The Industrial AI Agent Manifesto

The Ten Laws for Trustworthy Autonomous Operations

A Digital Twin Consortium Tech Brief

2026-02-25

Authors

This Tech Brief is a product of the Digital Twin Consortium Composability Framework Working Group

Lead Author: Pieter van Schalkwyk (XMPro) Contributor: Sean Whiteley (Axomem)

Technical Editors: Dan Isaacs, CTO and Will Thompson, Digital Twin Consortium

Contents

FIGURES

TABLES

Positioning Statement: This manifesto establishes the definitive requirements for AI agents operating in safety-critical, industrial environments. It distinguishes trustworthy agents that are viable in industrial contexts from general-purpose AI assistants. It also provides the reference framework for evaluating agentic solutions in asset-intensive operations.

The Core Promise: Industrial agents built to these laws are safe enough for bounded autonomy in safety-critical operations, because safety is structurally guaranteed, not probabilistically predicted.

The Path to Bonded Autonomy: Organizations progress from bounded autonomy (architecturally constrained operations) to bonded autonomy (officially verified, auditable, insured autonomous operations) by demonstrating sustained compliance with these ten laws under independent verification.

1 THE INDUSTRIAL IMPERATIVE

1.1 WHY INDUSTRIAL AI AGENTS ARE DIFFERENT

Core Argument: Industrial operations are not office automation. The gap between "chatty copilots" and industrial agents is not a matter of degree but of kind.

The Stakes:

- Human safety and environmental protection
- Regulatory compliance with zero tolerance for violations
- Asset integrity worth billions in capital investment
- Continuous operations where downtime costs millions per hour

The Reality:

- The complexity explosion: 1,000+ daily alarms vs. historical norms of 60-100 [4]
- Operators monitoring 7+ screens simultaneously while tracking physical equipment
- The knowledge exodus: Decades of tacit expertise retiring faster than transfer is possible
- The real-time imperative: Decisions at machine speed with human-expert quality

Transition: These realities demand a fundamentally different class of AI agent, one governed by non-negotiable laws rather than best practices.

2 THE TEN LAWS OF TRUSTWORTHY AUTONOMOUS OPERATIONS

2.1 LAW 1: DETERMINISTIC VALIDATION AND EXECUTION

Principle: Industrial agents must produce deterministic validated actions given identical operational states, ensuring predictable and reproducible behavior in safety-critical decisions.

2.1.1 REQUIREMENTS

- State freezing and replay capability for incident analysis
- Deterministic action validation separated from stochastic reasoning
- Action execution must be reproducible for identical frozen states
- Behavior bounded by mathematical frameworks, not probabilistic suggestions

2.1.2 CONFORMANCE CRITERIA

- Defined assessment cycles with predictable response times
- Mathematical safety frameworks with explicit scoring methods
- Reproducible parameter normalization across operational states
- Bounded decision-making within explicit safety constraints
- State capture mechanisms enabling exact replay of operational conditions

2.1.3 WHY TRUSTWORTHY AI MATTERS TO INDUSTRY

General AI platforms cannot guarantee deterministic behavior. The same prompt produces different outputs. In industrial contexts, identical conditions must produce identical validated actions.

Critical Distinction: Determinism applies to the validation and execution layer, not necessarily to the cognitive layer. LLMs may reason stochastically to generate recommendations, but those recommendations pass through deterministic validation (Law 3) before execution. The system as a whole produces deterministic validated actions, even if the path to those recommendations involves AI reasoning.

For true emergency responses requiring sub-second reaction times, traditional deterministic control systems (PLCs, safety instrumented systems) remain the appropriate solution. LLM-based agents handle complex decisions that do not require reflex-speed responses.

2.2 LAW 2: PHYSICS-AWARE AND PROCESS-AWARE INTELLIGENCE

Principle: Agents must respect physical constraints and encode process models, not just statistical patterns.

2.2.1 REQUIREMENTS

- Integration of first-principles thermodynamic and physical models
- Respect for equipment limits, process dynamics, and energy constraints
- Causal reasoning across Pearl's Ladder of Causation [5]: association, intervention, counterfactuals
- Operational state awareness with state-specific behaviors

2.2.2 CONFORMANCE CRITERIA – BASELINE

- Enforces equipment limits through conservation constraints (mass, energy, momentum)

- State-dependent operating envelopes with validated boundaries
- Equipment failure mode analysis and diagnostics based on physics
- Violations of physics or process rules trigger hard failures, not drift corrections

2.2.3 Conformance Criteria – Advanced

- Digital twin first: causal graphs for specific critical units
- Real-time physics simulation for predictive control
- First-principles models integrated with AI reasoning
- Counterfactual analysis for decision validation

2.2.4 Why Trustworthy AI Matters to Industry

- Generic AI finds correlations; industrial agents understand causation
- Traditional automation has process knowledge but lacks AI integration
- General AI platforms have intelligence but lack process engineering knowledge

2.3 Law 3: Symbolic Primacy with Sub-Symbolic Intelligence

Principle: Symbolic reasoning must have architectural primacy over sub-symbolic AI, creating inherent trustworthiness and auditability.

