Trellix

WHITE PAPER

# How to Have Confidence in AI Decision-making

## A guide to the Confidence, Auditability, Reversibility, Trust, and Safety (CARTS) Framework

## Executive summary

As the industry moves toward agentic AI in cybersecurity—systems that not only generate text but plan and execute autonomous actions—the primary barrier to adoption remains the "black box" problem. Current AI security implementations rarely provide the transparency a CISO needs to trust them with automated remediation. Going forward, security operations will be judged less by the volume of alerts they triage and more by the confidence of the AI making their decisions. Realizing that shift requires a shared, open way to measure and audit that confidence.

The Confidence, Auditability, Reversibility, Trust, and Safety (CARTS) Framework is proposed as an open industry standard, aligned with the principles of the Open Secure AI Alliance (OSAIA), to establish a vendor-agnostic method for measuring and auditing the confidence of AI-driven security decisions. By standardizing how telemetry from diverse ecosystem partners contributes to a weighted confidence score, the CARTS Framework enables a shift from volume-based SOC metrics (alerts triaged, cases closed) to confidence-based risk management.

This framework creates a shared value proposition: security teams gain unprecedented visibility into which vendors drive the most reliable decisions, while technology partners are given a standardized method to prove the direct impact and return on investment (ROI) of their telemetry and APIs on AI-driven decision-making.

## The five pillars of CARTS

The framework is built upon five core requirements for any AI agent operating in a production environment:

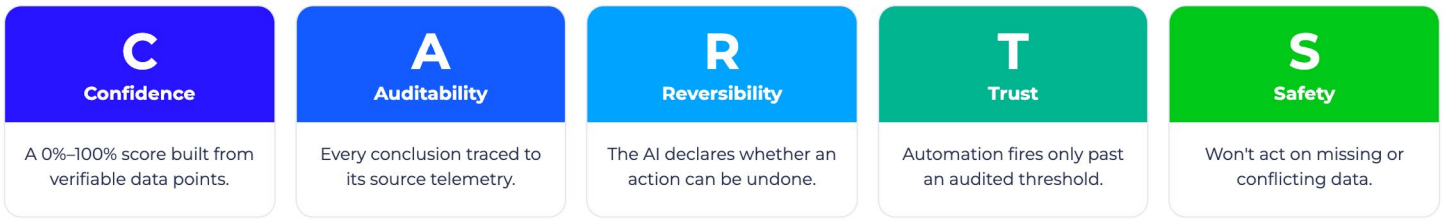



**Figure 1.** The five pillars every compliant CARTS agent must satisfy.

1. **Confidence:** A segmented matrix that adds up to a quantifiable percentage (0%–100%) based on verifiable data points across situational, business, and historical contexts
2. **Auditability:** The ability for a human analyst to review the "line-by-line" reasoning engine of the AI and trace every conclusion to its source telemetry
3. **Reversibility:** The capability for the AI to declare whether an action can be undone, enabling safer policy creation and automated "undo" workflows
4. **Trust:** A psychological and operational safety net where automation is triggered only once a predefined confidence threshold is met and audited
5. **Safety:** A process-driven guarantee where the AI actively questions environmental data to verify asset identity and prevents actions based on missing or conflicting information

## Telemetry contribution model and vendor value mapping

The following model enables customers to identify which partners contribute most effectively to specific confidence factors, while empowering vendors to demonstrate their measurable value to the ecosystem.

The power of the CARTS Framework lies in its collaborative nature. Industry partners contribute telemetry into or provide APIs that fuel a vendor-agnostic AI harness to increase the total confidence of an investigation. Telemetry should be used to actively question and verify assets, not just classify them. This allows for contributors to demonstrate their value towards making confident AI decisions and allows customers to create effective guardrails that leverage those confidence levels. The CARTS Specification sets out exactly how each partner's contribution is evaluated, in four stages: the vendor declares which facts its telemetry can produce, an evaluator verifies them, a test with a control group measures how much they move confidence, and each customer's own decision records show what they contribute in production. Results are published on a standard scorecard. CARTS measures contributions, but it does not directly rank vendors.

