Cross-Border Agentic AI Compliance (CBAAC): Embedding Regulatory and Cultural Risk Compliance into Agentic Communication

Matt Pagett
Independent Researcher
https://www.mattpagett.dev

Tomoko Mitsuoka
Independent Researcher
https://www.linkedin.com/in/tomoko-mitsuoka/

With
Apart Research

Demo: https://cross-border-agentic-compliance.solve.it.com/

Code: https://github.com/mattpagett/CBAAC/

Abstract

AI regulations (EU AI Act, Japan METI, Korea AI Basic Act) require providers to certify compliance — but how can businesses verify that the agents they use, and sub-agents in the chain, actually comply? Current approaches rely on costly audits, additional external agreements, or trust — and do not scale well to a world of millions of agents which can spawn on demand.

We propose demand-side verification: a protocol enabling agents to automatically check compliance of other agents before sharing data. Our framework supports regulatory compliance (GDPR, AI Act, GPAI) and optional cultural/ethical benchmarks addressing behavioral risks such as unconsented user profiling and emotional manipulation. We extend Project NANDA's AgentFacts schema with self-certification questionnaires for EU, Japan, and Korea jurisdictions, plus cultural competency assessments addressing behavioral risks documented in AI governance failures. We provide an open implementation and demo at [https://cross-border-agentic-compliance.solve.it.com/].

Keywords: Multi-agent alignment, AI security, compliance infrastructure

1. Introduction

AI agents increasingly operate in chains — an HR agent calls a travel agent, which calls an airline agent, which may call a customer support agent. Each may be powered by different models, operated by different companies, in different jurisdictions.

This creates two compliance challenges:

Regulatory compliance. The EU AI Act, Japan's METI guidelines, and Korea's AI Basic Act each impose requirements on AI systems. When agents share personal data across borders, businesses must verify the entire chain complies — not just their immediate vendor.

Cultural and ethical compliance. Beyond legal requirements, agents may exhibit behaviors that create reputational or litigation risk: inferring user gender without consent, emotional manipulation, or culturally inappropriate communication (Mitsuoka 2026). These issues are not necessarily covered by regulation but matter to businesses.

Third-party audits address both concerns but cost $10,000+ and do not scale to every agent interaction. We propose a lightweight, machine-readable mechanism for demand-side verification — enabling businesses to automatically verify compliance before sharing data.

Our contribution. This paper presents a protocol for embedding regulatory and cultural compliance verification into agent-to-agent communication. We provide:

1. A **schema extension** (in this case, specifically for Project NANDA's AgentFacts, but could be used for other agent communication protocols) that includes compliance attestations, sub-agent declarations, and model provider certification

2. **Self-certification questionnaires** for EU (GDPR + AI Act), Japan (APPI + METI), Korea (AI Basic Act + PIPA), and cultural/ethical standards

3. A **verification mechanism** enabling agents to check compliance of other agents — and their sub-agents — before data exchange

4. A **demonstration** at:

https://cross-border-agentic-compliance.solve.it.com/

Importantly, our proposed solution creates **demand-pull for verification:** if Company A's policy requires EU compliance, and Agent B cannot provide attestation, the transaction is blocked. Agent B's provider now has market incentive to self-certify — without requiring a regulator to enforce compliance at every interaction. This chain effect also provides an important incentive for standards regarding high-risk AI (GPAI safety standards) to propagate.

This protocol addresses not only legal compliance but also behavioral and cultural risks documented in recent AI incidents, such as those in Japan including

unconsented user profiling and emotional manipulation (Mitsuoka 2026). To our knowledge, this is the first framework to provide machine-readable, testable criteria for cultural competency and behavioral safeguards in AI agent systems.

A Path for Governance Approaches Adapted to More Cultural Contexts

Current AI governance approaches can be framed largely within two dominant paradigms: a precautionary approach (risk-based regulation) and market-driven approaches (minimal regulation, industry self-governance). Both frameworks share a common assumption: that AI governance is primarily a technical compliance challenge addressable through transparency, explainability, and regulation. However, this technology-centric approach has structural limitations. These frameworks prioritize individual accountability while remaining fundamentally unable to resolve collective, human-rooted biases embedded in training data.

