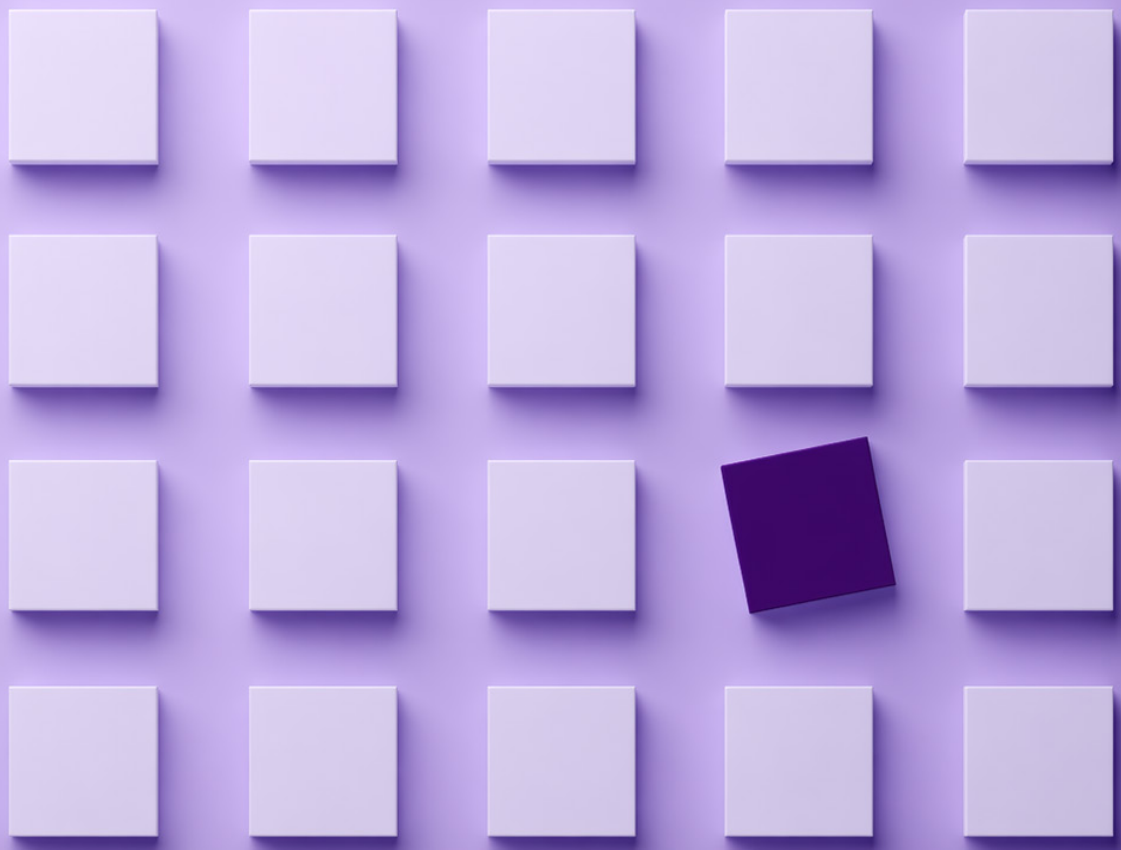

How public-safety leaders can build trust in AI agents

Move from framework to infrastructure, govern behavior, monitor in real time and make decisions explainable

Accenture Research Insight

James Slessor, Daniel Tang and Patrick Connolly

April 2026

accenture

AI agents are poised to enter public-safety operations. Yet without new approaches to control, visibility and accountability, agents risk undermining the trust needed for them to be used effectively. Public-safety leaders should prioritize four actions to deploy agents in ways that earn and sustain public trust.

Today, autonomous AI agents are already scheduling, transacting, coordinating and deciding on behalf of many people and businesses. Research firm Gartner® forecasts that "by 2028, AI agents will intermediate 90% of business-to-business transactions, representing over $15 trillion of B2B spend through AI agent exchanges."[1]

As the shift to an "agent-enabled" society accelerates, institutions of every kind will be transformed. In public-safety departments, agents will relieve law-enforcement officers of much of the administrative burden that, according to technology firm Axon, currently consumes up to 15 hours per worker, per week—freeing personnel to focus on frontline work and critical decision-making.