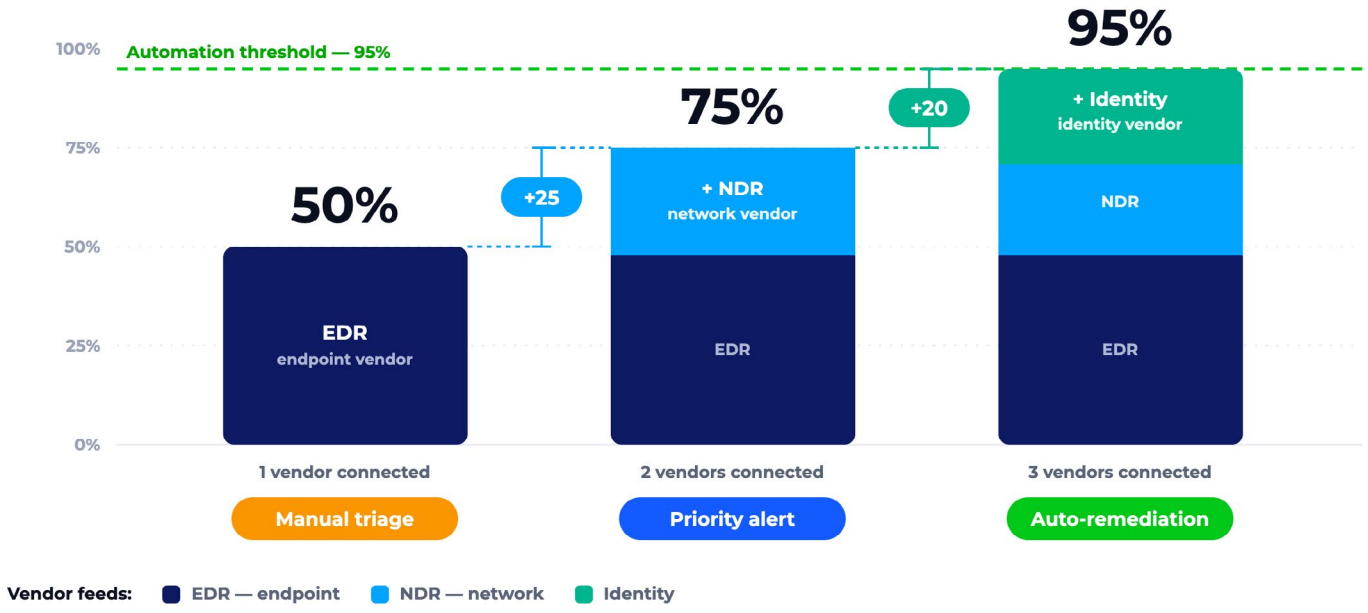



**Figure 2.** Confidence lift: Each connected feed raises the score and unlocks higher-tier action. The delta a vendor's telemetry adds is its measurable ROI.

| Partner category | Telemetry contribution | Impact on confidence factors |
|---|---|---|
| **Network (NDR)** | Netflow, port analysis, and lateral movement detection | **Situational Awareness:** Confirms what happened (e.g., port 443 identifies a web server) and which clients accessed it |
| **Identity providers** | User behavior, VIP status, and criticality levels | **Business Context**: Identifies if the affected user is a high-risk target or service account |
| **Endpoint (EDR)** | Process trees, script execution, and file signatures | **Situational Awareness:** Provides deep visibility into the activity of a local machine |
| **Vulnerability management** | Criticality of hardware and software inventory | **Business Context (and a capped Suspicion contribution):** Determines if authorized software and tools are vulnerable and / or present on that asset |
| **SIEM / data lake** | Historical logs and correlation across diverse sources | **Situational Awareness, Suspicion, and Historical Context:** Straddles situational awareness (what happened) and suspicion (abnormal activity patterns) |
| **Penetration testing / red team** | Simulated attack paths and validated vulnerability chains | **Suspicion and Behavioral Baseline:** Provides high-fidelity indicators of known attack methodologies |
| **ITSM / change-management** | Change records, approved maintenance windows, ticket state, and action rollback / undo history | **Reversibility and SOP Alignment:** Confirms whether an action is sanctioned by an approved change and provides the audit trail and rollback path required to safely automate |
| **Threat intelligence feeds** | IOC reputation, actor TTP mapping, and campaign / exploit-in-the-wild context | **Suspicion and Behavioral Baseline:** Corroborates whether observed indicators match known-malicious infrastructure or active adversary methodologies, raising or lowering suspicion |
| **CMDB / asset inventory** | Authoritative asset ownership, business criticality, environment (prod / dev), and service dependencies. | **Business Context:** Identifies who and what is affected, establishing asset criticality, ownership, and blast radius for the affected system |

