Binding agent intent and implementation with agentic data trust

An Intellyx Report for Trust3 AI by Jason English

There are a host of new "AI-powered" cybersecurity automation products on the market claiming to autonomously hunt down threats, patch vulnerabilities, and/or accelerate MTTR through a single pane of glass.

Whether due to market pressure or a fear of missing out, enterprises are moving quickly to get agentic AI into production. So much so, that many of them have gotten out over their skis.

Just this year, we've had an unpatched React frontend give out LexisNexis records, logins and personal data for more than 400 thousand users, a "rogue agent" inside Meta that helpfully opened up company records for all employees to help answer one engineer's issue, and a leak of Anthropic's entire source codebase, followed by remediation agents that took down thousands of Github repos.

Agentic AI projects may seem to be safely grounded on the right data in the pilot phase, but once multiple agents proliferate in production, they will find, access, and share sensitive data and secrets with each other in non-deterministic ways as they complete tasks.

©2026 Intellyx B.V.

The bigger they come, the harder they crash

Now that thousands of agents are interacting with each other both inside and outside the company firewall, Sev1 incidents become inevitable. Not because the developers at the company don't care about security, or that they lack talent. There is no easy way to predict the data sharing behaviors that could happen, much less control so many agents at once.

A ski instructor will tell you to 'trust your skis.' But when it comes to avoiding injury, the skis aren't as critical as the bindings. These complicated mechanisms must be tuned to keep your boots tightly locked in as you lean into the turns, until the moment there's too much tilt, and the skis need to be loosened and end the ride, according to your skill or risk tolerance level.

If we're going to safely control our descent downhill from high-level agent design concepts to complete a successful agentic AI run, we need to trust the binding between the business intent of our agents, and their grounding in the data the agents can access and change once they are in production.

Taking stock of agent and data inventory with discovery

In our previous chapter, my colleague Jason Bloomberg introduced Trust3 AI and their control plane for data and AI governance and security, with an underpinning metadata and relationship graph. To start building such a control plane, you need to know what you already have in place.

©2026 Intellyx B.V.

The first pillar of trust is focused on discovery of the agent fleet and the data landscape it operates in.

To answer this, Trust3 AI provides automated Agent DOS (Discovery, Observability and Security). Discovery, the first pillar, aids in conducting a comprehensive inventory of agents in production, identifying their developers and authorized business owners, and locating their source code and requirements documentation. Then Observability, the second pillar, surveys every data connection the production agents can use, whether through an API or MCP server, or a natural language prompt in a website or mobile app.

Not surprisingly, the process can produce a lot of unanticipated results. Since everyone wants to move fast, discovery will turn up rogue agents with uncertain owners, secrets or credentials that are logged in a shared file, and database records that are unencrypted or open to external queries. Time to tighten the bindings!

Security versus unpredictable agent data use

Security is the third pillar of Trust3 AI's agentic DOS control plane. What is the intended purpose of the agent, and what is the minimum scope of data it should have access to and retain, in order to reach its goals? What if one agent needs to talk to other agents – what data should be seen or shared by agents in production?

There are certain behaviors that should never be allowed, such as logging PII (personally identifiable information) in an open text file or Moltbook, a social media site for agents. Beyond that, there are a range of criteria for evaluating how trustable a given agent is, which will vary for different stakeholders in the organization. A compliance officer may be concerned with auditability for a given standard, while a data security officer will care about zero-trust data sharing and read/write privileges.

©2026 Intellyx B.V.

Trust3 AI recently released a Trustscore rating that reviews and measures agents on several factors such as data security, compliance and quality, in order to give each agent a rating that demonstrates its fitness for purpose and highlights potential concerns before it is placed within a business workflow.

Continuous observability and policy monitoring

Once agents and their enabling technologies such as MCP servers and RAG inference training databases are in production, we need continuous observability in order to monitor data in motion and at rest across the agentic AI estate.