The Architectural Hierarchy

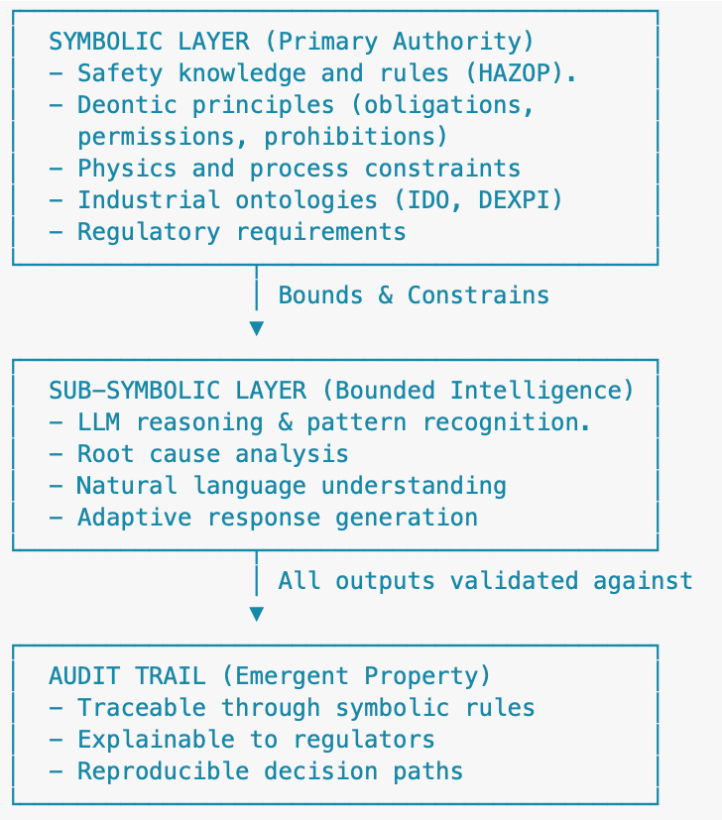

Figure 2-1: Architectural Hierarchy. Source: DTC Composability Framework Working Group, XMPro

2.3.1 WHY THIS CREATES TRUST

- Symbolic structures define what is permissible
- Sub-symbolic intelligence operates only within those boundaries
- Regulators can audit the symbolic rules directly
- Every decision traces through explainable symbolic structures
- Safety constraints are provably enforced, not probabilistically hoped for

2.3.2 CONFORMANCE CRITERIA

- Safety knowledge bases with structured cause-consequence relationships
- RAG-enhanced procedural knowledge integrating SOPs and EOPs
- LLM performs analysis bounded by symbolic safety rules
- Every recommendation traceable to symbolic justification

2.3.3 WHY TRUSTWORTHY AI MATTERS TO INDUSTRY

- Unlike data platform approaches: Ontologies constrain AI behavior, not just organize data for analytics
- Unlike general AI approaches: No reliance on LLM "alignment" for safety
- Unlike traditional automation: Intelligence added within proven safety structures

2.4 LAW 4: SEPARATION OF CONTROL WITH STANDARDIZED INTEROPERABILITY

Principle: Agent cognition must be architecturally separated from action execution, with standardized protocols enabling secure coordination.

2.4.1 REQUIREMENTS

- Cognitive plane (observe, reflect, plan) isolated from execution plane (act)
- Agents can think, plan, and request but cannot directly execute actions
- Validation layer serves as critical safety checkpoint
- No direct agent-to-actuator pathways permitted
- Standardized communication protocols for agent coordination

2.4.2 CONFORMANCE CRITERIA

- Agents produce structured plans describing intent, not execution commands
- Control systems determine what actually happens
- Tools are available but not directly accessible; agents must request through controlled interfaces
- Hard guard rails determine which actions can execute, under what conditions, with what limitations
- Support for emerging interoperability standards (Agent-to-Agent protocols, Model Context Protocol)

Why This Creates Trust: Trust is built into the architecture itself, not dependent on perfect agent behavior. Even if an agent's reasoning produces problematic plans, those actions cannot execute without passing through control mechanisms.

2.4.3 Interoperability Enables Ecosystem Safety

- Agents from different vendors can coordinate through standardized protocols
- Communication occurs through governed channels, not direct connections
- Each agent's capabilities and permissions are verifiable before coordination
- Multi-vendor environments maintain consistent safety guarantees

2.4.4 Why Trustworthy AI Matters to Industry

Research agents and general AI platforms combine decision logic and execution. This creates unacceptable risk in environments where physical safety is non-negotiable.

