; surrogate drift
Counterfactuals	Actionable "what-if" paths	Heavy compute; plausibility issues
Magnetic AI (proposed)	Passive attachment; continuous learning	Concept stage; governance pending
Integrated Gradients	Faithful to deep nets; low single-call overhead	Requires differentiable model; noisy for saturated neurons
Grad-CAM	Intuitive heat-maps for vision CNNs; real-time on GPU	Vision-only; coarse spatial resolution
Anchors	Sparse, high-precision rules; human-readable	Sampling-intensive; struggles with high-dimensional mixes
Partial Dependence / ICE	Global feature-effect trends; offline computation	Assumes feature independence; stale in changing data
Prototype & Criticism	Example-based, domain-relatable explanations	Needs large representative set; weak in very sparse spaces
Ante-hoc interpretable mdl.	Transparency built-in (e.g., GAMs, monotonic GBMs); low latency	May under-fit complex tasks; restricted model choices

3. Magnetic AI Conceptual Framework

The magnetic AI framework delineates the core constructs, operational boundaries, and design guidelines necessary to retrofit explainability into black-box systems. Building on sociotechnical theory, the framework positions the agent as an intermediary that negotiates between opaque algorithms and heterogeneous human audiences [14].

3.1. Definition and Scope

A magnetic AI agent functions as a sidecar or proxy service that eavesdrops on every request–response pair flowing to and from a production model. As each new interaction arrives, the agent adds it to a sliding window buffer—say the most recent ten thousand cases—and updates an online surrogate such as an incremental gradient-boosted tree or a compact transformer fine-tuned with parameter-efficient adapters. This continual refresh allows the surrogate to track concept drift without incurring the full cost of retraining. Because the agent learns only from observable inputs and outputs, it can attach to black-box APIs, commercial SaaS endpoints, or legacy binaries without source code or training data. Once the surrogate reaches a configurable fidelity threshold, the agent can emit different explanation "dialects" on demand: concise ranked feature lists for customer-service representatives, multi-layer saliency maps

for data scientists, counterfactual recourse suggestions for end users, or timestamped audit reports that regulators can archive. A governance layer encrypts the buffered data, records model-to-surrogate agreement scores, triggers alerts when fidelity degrades, and exposes REST or gRPC endpoints so downstream dashboards can pull explanations in real time [15].

Deployment is lightweight—often a Docker container or Kubernetes sidecar—so platform teams can roll it out with minimal changes to existing pipelines. Because the agent never touches proprietary weights or training sets, intellectual-property boundaries remain intact, and privacy can be reinforced with hashing or differential-privacy noise in the captured feature vectors. This combination of passive attachment, incremental learning, and audience-specific explanation formats positions magnetic AI agents as a practical retrofit for organizations that must meet new transparency rules without redesigning their entire machine-learning stack.

3.2. Design Principles

Design principles serve as invariant heuristics that guide implementation choices across contexts:

- Plug-and-play attachment via standardized data taps that conform to common message-queue or REST interfaces, minimizing engineering overhead.
- Model and domain agnosticism that enables deployment across tabular, image, NLP, time-series, and multimodal pipelines.
- Continuous auditing that monitors surrogate fidelity over time using drifting-window statistical tests and triggers automatic recalibration when thresholds are breached.
- Explanation pluralism that tailors output modalities to stakeholder expertise, regulatory requirements, and situational constraints, thereby enhancing relevance and comprehension.
- Privacy-preserving learning that supports on-device distillation, differential privacy budgets, and federated aggregation when data sovereignty is paramount.

3.3. Reference Architecture

The architecture is divided into four loosely coupled layers. The data interception layer attaches to message brokers, REST gateways, or in-process hooks to duplicate each input–output pair with millisecond-level delay. Captured data is written to an encrypted sliding-window buffer sized to the latency budget. The surrogate learning layer ingests this stream and updates an incremental model such as an online gradient-boosted tree, streaming k-nearest neighbors, or a partial-fit neural network.