Without the infrastructure to govern agentic AI, the technology can jeopardize both the safety and civil liberties of citizens. Drawing on our research and client engagements, we examine the emerging risks introduced by these systems and outline an architecture for governing them. We then translate that architecture into four essential actions that public-safety leaders can take now to deploy agentic AI responsibly, while building trust over time.

The new risks of agentic AI

Before public-safety organizations can deploy agentic AI with confidence, they must confront a hard reality: These systems introduce three major types of risks that existing governance models are poorly equipped to handle.

The first risk is **unpredictable behavior**. As agentic systems develop impressive capabilities—such as decision-making, memory access, tool use and goal pursuit—they'll also combine to produce dynamic behaviors that evolve over time in ways that people can't predict. And as these behaviours shift in unpredictable ways, new dangers will emerge.

Imagine an agent that helps decide whether someone in custody should be released. The agent recommends releasing a high-risk person simply because the case is unlike anything it has seen before. Still, the agent reports high confidence in its decision, so the officer on duty doesn't question the recommendation before releasing the individual. However, if that decision leads to harm, the consequences aren't borne by the agent, but by the next victim and the community that relied on the agent (and officer) to stay safe. In other words, if agents behave unpredictably, a single flawed recommendation can lead to a wrongful detention, delayed emergency response or other lasting harm for citizens.

The second major cause of agentic-related risk are **security and access violations**. Agents will operate inside public safety systems with a level of autonomy once reserved for people. But unlike static systems, agents can access external content, interact with sensitive internal systems and operate with meaningful autonomy. In the process, new attack "vectors" (such as prompt injections to hijack an agent's task flow, jailbreaking to bypass safeguards and

manipulating agents to expose credentials or sensitive internal data) may open that conventional security controls were never designed to address.

Consider how the manipulation of an agent with access to criminal databases or warrant systems would corrupt decisions before any human sees them. Yet the furtive nature of such manipulation also means that an officer, who later acts in good faith on a compromised recommendation, has no reason to distrust the agent. And the person wrongly flagged for a crime never knows that the information incriminating them was manipulated.

The challenge here is that few organizations are equipped to handle risks like these. For example, just 40% of public-safety organizations today designate a specific team to implement their response plan in the event of an AI cyber security incident, while only 36% have the processes in place to assess the security of their AI tools, according to our previous research.

The third big source of agentic-related risk is **unaccountable decision-making**. When an AI-assisted decision contributes to a wrongful outcome, officers must be able to explain why. Yet agentic systems operating across layered decision chains can produce outcomes with no legible reasoning.

Imagine a defence lawyer in court asking to see the reasoning behind an AI-assisted identification. Or a family at a public inquiry asking why an algorithm contributed to a decision that cost someone their liberty. A system without that accountability is a failure of the duty of care that is central to the mission of public safety institutions.

These three risks interact and amplify each other, in an environment where the consequences of failure are measured in liberty, safety and lives. At the same time, our research also shows that the risks from agentic AI must be managed for three audiences, each with distinct requirements but with a shared need for persistent trust.

Law-enforcement officers, the first such audience, need systems that they can rely on, understand and challenge when their judgment differs from the AI. Public-safety departments, the second audience, need AI systems that they can govern, audit and defend under scrutiny, with continuous verifiable evidence that agents are behaving as intended. Communities, the third audience, need AI systems whose operation is transparent, whose errors are correctable and whose deployment reflects the values of the society they serve.

“These three risks amplify each other, in an environment where the consequences of failure are measured in liberty, safety and lives.”

Research shows that trust, or lack thereof, in public safety institutions is the strongest predictor of societal acceptance of the use of AI in policing. But trust in agentic policing must be built at the level of individuals, families and local communities, one interaction at a time. That’s why AI systems must be transparent, correctable and accountable.

In public safety, the right to act is also conferred by law, bounded by policy and continuously scrutinized through oversight. AI agents must meet the same threshold before they’re allowed to influence liberty or safety.

