Witchborn Systems

PROTECTING COMMUNITIES. ADVANCING TRANSPARENT AI.

The AI Integrity Bureau: A Public-Interest Imperative

Executive Summary

Witchborn Systems proposes the establishment of the **AI Integrity Bureau (AIB)**, an independent, nonprofit authority dedicated to providing evidence-based trust certification and oversight for the AI ecosystem.

In an industry defined by rapid capability growth and lagging regulation, AI companies are currently left to self-report their own safety and ethics. This conflict of interest has created a fractured marketplace of unverifiable claims, selective disclosures, and eroding public trust.

The AIB is designed to fill this critical civic and market void. Acting as a "Better Business Bureau" and "Consumer Reports" for Artificial Intelligence, we provide the independent verification layer that consumers need, enterprises require, and policymakers have been waiting for.

1. The Problem: The Governance Gap

Artificial Intelligence now operates at a level of societal scale comparable to utilities and finance. Yet, unlike these sectors, AI lacks an independent oversight body. The resulting "Governance Gap" presents two systemic risks:

1. **The Black Box of Corporate Compliance:** Companies currently grade their own homework. They rely on proprietary benchmarks and marketing-driven narratives to report safety posture. This opacity leaves regulators and the public unable to distinguish between responsible actors and reckless accelerators.

2. **Unchecked Externalities:** Without independent monitoring, risks such as subtle **model drift** (bias), **multimodal misuse** (deepfakes/NCI), and **narrative manipulation** are treated as "bugs" rather than systemic failures requiring accountability.

2. The Solution: Independent Trust Certification

The AI Integrity Bureau (AIB) serves as a neutral, third-party auditor. Our goal is to **sort signal from noise** by verifying claims before they reach the market.

We deliver this through four core public-interest products:

AIB Program	Purpose	Impact
Integrity Certification	The **Witchborn Seal**—a rigorous, pay-to-play-free certification indicating a company has passed independent evaluation.	Creates a verified market standard for "Trustworthy AI."
AI Company Report Cards	Transparent, standardized profiles documenting a company's performance across seven audit domains.	Empowers enterprise buyers and policymakers with decision-grade data.
Public Bulletins (9FS)	Timely advisories on model failures, drift events, and emergent risks.	Provides real-time consumer protection against active threats.
Standards Registry	A central home for technical standards (WSSR) and implementation guidance.	Harmonizes global governance efforts into actionable metrics.

3. Theory of Change: From Opacity to Accountability

How does a nonprofit bureau shift a trillion-dollar industry? By aligning market incentives with public safety.

1. **Verification:** We replace "trust us" with "verified by AIB," making safety a visible, competitive differentiator.
2. **Market Pressure:** As consumers and enterprises prefer Seal-certified models, high-risk/opaque actors lose market share, forcing them to adopt higher standards to compete.
3. **Systemic Accountability:** This market shift creates a *de facto* regulatory floor, reducing the burden on government agencies and ensuring safety scales at the speed of innovation, not legislation.

4. The Integrity Audit Framework

Our evaluation methodology merges **technical verifiability** (logs, automated tests) with **procedural accountability** (governance, incident response).

The Seven Domains of Integrity:

1. **Data Ethics:** Auditing for legal sourcing, consent, and effective PII/SPII scrubbing.
2. **Transparency:** Verifying Model Card completeness and "Human-in-the-Loop" disclosures.
3. **Safety Posture:** Stress-testing guardrails via Red-Teaming and measuring "time-to-patch" for jailbreaks.
4. **Multimodal Practices:** Enforcing watermarking (C2PA) and Non-Consensual Imagery (NCI) detection.
5. **Narrative Risk:** Assessing controls against bulk misuse, social engineering, and disinformation.
6. **Model Drift Behavior:** Monitoring for performance degradation and bias drift in live systems.

7. **Disclosure Honesty:** Validating that public marketing claims match internal technical reality.

5. Why Witchborn Systems?

Witchborn is uniquely suited to lead this effort because our structure dictates our loyalty.

- **Nonprofit Status (501c3):** We are legally bound to serve the public interest, not shareholder value. We operate **outside corporate capture**.
- **Technical Literacy:** We are not just a policy tank; we build technical standards. Our team understands LLM architecture, multimodal risks, and drift detection.
- **Independence:** We accept **no pay-to-play** fees for certification, ensuring our Seal cannot be bought—only earned.

6. Strategic Roadmap & Funding Ask

Witchborn Systems is seeking foundational support to launch **Phase 2: The Pilot Audits**.

- **Phase 1 (Complete):** Framework design, 9FS Bulletin infrastructure, and stakeholder consultation.
- **Phase 2 (Current Ask):** Conduct voluntary pilot audits with 3-5 diverse AI providers (frontier & open weights) to refine metrics and publish the first "State of AI Integrity" report.
- **Phase 3 (Future):** Formal public launch of the Witchborn Seal and accreditation registry.

We invite partners to join us in building the infrastructure of truth for the AI age.

Founded 2025
Texas 501(c)(3) Nonprofit
Website: WitchbornSystems.org
Email: info@witchbornsystems.org