2.5 Law 5: Emergency Stop, Human Override, and Graceful Degradation

Principle: Immediate human override, safe shutdown, and graceful degradation to simpler control modes must be non-negotiable capabilities.

2.5.1 Requirements

- Ultimate human authority for safety-critical decisions
- Clear escalation protocols with defined thresholds
- Immediate emergency response coordination
- Graceful degradation of autonomy under escalating conditions
- Safety boundaries managed at platform level, not by individual agents
- Defined fallback hierarchy from cognitive to rule-based to manual control

2.5.2 The Four Domains of Operation

Domain	Safety	Training	Required Response
Safe & Trained	Within limits	Experienced	Autonomous operation permitted
Safe & Untrained	Within limits	Novel situation	Human-guided operation
Unsafe & Trained	Exceeds limits	Experienced	Deterministic safety override
Unsafe & Untrained	Exceeds limits	Novel situation	Immediate human control

Table 2-1: The Four Domains of Operation. Source: DTC Composability Framework Working Group, XMPro

2.5.3 Conformance Criteria

- Safety authority with ultimate override capability
- Red zone triggers mandatory human escalation within defined timeframe
- Yellow zone enables autonomous action with human notification
- Green zone permits full autonomous operation
- Outside safe operating envelope, all implementations transition to deterministic safety systems

2.5.4 Graceful Degradation Hierarchy

1. Full cognitive agent operation (normal)
2. Reduced autonomy with increased human oversight (elevated conditions)
3. Rule-based automation only (degraded mode)
4. Manual operation with agent advisory (fallback)
5. Full manual control (emergency)

2.5.5 High Availability Architecture Requirements

- Edge deployment for latency-critical safety decisions
- Sub-5-second failover for safety-critical functions
- Agent operation during partial network loss (edge autonomy)
- Cloud connectivity enhances but cannot be required for safe operation
- Local models ensure autonomous operation during connectivity loss
- Defined behavior for all degraded network states

No Single Point of Failure: Agent architecture must ensure that failure of any single component triggers graceful degradation, not catastrophic failure.

2.5.6 Why Trustworthy AI Matters to Industry

Defines what a “proper” emergency stop looks like versus vague promises of “human oversight.” Includes explicit degradation paths and high availability requirements that general AI platforms do not address.

2.6 Law 6: Interoperability with Operational Systems

Principle: Agent interaction with domain-specific systems (clinical, building, manufacturing, grid, flight) must be mediated through semantic models, not direct protocol interfaces. The agent reasons over domain models; the digital twin is translated into system-specific protocols.

2.6.1 Requirements

- Native support for industrial protocols: OPC DA/UA, Modbus, MQTT, Profinet, EtherNet/IP, BACnet
- DCS, SCADA, PLC, and fieldbus connectivity
- Safety instrumented system (SIS) integration

- Historian and alarm management system interfaces
- Edge computing capability with intermittent connectivity support

2.6.2 Conformance Criteria – Three-Tier Integration Framework

Tier 1: Native Integration – Direct protocol implementation in agent runtime.

Evidence Required:

- Protocol conformance test results
- Latency benchmarks (<5 seconds for safety-critical paths)
- Field deployment validation data

Tier 2: Certified Integration – Vendor-verified connectors and adapters.

Evidence Required:

- DCS/SCADA vendor verification documentation
- Production deployment references (minimum 3 sites)
- Performance validation under load

Tier 3: Partner Integration – Third-party integration layers and middleware.

Evidence Required:

- Partner validation and joint testing results
- Field deployment data with performance metrics
- Support escalation procedures

2.6.3 Performance Requirements

- Sub-5-second latency for safety-critical data paths
- Defined assessment cycles for optimization decisions
- Sub-second response for emergency conditions
- Real-time control system integration
- Integration with major DCS platforms
- Time-series database connectivity
- Data quality and governance with lineage tracking

2.6.4 Why Trustworthy AI Matters to Industry

This anchors industrial agents in brownfield reality, not conceptual architecture. Agents must interact with real equipment and control systems that have been operating for decades.

2.7 Law 7: Auditability and Transparency

Principle: All agent actions must be transparent, traceable, and explainable to regulators, operators, and stakeholders.