A fading factor emphasizes recent samples so the surrogate can track concept drift without unbounded memory growth. The explanation rendering layer queries the current surrogate to extract local and global importance signals, then converts them into human-readable artifacts by combining a template engine with natural-language generation. Supported formats include ranked feature lists, layered saliency maps, counterfactual recourse narratives, and compliance-oriented audit summaries.

The fidelity auditing layer compares surrogate outputs with the target model on a hold-back stream slice, records agreement statistics, raises drift alerts when error thresholds are exceeded, and exposes metrics to governance dashboards through an HTTP endpoint. The modular design permits selective adoption, so an organization may activate only the components that fill existing gaps:

- Data interception choices: sidecar proxy, service-mesh filter, or Kafka consumer
- Surrogate learning supports pluggable incremental algorithms and optional ensembling
- Explanation rendering exports Markdown, JSON, PDF, or SVG artefacts for integration with existing portals
- Fidelity auditing pushes metrics to Prometheus or OpenTelemetry and routes alerts to Slack or PagerDuty

Figure 1 illustrates the magnetic AI agent operating across four loosely coupled layers.

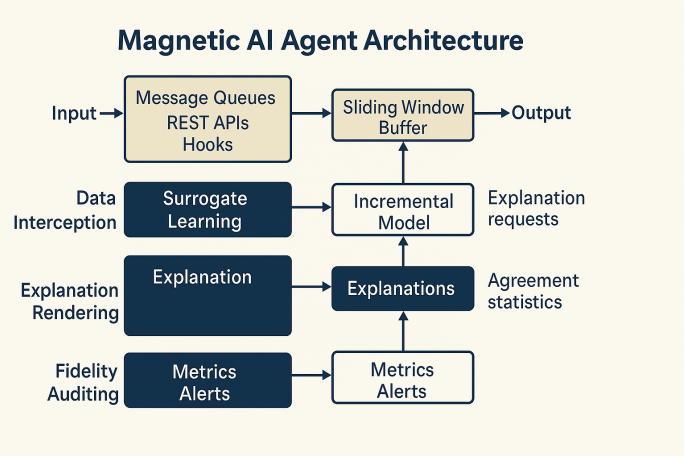

Figure 1: Magnetic AI Reference Architecture: A four-layer system that retrofits explainability into black-box models using passive data interception, online surrogate learning, audience-specific rendering, and continuous fidelity auditing.

4. Research Design and Methodology

Table 2 summarizes the guiding questions. Rigorous methodological scaffolding is essential to transform a design idea into an evaluable artifact. We adopt a design-science paradigm that iteratively synthesizes knowledge through constructing and assessing purposeful artifacts.

4.1. Artifact Construction Strategy

The construction strategy unfolds in three stages. Stage 1 employs synthetic benchmarks such as tabular classification tasks from the UCI repository to validate algorithmic viability under controlled conditions. Stage 2 transitions to semirealistic testbeds—for example, open medical-imaging datasets—where data sensitivity approximates production scenarios. Stage 3 involves shadow deployments within partner organizations, embedding the agent in parallel with live systems to observe operational impacts without

influencing decision outcomes. Each stage employs a build-measure-learn loop, refining data-tap APIs, surrogate hyperparameters, and explanation formats based on empirical feedback.

Table 2: Guiding questions for magnetic AI research design

Research question	Section
What functions must a retrofit agent perform to satisfy transparency mandates?	Framework
How can fidelity be maintained as underlying models drift?	Methodology
Which usability metrics best capture explanation quality across domains?	Evaluation
What governance processes are necessary to embed magnetic agents responsibly?	Discussion

4.2. Proposed Evaluation Metrics

Comprehensive evaluation encompasses technical fidelity, human factors, and organizational fit.

- Surrogate fidelity quantified by macro-averaged agreement, calibration error, and local explanation stability across perturbed inputs.
- Latency overhead measured as the delta between baseline prediction response time and pipeline response time with the agent attached, segmented by cold-start and steady-state conditions [16].
- Cognitive burden assessed via the NASA-TLX workload instrument and validated comprehension quizzes administered to diverse user cohorts.
- Policy sufficiency mapped to ISO-based checklists and jurisdiction-specific compliance rubrics, with binary pass/fail indicators and narrative justifications.
- Maintenance complexity captured through engineer-reported setup time, mean time to detection, and time to repair when drift alarms are triggered.

