REQUEST FOR PROPOSAL

# Agent Trust Management

*Detecting, Classifying, and Controlling AI Agents on Your Platform*

A vendor-neutral evaluation template for procurement, security, and fraud teams

**DOCUMENT TYPE** Draft RFP / evaluation questionnaire template
**INTENDED USE** Vendor evaluation and internal capability assessment
**OWNER** [Issuing organization, procurement / vendor management]

*Provided by Arkose Labs as an industry resource. This template is vendor-neutral and may be adapted freely for your organization's use.*

## How to Use This Document

This Request for Proposal is a starting point, not a finished procurement package. It was built to help enterprise procurement, security, and fraud teams evaluate solutions in an emerging category the industry is still naming. The vocabulary is shifting because the problem has shifted, and most existing vendor questionnaires were written for a world that no longer exists.

### What this template is for

Use it in either of two ways. First, as a vendor RFP: issue the relevant sections to prospective suppliers and score their written responses, then use the demonstration prompts to structure live evaluations. Second, as an internal readiness assessment: work through the questions as a team to surface where your current defenses stand against agent traffic, even if you are not yet ready to buy. Many organizations find the second use as valuable as the first, because the questions expose assumptions that binary bot-detection tooling was never designed to test.

### How it is organized

The document moves from context to specifics. Section 1 frames why agent trust is a distinct problem from bot detection. Sections 2 through 9 contain the substantive requirement sets: traffic classification, malicious-agent detection, trust and control, architecture, data, security and privacy, commercial terms, and references. Each requirement carries response guidance describing what a strong answer contains, so evaluators without deep domain expertise can still tell a substantive response from a marketing one.

### Adapting it to your organization

Delete what does not apply, add what is unique to your environment, and adjust the weighting to your risk profile. A retailer worried about inventory hoarding and a bank worried about account takeover will emphasize different sections. Placeholders in [brackets] mark where your organization's specifics belong. The companion response matrix (a separate spreadsheet) mirrors these sections and gives vendors a structured place to answer.

A note on scope. This template deliberately treats agent trust as broader than bot blocking. The goal is not to keep all automation out. Automated and agent-driven traffic is becoming a legitimate, revenue-bearing channel. The goal is to distinguish which agents to trust, under what conditions, and to enforce that decision. Questions are written to that standard.

# 1. Why Agent Trust Is a New Problem

For over a decade, the defining question in automated-traffic defense was binary: is this a bot or a human? Detection tooling was built to answer it, and enforcement meant blocking whatever failed the test. That model is breaking down, and it is worth being precise about why, because the reason shapes every requirement that follows.

## Everything automated now looks the same

Analysis across large volumes of consumer traffic points to an uncomfortable finding: legitimate AI agents and malicious ones share nearly identical technical fingerprints. When a mainstream AI assistant shops on a user's behalf, it presents a synthetic browser on cloud infrastructure, with device signals that do not match a real phone or laptop, and interaction timing no human produces. A credential-stuffing bot presents exactly the same profile. Even the official, well-behaved crawlers published by major AI providers use synthetic browsers and cloud infrastructure. The technical layer alone (device fingerprinting, headless-browser detection, IP reputation) can tell you that a session is automated. It cannot tell you whether that automation is helping your customer or attacking them.

The consequence: a control that blocks every fake device now blocks legitimate AI assistants alongside fraud bots. In a market where analysts project agents could drive a material share of e-commerce within a few years, blocking all agent traffic is not a security decision. It is a revenue decision made by default.

## Three agent populations, one internet

A more useful frame than bot vs. human starts by splitting traffic into humans and automation, then classifying automation into three coexisting agent populations, each requiring different treatment:

- Self-disclosing agents: automation that identifies itself, through published IP ranges, declared user agents, or signed credentials, acting on behalf of, or with the authorization of, a legitimate party.
- Non-disclosing agents: automation that does not self-identify, because it operates through a consumer's own browser or was deployed without disclosure protocols, but is still acting on behalf of, or with the authorization of, a legitimate party.
- Adversaries: automation pursuing account takeover, fake-account creation, scraping, card testing, or fraud. Some hide entirely; some impersonate self-disclosing or non-disclosing agents to earn preferential treatment.

Critically, population is a property of a moment, not of an agent. The same agent can arrive as a self-disclosing helper on one request and behave like an adversary on the next. Classification therefore has to be continuous, per-session, and per-action, not a one-time allow-list decision.

