PUBLIC SUBMISSION

Received: May 12, 2025
Tracking No. mam-tv4z-v3vx
Comments Due: May 28, 2025 **Submission Type:** Web

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0074
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Sustainable Future Tech Inc

General Comment

See attached comments

Attachments

Response to the RFI on the 2025 National AI R and D Strategic Plan-Willis J.M.-20250513

Response to the 2025 National AI R&D Strategic Plan RFI (90 FR 17835)

Embedding Interpretability, Risk Governance, and Economic Precision into the Foundations of U.S. AI Leadership

Submitted by:
John M. Willis
Founder & Chief Innovation Officer
Sustainable Future Tech Inc
Gaithersburg, MD USA
May 13, 2025
Public record of this response: DOI:10.5281/zenodo.15398521

About the Author
John M. Willis is the Founder and Chief Innovation Officer of Sustainable Future Tech Inc, where he leads development of AI, cybersecurity, and quantum-aligned technologies with a focus on lifecycle risk governance and national resilience. This response draws upon original internal innovation frameworks including:

- **QIXAI** – *Quantum-Inspired Explainability Framework*
  arXiv:2410.16537 – Entropy-based transparency methods for classical and quantum neural networks.
- **QILIS** – *Quantum-Inspired Lifecycle Interpretability System*
  Patent pending (filed Dec 2024) – An architecture for embedding feature relevance, entropy telemetry, and semantic drift detection across the AI lifecycle.

---

1. Executive Summary

This response is submitted in reply to the April 29, 2025 **Request for Information on the National Artificial Intelligence (AI) Research and Development Strategic Plan** [Federal Register Document 2025–07332]. It addresses multiple priorities called for in the RFI, including foundational research, evaluation testbeds, trustworthy AI, lifecycle governance, and national security-aligned AI architecture.

The United States must maintain sovereign control over critical AI infrastructure, models, and inference pipelines—ensuring defense autonomy, export control integrity, and supply chain independence.

Strategic AI funding should support not only R&D but the development of a secure, U.S.-based AI workforce and compute infrastructure—ensuring domestic control of both innovation and execution.

This response proposes that the revised National AI R&D Strategic Plan elevate three technical pillars essential to the economic, regulatory, and operational success of U.S. AI systems:

1. **Embedded interpretability** that is architecturally encoded—not applied post hoc.

2. **Lifecycle-structured governance**, including risk visibility from development through decommissioning.
3. **Decision assurance architectures**, enabling trust, auditability, and safe delegation in mission-critical domains.

We argue that these are not just safeguards—they are **financial and strategic enablers**. AI systems that emit structured inference telemetry (e.g., entropy, attribution, confidence signals) can reduce capital costs, accelerate oversight, and unlock profitable automation. Opaque systems, by contrast, increase regulatory exposure, operational error, and rework cost.

This response draws upon internal development work (e.g., QILIS), lifecycle risk modeling frameworks, and applied research in cybersecurity, national defense, and socially embedded systems. While no proprietary materials are submitted here, the analysis reflects practical implementation insight.

2.1 Foundational Research for Interpretable and Trustworthy AI (per RFI Priority #1)

AI systems deployed in high-stakes sectors must be designed to explain, trace, and justify their behavior—mathematically and semantically. Post-hoc interpretability tools, while helpful, cannot provide real-time assurance, nor can they influence training dynamics. We recommend shifting federal research toward embedded interpretability and simplification-aware model architectures.

A. Precision Learning through Embedded Interpretability

Invest in architectures that track **feature relevance, uncertainty, and semantic signal flow** throughout the learning process. Approaches such as entropy-normalized attribution, cosine-aligned relevance scoring, and phase coherence analysis offer quantifiable metrics for precision.

This design paradigm:

- Supports trustable automation by flagging ambiguity in real-time.
- Enables low-cost governance by aligning training outputs with audit requirements.
- Reduces downstream corrective costs through better alignment with operational context.

Quantum-inspired models—such as those being explored internally—show how hybrid systems can embed transparency into both classical and non-classical components.

B. Oversimplification as Structural and Financial Risk

Common model compression and preprocessing techniques often erase variance that reflects **policy-relevant, risk-bearing, or demographically unique** signals. These include:

- Imputation that masks income irregularity in gig workers
- Aggregation that reduces diagnostic specificity in healthcare
- Outlier removal that deletes emerging fraud patterns in financial systems

These practices introduce measurable financial error, increase exclusion, and weaken model generalizability. Instead, support methods that:

- Preserve **causal and semantic structure**
- Offer **fairness-aware dimensionality reduction**
- Include **relevance decay scoring** during compression

Summary Recommendations

- Fund foundational research into **metric-driven interpretability** (entropy, attribution, alignment).
- Avoid reductive model assumptions; support **structure-preserving learning**.
- Develop hybrid systems that embed transparency for both classical and quantum models.
- Treat interpretability as a **precision optimization**, not a reporting add-on.