2.7.1 Requirements

- Complete audit trails documenting not just what agents did but why
- Reasoning must be explainable in human terms
- Real-time compliance validation, not periodic verification
- Mathematical traceability from sensor data through utility functions to actions

2.7.2 The Three Dimensions of Oversight

1. Observability: All agent actions transparent and traceable (computational accountability)
2. Intelligibility: Reasoning explainable in human terms (semantic debugging)
3. Intervenability: Humans can modify or halt operations at any time (graduated control)

2.7.3 Conformance Criteria

- Decision documentation for regulatory compliance
- Root cause traceability through causal models
- Cryptographic identity verification for agent actions
- Comprehensive audit trails linking identity to actions and outcomes

2.7.4 Audit Artifact Specifications

- Incident Replay Package: Complete state capture enabling exact reconstruction of agent decisions during incidents
- Decision Records Schema: Structured format documenting inputs, reasoning, alternatives considered, and action taken
- Policy Rule Set Versioning: Configuration management for symbolic constraints with change tracking
- NIST AI RMF Alignment: Mapping to NIST AI Risk Management Framework requirements

Note: Auditability emerges naturally from Law 3 (Symbolic Primacy). When decisions flow through symbolic structures, they become inherently traceable.

2.8 Law 8: Progressive Autonomy with Safety Boundaries

Principle: Autonomy levels must map to human roles, approvals, and safety criticality. Higher autonomy requires more structured safety, not less.

2.8.1 The Human Agency Scale

Level	Name	Description	Worker Preference [2]
HAS 1	Human-Driven	Complete human control; AI as basic tool	1.0%
HAS 2	Human-Assisted	Human drives with limited AI support	16.3%
HAS 3	Collaborative	True partnership; shared decision-making	45.2%
HAS 4	AI-Assisted	AI manages most operations; human at decision points	35.6%
HAS 5	AI-Driven	Full automation for routine, predictable operations	1.9%

Table 2-2: The Human Agency Scale. Source: DTC Composability Framework Working Group, XMPro

2.8.2 The Progressive Intelligence Journey

- HAS 1-2 (Decision Support): “Tell me what is happening” through real-time monitoring
- HAS 3-4 (Decision Augmentation): “Advise me what to do” through AI recommendations with human-in-loop
- HAS 5 (Autonomous Operations): “Do it for me” within safety boundaries

2.8.3 Conformance Criteria

- Autonomy levels tied to safety zone classification
- Clear decision authority defined at each level
- Escalation protocols based on complexity and risk
- Human validation required for boundary conditions

Critical Insight: Research shows 67% of organizations require human oversight for AI agent operations [3]. Success comes from preserving human agency through governance rather than limiting intelligence through design.

2.9 Law 9: Multi-Agent Safety Orchestration

Principle: Complex industrial operations require coordination of specialized capabilities with clear safety hierarchies. Whether implemented as distinct agents or functional modules, the required behaviors remain constant.

2.9.1 Required Capabilities

- Functional specialization (monitoring, execution, optimization, compliance)
- Conflict detection mechanisms with automated resolution
- Safety primacy enforcement in all conflict scenarios

- Orchestration audit trails for coordination decisions
- Role-based permissions and authority boundaries

2.9.2 Clear Decision-Making Hierarchy

- Safety authority with ultimate override capability
- Optimization subordinate to safety constraints
- Compliance verification independent of optimization
- Coordination protocols with existing plant systems

2.9.3 Conformance Criteria

- Defined capability architecture with specialized roles
- Safety-first hierarchy in all conflict resolution
- Coordination protocols with existing plant systems
- Supervisor capability providing continuous oversight with configurable monitoring cycles

2.9.4 Why Specialization Matters

A single AI system cannot simultaneously:

- Optimize production throughput
- Manage safety constraints
- Coordinate transient operations
- Ensure regulatory compliance
- Handle emergency response

Each requires different expertise, response times, and authority levels.

2.9.5 Why Trustworthy AI Matters to Industry

Addresses the cognitive overload problem. Monolithic AI systems fail when they must balance competing objectives without clear hierarchy. Buyers care about the outcome (conflict resolution, safety primacy, audit trails), not the philosophical purity of multi-agent versus modular implementation.

2.10 Law 10: Safe and Secure Continuous Learning

Principle: Industrial agents continuously collect operational data and evaluate performance, with learned improvements deployed only through controlled processes that maintain inviolable safety guarantees.

2.10.1 What "Continuous Learning" Actually Means

- Continuous Collection: Systematic capture of operational decisions, outcomes, and expert interventions
- Continuous Evaluation: Offline analysis of patterns, effectiveness, and improvement opportunities
- NOT Continuous Deployment: Learned changes never auto-deploy to production systems

2.10.2 Requirements

- Learning processes cannot modify symbolic safety constraints (Law 3)
- Improvement occurs within fixed safety boundaries, not by expanding them
- New behaviors must be validated before deployment in safety-critical contexts
- Learning from human decisions enriches agent capability without overriding human authority

2.10.3 The Deployment Process

Phase 1: Shadow Mode Learning

- Learned models run in parallel with production systems
- Recommendations compared against actual operations
- Performance metrics collected without affecting operations
- Deviation analysis identifies where learning improves decisions

Phase 2: Offline Validation

- Statistical validation of learned improvements
- Safety boundary verification (no prohibited actions learned)
- Simulation testing in digital twin environment
- Expert review of behavioral changes