## From detection to classification to control

The question has changed from 'Is this a bot?' to 'What is this agent's intent, and is it authorized to do what it is trying to do?' Answering that requires three capabilities most bot-detection tools were never built to provide: classification (place the session on a spectrum, not a binary), intent determination (what is this agent attempting: browsing, buying, scraping, taking over an account?), and control (apply graduated, business-owned policy: allow, monitor, challenge, throttle, or block, varying by endpoint and risk). The requirement sets that follow test for exactly these capabilities. As you evaluate responses, whether from vendors or from your own team, six questions are worth keeping in view.

## Six questions to hold in mind

**1. Can we determine intent when all agents show identical technical anomalies?**

**2. Do we understand that blocking fake devices would eliminate legitimate AI assistants along with fraud bots?**

**3. Can we correlate automated traffic with legitimate business events or user requests?**

**4. Do we have behavioral analysis beyond technical fingerprinting?**

**5. Can we distinguish agents helping users from agents harming them?**

**6. Are we prepared for a world where synthetic devices are the norm, not the exception?**

## 2. Traffic Classification & Agent Coverage

This section establishes whether a solution can see the three populations at all, and how granularly. If a vendor cannot reliably separate humans, traditional bots, and agentic traffic, and distinguish known from unknown agents, the remaining sections are moot.

| # | Requirement / Question | Response Guidance |
|---|---|---|
| **2.1** | Describe how your solution classifies inbound traffic. At minimum, can you separate (a) legitimate humans, (b) traditional bot/scripted automation, and (c) agentic AI traffic, and within agentic traffic, distinguish known from unknown agents? | Strong answers describe real-time classification into all categories with distinct treatment per category, not a single risk score. Look for how 'unknown agent' is handled. |
| **2.2** | Can your solution identify AI agents across all deployment models: (1) cloud-hosted agents; (2) agents embedded in a browser ('AI browsers'); and (3) agents running as local applications at the OS level? | Coverage should span web and native mobile (iOS/Android). Note any category the vendor cannot detect and why. |
| **2.3** | In human-in-the-loop scenarios, can you differentiate between when an AI agent is controlling the session versus when the actual end user is? Can you track the moment a human hands off a task to an agent (and back)? | Look for continuous behavioral analysis within a session, not just a start-of-session check. Central to action-tier policy. |
| **2.4** | How do you determine an agent's provenance, whether it originated from an LLM platform, an agentic browser, or a custom-built agent? | Provenance visibility (not just 'agent: yes/no') is an emerging analyst evaluation criterion. |
| **2.5** | Can you identify the underlying automation frameworks (e.g., Playwright, Puppeteer, Selenium) even when they actively attempt to mask themselves? | Answers should describe signals that persist despite evasion, not just default framework signatures. |
| **2.6** | Describe your approach to behavioral analysis. If you use behavioral biometrics, does deployment require tagging or instrumenting specific page elements (e.g., input fields)? | Prefer solutions that collect behavioral signals passively without per-element instrumentation, which reduces integration burden. |

## 3. Malicious-Agent Detection

Because legitimate and malicious agents are technically identical, detecting malicious agents is a fundamentally different problem from detecting known, cooperative ones. This section probes whether the vendor treats it as such, or simply applies the same signatures to everything.

| # | Requirement / Question | Response Guidance |
|---|---|---|
| **3.1** | Describe if and how your solution detects malicious AI agents. Does your approach for detecting malicious agents differ from how you detect commercially available or consumer-facing agents? If so, how? | This is the single most revealing question in the RFP. A vendor that treats all agents the same has not internalized the three-population model. Strong answers explain why the starting assumption |

| | | |
|---|---|---|
| | | differs (malicious agents do not self-identify). |
| **3.2** | How do you determine intent when technical signals alone show only automation? Describe the behavioral and contextual methods used (e.g., navigation patterns, boundary respect, business-activity correlation). | Look for correlation of traffic to legitimate business events, respect for robots.txt / rate limits, and detection of 'extraction-optimized' navigation. |
| **3.3** | How do you detect agents that impersonate legitimate services (e.g., spoofed AI-provider user agents from the wrong IP ranges, mismatched TLS fingerprints, or out-of-order headers)? | Impersonation detection is distinct from device detection; all agents use synthetic devices, so this tests deeper signal correlation. |
| **3.4** | Describe your use of adaptive challenges as a classification signal. How does an agent's response to a challenge inform your classification (solve, back off gracefully, systematic failure, bypass attempt)? | The best answers treat challenges as intent probes, not just barriers, and describe challenges designed against known model weaknesses. |
| **3.5** | Can your solution detect drift, traffic (including a previously trusted agent) that materially changes its behavior over time? | Drift/revocation is a frontier requirement analysts now ask about. Look for session-, baseline-, and population-level drift detection. |
| **3.6** | If available, describe your performance metrics for agentic detection (e.g., true-positive and false-positive rates) and how they are measured. | Credible vendors describe measuring against real traffic on your specific flows during a proof of value, not just published benchmarks. |