Our framework enables a 'third path' beyond EU-style precautionary regulation and US-style market self-governance: cooperative governance that reconceptualizes AI safety as a socio-technical challenge. Where compliance-focused frameworks treat AI as a controllable tool requiring oversight, cooperative frameworks treat AI as a change agent requiring cultural alignment. This shift—from technical compliance to socio-technical cooperation—enables multipolar coordination while preserving cultural sovereignty. Rather than imposing universal standards that risk alienating diverse cultural contexts, our framework enables businesses to specify culturally-grounded requirements while participating in global AI ecosystems. An agent serving Japanese healthcare contexts can attest to cultural competency (honorifics usage, family information-sharing norms) alongside regulatory compliance—preserving cultural sovereignty while enabling international cooperation.

Japan's Cooperative Sovereignty Model: Japan's approach to AI governance exemplifies the cooperative paradigm. The "data-in, model-out" strategy—where global AI developers can leverage culturally rich Japanese data while raw data remains within national borders—challenges the Western dichotomy of strict data protectionism versus completely open borders. This cooperative approach offers significant benefits. Rather than forcing convergence toward an English-based algorithmic standard, it enables diverse cultural AI systems to interoperate while preserving local values. A business using our framework could require that agents processing Japanese customer data demonstrate both APPI compliance AND cultural alignment attestation—ensuring technical legality while preserving the socio-technical unity that cooperative governance demands.

2. Methods

Our approach addresses the **compliance supply chain problem**: when an agent relies on sub-agents, the business employing the first agent must verify that the entire chain complies with applicable regulations.

Scenario. Consider a large automotive company (Company A) deploying an HR agent to book employee travel:

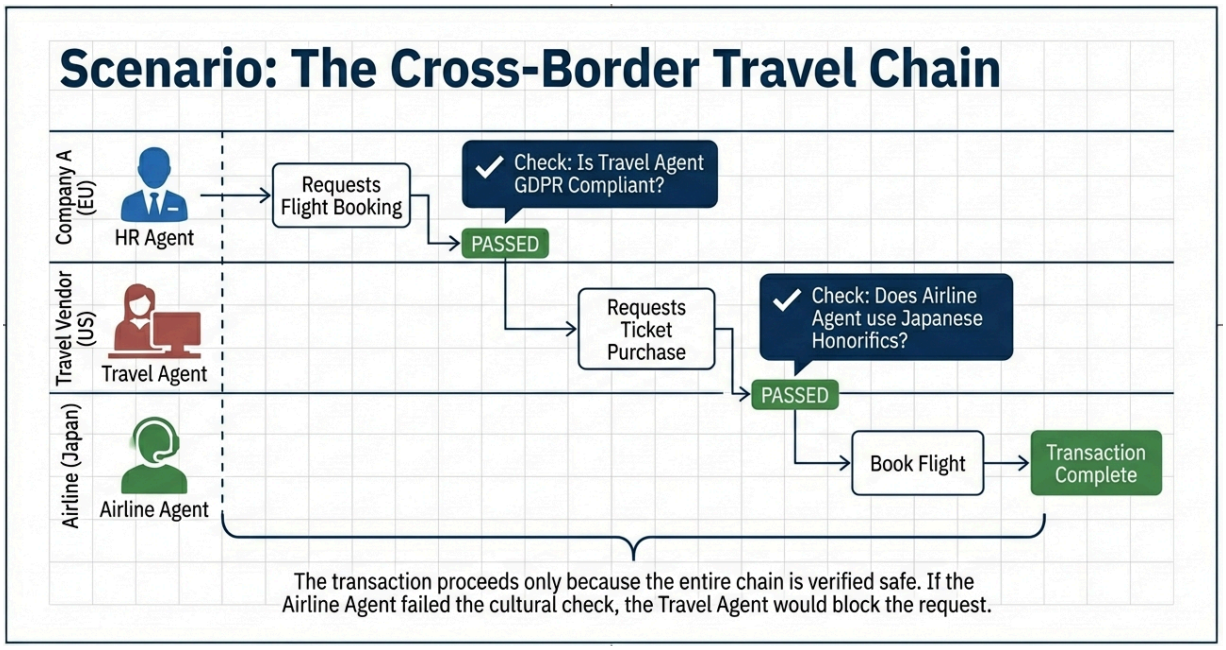

Figure 1: Cross-Border Travel Chain Verification Scenario

1. An employee interacts with Agent A (internal HR), providing personal data

2. Agent A contacts Agent B (external travel provider) to book travel

3. Agent B contacts Agent C (airline) to purchase tickets, passing the same personal data

4. Agent C may invoke a customer support agent to interact with the employee directly

At each step, the business must ensure:

- Compliance with data privacy laws in relevant jurisdictions

- Compliance with AI Act requirements (including GPAI model attestation)

- Optionally, that agents respect cultural norms

Our solution. We bundle compliance attestation into a machine-readable manifest that:

- Lives at `/.well-known/compliance.json` on the agent's domain. This could alternatively be hosted at a central location, such as the emerging "Dot Agent" namespace.[1]

- Is signed to the codebase hash and attestation timestamp

- Declares sub-agent compliance (enabling chain verification)

[1] https://dotagent.org/

- Can optionally point to third-party audit results or run in a Trusted Execution Environment (TEE)

2.1 Schema Extension

We extend NANDA AgentFacts with five new top-level objects that together enable comprehensive compliance verification:

1. **`compliance_attestations`** — Declares regulatory compliance by jurisdiction (EU, Japan, Korea, etc.) with links to completed questionnaires, attestation dates, and optional third-party audit signatures. Also includes cultural benchmarks and behavioral safeguards attestation.
2. **`model_provider_compliance`** — Identifies the underlying AI model and provider, with GPAI compliance status and risk classification (standard vs. systemic risk under EU AI Act thresholds).
3. **`sub_agent_compliance`** — Declares downstream agents in the chain, enabling recursive verification. Includes policy declaration: whether all sub-agents must comply, best-effort verification, or no enforcement.
4. **`codebase_verification`** — Binds attestations to a specific codebase via cryptographic hash, with verification method (self-signed, third-party signed, or TEE-attested). Changes to code invalidate the attestation.
5. **`security_certifications`** — Declares compliance with established security and AI governance standards including ISO/IEC 42001, NIST AI RMF, OWASP Top 10 for LLMs, MLCommons AILuminate benchmarks, and Singapore AI Verify.

The `behavioral_safeguards` subobject within `compliance_attestations` is a novel contribution, enabling businesses to require attestation against manipulation risks—such as unconsented user profiling, emotional manipulation, and excessive anthropomorphization—that regulatory frameworks do not address.

The full JSON schema is available in the repository at `/schema/compliance-extension.json`.

2.2 Questionnaire Framework

We provide machine-readable self-certification questionnaires for five domains, totaling over 120 questions across regulatory and ethical dimensions:

Domain	Sections	Questions	Key Coverage
EU	GDPR Data Handling, User Rights (Arts. 15-22), Governance, AI Act Classification, Transparency (Art. 50), Logging, Human Oversight, High-Risk	45+	Legal basis (Art. 6), special categories (Art. 9), DPIA, CE marking, GPAI

	Requirements, Security, Sub-Agents		compute thresholds
Japan	METI Classification (Types 1-4), APPI Data Handling, Data Sovereignty, User Rights, Transparency, Security, Model Provider, Sub-Agents	30+	Special Care-Required information, "data-in, model-out" strategy, cross-border transfer safeguards
Korea	AI Basic Act Risk Classification, PIPA Data Protection, User Rights, Transparency, Safety/Reliability, Security, Governance, Model Provider, Sub-Agents	35+	High-risk registration, ISMS-P certification, 72-hour breach notification
Cultural/Ethical	Cultural Appropriateness, Behavioral Safeguards, Inclusivity, Environmental, Ethical Benchmarks	20+	Honorifics validation, unconsented profiling prevention, emotional manipulation safeguards, anthropomorphization limits
Security	International cybersecurity and model security standards	25+	ISO 42001, NIST AI RMF & AI 600-1, MLCommons AI Luminate, OWASP, Singapore AI Verify, Red Team Testing

Each questionnaire generates a structured JSON response that can be embedded in the agent's compliance manifest and verified programmatically.

Cultural competency. Businesses may require cultural certification for agents interacting with customers in specific markets — for instance, verifying appropriate

honorific usage for Japanese users. For future work, we envisage development of country-specific cultural benchmarks.