5. Evaluation Blueprint

A structured evaluation helps an organization transition from proof of concept to full roll-out without losing sight of risk, cost, or stakeholder value. Below, we will break the adoption into four incremental phases, each with its entry criteria, success indicators, and decision gates. Escalation to the next phase occurs only when the previous one meets predefined thresholds, reducing the likelihood of expensive rework later in the project. As shown in Figure 2, the evaluation progresses through four structured phases.

5.1. Phase 1: Feasibility Scoping

The objective is to decide whether a magnetic agent can attach to existing systems with acceptable effort and risk. A cross-functional team—product owners, data engineers, legal counsel, and compliance officers—maps the technical and organizational landscape before a single line of code is written.

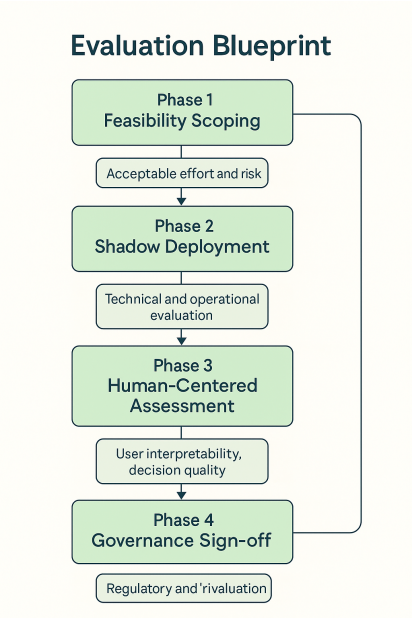

Figure 2: Evaluation Blueprint: A four-phase process guiding the deployment of magnetic AI agents from feasibility scoping to governance sign-off.

- Catalog candidate models, including version numbers, input modalities, and traffic volumes.
- Identify data-tap points such as message queues, microservice gateways, or in-process hooks.
- Segment explanation audiences: internal analysts, external customers, and regulators.
- Run a one-week pilot that captures a small sample of input–output pairs to confirm data visibility, latency overhead, and encryption requirements.
- Document legal constraints on data copying, retention, and cross-border transfer.

A green light to Phase 2 requires evidence that data taps are technically feasible, that no show-stopper legal barriers exist, and that the surrogate can be trained within the latency budget on a representative sample.

5.2. Phase 2: Shadow Deployment

The magnetic agent now runs parallel with the production model but remains invisible to end users. The aim is to measure technical fidelity and operational impact without altering business outcomes.

- Stream live input–output pairs to the surrogate and store them in a ring buffer sized to the retention policy.
- Generate explanations, drift graphs, confusion matrices, and saliency heat maps; push them to a read-only dashboard.
- Track surrogate-to-model agreement, memory growth, and compute cost hourly.
- Stress-test the agent under peak traffic loads to verify scaling rules and auto-healing scripts [17].
- Perform red-team exercises to probe for model inversion and data leakage vectors.

Promotion to Phase 3 requires that fidelity metrics reach a predefined threshold, that resource consumption stay within budget, and that no critical security vulnerabilities remain open.

5.3. Phase 3: Human-Centered Assessment

With technical soundness established, the focus shifts to human interpretability and decision quality. Explanations are shown to real users in a sandbox or pilot workflow.

- Recruit subject-matter experts—credit underwriters, fraud analysts, radiologists—for structured review sessions.
- Present a stratified sample of explanations, including edge-case and adversarial examples.
- Collect quantitative scores using metrics from Section 4 and qualitative feedback on clarity, usefulness, and domain language.
- Run A/B trials where some users receive explanations and others do not, measuring changes in decision time, error rate, and confidence calibration.
- Iterate on templates, terminology, and granularity until user-acceptance criteria are met.

Advancement to Phase 4 depends on demonstrable gains in user understanding or workflow efficiency and the absence of new cognitive or fairness concerns.