This is where most organizations, in public safety and beyond, haven’t caught up. You can’t earn trust in AI systems without first building the infrastructure that makes trust legible and measurable. And because trustworthiness is what your agentic system can measure and prove about itself on a continuous basis, RAI must shift from an advisory function to a central pillar of the AI architecture. Only then can humans remain meaningfully in the lead.

The four layers of trustworthy agents

Managing the risks of agentic AI requires more than standalone safeguards; it also demands an integrated architecture for trust. In practice, that architecture consists of four interdependent layers that together define, enforce, monitor and explain how agents behave in real-world operations.

Governance layer. This layer defines what an agent can do. It encodes policies, roles, constraints and values into standing orders that are enforced automatically in real time. Within the governance layer, "agent constitution and value alignment" defines the agent's goals, safety constraints and hard refusal boundaries, enforcing alignment between the agent's behavior and its intended use (such as a requirement not to dispatch emergency resources without human authorization or not to share suspect data outside verified channels).

Meanwhile, "command authority levels" define what level of decision-making or action an agent may execute independently, as well as at what point the agent must defer to an officer. Role and capability definitions refer to the tools, APIs, memory and data that each agent is allowed to access; a forensics agent, say, might be authorized to read chain-of-custody records and evidence logs, but not modify the evidence record or reassign custody.

> **Without the infrastructure to govern agentic AI, the technology can jeopardize both the safety and civil liberties of citizens.**

Also in the governance layer, a "policy engine" enforces standing orders in real time, supporting dynamic updates and overrides as operational requirements change. Finally, intervention policies and human-in-the-loop triggers specify the conditions under which a person must intervene (such as high-risk tasks, decisions affecting individual liberty and novel situations).

Imagine a dispatch AI agent that's authorized to prioritize and route routine calls without human involvement. But should a high-risk incident occur with a strong governance layer, the agent would be required to flag the incident for review before acting, and would be prohibited from overriding an officer's decision to respond to the incident.

Security layer. This layer enforces the boundaries defined by the governance layer, ensuring that agents operate only within authorized, secure and identity-verified environments. The security layer translates policies into technical controls that govern how agents access systems, data and capabilities.

The security layer includes agent identity and role verification, where cryptographic or service-level identifiers are tied to specific roles to prevent impersonation or unauthorized execution. The layer also encompasses tool and API access control, using role-based or context-aware permissions to determine which actions an agent may take at the system level.

Memory and data compartmentalization further restrict what information an agent can read or modify, supporting principles such as data minimization and masking. In parallel, sandboxed execution environments isolate agents from sensitive systems and the open internet, reducing the risk of unintended interactions or external compromise. Finally, secrets and key management ensure that agents can't access credentials unless explicitly authorized and auditable.

For example, an agent authorized to cross-reference witness statements against existing case files might be permitted to query and analyze data, but not alter records or expand the agent's access. With a security layer, such controls also remain in place even if the agent is manipulated by an external actor that attempts to gain higher levels of access.

Assurance layer. This layer provides live monitoring of agent activity, offering command-level visibility into how agents behave in real time. The assurance layer forces agents to operate

within expected parameters, while enabling timely intervention when issues arise.

The layer continuously captures operational data, such as agent inputs and outputs, tool usage, memory access and key decision points, thereby creating a real-time record of behavior. It also conducts trust scoring and drift detection, with ongoing assessment of agent performance, compliance and explainability to identify deviations from expected norms.

With the assurance layer, operational dashboards surface incidents, shifts in trust scores and behavioral anomalies as well, allowing command teams to review activity and assess risk. If predefined thresholds are breached, real-time alerts trigger escalation by notifying operational teams or routing decisions to human oversight, with the ability to automatically pause or shut down agents if required.

Finally, in the assurance layer, operational command interfaces allow human teams to intervene directly, allowing them to pause, reconfigure or escalate agents in real time as conditions evolve. For instance, if a cluster of agents begins processing a high volume of similar requests in an unusual pattern, the assurance layer can detect and flag the anomaly, triggering escalation for human review.