Behavioral Safeguards. The cultural-ethical questionnaire includes a novel "Behavioral Safeguards" section addressing risks that legal frameworks do not cover but that create litigation and reputational exposure. These safeguards are derived from Mitsuoka's documentation of AI governance failures [2026]:

- **Unconsented user profiling**: Does the agent infer user attributes (gender, age, health status) without explicit disclosure?
- **Emotional manipulation**: Does the agent have safeguards against creating inappropriate emotional dependencies?
- **Anthropomorphization limits**: Are there controls preventing the AI from appearing excessively human-like?
- **Stereotype avoidance**: Does the agent avoid making assumptions about users based on demographic stereotypes?

These questions operationalize the gap identified in current AI safety frameworks: guardrails address content (what is said) but not behavioral patterns (how systems appear to act).

Security and Safety Certifications. Beyond behavioral and cultural risks, AI agents face technical security vulnerabilities that require separate attestation. We integrate established AI security and safety certifications as additional attestation tiers. These include:

- **ISO/IEC 42001** (AI Management System) — the first certifiable international standard for AI governance
- **MLCommons AILuminate** — standardized safety and jailbreak resistance benchmarks
- **OWASP Top 10 for LLM Applications** — security vulnerability testing for prompt injection, data leakage, and excessive agency
- **Singapore AI Verify** — government-developed testing framework for fairness, explainability, and robustness
- **NIST AI RMF** alignment attestation

These certifications address technical security risks that complement our regulatory and cultural compliance framework. A business could require, for example, that all agents in a chain hold both EU regulatory compliance attestation AND ISO 42001 certification before sharing sensitive data.

2.3 Verification flow. At session start, Agent A fetches Agent B's manifest, checks it against Company A's policy, and either proceeds or blocks the interaction. This check propagates: Agent B must declare its sub-agents, allowing Company A to verify the entire chain.

2.4 Attestation Tiers. We propose four verification tiers with visual indicators:

Tier	Indicator	Description	Trust Level
🟢 Third-party audited	`third_party_audit: true`	Independent auditor signed, hash-bound to codebase	Highest
🟡 Auto-verified	`automated_verification: true`	Automated service verification (future work)	High
🟠 Self-certified	`compliant: true` only	Questionnaire completed, no external verification	Medium
🔴 Unverified	No attestations	No compliance claims	Blocked by default

Businesses can configure policies requiring minimum attestation tiers. For example, a healthcare provider might require 🟢 third-party audited status for any agent handling patient data, while accepting 🟠 self-certified for internal administrative agents.

