# TSI MANIFESTO

*Trust and Safety Institute*

## PART I: THE WORLD AS IT IS

People are deploying AI faster than they can govern it. This is the current state of affairs, spoken plainly, observed across industries, and confirmed by those building, buying, and regulating artificial intelligence at every level.

Autonomous agents are being activated without defined authority. Without explainability. Without human checkpoints. And without a credible answer to the question every executive is quietly asking: what happens when this goes wrong, and who is accountable?

The concern is specific. AI agents, autonomous systems that access data, trigger workflows, move resources, and make decisions, can act outside their intended boundaries with no mechanism for a human to understand, reconstruct, or override what happened. This is the architecture most organizations are currently operating under.

Gartner projects that more than 40% of agentic AI projects will be canceled by end of 2027 due to inadequate risk controls, unclear business value, and costs that outpaced governance. The tools arrived before the infrastructure to govern them.

The knowledge gap is real. Most leaders making AI deployment decisions lack the technical foundation required to govern what they are releasing. Governance is being treated as a compliance layer, applied after the fact and bolted onto systems already in motion.

In healthcare and beyond, AI is making consequential decisions without transparency. When the reasoning behind a decision is invisible, accountability disappears with it.

Workers across every sector are watching this unfold. The fear of displacement is a reasonable conclusion when no framework exists to define the human role or put human judgment in the loop with real authority.

This is a governance problem. The absence of governance is an active underuse of the most important resource in the intelligence equation: human purpose, human judgment, and human accountability.

## PART II: THE WORLD WE ARE DECLARING INTO EXISTENCE

The Trust and Safety Institute exists to make collective intelligence legible. As a field. With vocabulary, standards, protocols, events, university chapters, assessments, and reference implementations that any organization, government, or community can apply. That is what TSI is building, and that work begins now.

The foundation of this field rests on one equation: **AI + BI = CI. Artificial Intelligence plus Brain Intelligence equals Collective Intelligence.**

Artificial Intelligence brings scale, automation, pattern recognition, and simulation. Brain Intelligence, the human side of the equation, brings judgment, ethics, purpose, context, creativity, and accountability. These are qualities that cannot be automated. They can only be honored, protected, and designed into the systems we build. When both forces operate together through explainable, privacy-preserving, and governed networks, intelligence compounds safely. Humans are the ceiling of this equation.

The human is the most trustworthy part of any AI system because the human carries the why. Purpose lives with people. Consequence is felt by people. When an autonomous agent makes a decision, a human somewhere bears responsibility for that outcome. TSI declares that this relationship must be designed in from the beginning, with the human in the loop as the foundational structure.

This future is built on five principles. **1)** AI must begin with human purpose. **2)** Every system must be grounded in a verifiable identity and a clear boundary of authority. **3)** Privacy must be preserved by protocol. **4)** Every decision made by an agent must be explainable and reconstructable. **5)** Before any mission-critical system is deployed, it must be tested through simulated intelligence against real-world conditions.

These principles are the architecture of the world we are creating. A world where AI progress expands human opportunity. Where agents operate under clear authority. Where decisions are explainable at every layer. Where trust is earned through design, verified through evidence, and sustained through accountability.

## PART III: THE ARCHITECTURE AND OUR COMMITMENTS

Declaring a future is the beginning. Meeting it requires architecture and action. TSI puts forward the Collective Intelligence Trust Stack as the operational framework for governing AI in the age of agentic systems. It is a seven layer model, and every layer carries equal weight in the structure.

**1)** Human Purpose, the foundational question every AI system must answer before anything else: what human benefit is this system designed to advance, and what is its intended impact on the people it touches? **2)** Identity and Authority, which establishes verifiable identities and explicit authority boundaries for every human, agent, model, tool, and dataset operating within a system. **3)** Data Stewardship and Privacy, built on privacy-preserving protocols, informed consent, data minimization, federated learning, and confidential computing. **4)** Explainable Execution, a full and reconstructable record of what was requested, what authority was granted, what data was accessed, what reasoning was followed, what humans approved, and what outcome occurred. **5)** Simulated Intelligence, the practice of testing systems against simulated conditions, adversarial scenarios, and ethical constraints before real-world deployment. **6)** Runtime Governance and Incident Learning, which covers self-monitoring systems, escalation protocols, failure capture, and the institutional memory that grows from every incident. **7)** Ecosystem Accountability, encompassing shared definitions, benchmarks, maturity models, certification standards, and the public narrative that holds the entire field to account.

This is the stack. Every organization deploying agentic AI should be able to locate themselves within it, measure their maturity against it, and use it as a roadmap forward. TSI commits to making that possible.

We will develop shared standards and vocabulary so that trust and safety practitioners across industries, governments, and research institutions are speaking the same language. We will build a practitioner community that brings together technologists, ethicists, policymakers, economists, and domain experts to advance this field together. We will host convenings, webinars, and events that make the work of governing AI visible and accessible. We will establish university chapters to bring the next generation of practitioners into this field with the right foundations. We will create maturity assessments and certification pathways so that organizations can measure, demonstrate, and improve their governance posture. And we will run real-world pilots with enterprises and governments to test, validate, and refine the trust stack in practice.