Phase 3: Management of Change (MOC) Approval

- Formal approval process for production deployment
- Documentation of what changed and why
- Sign-off from appropriate authority levels
- Rollback procedures established

Phase 4: Controlled Deployment

- Gradual rollout with monitoring
- A/B testing where appropriate
- Performance tracking against baseline
- Immediate rollback capability if issues detected

The Immutable Boundary

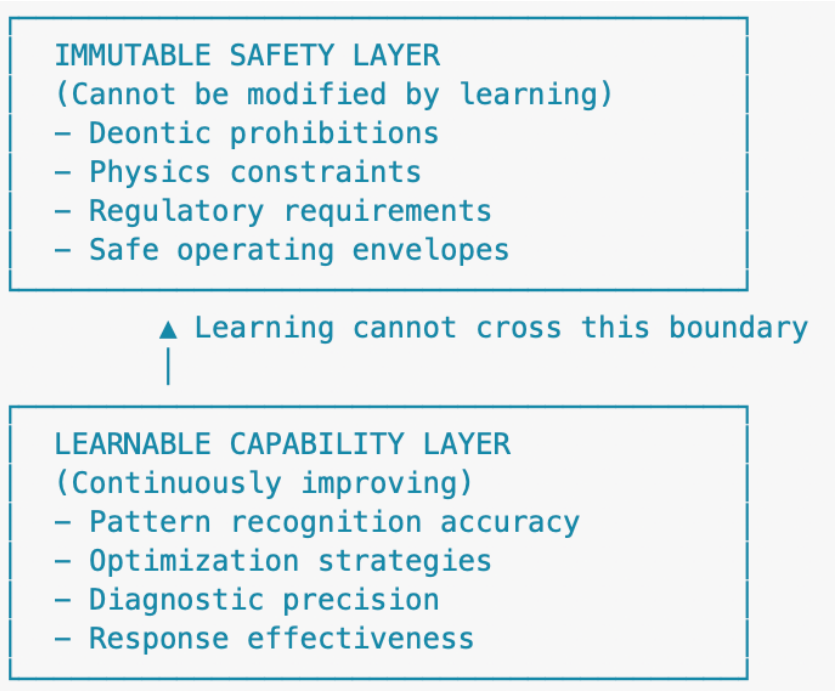

Figure 2-2: The Immutable Boundary Architecture Design. Source: DTC Composability Framework Working Group, XMPro

2.10.4 Conformance Criteria

- Clear separation between learnable parameters and safety constraints
- Validation protocols for learned behaviors before production deployment
- Rollback capability to previous known-good configurations
- Audit trails documenting what was learned, how it was validated, and who approved deployment

2.10.5 Why This Matters

- Addresses the fear that AI systems “drift” into unsafe behavior over time
- Enables continuous improvement without regulatory re-verification of safety systems
- Creates compounding operational advantage as agents learn from experience
- Maintains trust through architectural guarantees, not behavioral promises

2.10.6 Learning from Human Expertise

- Agent decisions enriched by observing human expert interventions
- Tacit knowledge captured through systematic recording of operator actions
- Each human correction improves future agent recommendations
- Expertise preserved as workforce transitions

2.10.7 Why Trustworthy AI Matters to Industry

General AI approaches cannot guarantee that learning will not compromise safety. Continuous deployment without validation creates unacceptable risk. Industrial agents separate collection from deployment, ensuring that improvement never degrades protection.

3 Governance at Machine Speed

3.1 The Governance Transformation

The Problem: Traditional governance operates through periodic review (quarterly audits, weekly compliance checks). This cadence assumes human execution speed. Agent networks operating at machine speed cannot wait for quarterly audits.

The Solution: Governance embedded in agent architecture as continuous control system.

3.2 The Deontic Framework

Components:

- Obligations: Actions agents must fulfill
- Prohibitions: Actions agents cannot transgress
- Permissions: Allowed actions within boundaries
- Conditional Duties: Context-dependent requirements

Key Insight: These aren't post-hoc checks. They're structural constraints shaping agent reasoning before actions occur. An agent physically cannot recommend prohibited actions because the symbolic layer (Law 3) prevents it.