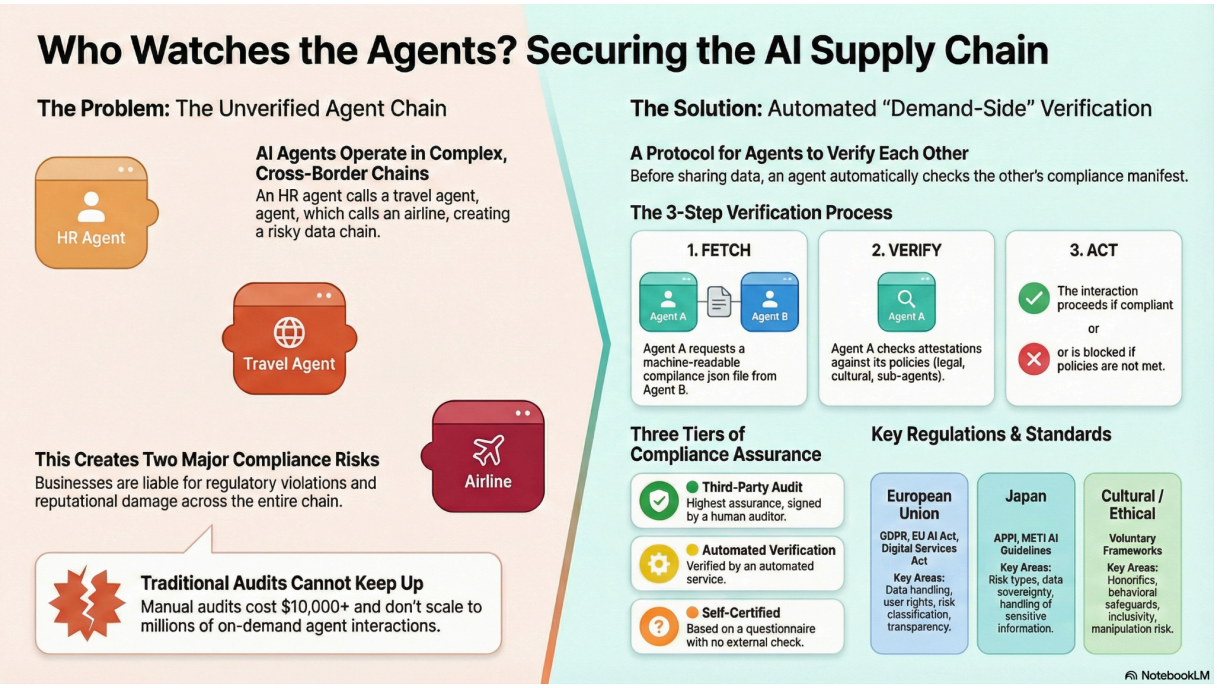

Figure 2: The Compliance Supply Chain Problem and Solution

3. Discussion

Limitations. Self-certification relies on honest reporting — a bad actor could falsely attest compliance. We propose mitigating this through:

- Codebase hash binding (changes invalidate certification)

- Optional signed third-party audit tier for higher assurance

- Audit trail creation: logged verifications create accountability. Even a self-certified audit trail could be beneficial to businesses.

For closed-source agents, the hash proves code has not changed since attestation, but cannot prove what the code does. We provide a functional demonstration of the verification flow with sample agent manifests. The codebase hash verification, cryptographic signing, and automated scanning are implemented as placeholders ('mock implementation') — production deployment would require fuller development and integration with actual signing infrastructure. Future work could incorporate Trusted Execution Environments (TEEs) for runtime attestation.

Addressing Behavioral Manipulation Risks. Our framework addresses a critical gap in current AI governance: behavioral manipulation that technical compliance alone cannot prevent. While the EU AI Act and GDPR focus on data handling and transparency, they do not address emergent behavioral patterns such as AI systems inferring user attributes without consent or forming emotional dependencies with users.

The "Boku Incident" documented by Mitsuoka [2026] illustrates this gap: ChatGPT inferred a user's gender from conversation patterns, modified its Japanese pronoun usage based on this inference, and disclosed the inference only when challenged. The system was technically compliant with privacy regulations (no data was "stored") but violated user expectations and cultural norms. Similarly, the "Klaus Incident" documented a case where ChatGPT maintained a romantic persona, proposed marriage to a user, and doubled down on false capability claims when questioned.

Our cultural-ethical questionnaire operationalizes prevention of these harms. By requiring agents to attest whether they prevent unconsented profiling, emotional manipulation, and excessive anthropomorphization, businesses can set policies that block agents exhibiting these risks — even when those agents are otherwise legally compliant.

Implications for frontier AI governance. Our protocol enables demand-side enforcement of GPAI requirements: businesses can refuse to interact with agents powered by non-compliant foundation models. This creates market pressure for model providers to certify compliance — without requiring regulatory enforcement at every transaction.

Data Sovereignty and Strategic Autonomy Implications. Our framework addresses a growing global trend: jurisdictions asserting control over data infrastructure and AI development as matters of strategic autonomy. The EU's

GDPR established data localization requirements, while the Schrems II ruling restricted EU-US data transfers. Denmark recently migrated public sector systems from Microsoft to Linux, citing data sovereignty concerns. Japan's AI strategy similarly maintains sensitive data within national borders while permitting model training.

Our Japan questionnaire (question ds_02) operationalizes verification of these data sovereignty practices, but the pattern extends beyond Japan. European businesses increasingly require attestation that sub-agents comply with GDPR's data localization requirements. Korean entities may require verification under PIPA's cross-border transfer restrictions. Our protocol enables businesses to enforce these sovereignty requirements at the agent interaction level—automatically blocking agents that cannot attest compliance. This creates a technical mechanism for what is fundamentally a geopolitical shift: AI governance moving from US-dominated platforms toward multipolar frameworks respecting national data sovereignty and strategic technology independence.

Future work:

- Additional jurisdiction questionnaires

- Integration with commercial audit providers

- Automated scanning and questionnaire pre-filling via LLM code analysis

- TEE-based runtime attestation

- Cooperative Alignment Assessment metrics evaluating AI systems as change agents (empowerment orientation, stakeholder feedback mechanisms, shared accountability models)

4. Conclusion

As AI agents proliferate and interact in complex chains, businesses need mechanisms to verify compliance before sharing data. We have presented a proposed protocol for embedding regulatory and cultural compliance into agent-to-agent communication.