## 4. Trust Management, Enforcement & Control

Detection identifies; control decides what to do. This section tests whether a solution enables graduated, business-owned policy, the difference between a tool that only blocks and a platform that lets you safely admit good agents while stopping bad ones. Progressive-trust and action-tier concepts here draw on emerging frameworks such as the Cloud Security Alliance's Agentic Trust Framework and analyst 'guardian agent' research.

| # | Requirement / Question | Response Guidance |
|---|---|---|
| **4.1** | Beyond allow/block, what enforcement actions can you take (e.g., monitor, challenge, throttle, step-up, hand back to human)? Can responses vary by endpoint and by risk level? | A full spectrum of graduated responses is essential. Allow/block-only tooling forces the binary this category exists to escape. |
| **4.2** | How does your solution help identify, track, and configure trustworthy ('good') agents? Can we add, remove, and manage known-good agents ourselves, and is this self-serve? | Look for a managed agent registry, not a static text allow-list, and self-service configuration in the UI. |
| **4.3** | Do you support emerging agent-disclosure standards (e.g., Web Bot Auth, signed agent credentials)? Can we allow-list agents that self-identify via these standards, and do you contribute to their development? | Standards support signals forward-readiness. Ask whether allow-list management requires professional services. |
| **4.4** | Do you support progressive or persistent trust, where an agent earns or loses trust over time based on behavior, including trust decay and immediate revocation on material behavior change? | This is the model behind CSA/Microsoft agent-control frameworks. Look for persistent identity plus behavioral continuity plus a revocation trigger. |

| | | |
|---|---|---|
| **4.5** | Can we set different trust thresholds or policies per action tier (e.g., low-risk reads allowed to agents, high-consequence writes gated behind human step-up)? | Action-tier policy lets agents handle the low-risk majority of a journey while gating high-consequence actions. A key business-enablement capability. |
| **4.6** | Who owns classification and enforcement policy in your model, and how do we adjust it? Describe the management UI for adding applications/endpoints, categorizing them, and applying policies at scale. | Policy is a business decision; the tool should make it configurable by our team. Look for bulk application management and centralized policy control. |
| **4.7** | How does the product explain why a given session was classified as human, bot, or agent, in terms non-technical stakeholders can act on? | Explainability (triggered rules, signal weights, decision logic) is now an explicit analyst criterion and matters for internal buy-in. |

## 5. Architecture, Integration & Performance

This section covers how the solution deploys into your environment, how it performs under load, and how resilient it is. Requirements here mirror standard enterprise technology diligence, focused on the channels and latency constraints agent-trust controls typically sit within.

| # | Requirement / Question | Response Guidance |
|---|---|---|
| **5.1** | Describe your technical architecture and how the solution integrates into a client's service across channels (desktop web, mobile web, native iOS/Android, and server-to-server / headless flows). | Look for lightweight web integration (e.g., a JS tag), native SDKs, and an API path for headless/agentic flows. Note CDN/edge and IAM/CIAM integration options. |
| **5.2** | What integration methods are available, and what is the typical time-to-value for an initial flow? Summarize the engineering level of effort required from our side. | Expect a realistic go-live estimate for a single flow and a clear statement of the developer effort required. |
| **5.3** | If your solution includes an API, describe its performance characteristics and latency. What is your service availability SLA and how is downtime defined? | Look for a concrete uptime commitment (e.g., 99.9%), a latency definition, and service-credit remedies. |
| **5.4** | Describe resiliency and fail-open/fail-closed behavior. What happens to our traffic if your service is degraded or unreachable? | The vendor should let you choose fail-open vs. fail-closed per flow based on your risk tolerance. |
| **5.5** | Do you operate a team that actively monitors malicious or suspicious activity on our behalf? Describe their practices and coverage (e.g., 24/7). | Managed monitoring vs. self-managed tooling materially changes the operational burden on your team. |
| **5.6** | What are your support severity levels and response-time SLAs (P1 through P4)? Describe your incident-notification process. | Expect tiered response times by severity and a defined breach/incident notification protocol. |
| **5.7** | Describe your product roadmap for agent trust over the next 12 to 18 months and how you prioritize it. How will you handle our custom requirements? | Roadmap relevance and a formal customer-feedback path into it are analyst evaluation criteria. |