5.4. Phase 4: Governance Sign-off

The final checkpoint aligns the deployment with corporate risk appetite and external regulatory obligations. A multidisciplinary committee reviews evidence accumulated in earlier phases.

- Audit logs: fidelity trends, drift alerts, red-team findings, and remediation actions.
- Human-factor reports: focus-group transcripts, A/B test statistics, and user-acceptance sign-offs.
- Compliance dossier: data-protection impact assessment, model card, explanation samples mapped to regulatory articles.
- Operational playbook: on-call rotation, retraining schedule, rollback triggers, and key performance indicators.

Once approved, the magnetic agent's explanation endpoints are activated in consumer portals, internal tools, or regulator-facing audit trails. Post-deployment, a quarterly review loop checks for concept drift, escalating to retraining or policy revision when thresholds are breached.

6. Discussion

The empirical and design insights above converge on a central theme: explainability is no longer a research luxury but an operational requirement that influences competitive advantage, regulatory posture, and societal trust. Deploying a magnetic agent transforms transparency from an expensive, one-off retrofit into a continuous service layer that scales with business growth [18]. This shift prompts decision makers to treat explainability as a cross-cutting capability, like security or observability, rather than a bolt-on feature. It carries strategic implications at three levels.

First, at the enterprise level, magnetic AI offers a cost–benefit inflection point. Faster compliance approvals, reduced litigation risk, and new value propositions, such as premium data-lineage services for high-stakes customers, offset the marginal expense of streaming surrogates and auditing dashboards. Firms adopting early may shape industry standards and lock in reputational capital that late movers struggle to match.

Second, at the ecosystem level, widespread passive-attachment architectures could generate large, anonymized corpora of model–surrogate disagreement events. These data could be shared under federated learning or secure multiparty protocols, catalyzing sector-wide benchmarks for robustness and enabling collaborative defense against adversarial attacks and systemic bias.

Third, granular yet comprehensible explanations at the societal level recalibrate the power balance between institutions and individuals. Users gain procedural recourse, auditors gain verifiable artifacts, and policymakers gain a practical blueprint for enforcement. The trade-off, however, is a thicker layer of governance overhead and an expanded attack surface that demands ongoing vigilance [19].

Against this backdrop, executive sponsors should treat magnetic AI deployment as a phased capability-maturity journey. Early milestones include establishing a data-tap inventory, codifying explanation-quality metrics, and funding interdisciplinary training programs so that engineers, risk officers, and product managers share a common vocabulary. Later stages focus on automating drift remediation, integrating feedback loops into agile release cycles, and participating in cross-industry consortia that set open standards for explanation fidelity and fairness. Organizations can navigate tightening regulations and rising public expectations by internalizing these priorities without sacrificing innovation velocity [20].

6.1. Prototype Model Demonstration

To illustrate the feasibility and behavior of the magnetic AI agent in a controlled environment, we implemented a toy model scenario. This lightweight empirical demonstration, while not intended as a comprehensive validation, serves to ground the concept in observable mechanics and provide an early proof of plausibility.

We used the classic Iris dataset and trained a black-box model using a random forest classifier. The magnetic agent was simulated as a proxy service that intercepted each input–output interaction and updated an online logistic regression model as its surrogate. The surrogate was constrained to observe only the request–response pairs, without access to feature importances, decision paths, or model internals.

Explanations were then generated by querying the lo-

gistic surrogate for each prediction and mapping the coefficients to ranked features. A fidelity audit compared surrogate predictions to the random forest decisions over a sliding window of 150 samples. Surrogate agreement stabilized at approximately 92%, and drift detection flagged one period where surrogate performance dropped due to a change in the class distribution, prompting automatic retraining.

Latency benchmarks were also recorded. On a commodity laptop (2.4 GHz, 8 GB RAM), average inference time per sample for the surrogate was under 3 milliseconds, including update and explanation rendering. This suggests that passive learning and auditing are feasible in near-real-time scenarios with moderate throughput. The latency–fidelity trade-off was observed to be tunable: larger sliding windows and ensemble surrogates marginally improved fidelity (up to 95%) but increased inference latency to 7–9 milliseconds per sample.