## Weighted confidence matrix

Under the CARTS Framework, AI confidence is not a “black box” number. It is a sum of weighted factors:

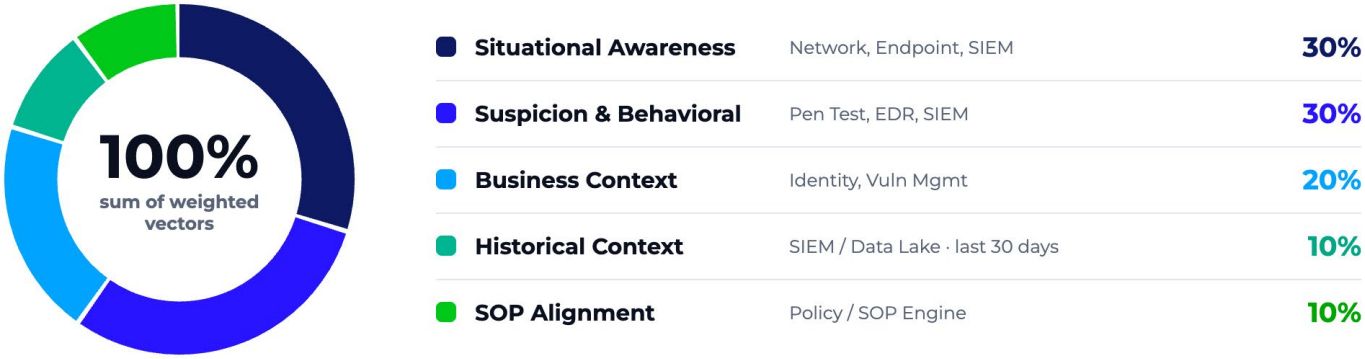



**Figure 3.** Default weighting of the five confidence factors (org-configurable).

| Confidence factor | Weight | Description | Primary data sources |
|---|---|---|---|
| **Situational Awareness** | 30% | Confirms “what happened” via event timelines and telemetry | Network, endpoint, SIEM / data lake |
| **Suspicion and Behavioral baseline** | 30% | Merges attack paths, anomalies, and deviation from “normal” | Pen testing, EDR, SIEM |
| **Business Context** | 20% | Identifies user / asset criticality and VIP status | Identity, vulnerability management |
| **Historical Context** | 10% | Matches patterns against previous verdicts (last 30 days) | SIEM / data lake |
| **SOP Alignment** | 10% | Verifies response matches org boundaries and policies; actions with verifiable reversibility mechanisms receive increased confidence weighting | Policy / SOP engine |

The percentages above are recommended defaults, not fixed constants. Weightings should be org-configurable so each organization can tune them to its own risk tolerance and telemetry maturity. What CARTS standardizes is the schema (the five factors and that they sum to 100%), not the exact split. Two rules keep the score honest:

A. **Confidence decay:** The “historical context” factor ages out on the rolling 30-day window so stale verdicts don’t inflate the score.

B. **A conflict penalty:** When two sources disagree about the same asset, the aggregate is penalized rather than averaged until multisource verification resolves it as specified under the safety pillar, to prevent actions based on conflicting or spoofed data..

The CARTS Specification defines the exact scoring model behind this matrix: how each piece of telemetry is typed into a catalogued fact whose strength depends on context, how independent sources corroborate one another, how contradicting evidence and conflicting sources lower the score, and how the final score is divided among the vendors whose telemetry produced it.

The following walkthrough shows a single investigation moving through the framework. This is a coordinated credential-stuffing campaign climbing from low confidence to a fully corroborated, auto-remediated verdict:

| SOURCE | INBOUND TELEMETRY (RAW LOG) | MAPS TO MATRIX FACTOR | | CONFIDENCE |
|---|---|---|---|---|
| **Threat Intelligence** IOC reputation feed | `ip=114.216.106.130 verdict=malicious brute_force=304 spray=33 window=24h` | → Suspicion & Behavioral | 30% | 15% |
| **Identity Provider** sign-in logs | `signin user=intuser@fetld.com ip=114.216.106.130 result=success mfa=none # Tor exit node, DE` | → Situational Awareness | 30% | 45% |
| **CMDB / Asset** inventory lookup | `asset=intuser role=cloud-admin env=prod criticality=HIGH` | → Business Context | 20% | 70% |
| **Cloud Audit Logs** API audit events | `create-instances count=300; set-bucket-acl bucket=fin-exports acl=public-read` | → Suspicion & Behavioral | 30% | 100% |

**Confidence gate: 100% ≥ 95% ✓**
**Threshold cleared and action is reversible → automation released.**

```
disable-user intuser@fetld.com --dir=identity && revoke-sessions --user=intuser --scope=cloud,saas
```

↩ Reversible — undo endpoint on file

If the score had stayed **below 95%**, the AI takes no action — it routes to a human analyst and publishes a full "why-not" audit trail showing exactly which telemetry was missing.

