```
<!-- Page metadata for Replit build -->
```

Suggested URL slug: collective-intelligence-trust-stack Meta description: A seven-layer reference architecture for evaluating whether an AI or agent system is actually ready for real-world deployment.

The Collective Intelligence Trust Stack: A Reference Architecture

As artificial intelligence (AI) moves from single-model tools to networks of autonomous agents, organizations need more than good intentions to manage the risk. They need a shared map of where trust is created and where it can fail. The Collective Intelligence Trust Stack is TrustandSafety.Institute's (TSI's) working reference architecture for that purpose: seven layers that apply to any AI or agent system, regardless of industry or use case.

1. Human Purpose

Every layer of the stack rests on one question: does this system expand human agency, creativity, and opportunity, or does it simply reduce headcount? A system built purely to cut labor costs is a fundamentally different proposition from one built to help people make better decisions, work more safely, or create new kinds of work. TSI treats a clear answer to this question as a precondition for trust, not an afterthought.

2. Identity and Authority

In a system with autonomous agents, every actor, human or machine, needs a verifiable identity and a defined scope of authority. An agent should never act simply because it is technically capable of acting; it should act because it has been authorized to, within boundaries that can be monitored and revoked. Efforts such as Project NANDA, a Massachusetts Institute of Technology (MIT)-affiliated initiative building open standards for how agents discover and verify one another, are early examples of infrastructure at this layer (Project NANDA).

3. Privacy and Data Stewardship

Agents often need to collaborate on sensitive information. That collaboration has to happen without exposing data beyond what is necessary, and without violating the consent under which the data was originally collected. Privacy needs to be enforced through the design of the system, not promised through a policy document that no one reads.

4. Explainable Execution

Every consequential action an agent takes should leave a record: what was requested, what authority was granted, what data was used, what reasoning path was followed, and what a human ultimately approved. Without this record, accountability after the fact is close to impossible.

5. Simulated Intelligence

Before a system takes on high-stakes, real-world decisions, it should be tested against alternative scenarios, edge cases, and adversarial conditions in simulation. Simulation does not eliminate uncertainty. It reveals where a system is fragile before that fragility becomes a real incident.

6. Runtime Governance

Once live, systems need the ability to monitor themselves, escalate uncertain situations to a human, pause when something looks wrong, and learn from failures instead of repeating them.

7. Ecosystem Accountability

The final layer extends beyond any single organization. Definitions, benchmarks, incident reporting, and standards need to be shared across the community building and deploying these systems. Trust and safety improve faster when organizations are not solving the same problems in isolation from one another.

Using the Stack

Organizations can use the Trust Stack as a self-assessment tool. For any AI or agent system under development, the standard-setting question at each layer should be answered before deployment, not after. A system that cannot answer the identity and authority question, for example, is not ready for production, regardless of how well its underlying model performs on benchmarks.

TSI will continue to publish and refine this framework as agentic AI matures, alongside reference implementations and assessment tools organizations can apply directly to their own systems.