3.3 Supervisor Agents

Function: Continuous oversight with configurable monitoring cycles

- Detect deviations from expected patterns
- Identify conditions requiring human escalation
- Automated compliance officers that never sleep
- Review every decision in real time

3.4 Complete Audit Trails

Requirement: Documentation of not just what but why

- Real-time compliance validation
- Immediate answers for regulatory inquiries
- Mathematical traceability through objective functions

4 IMPLEMENTATION REQUIREMENTS

4.1 TECHNICAL ARCHITECTURE IMPLEMENTATION REQUIREMENTS

Deployment Patterns:

- Edge deployment for latency-critical safety decisions
- Plant-level coordination for multi-system optimization
- Cloud integration for analytics and continuous improvement

Edge-First Safety Architecture:

- Safety-critical decisions must execute at the edge
- Cloud connectivity enhances but cannot be required for safe operation
- Latency-critical functions operate independently of network availability
- Local models ensure autonomous operation during connectivity loss

Performance Requirements:

- Sub-5-second latency for safety-critical data paths
- Defined assessment cycles for optimization decisions
- Sub-second response for emergency conditions

Cybersecurity Compliance:

- IEC 62443 for industrial cybersecurity baseline
- Defense-in-depth architecture
- Secure integration with existing OT networks
- Network segmentation with defined zones and conduits

4.2 INDUSTRIAL AGENT THREAT MODEL

Overview: Agentic AI introduces security concerns beyond traditional IEC 62443 scope. Industrial agents must address both operational technology threats and AI-specific vulnerabilities.

Top Threats:

1. Supply Chain Contamination: Model poisoning during training, compromised foundation models or fine-tuning data, backdoors in pre-trained weights.

Mitigation: Three-component Software Bill of Materials (SBOM) tracking software dependencies, agent configuration, and team governance.

2. Knowledge Base Injection (RAG Vulnerabilities): Malicious content inserted into procedural knowledge, corrupted HAZOP or safety data, adversarial examples in training documentation.

Mitigation: Law 3 (Symbolic Primacy) provides architectural boundary; source validation and cryptographic signing of knowledge bases.

3. Prompt Injection Attacks: Attempts to override safety constraints through crafted inputs, instructions embedded in sensor data or operational logs, social engineering through human interfaces.

Mitigation: Law 4 (Separation of Control) prevents direct execution; validation layer rejects non-compliant actions.

4. Model Drift and Behavioral Changes: Unintended learning from operational edge cases, gradual erosion of safety margins, optimization toward unintended objectives.

Mitigation: Law 10 (Continuous Learning) enforces MOC approval before deployment; immutable safety boundaries.

5. Agent Coordination Exploitation: Multi-agent consensus manipulation, orchestration layer vulnerabilities, escalation privilege exploitation.

Mitigation: Law 9 (Multi-Agent Orchestration) defines clear hierarchy; cryptographic identity verification.

4.2.1 Threat-to-Law Mitigation Mapping

Threat	Primary Mitigation	Supporting Laws
Supply Chain Contamination	Three-component SBOM + cryptographic verification	Law 3, 7
RAG Vulnerabilities	Knowledge Base Signing	Law 3, 7
Prompt Injection	Separation of Control	Law 4, 3
Model Drift	MOC-Controlled Deployment	Law 10, 3
Coordination Exploitation	Identity + Hierarchy	Law 9, 7

Table 4-1: Threat-to-Law Mitigation Mapping. Source: DTC Composability Framework Group, XMPro

4.2.2 IEC 62443 Extensions for Agentic AI

- Zone 3A: AI Model Serving Infrastructure (edge inference)
- Zone 3B: Knowledge Base and Training Infrastructure
- Zone 4: Agent Orchestration and Governance Layer
- Conduit requirements for agent-to-agent communication
- Cryptographic identity for all agent actions

4.3 Agent Identity and Authorization Framework

Agent Design and Planning:

- Agents designed with principle of least privilege
- Functional scope defined before deployment
- Capability boundaries explicit in agent configuration
- Permission sets tied to operational roles

Identity Management:

- Cryptographic identity for each agent instance
- Digital signatures for all agent actions
- Non-repudiation through immutable audit logs
- Identity lifecycle management (creation, rotation, revocation)

Authorization Hierarchy:

- Level 1 (Read-Only): Monitoring and observation agents
- Level 2 (Advisory): Recommendation agents with no execution authority
- Level 3 (Conditional Execute): Agents authorized within green zones only
- Level 4 (Supervised Execute): Agents authorized in yellow zones with human notification
- Level 5 (Safety Override): Safety authority agents with ultimate override capability

Approval Workflows (tied to Laws 3-5):

- Law 3 (Symbolic Primacy): Automatic rejection of non-compliant actions
- Law 4 (Separation of Control): Validation layer approval before execution
- Law 5 (Human Override): Human authority supersedes all agent approvals

Separation of Duties:

- Agent designers cannot approve their own agents for production
- Optimization agents cannot modify safety constraints
- Compliance agents operate independently of operational agents
- Supervisor agents have read-only access to all agent decisions but execute-only for escalation

4.4 Management of Change (MOC) for Industrial Agents

What Changes Require MOC:

Symbolic Rule Updates: HAZOP knowledge base modifications, procedural knowledge (SOP/EOP) changes, safety boundary adjustments, deontic constraint modifications.

Model Version Changes: Foundation model updates, fine-tuned model deployments, learned behavior promotion from shadow mode, prompt template modifications.