Safety and interpretability must be embedded in the architecture—not to slow innovation, but to enable faster deployment by eliminating post-hoc validation cycles, oversight bottlenecks, and public trust barriers.

2.2 Evaluation Infrastructure: Testbeds and Synthetic Data (per RFI Priority #2)

To evaluate AI systems intended for real-world impact, testbeds must move beyond measuring accuracy. They must assess semantic stability, inference traceability, and economic risk alignment. Synthetic data environments should be designed not only to simulate data—but to preserve explanatory fidelity and stress-test interpretability.

A. Modernizing Testbeds for Trust and Risk Awareness

We recommend investment in testbeds that:

- Evaluate **inference confidence** and feature attribution under distribution shift
- Include metrics for **semantic drift** and **relevance degradation**
- Quantify **cost of misclassification**, especially in high-stakes domains

These capabilities help assess model robustness in dynamic settings like healthcare triage, financial lending, or mission logistics—where **failure modes incur tangible economic or reputational cost**.

QILIS-like frameworks demonstrate how feature-level telemetry can simulate cost exposure by mapping attribution vectors to domain-specific outcome metrics.

B. Designing Synthetic Data for Interpretability Preservation

Synthetic datasets often fail to preserve meaningful structure, making them unreliable for training interpretable or auditable models. We recommend:

- Embedding **relevance tags** or **salience annotations** during data generation
- Supporting synthetic benchmarks that test **causal continuity** and **fairness resilience**
- Auditing synthetic data generators for **semantic distortion** and **feature masking bias**

This ensures that models trained on synthetic data remain **explainable in operational use**—especially in sectors with regulatory or ethical scrutiny.

C. Deployment-Aware and Edge-Compatible Evaluation

Testbeds should also simulate **low-resource environments** common in edge computing (e.g., mobile health, field surveillance, IoT security). We encourage:

- Evaluation of **pruned or quantized models** for interpretability preservation
- Testing of **real-time telemetry extraction** under constrained conditions
- Benchmarking of **drift detection latency** for lightweight monitoring systems

B. Deployment-Aware and Pilot-Ready Evaluation

Testbeds should assess not only semantic fidelity and drift detection but also the viability of telemetry-rich architectures. We encourage federal evaluation of **QILIS-style models**, which emit entropy scores, attribution vectors, and relevance decay logs to support real-time traceability.

QILIS is currently in prototype development, with classical components available for integration. Pilot use cases may include public health triage, financial risk scoring, and procurement optimization—domains where inference transparency reduces oversight cost and operational risk.

> *This is proposed as a technical contribution to support public testbed evaluation, not as a commercial offering. This offer is made solely to support public-sector technical evaluation. No rights to source code, IP, or commercial use are granted or implied. Any formal collaboration would be subject to separate agreement.*

Summary Recommendations

- Build testbeds that evaluate **semantic relevance, cost-aware misclassification, and traceable inference**.
- Fund SDG architectures that embed **interpretable structure**, not just data realism.
- Extend testbed evaluation to **resource-constrained and dynamic deployment settings**.
- Treat testbed interpretability as a **deployment-critical property**, not a research luxury.

2.3 Robust and Assured Decision-Making Systems (per RFI Priority #3)

In high-stakes domains—such as finance, cybersecurity, healthcare, and defense—**accuracy alone is insufficient**. AI systems must be engineered for **decision assurance**: the ability to explain, quantify, and justify inferences in context. This capability reduces downstream review costs, improves strategic reliability, and unlocks safe automation.

A. Embedded Decision Telemetry Enables Shared Authority

We recommend federal support for AI architectures that:

- Emit **inference-level telemetry**: entropy scores, attribution maps, and abstention thresholds
- Allow real-time judgment on **when to delegate**, **when to escalate**, and **when to reject**
- Store **version-aware behavior logs** for audit and model governance

Such systems allow confident delegation without full automation—essential for domains like **fraud prevention**, **clinical triage**, and **logistics optimization**. QILIS-style pipelines demonstrate how structured outputs can flag degraded signal states, prompting fallback or human review.

B. National Security: Cyber Defense and Autonomous Engagement Assurance

Three critical defense domains require embedded AI trust:

1. Cybersecurity systems must:

- Distinguish meaningful anomalies from false positives
- Explain triggers to SOC teams and support traceable escalation
- Integrate with Zero Trust policies and forensic response workflows

2. Targeting systems must:

- Quantify signal ambiguity (e.g., occlusion, sensor fog, adversarial camouflage)
- Justify decisions under time constraint and operational uncertainty
- Prevent accidental escalation or friendly fire through abstention and rationale signaling

3. Autonomous defense systems, including missile defense and active threat interception, must:

- Emit real-time confidence levels before engagement
- Detect ambiguous, spoofed, or degraded inputs
- Allow for abstention, escalation, or override pathways when signal integrity is uncertain
- Be resilient to adversarial AI, including spoofed signals, sensor deception, and adversarial perturbations.