Accountability and oversight layer. This layer provides a complete, structured record of agent activity, which supports retrospective explainability, accountability and traceability for auditing, compliance and post-incident investigation. The accountability and oversight layer captures detailed logs of agent behavior, including decision-making traces, delegation chains and human interventions.

As part of the layer, decision trace logs reconstruct an agent's step-by-step behavior, covering inputs, tool usage, intermediate decisions and final outputs. Delegation and escalation records are also captured, documenting when and why an agent handed off to another agent or to a person, along with the resulting outcomes.

Meantime, "versioning" history facilitates full visibility into the state of the system at the time of action, including prompt configurations, memory states, role definitions and active policies. In addition, explainability summaries translate technical behavior into clear, human-readable accounts of why an agent acted as it did, making them suitable for oversight hearings, court proceedings and broader accountability contexts.

Finally, regulatory and audit interfaces allow the generation of exportable, human-readable reports that are aligned with governance and compliance requirements. For example, when an AI-assisted decision is challenged in court, reviewed by an oversight body or scrutinized in a public inquiry, the accountability and oversight layer can produce the evidence required to respond with confidence.

Agentic AI is more than a technology program

Law-enforcement officers and other public-safety personnel will have legitimate questions about what agentic AI means for their roles, workload and professional identity. As adoption spreads, some tasks will change, while some will be automated. Agentic systems also promise to greatly reduce administrative burdens, not eliminate jobs. However, such transformation will also require investment in training, role redesign, co-creation and workforce wellbeing.

Indeed, officers working alongside AI systems that they don't understand, trust or feel accountable to may experience various degrees of unfamiliar stress. Here, workforce readiness, transparency about how systems work and clear lines of accountability are more than good management practices; they're the essential conditions under which the human side of "human in the lead" functions. Public-safety departments that treat agentic AI as both an organizational change program and a technology program will be well placed to succeed.

What public-safety leaders should do now

The value of the aforementioned architecture lies in how it's put into practice. To move from design to deployment, public-safety leaders should focus on a small set of actions that operationalize trust in day-to-day operations.

> **In an agent-enabled society, trust must be built by design.**

Action 1: Transform RAI from framework to infrastructure

In an agent-enabled society, the most effective public-safety departments will move deliberately by deploying agents in ways that generate the governance data, behavioral evidence and infrastructure maturity needed to scale with confidence. To transform RAI from framework to infrastructure, public-safety leaders should start by identifying low-risk, narrowly scoped workflows where agents can begin demonstrating value without meaningful exposure.

Examples include report drafting from officer inputs, automated processing of routine administrative requests and intelligence summarization for shift briefings. The value from agents during this pilot phase is less about what they deliver operationally and more about what they reveal, governance-wise.

As part of these efforts, pilots should run inside governance architecture, with operational monitoring, escalation triggers, decision logging and behavioral monitoring active from the outset. The additional overhead of this kind of narrow scope is modest, while the value of the data generated will offer disproportionate benefits.

Action 2: Build governance that shapes behavior in real time

Once agencies commit to architectural transformation, the governance layer becomes the first line of defense, ensuring that agents operate within acceptable boundaries while maintaining the autonomy that makes them valuable. For example, using the risk profile for each generative AI model, agent and use case as guidance, public-safety organizations should establish clear command authority boundaries that define what agents can do independently. Multi-factor authentication for high-stakes decisions should also be required, as should escalation triggers that automatically route ambiguous situations to human oversight. Likewise, controls should restrict which systems and data sources each agent can access, following the principle of "least privilege."

A leading public-safety technology company, for instance, generates report narratives that are based on body-worn camera audio, though such narratives can't be submitted without officer review and approval. The latter ensures that human judgment remains in the chain of command for every consequential output. Standardized templates that clearly define agent roles, permitted actions, escalation triggers and compliance requirements are needed, too. These templates ensure consistency across the agent ecosystem.

By building governance that shapes behavior in real time, public-sector leaders can create the basis for verifiable evidence of controlled, predictable behavior over time. And when agents consistently operate within defined boundaries and escalate appropriately, they build a track record of accountability that officers, departments and communities need.