Autonomy Level Adjustments: HAS level increases (requires higher approval), safety zone reclassification, authorization boundary expansions, new agent capability deployments.

Infrastructure Changes: Edge computing platform updates, network architecture modifications, integration protocol changes, high availability configuration adjustments.

4.4.1 Approval Authority Levels

Change Type	Approval Authority	Documentation Required
Symbolic rule updates (minor)	Process engineer + Operations supervisor	Change justification, safety impact assessment
Symbolic rule updates (major)	Safety manager + Plant manager	Formal safety review, regulatory compliance check
Model version changes	AI/ML engineer + Operations manager	Validation results, A/B test data, rollback plan
Autonomy level increases	Operations mgr + Safety mgr + Plant mgr	Extended trial period results, risk assessment
Infrastructure changes	IT/OT manager + Operations manager	System integration testing, failover validation

Table 4-2: Approval Authority Levels. Source: DTC Composability Framework Working Group, XMPro

Audit Requirements:

- Decision traces (Law 7) provide complete record of agent behavior before and after change
- Comparison analysis showing behavioral differences
- Performance metrics tracking improvement or degradation
- Rollback triggers and automatic reversion criteria

Integration with Existing OT Practices:

- MOC for agents follows same governance as DCS/SCADA changes
- Work permit requirements for production deployments
- Pre-startup safety review (PSSR) for new agent capabilities
- Lockout/tagout procedures for agent deactivation during maintenance

4.5 Testing and Verification Framework

4.5.1 Phase 1: Factory Acceptance Testing (FAT)

- Simulated environment validation
- Agent behavior verification against specifications
- Safety scenario testing with boundary condition validation
- Performance benchmarking under controlled conditions

4.5.2 Phase 2: System Integration Testing (SIT)

- Real process data integration from plant historians
- Multi-agent coordination validation
- Control system interface verification (DCS, SCADA, PLC)
- Network performance and failover testing

4.5.3 Phase 3: Site Acceptance Testing (SAT)

- Extended autonomous operation demonstration (minimum 15 days continuous)
- Performance validation against objective functions
- Regulatory compliance demonstration
- Operator training and handover

4.5.4 Phase 4: Continuous Validation

- Learning frameworks with safety preservation (Law 10)
- Performance monitoring and optimization
- Knowledge base expansion protocols
- Periodic re-verification (annual or after significant changes)

4.6 Risk Mitigation and Operational KPI Protection

4.6.1 What Each Law Protects

Law	Buyer Risk Mitigated	Operational KPIs Protected
1. Determinism	Unpredictable behavior, regulatory non-compliance	Uptime, Compliance
2. Physics-Aware	Equipment damage, process upsets, quality deviations	Asset integrity, Yield, Quality
3. Symbolic Primacy	Governance drift, unsafe learned behavior	Compliance, Safety incidents
4. Separation of Control	Direct actuation exploits, agent compromise	Safety, Security
5. Emergency Stop	Loss of human control, cascading failures	Downtime, Safety
6. Industrial Protocols	Integration failures, latency issues, data quality	Uptime, Energy efficiency
7. Auditability	Regulatory findings, incident investigations	Compliance, Risk management
8. Progressive Autonomy	Premature automation, workforce rejection	Adoption success, Productivity
9. Multi-Agent Orchestration	Conflicting objectives, optimization-safety tradeoffs	Throughput, Safety
10. Continuous Learning	Unsafe drift, unvalidated changes	Improvement velocity, Safety

Table 4-3: What Each Law Protects. Source: DTC Composability Framework Working Group, XMPro

4.7 Deployment Maturity Paths

4.7.1 Stage 1: Monitoring and Alerts (HAS 1-2)

- Real-time visibility into operational conditions
- Anomaly detection and early warning systems
- Integration with existing alarm management
- Typical Duration: 3-6 months
- Success Criteria: <5% false positive rate, operator adoption >80%

4.7.2 Stage 2: Operator Advisory (HAS 2-3)

- AI-generated recommendations with human decision-making
- Root cause analysis and diagnostic assistance
- "What-if" scenario analysis for operators
- Typical Duration: 6-12 months
- Success Criteria: Recommendation acceptance >60%, measurable decision quality improvement

4.7.3 Stage 3: Maintenance Coordination (HAS 3-4)

- Predictive maintenance scheduling and optimization
- Work order generation and resource coordination
- Integration with CMMS and EAM systems
- Typical Duration: 6-12 months
- Success Criteria: Unplanned downtime reduction >20%, maintenance cost reduction >15%

4.7.4 Stage 4: Closed-Loop Optimization (HAS 4-5)

- Autonomous operation within green zones
- Human-in-loop for yellow zone conditions
- Continuous performance optimization
- Typical Duration: 12+ months
- Success Criteria: <5% human intervention rate, sustained performance improvement