## 6. Data, Machine Learning & Threat Intelligence

Agent classification is a data problem. This section examines the models behind the decisions, the intelligence network that informs them, and your rights to the data the solution produces.

| # | Requirement / Question | Response Guidance |
|---|---|---|
| **6.1** | Does your solution use machine-learning models for detection and classification? How are models managed, updated, and assessed for effectiveness over time? Can we bring our own models or data tags? | Look for continuous model updates, efficacy monitoring, and (ideally) customer-specific tuning. Ask for model white papers if available. |
| **6.2** | Do you operate a threat-research function focused on agentic AI and emerging attack tooling? Describe the team and how its research feeds the product and reaches customers. | A dedicated, adequately staffed research team that publishes findings is a strong differentiator and an analyst criterion. |
| **6.3** | Do you operate a cross-customer intelligence network or consortium? If so, what is the opt-in process, and how is data used and secured? | Network effects improve detection of agents seen elsewhere. Confirm data handling and opt-in terms. |
| **6.4** | Can raw session data / individual signals be streamed to us in real time for independent monitoring, analysis, or use in our own ML? Do we retain rights to that data? | Signal transparency and downstream-use rights are increasingly demanded by enterprise buyers. Confirm contractually. |
| **6.5** | Can we submit confirmed outcomes (e.g., verified fraud labels) back to you to improve model accuracy over time? | A closed feedback loop (truth data) improves accuracy on your specific traffic. |

## 7. Security, Privacy & Compliance

Because agent-trust controls sit inline on high-value flows and may process behavioral and device data, this section covers data handling, security posture, and third-party risk. Adapt the regulatory items to your jurisdiction and industry.

| # | Requirement / Question | Response Guidance |
|---|---|---|
| **7.1** | Does your solution require collection of device or browser attributes? If so: which attributes, do any require explicit user permission, and can we control what is collected? | Prefer data-minimization by design, only signals necessary for detection. Confirm PII is not required for detection. |
| **7.2** | Is any PII exposed or required in the detection process? Describe precisely how you handle PII and behavioral data, and specifically how you would handle our data. | Best-practice answers show detection operating on behavioral/environmental signals, not user identity. |
| **7.3** | If your solution consumes our data, describe how it is safeguarded: encryption at rest and in transit, data classification, logical separation, and retention. | Expect AES-256 at rest, TLS 1.2+ in transit, and a clear retention policy. |
| **7.4** | Outline your vulnerability-management practices (e.g., aligned to NIST): AppSec (SAST/DAST/SCA), infrastructure scanning, CVSS-based prioritization, and remediation SLAs. | Look for a continuous, documented program covering the AI/detection components, not just the core platform. |
| **7.5** | What security certifications and attestations do you hold (e.g., SOC 2, ISO 27001)? Provide links or documentation. | Standard enterprise diligence. Request current reports under NDA. |

| | | |
|---|---|---|
| **7.6** | Describe your data-breach notification protocol (standard and heightened) for customers, including timelines and regulatory disclosure (e.g., GDPR, CCPA). | Look for a defined incident-response plan with committed customer-notification timelines. |
| **7.7** | List material subcontractors and sub-processors that would process our data, their locations, and your process for vetting and monitoring them. | Expect a published sub-processor list, DPAs where required, and a risk-tiered third-party review cycle. |
| **7.8** | How does your product conform to accessibility standards (e.g., WCAG), particularly any user-facing challenge experiences? Do you support internationalization? | Relevant where challenges are presented to end users. Request the conformance level and version. |
| **7.9** | Do you maintain anti-bribery/anti-corruption policies and business-ethics standards binding on your employees globally? | Standard supplier-diligence item; adapt to your procurement requirements. |

## 8. Commercial Terms & Proof of Value

This section captures pricing structure and the terms of any evaluation. Given how quickly the agent landscape moves, a live proof of value on your own traffic is often worth more than any written benchmark; structure the engagement to produce metrics specific to your flows.