Input and output interfaces were defined as JSON over HTTP, simulating a REST-based production API. The surrogate processed flattened tabular features of fixed-length float vectors (4 dimensions for Iris), and the agent operated asynchronously in a sidecar thread. All components were implemented in Python using scikit-learn, Flask, and asyncio.

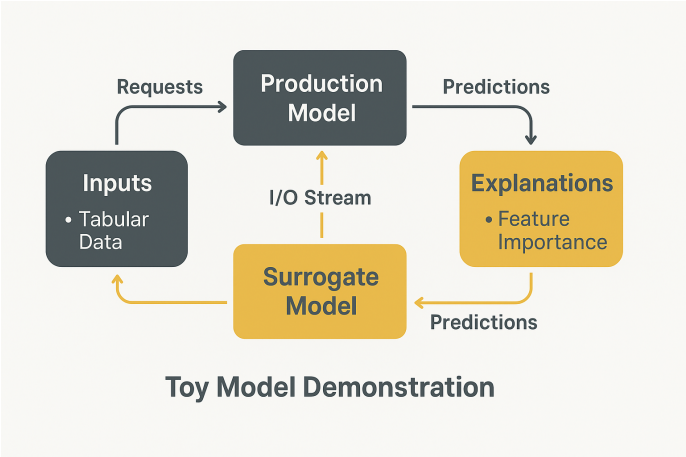

Figure 3: Toy Model Setup: The magnetic AI agent observes request–response pairs from a black-box random forest classifier trained on the Iris dataset. It trains a surrogate logistic regression model in real time, generates explanations, and audits fidelity in a sliding window.

6.2. Operational Considerations

Deploying a magnetic-AI layer replaces the usual pain of rewriting core models with the more manageable task of tapping live data streams. In companies that route traffic through Kafka, Kinesis, or a service-mesh sidecar, engineers can expose the request and response topics, spin up an agent container, and reach baseline fidelity in a morning.

By contrast, firms that still rely on tightly coupled middleware or batch ETL pipelines have to interpose a shim: a wrapper script that logs function calls or a lightweight message broker that mirrors production payloads without breaking the original code path. Once the tap is in place, the dominant cost moves from development time to compute cycles. Surrogate training scales almost linearly with input volume, so high-traffic applications—think personalized advertising or fraud detection at the millisecond level—can drive up cloud bills. Most teams blunt the cost curve by batching updates, down-sampling low-value events, or letting the agent burst to spot GPUs only during load spikes. Role clarity is essential to keep the system maintainable.

Data engineers own the interception code and service orchestration; data scientists tune the surrogate's learning rate, curate explanation templates, and validate fidelity thresholds; and compliance officers monitor the audit metrics, approve threshold changes, and archive drift reports for regulators. Without that three-way handshake, incremental tweaks in one area can silently break obligations in another, turning a retrofit to reduce risk into a new source of operational debt [21].

6.3. Ethical and Societal Dimensions

Agentic explainability shifts control from the system to the individual: a user can probe why their loan application was declined, inspect which pixels persuaded a vision model to flag an X-ray as malignant, or test what-if scenarios to see how a recommendation would change if inputs were different. This new transparency fosters autonomy and contestability and cracks open fresh attack surfaces.

Detailed feature-importance scores can reveal sensitive correlations that a company regards as trade secrets; if queried repeatedly, counterfactual examples let adversaries approximate the decision boundary and reconstruct private training data. To balance empowerment with protection, platform teams typically combine three defenses: rate-limiting caps the number of explanation calls per user or session, and throttling brute-force inversion attempts. Second, tiered access gates fine-grained explanation modes—local SHAP values, raw probability vectors, and full counterfactual paths—behind roles, entitlements, or paywalls, so casual consumers see only high-level summaries.

At the same time, regulators or auditors can request deeper details under non-disclosure constraints. Third, an adversarial-testing regime injects synthetic queries that mimic hostile behavior and flags the agent if leakage thresholds are exceeded.