*In all three cases, opaque models increase strategic fragility. Trustworthy systems must signal actionable confidence **prior to engagement**, not attempt justification **after mission risk is realized**—particularly in life-critical or geostrategic contexts.*

Summary Recommendations

- Prioritize architectures with **structured, real-time inference telemetry** and audit-friendly outputs.
- Require decision assurance capabilities for AI used in **federal, financial, and security-sensitive deployments**.

- Fund confidence-calibrated architectures that **support shared authority**, rather than assume full automation.
- Treat decision transparency as a **precondition for scalable AI trust**, especially in systems affecting public safety or national operations.

Missile defense and cyber deterrence systems must be hardened against adversarial AI, signal spoofing, and data poisoning—requiring telemetry-rich models capable of autonomous verification under threat.

2.4 Lifecycle Risk Management and Governance (per RFI Priority #4 and #5)

AI risk cannot be managed solely through external oversight—it must be measured, surfaced, and aligned from within the model lifecycle. While the NIST AI Risk Management Framework provides essential structure, it lacks phase-specific operationalization and does not address the technical telemetry required to embed governance into model design.

We recommend governance architectures that minimize regulatory friction by automating risk telemetry—reducing the need for manual audits, compliance delays, or government-imposed retraining.

A. Lifecycle-Aware Governance: From Objectives to Archival

We recommend extending the AI RMF by adopting a lifecycle-stage risk taxonomy, with embedded technical indicators. Key phases and risks include:

- **Objective misalignment**: models trained to optimize throughput may conflict with fairness or compliance goals.
- **Data preprocessing risk**: imputation, aggregation, and filtering can hide demographic or causal variance.
- **Deployment and drift risk**: absence of semantic change detection increases error and oversight cost.
- **Decommissioning risk**: legacy models lacking version traceability pose legal, security, and reproducibility liabilities.

AI architectures should emit governance artifacts natively—such as **relevance logs, entropy flags, and behavioral changelogs**—supporting model continuity across updates, teams, and jurisdictions.

B. Embedded Governance Reduces Oversight Cost

Manual audits, retrofitted explainability, and third-party compliance reviews add significant delay and expense. Systems with **built-in telemetry** enable:

- **Fast audit response** via inference tracebacks
- **Continuous risk scoring** tied to operational use
- **Reusable models** with lifecycle-aligned metadata and policy linkage

As implemented in QILIS, inference logs can map relevance decay and entropy changes across retraining cycles—allowing automated version audits and forensic replay.

Governance should be **a design property**, not a post-deployment patch.

Summary Recommendations

- Extend the NIST AI RMF with **phase-aligned risk indicators** and embedded technical governance standards.
- Fund model architectures that produce **traceable, auditable inference data** throughout the lifecycle.
- Require governance capabilities for federally deployed or regulated AI systems, especially in critical sectors.
- Treat embedded governance as **infrastructure for compliance, accountability, and long-term cost control**.

We recommend governance architectures that minimize regulatory friction by automating risk telemetry—reducing the need for manual audits, compliance delays, or government-imposed retraining.

2.5 Economic Drivers and Financial Relevance of Interpretable AI (per RFI Priority #7)

Interpretability is not just a trust feature—it is a **cost optimization and risk management capability**. When built into AI models from the ground up, it enables **more precise automation**, **faster oversight**, and **lower financial exposure** across domains.

Opaque models force organizations to:

- Maintain **high reserve capital** to offset model risk (e.g., finance)
- Rely on **human-in-the-loop review**, slowing throughput (e.g., logistics, defense)
- Accept **regulatory audit delays**, increasing compliance cost

In contrast, interpretable systems:

- Support **confidence-weighted delegation**, optimizing human-machine collaboration
- Enable **automated dispute resolution** via attribution logs and explanation trails
- Reduce **retraining and override frequency** by exposing causal fragility before deployment

---

Use Case Highlights

- In **finance**, relevance scoring enables cost-aligned loan approvals, improves fairness, and minimizes override.
- In **procurement**, explainable forecasts reduce order variance and overstocking penalties.
- In **public health**, feature transparency improves diagnostic trust and reduces malpractice risk.

QILIS-like frameworks link inference telemetry directly to decision impact—helping institutions **quantify return on model performance, not just prediction accuracy**.

Summary Recommendations

- Frame interpretability as a **capital efficiency lever**, not just a governance requirement.
- Fund models that output **telemetry for confidence, cost estimation, and decision traceability**.
- Encourage architectures that **reduce compliance overhead** while increasing automation confidence.
- Align interpretability R&D with **financial outcomes**—including error avoidance, oversight speed, and resource allocation precision.