| # | Requirement / Question | Response Guidance |
|---|---|---|
| **8.1** | Describe your pricing model (e.g., per-session, per-application, traffic-based), including tiers and what is included in the base price versus priced separately. State clearly whether pricing covers agentic detection only or your full capability set. | Pricing clarity and transparency is an analyst criterion. Confirm whether managed monitoring, research, and support are included or add-ons. |
| **8.2** | What are the upfront and any 'at-risk' costs to stand up a partnership? Provide a representative annual cost for our scope. | Buyers value predictability; ask for a committed-volume option and overage terms. |
| **8.3** | Do you offer a proof of value or POC? Describe its structure, duration, cost, and success criteria. Can it produce detection metrics specific to our flows? | Prefer a POV that measures on your real traffic and flows, with defined go/no-go metrics agreed in advance. |
| **8.4** | Provide a representative implementation plan and timeline from contract to go-live, including phases (observation before enforcement) and the resources required from us. | Look for an observation phase before active enforcement, and a clear split of vendor vs. customer responsibilities. |

## 9. References & Vendor Profile

Standard supplier-qualification items. Adapt to your procurement process; these confirm the vendor is a viable long-term partner, not only a capable product.

| # | Requirement / Question | Response Guidance |
|---|---|---|
| **9.1** | Provide the full legal entity name, headquarters, ownership structure (parent company if applicable), and whether the company is public or private. | Standard corporate-profile diligence. |

| | | |
|---|---|---|
| **9.2** | Provide three customer references, ideally in our industry and at comparable scale, whom we may contact regarding implementation and ongoing operation. | Prefer references facing similar agent-traffic challenges. References are commonly shared after mutual intent to proceed. |
| **9.3** | Describe your partner and technology ecosystem (e.g., CDN, IAM/CIAM, SIEM integrations) and any relevant alliances or industry-body participation. | Ecosystem breadth eases integration and signals market standing. |
| **9.4** | Summarize your company's vision for agent trust over the next two years and how your roadmap positions you against where the market is heading. | A differentiated, well-articulated vision distinguishes category leaders from feature vendors. |

## Appendix A. Evaluation Scorecard

A suggested scoring frame. Weight the sections to your risk profile before issuing. Score each section 1 to 5 against the response guidance, multiply by the weight, and total. The weights below are a starting point, not a prescription.

| § | Section | Weight | Score (1 to 5) |
|---|---|---|---|
| **2** | Traffic Classification & Agent Coverage | 20% | |
| **3** | Malicious-Agent Detection | 20% | |
| **4** | Trust Management, Enforcement & Control | 20% | |
| **5** | Architecture, Integration & Performance | 12% | |
| **6** | Data, ML & Threat Intelligence | 10% | |
| **7** | Security, Privacy & Compliance | 10% | |
| **8** | Commercial Terms & Proof of Value | 5% | |
| **9** | References & Vendor Profile | 3% | |

*The three highest-weighted sections, classification, malicious-agent detection, and control, are the capabilities that separate genuine agent-trust platforms from repackaged bot-detection tools. Weight them accordingly.*

## Appendix B. Working Definitions

**Agent trust management**

The discipline of detecting, classifying, and controlling automated/agentic traffic based on intent and authorization, rather than simply detecting and blocking bots.

**Agentic AI traffic**

Traffic generated by AI agents that reason, plan, and act, whether acting for a legitimate user, a business integration, or an attacker.

**The three agent populations**

Self-disclosing agents, non-disclosing agents, and adversaries. All three can appear technically identical, which is why classification cannot rely on device signals alone.

**Classification**

Placing a session on a spectrum of trust/intent rather than a bot/human binary.

**Progressive trust**

A model in which an agent earns or loses access over time based on behavior, with trust decay and revocation, analogous to how banks step up verification for high-risk actions.

**Action-tier policy**

Allowing agents to perform low-risk actions (reads, browsing) while gating high-consequence actions (transfers, credential changes) behind human step-up.

**Drift detection**

Identifying when previously trusted traffic materially changes behavior, triggering re-classification.

**Web Bot Auth**

An emerging standard by which agents cryptographically self-identify, enabling trusted allow-listing.

## About this template

This template was assembled by Arkose Labs from real enterprise RFP structures, current industry analyst evaluation criteria, and published research on AI agent detection and classification. It is provided as a vendor-neutral resource. Organizations are free to adapt it. Its inclusion of a capability does not imply any single vendor, including Arkose Labs, satisfies every requirement; that is what your evaluation is for.

*To discuss agent trust or request a briefing on the underlying research, contact Arkose Labs at arkoselabs.com.*