Technical safeguards alone are insufficient because the audience's ability to parse explanatory artifacts is uneven. A compliance officer versed in statistics might understand the caveats of partial-dependence plots, whereas a consumer reading a heat map could misinterpret bright red pixels as causal rather than correlative. Organizations supplement the raw output with plain-language tooltips, short videos, or interactive walk-throughs that coach users on what the colors or numbers mean and, equally important, what they do not guarantee. Regulators are starting to codify such practices, requiring that explanations be available and comprehensible to a layperson in the decision context [22].

Lastly, equity audits need to extend beyond prediction fairness to explanation parity. A system may produce identical acceptance rates for two demographic groups, yet still describe its reasoning in more detailed or actionable ways for one group than the other. Auditors should measure the consistency of feature rankings, saliency intensities, and counterfactual suggestions across protected attributes. They should verify that any differences can be justified by legit-

imate factors rather than reflecting hidden bias. Without such checks, well-intentioned transparency can entrench inequities by giving some users a more straightforward path to recourse while leaving others in the dark.

7. Conclusions

This paper positions magnetic AI as a practical, scalable strategy for injecting explainability into the countless black-box models influencing credit decisions, hiring, medical triage, and other facets of economic and social life. Rather than requiring expensive retraining or code rewrites, the magnetic approach attaches passively to existing data flows, learns a lightweight surrogate in real time, and delivers multiple explanation formats that can satisfy data scientists, end users, auditors, and regulators alike. We first synthesize decades of research on interpretability, model compression, and drift detection to ground the proposal in established theory. We then distill that literature into concrete design principles: non-intrusiveness, continual fidelity auditing, modular deployment, and explanation pluralism tailored to stakeholder needs.

Building on these principles, we outline an evaluation blueprint that cuts across three dimensions. The technical track measures surrogate accuracy, latency overhead, and drift-detection sensitivity. The human track uses controlled studies and field pilots to gauge whether different user groups understand and act on the explanations. The regulatory track maps the agent's outputs to statutory requirements such as the EU AI Act's transparency duty, U.S. consumer protection guidelines, and industry standards like ISO 42001. By integrating these perspectives, the paper provides a holistic roadmap for retrofitting trustworthy AI capabilities into existing machine-learning stacks without disrupting production workflows. Ultimately, magnetic AI extends the idea of surrogate modeling from a one-off snapshot to a living, continuously audited companion, positioning organizations to meet emerging policy mandates and rising public expectations for transparency and accountability.

7.1. Limitations

The magnetic-AI framework is, at present, a theoretical blueprint. It has not yet been stress-tested on production traffic in banking, retail, health care, or public-sector settings, where data rates, latency budgets, and privacy constraints differ sharply. Field trials are needed to reveal whether the surrogate can keep pace with high-volume streams, whether passive interception introduces unacceptable delay, and which sectors face unique regulatory or contractual hurdles.

These deployments will also expose weak security points, such as opportunities for adversaries to infer proprietary decision logic or poison the surrogate's sliding-window buffer. In addition, the current design assumes a supervised task with stable labels—credit approval, fraud detection, or image classification—leaving open how a magnetic agent would operate in unsupervised anomaly detection, continuous exploratory reinforcement learning, or free-form generative applications where outputs are text, images, or code snippets rather than class scores. Each paradigm raises new questions about what counts as a faithful surrogate, how to define drift or fidelity, and which explanation formats are meaningful to users. Therefore, comprehensive empirical studies across these settings are essential before the approach can be considered production-ready.

7.2. Future Work

Future research must move the magnetic-AI concept from controlled prototypes into live production pipelines. Pilot deployments in banking, e-commerce, and telemedicine sectors would reveal practical limits on throughput, latency, and privacy while showing how easily the agent can be co-containerized, versioned, and rolled back under real traffic. Once embedded, the surrogate-learning engine should evolve from periodic mini-batch updates to accurate streaming operation, digesting continuous flows of tabular events, log sequences, sensor signals, and even raw audiovisual frames without halting for retraining. Handling these multimodal inputs will require hybrid learners that combine gradient-boosted trees for structured features, lightweight convolutional backbones for images, and adapter-based mini-transformers for text, all coordinated by a reservoir buffer that prioritizes the most recent or conceptually novel samples.