2.6 Equity, Context Sensitivity, and Socio-Structural Risk (per RFI Priority #8)

AI systems that rely on **overly reductive preprocessing and compression techniques** risk excluding the very complexity that makes real-world decisions robust and fair. Practices such as mean imputation, outlier suppression, or excessive dimensionality reduction can eliminate **social, causal, and economic nuance**—resulting in misclassification, error amplification, and public backlash.

Oversimplification obscures critical edge-case behavior and reduces system resilience—creating hidden operational and financial liabilities..

A. Fragility in Policy- and Finance-Relevant Contexts

Simplified models frequently misclassify:

- Non-salaried workers in credit scoring
- Minority subgroups in clinical decision-making
- Small or local suppliers in procurement prioritization

These misclassifications result in:

- **Exclusion of high-variance populations**
- **False policy compliance**
- **Strategic fragility and litigation risk**

Such models appear performant but fail under stress—creating hidden financial and reputational costs.

B. Funding Context-Aware Alternatives

Federal R&D should support:

- Model compression techniques that preserve decision-relevant variance across diverse user contexts and operating conditions
- Relevance-preserving imputation and pruning, guided by entropy thresholds to avoid masking edge-case signals
- Benchmarking approaches that stress-test model performance on low-frequency events, rare categories, and long-tail inputs

QILIS-style telemetry helps detect when **simplification suppresses risk-sensitive signals**, allowing correction before models are deployed at scale.

Summary Recommendations

- Recognize oversimplification as a **strategic and financial risk**, not just a modeling shortcut.
- Fund methods that preserve **semantic and causal complexity** essential to trust and generalization.
- Align modeling pipelines with **social and economic diversity**, reducing hidden failure and exclusion costs.
- Treat contextual fidelity as a **core requirement** for AI in public-facing and regulated domains.

2.7 Model Retirement, Traceability, and Lifecycle Closure (per RFI Priority #9)

AI governance cannot end at deployment. The **retirement and archival phase** is an often-overlooked source of institutional vulnerability. Models that lack version control, traceable behavior histories, or formal deactivation protocols expose organizations to **legal, operational, and reputational risk**.

Dormant models may persist in:

- Embedded systems or local environments
- Legacy workflows without centralized tracking
- Shadow AI usage by third-party teams or integrators

Without verifiable retirement, these systems may continue to influence decisions, produce untraceable outputs, or trigger liability years after their design intent has expired.

A. Lifecycle Closure as a Risk Control

We recommend support for:

- **Standardized model retirement documentation**, capturing usage context, behavioral history, and rationale for deactivation
- **Archival formats** that preserve attribution metadata, training context, and inference logs
- **Retirement verification tooling**, confirming removal from production workflows and access endpoints

These protocols are especially critical in **federally deployed**, **regulated**, or **security-sensitive systems**, where untracked model behavior could have mission or compliance consequences.

Without lifecycle closure, there is no audit continuity. Without explainability artifacts, there is no retrospective defense.

Summary Recommendations

- Establish decommissioning and archival as a **formal phase in the AI lifecycle**.
- Fund tools and formats for **traceable model shutdown, storage, and policy handoff**.
- Require lifecycle closure documentation for AI systems in **public, regulated, or high-consequence domains**.
- Treat retirement risk as **financially material** and **legally consequential**, not an afterthought.

---

3. Conclusion and Closing Remarks

To fulfill the priorities set forth in the 2025 RFI and strengthen U.S. competitiveness, national security, and responsible innovation, the AI research ecosystem must support systems that are architecturally transparent, economically scalable, and governable by design.

Systems that explain themselves in real time—quantitatively and contextually—enable:

- **Cost-effective oversight**
- **Faster compliance response**
- **Confidence-weighted automation**
- **Stronger alignment with policy and public trust**

By contrast, opaque or oversimplified systems introduce:

- Higher regulatory and audit costs
- Increased model risk capital
- More frequent override and retraining cycles
- Delayed or failed deployments in sensitive sectors

This response has outlined priority areas where federally supported AI R&D can reduce these costs and improve deployment confidence:

- Embedding interpretability during training—not bolting it on afterward
- Building governance telemetry into the architecture—not managing it externally
- Supporting decision traceability from inception to decommissioning—not inferring rationale post hoc

These capabilities are not simply desirable—they are **economic enablers**, **compliance accelerators**, and **risk suppressors**.

While the recommendations herein focus on deployable near-term systems, the principles of embedded interpretability and lifecycle control are also foundational safeguards against potential future risks from more advanced, general-purpose AI systems.

We encourage OSTP and the National AI Initiative Office to incorporate these priorities into the 2025 Strategic Plan, and we welcome the opportunity to support the operationalization of these principles through technical input, policy alignment, or collaborative implementation.