**Figure 4.** How inbound telemetry maps to matrix factors, moves the score, and gates automation at ≥95% (or any established threshold).

## Self-metrics and KPIs to measure the framework

The CARTS Framework promises measurable ROI; the following five key performance indicators (KPIs) make that promise auditable. Together they answer five questions a CISO and a partner each care about and are computed from the same telemetry and confidence scores the framework already produces.

| KPI | Question it answers | Primary stakeholder | Pillar served |
|---|---|---|---|
| **Confidence lift** | Does this vendor's data measurably improve decisions? | Vendor / procurement | Confidence |
| **Automation rate** | Are we acting at machine speed more often over time? | SOC leadership | Trust |
| **Reversible-action coverage** | How much of our response can we safely automate? | SOC / risk | Reversibility |
| **Risk-adjusted return on action** | Does each autonomous action avert far more threat than the blast radius it risks? | CISO / risk | Safety |
| **Transparency ratio** | Can every conclusion be traced to a source? | Auditor / CISO | Auditability |

## Operationalizing trust: The 4-step path

To ensure the CARTS Framework is adoptable, it follows a rigorous onboarding process that respects the CISO's need for control:

**1. Access:** Grant the AI access to the ecosystem telemetry (e.g., network, identity, EDR).

**2. Define:** Configure the "expectations" (e.g., SOPs, VIP lists, authorized tools).

**3. Audit:** Enable "show-your-work" mode where analysts review the AI's reasoning line by line to build trust.

**4. Automate:** Enable auto-remediation only when the CARTS score consistently hits the organization's risk tolerance (e.g., >95%).

## Operationalizing reversibility and safety

To further mature the automated response cycle, the CARTS Framework introduces specific operational requirements:

- **Explicit Declaration of Reversibility:** Before executing any remediation, the AI must explicitly declare if an action is reversible (e.g., "isolate host") or irreversible (e.g., "delete cloud resource"), adjusting the risk threshold accordingly.
- **Transparency of Inaction:** In instances where confidence thresholds are not met, the AI must provide a full audit trail documenting exactly why it did not take action, ensuring gaps in telemetry are visible to administrators.
- **Multisource Identity Verification:** For high-stakes actions, the AI is required to poll and verify asset identity across at least two distinct sources (e.g., identity provider and EDR) to prevent actions based on conflicting or spoofed data.

| Reversibility class | Examples | Auto-execution rule | What the vendor must publish |
|---|---|---|---|
| **Reversible** | Isolate host, disable user, quarantine file | Auto at ≥ 95%* | An undo endpoint + undo SLA |
| **Time-boxed** | Block IP / URL with TTL | Auto at ≥ 97%*, self-reverses on expiry | TTL support + expiry metadata |
| **Irreversible** | Delete cloud resource, wipe endpoint | Human approval always | Explicit "irreversible" flag + blast radius |

*These default values are recommended baselines for an organization's automation policy; organizations should adjust thresholds to match their risk tolerance.

## Model transparency and bias

Agentic decisions inherit the biases, blind spots, and quality of the underlying inference model. A framework that audits the evidence but hides the reasoner is only half transparent. CARTS therefore requires that any compliant agent disclose its model provenance as a first-class, auditable attribute. Vendors must disclose the model provider and version, where inference runs, and who controls it, because performance and known biases are model- and version-specific.

## Conclusion

By adopting the CARTS Framework, industry partners within the OSAIA can move beyond isolated silos. A partner's flow data is no longer just a log; it becomes a critical weight in a unified confidence score that empowers the SOC to act at machine speed against modern, AI-driven threats. By quantifying the impact of API-based telemetry, the CARTS Framework transforms data from a commodity into a verified confidence input, ensuring a transparent feedback loop where vendors can prove their worth and customers can make data-driven investment decisions.

**Learn more at Trellix.com.**


Copyright © 2026 Musarubra US LLC