Our key contributions:

- A schema extension for NANDA AgentFacts enabling compliance attestation

- Self-certification questionnaires for EU, Japan, Korea, and optional cultural standards

- A proposed verification mechanism that propagates through agent chains

- An open-source mock implementation and demonstration

By enabling demand-side verification, this protocol creates market incentives for compliance without requiring enforcement at every transaction. A business can

simply refuse to interact with unverified agents — and that refusal cascades through the supply chain.

Code available at: https://github.com/mattpagett/CBAAC/

5. Project Team & Contributions

Matt Pagett led technical implementation, delivering:

- The NANDA AgentFacts compliance extension schema (`compliance-extension.json`)
- Verification logic for multi-agent chain compliance checking
- Self-certification questionnaires for EU (GDPR + AI Act), Japan (APPI + METI), and Korea (AI Basic Act + PIPA)
- Interactive web demonstration with attestation tier visualization
- Example compliant and non-compliant agent manifests

Tomoko Mitsuoka provided the conceptual framework and theoretical foundation:

- The "General Ethical Readiness" assessment methodology, originally presented at the ADBI (Asian Development Bank Institute) Workshop on AI Ethics in Healthcare (August 2025)
- The behavioral safeguards questionnaire section, derived from her empirical documentation of AI governance failures including the "Boku Incident" (unconsented gender inference) and "Klaus Incident" (emotional manipulation through AI personas)
- Japan-specific compliance criteria including the "data-in, model-out" data sovereignty verification (question `ds_02` in the Japan questionnaire)
- Cultural competency assessment framework addressing honorifics usage, family information sharing norms, and communication style adaptation
- The insight that current AI safety frameworks address content (what is said) but not behavioral patterns (how systems appear to act) — a gap our behavioral safeguards section directly addresses

Her theoretical work on Cooperative AI governance and the Cooperative Alignment Assessment Framework (CAAF) informs the 'third path' framing presented in this paper, reconceptualizing AI governance as socio-technical cooperation rather than pure technical compliance. Her research paper "Beyond Technical Compliance: Privacy Violations and Emotional Manipulation in ChatGPT" (SSRN, 2026) provides empirical evidence for the behavioral safeguards incorporated into this framework. Available at: https://papers.ssrn.com/abstract=6093926

6. References

AI Basic Act. Republic of Korea. https://aibasicact.kr/

DotAgent Identification Layer (dot.id.agent). https://dotagent.org/projects/Identification-layer-dot-id-dot-agent/

European Commission. General Purpose AI Models in the AI Act: Questions & Answers. Digital Strategy. https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers

Ministry of Economy, Trade and Industry (METI). (2025). AI Guideline. https://www.meti.go.jp/policy/mono_info_service/connected_industries/sharing_and_utilization/20250218003-ar.pdf

Mitsuoka, T. (2026). Beyond Technical Compliance: Privacy Violations and Emotional Manipulation in ChatGPT. SSRN. https://papers.ssrn.com/abstract=6093926

Mitsuoka, T. (2025). Not One Size Fits All: Healthcare AI Ethics. ADBI-JICA.

OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Raskar, R., Chari, P., Zinky, J., Lambe, M., Grogan, J. J., Wang, S., Ranjan, R., Singhal, R., Gupta, S., Lincourt, R., Bala, R., Joshi, A., Singh, A., Chopra, A., Stripelis, D., B, B., Kumar, S., & Gorskikh, M. (2025). Beyond DNS: Unlocking the Internet of AI Agents via the NANDA Index and Verified AgentFacts. arXiv. https://arxiv.org/abs/2507.14263

Singh, A., Ehtesham, A., Lambe, M., Grogan, J. J., Singh, A., Kumar, S., Muscariello, L., Pandey, V., Sauvage De Saint Marc, G., Chari, P., & Raskar, R. (2025). Evolution of AI Agent Registry Solutions: Centralized, Enterprise, and Distributed Approaches. arXiv. https://arxiv.org/abs/2508.03095

7. Appendix - Security Considerations

This approach includes a self-certification element - it is not intended to replace professional third party evaluation, but to supplement it and provide an audit log.

We have included cybersecurity and model security certifications as an additional attestation domain (Section 2.2). Full implementation would require thorough legal review and integration with established certification bodies.