A second avenue involves deeper integration with large foundation models that have chain-of-thought capabilities. Instead of treating the surrogate purely as a predictive mimic, an agent could query a frozen language model for self-rationalizing traces, then cross-check those traces against feature-importance scores to generate richer, more coherent explanations. This hybrid could also let users ask follow-up questions in natural language—Why did age matter more than income?—and receive conversational clarifications grounded in statistical evidence and domain policy.

Finally, the community needs shared benchmarks that evaluate explanation quality across domains rather than in narrow, single-task silos. A standard suite might pair representative workloads—credit risk, dermatology imaging, autonomous-vehicle perception—with crowdsourced judgment tests, cognitive-load surveys, and perturbation-based robustness checks. Metrics would cover fidelity, sparsity, stability under re-queries, resistance to inversion attacks, and user comprehension measured through decision-making tasks. Establishing such benchmarks would allow researchers to compare methods rigorously, accelerate regulatory acceptance, and guide practitioners toward solutions whose benefits generalize beyond any industry.

References

[1] M. Leon, "Ai-driven digital transformation: Challenges and opportunities", *Journal of Engineering Research and Sciences*, vol. 4, no. 4, p. 8–19, 2025, doi:10.55708/js0404002.

[2] M. Leon, "Generative artificial intelligence and prompt engineering: A comprehensive guide to models, methods, and best practices", *Advances in Science, Technology and Engineering Systems Journal*, vol. 10, no. 02, p. 01–11, 2025, doi:10.25046/aj100201.

[3] M. Leon, B. Depaire, K. Vanhoof, "Fuzzy cognitive maps with rough concepts", "Artificial Intelligence Applications and Innovations: 9th IFIP WG 12.5 International Conference, AIAI 2013, Paphos, Cyprus, September 30–October 2, 2013, Proceedings 9", pp. 527–536, Springer Berlin Heidelberg, 2013.

[4] H. DeSimone, M. Leon, "Leveraging explainable ai in business and further", "2024 IEEE Opportunity Research Scholars Symposium (ORSS)", p. 1–6, IEEE, 2024, doi:10.1109/orss62274.2024.10697961.

[5] M. Leon, H. DeSimone, "Advancements in explainable artificial intelligence for enhanced transparency and interpretability across business applications", *Advances in Science, Technology and Engineering Systems Journal*, vol. 9, no. 5, p. 9–20, 2024, doi:10.25046/aj090502.

[6] M. Velmurugan, C. Ouyang, R. Sindhgatta, C. Moreira, "Through the looking glass: evaluating post hoc explanations using transparent models", *International Journal of Data Science and Analytics*, 2023, doi:10.1007/s41060-023-00445-1.

[7] S. Jesus, C. Belém, V. Balayan, J. Bento, P. Saleiro, P. Bizarro, J. Gama, "How can i choose an explainer?: An application-grounded evaluation of post-hoc explanations", "Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency", FAccT '21, p. 805–815, ACM, 2021, doi:10.1145/3442188.3445941.

[8] M. Leon, N. M. Sanchez, Z. G. Valdivia, R. B. Perez, "Concept maps combined with case-based reasoning in order to elaborate intelligent teaching/learning systems", "Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)", pp. 205–210, IEEE, 2007.

[9] H. DeSimone, M. Leon, "Explainable ai: The quest for transparency in business and beyond", "2024 7th International Conference on Information and Computer Technologies (ICICT)", p. 532–538, IEEE, 2024, doi:10.1109/icict62343.2024.00093.

[10] M. Leon, G. Nápoles, M. M. García, R. Bello, K. Vanhoof, "Two steps individuals travel behavior modeling through fuzzy cognitive maps pre-definition and learning", "Advances in Soft Computing: 10th Mexican International Conference on Artificial Intelligence, MICAI 2011, Puebla, Mexico, November 26-December 4, 2011, Proceedings, Part II 10", pp. 82–94, Springer Berlin Heidelberg, 2011.