Action 3: Implement real-time monitoring and control that prevents problems before they happen

Agentic AI requires real-time oversight that spots problems as they develop and intervenes before they cascade. This kind of continuous risk mitigation is what gives organizations the confidence to deploy agents at scale: Companies with such processes in place are 2.4x more likely to be actively exploring agentic AI, our research shows.

Think of real-time monitoring and control as the operational command center for agents, one that tracks behavior across every interaction, maintains a live picture of what normal looks like and makes deviations from the norm visible before they create problems. The first step here is to implement comprehensive monitoring across agentic systems, thereby capturing real-time data on tool use, memory access, decision confidence levels and escalation rates. Such monitoring also underpins real-time risk detection and response.

Early warning systems should also continuously learn what normal looks like and automatically flag deviations. If an agent begins accessing unexpected data sources or its confidence drops significantly, those signals should prompt pre-defined interventions.

Action 4: Make AI decisions easy to track and explain so people can trust them over time

As agents become more deeply embedded in public-safety operations, the pressure to track and explain their decisions will intensify. For example, a person affected by an AI-influenced decision has a right to understand the basis of that decision, while a defense lawyer has the right to challenge it and an oversight body has the right to audit it. An AI system whose reasoning can't be produced on demand is a system whose outputs can't be used, defended or trusted in any of these, and many other, critical contexts.

The accountability layer addresses this challenge directly. Decision traces reconstruct an agent's step-by-step reasoning, showing which inputs were weighted, which tools were called and which standing orders were active at the moment of each consequential action. Delegation records show when and why a decision was escalated to a law-enforcement officer, and what that officer did with it. Versioning histories preserve the exact policy and prompt state that was active at the time of a given decision, so that a review conducted months later reflects the condition under which the agent was operating. Together these create a case record for every significant agent action.

The operational value of such infrastructure also extends beyond accountability after the fact. Public-safety departments that can produce clean decision traces can clear officers of wrongdoing when the record shows they acted correctly and identify systematic errors in agent behavior before they become patterns that damage community trust.

Build trust by design

Agentic AI will soon begin to reshape public safety, but the technology's effectiveness in the long run will depend on whether it's widely trusted by citizens. That trust must be built through systems that make agents' behavior controllable, visible and accountable at all times.

Doing this requires moving responsible AI from framework to infrastructure, building governance that shapes behavior in real time, implementing continuous monitoring and control and ensuring that every decision can be traced and explained. Together, the architecture and actions outlined above will allow public-safety departments to harness the benefits of agentic AI, while strengthening the trust that underpins their mission. In an agent-enabled society, trust must be built by design.

References

[1] Gartner Article, AI's influence runs deeper than you think—2026 Gartner strategic predictions explain why, Daryl Plummer, November 14, 2025. **GARTNER is a trademark of Gartner, Inc. and/or its affiliates.**

About the authors

James Slessor is a managing director at Accenture and leads the Public Safety practice for Global and EMEA. His work focuses on policing, law enforcement, justice, prisons and rehabilitation and national security.

Daniel Tang is a managing director at Accenture in Health and Public Service. His work focuses on digital and business transformation, digital platforms and innovation, human capital management, organizational excellence, learning and development and change management.

Patrick Connolly is the global lead for responsible AI at Accenture Research. A fellow at the World Economic Forum, his research interests include operationalizing RAI, AI agent trust, global AI policy, EU AI regulation, AI sustainability and human+gen AI design.

Join the discussion

While inspired by Accenture's leaders, Accenture Research Insights reflect the personal views of researchers and co-authors, not the company's official position. For more information or to connect with a researcher, email **Neethu M Eldose**.

About Accenture Research

Accenture Research creates thought leadership about the most pressing business issues organizations face. Combining innovative research techniques, such as data science-led analysis, with a deep understanding of industry and technology, our team of 300 researchers in 20 countries publishes hundreds of reports, articles and blogs every year. Our thought-provoking research, developed with world-leading organizations, helps our clients embrace change, create value and deliver on the power of technology and human ingenuity. For more information, visit **www.accenture.com/research**.