Here's where policy design and enforcement are key. Policies should set forth clear guardrails with fine-grained access controls, so that any out-of-bounds agent data handling can trip off an automated remediation or block action, or alert the responsible owners or teams of the potential violation.

That 'fine-grained controls' part makes building policies hard. If someone had to write regex or navigate a cascading set of restrictions in a conventional IAM or security tool for every instance of every agent interacting with every type of data it could access, it would never get done!

Here's where conversational AI really shines. Trust3 AI allows even non-engineers to self-service define and adjust policy across groups of agents and data sources by importing documentation from any compliance regime (GDPR, EU AI Act, HIPAA, PCI, etc.) and then asking for adjustments in plain English. Then, these policies can be monitored and enforced through an observability dashboard.

©2026 Intellyx B.V.

Real-world use cases for binding agent policy

Just like our skiing metaphor for bindings, there is no 'one size fits all' when it comes to governing agents. That is why we need to separate the concerns of setting agent data policies from the development and operation of the agents themselves.

For a financial use case, let's consider a loan app with 3 agents, though there would likely be more:

- **Loan pre-check agent** that can only gather basic income and credit information and verify that the customer's identity exists from public sources.
- **Loan quoting agent** that takes in secure PII and only uses it as an in-process secret to check credit services and validate attestations of income, job history, home ownership or business registrations, and then sends genericized info to a quoting service to return pricing details for a quote.
- **Loan application agent** that handles lots of customer data input, communicating only required least-privilege data through secure encrypted external agents and sources to complete and produce the loan terms and documentation for the lessor's agent and the customer to approve. This agent would be under the tightest governance.

You can extrapolate from there, for instance a healthcare app for a chain of clinics or pharmacies would want to start from HIPAA guidelines. A patient may interact with an admitting agent, insurance agent, and a medical records agent to schedule visits or treatments without knowing how little data each of the agents in the process was allowed to access.

©2026 Intellyx B.V.

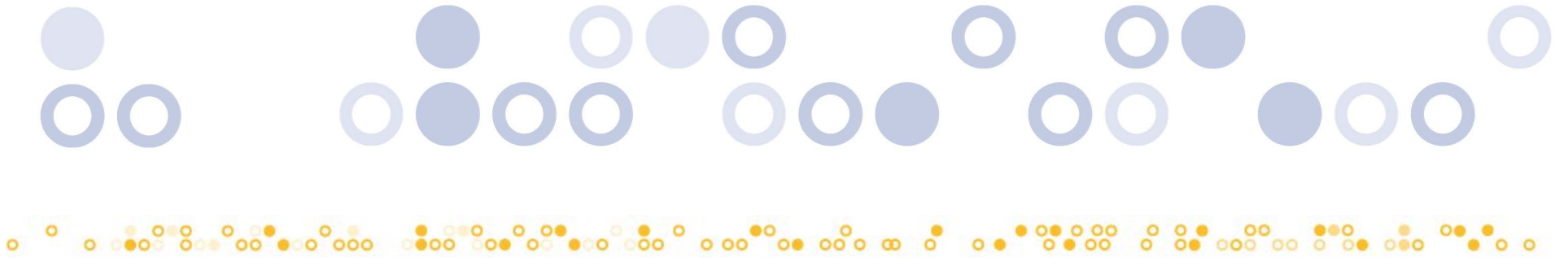
The Intellyx Take

If you believe today's AI vendors, data platform providers, and open source projects will responsibly govern your data, I have a bridge [err, gondola] to sell you.

We need to protect data from non-deterministic agent behaviors. We also need to make agentic AI data governance easy, or it won't happen.

Binding the intent of an agent to its actual use once implemented isn't just a data governance or compliance concern, because it demonstrates the business utility of the agent within the application environment – which is critical for measuring value and return on investment beyond the agentic hype.

©2026 Intellyx B.V. Intellyx is editorially responsible for this content, and no AI was used to write it. At the time of writing, Trust3 AI is an Intellyx customer. None of the other organizations mentioned here are Intellyx customers. Image Source: Adobe Image Express.



©2026 Intellyx B.V.