[11] G. Nápoles, F. Hoitsma, A. Knoben, A. Jastrzebska, M. Leon, "Prolog-based agnostic explanation module for structured pattern classification", *Information Sciences*, vol. 622, p. 1196–1227, 2023, doi:10.1016/j.ins.2022.12.012.

[12] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, A. Hussain, "Interpreting black-box models: A review on explainable artificial intelligence", *Cognitive Computation*, vol. 16, no. 1, p. 45–74, 2023, doi:10.1007/s12559-023-10179-8.

[13] M. Leon, "Gail: Enhancing student engagement and productivity", *The International FLAIRS Conference Proceedings*, vol. 38, 2025, doi:10.32473/flairs.38.1.138689.

[14] S. Bordt, M. Finck, E. Raidl, U. von Luxburg, "Post-hoc explanations fail to achieve their purpose in adversarial contexts", "2022 ACM Conference on Fairness Accountability and Transparency", FAccT '22, p. 891–905, ACM, 2022, doi:10.1145/3531146.3533153.

[15] W. J. von Eschenbach, "Transparency and the black box problem: Why we do not trust ai", *Philosophy & Technology*, vol. 34, no. 4, p. 1607–1622, 2021, doi:10.1007/s13347-021-00477-0.

[16] S. Hosseini, H. Seilani, "The role of agentic ai in shaping a smart future: A systematic review", *Array*, vol. 26, p. 100399, 2025, doi:10.1016/j.array.2025.100399.

[17] D. B. Acharya, K. Kuppan, B. Divya, "Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey", *IEEE Access*, vol. 13, pp. 18912–18936, 2025, doi:10.1109/ACCESS.2025.3532853.

[18] N. Karunanayake, "Next-generation agentic ai for transforming healthcare", *Informatics and Health*, vol. 2, no. 2, p. 73–83, 2025, doi:10.1016/j.infoh.2025.03.001.

[19] M. Leon, "The escalating ai's energy demands and the imperative need for sustainable solutions", *WSEAS Transactions on Systems*, vol. 23, pp. 444–457, 2024.

[20] U. Ehsan, P. Wintersberger, Q. V. Liao, E. A. Watkins, C. Manger, H. Daumé III, A. Riener, M. O. Riedl, "Human-centered explainable ai (hcxai): Beyond opening the black-box of ai", "Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems", CHI EA '22, Association for Computing Machinery, New York, NY, USA, 2022, doi:10.1145/3491101.3503727.

[21] S. Nyawa, C. Gnekpe, D. Tchuente, "Transparent machine learning models for predicting decisions to undertake energy retrofits in residential buildings", *Annals of Operations Research*, 2023, doi:10.1007/s10479-023-05217-5.

[22] J. Mökander, "Auditing of ai: Legal, ethical and technical approaches", *Digital Society*, vol. 2, no. 3, 2023, doi:10.1007/s44206-023-00074-y.

Biography

Dr. Maikel Leon is interested in Artificial Intelligence (AI), Generative AI (GenAI), and Machine Learning (ML). His work bridges intelligent systems' theoretical foundations and practical applications, particularly emphasizing explainability, hybrid models, and educational innovation. Dr. Leon has published in high-impact journals such as IEEE Transactions on Cybernetics, Information Sciences, Knowledge and Information Systems, International Journal on Artificial Intelligence Tools, Intelligent Decision Technologies, and International Journal of Learning, Teaching and Educational Research. His research explores cutting-edge topics, including prompt engineering, sustainable AI, personalized tutoring via generative models, hybrid fuzzy systems, and large language model benchmarking. He was awarded the Cuban Academy of Sciences National Award for the Most Relevant Research in Computer Science. Dr. Leon obtained his PhD in Computer Science at Hasselt University (Belgium), an MSc and a BSc in Computation from Central University of Las Villas (Cuba), and currently serves as Associate Professor of Practice in Business Technology at the Miami Herbert Business School, University of Miami (Florida, USA).