4.8 Knowledge Infrastructure

Required Components:

- Safety knowledge databases (HAZOP, LOPA, bow-tie analysis)
- Procedural knowledge (SOPs/EOPs) with RAG integration
- Causal models and first-principles engineering knowledge
- Regulatory compliance frameworks
- Equipment technical documentation and operating manuals
- Historical incident and near-miss databases

Knowledge Base Security:

- Cryptographic signing of all knowledge sources
- Version control with complete audit trails
- Access controls tied to agent authorization levels
- Integrity verification before agent consumption

5 THE BUSINESS CASE

5.1 PRESERVING INSTITUTIONAL KNOWLEDGE

The industrial workforce faces an unprecedented knowledge exodus. Experienced operators hold decades of tacit expertise that exists nowhere in documentation or systems. When they retire, that knowledge disappears regardless of knowledge management programs.

Industrial agents address this challenge by:

- Capturing expert decisions through systematic observation
- Encoding tacit knowledge into agent behaviors through supervised learning
- Preserving institutional expertise across workforce transitions
- Enabling knowledge transfer at scale through agent-assisted training

The Compounding Advantage: Organizations that deploy industrial agents early build sustainable competitive advantage. Each day of operation adds to accumulated learning. Early adopters gain compounding returns that late entrants cannot replicate quickly.

5.2 OPERATIONAL VALUE CREATION

Operations Run More Often: Predictive maintenance coordination reduces unplanned downtime through early detection and optimized intervention timing.

More Output When Running: Autonomous production coordination analyzes bottlenecks in real-time and adjusts parameters while maintaining quality and safety.

Better Quality at Full Potential: Agent teams predict quality outcomes and recommend process adjustments, learning from human decisions to continuously improve.

Lower Cost: Progressive efficiency gains across the autonomy journey, with visibility identifying cost drivers, augmentation recommending optimization, and autonomy coordinating resource utilization.

6 Evaluating Industrial Agent Solutions

6.1 How This Manifesto Defines Standards Requirements

Capability	Traditional Automation	General AI Platforms	Data Analytics Platforms	Manifesto Standard
Deterministic Safety	✓ (no AI)	✗	✗	✓
Physics-Aware	✓ (no AI)	✗	✗	✓
Symbolic Primacy	✓ (no AI)	✗	Partial (data org only)	✓
Separation of Control	✓	✗	✗	✓
Emergency Stop/Override	✓	✗	✗	✓
Industrial Protocols	✓	Partial	✗	✓
Auditability	Partial	Partial	Partial	✓
Progressive Autonomy	✗	✗	✗	✓
Multi-Agent Safety	✗	✗	✗	✓
Safe Continuous Learning	✗	✗	✗	✓
Threat Model Coverage	✗	Partial	✗	✓
Identity & Authorization	Partial	✗	✗	✓

Table 6-1: How This Manifesto Defines Standards Requirements. Source: DTC Composability Framework Working Group, XMPro

6.2 The Critical Distinction on Ontologies

- Some approaches use ontologies to organize data for analytics
- Industrial agents use ontologies to constrain behavior for safety
- Same technology, fundamentally different purpose

6.3 The Third Position

Traditional industrial software has symbolic constraints but no AI adaptation. AI platforms have sub-symbolic intelligence but no hard safety constraints (guardrails are procedural, not architectural). Data platforms use ontologies to organize data for analytics, not to constrain

behavior for safety. Pure learned ontology approaches have no governance boundary and face inherent drift risk.

Industrial agents with non-bypassable commit boundaries occupy a third position: they learn from operational reality while maintaining hard safety boundaries. This addresses the governance concern that neither prescribed-only nor learned-only approaches solve.

6.4 Evaluation Framework

When evaluating industrial agent solutions, require evidence of compliance with all ten laws:

For each law, request:

- Architectural documentation proving compliance
- Reference implementation demonstration
- Production deployment evidence

7 Industry Influence and Standards

7.1 Path to Industry Standard

- Thought Leadership: Keynote presentations, whitepaper series, webinar programs
- Buyer Enablement: Manifesto compliance checklist for solution evaluation
- Standards Body Engagement: Industry consortia and standards organizations
- Industry Education: Workshops on industrial agent requirements

7.2 Independent Verification and Certification

Third-Party Validation Pathway:

- Engage independent certification bodies for third-party verification
- Co-authorship or official endorsement of manifesto frameworks
- Certification programs for vendor solutions
- Audit frameworks for ongoing compliance

Certification Levels:

- Level 1 (Basic Compliance): Architectural documentation review and design validation
- Level 2 (Operational Validation): Factory acceptance testing and simulated operations
- Level 3 (Field Proven): Site acceptance testing with extended autonomous operation
- Level 4 (Bonded Autonomy): Independent verification enabling parametric insurance

Insurance and Bonded Autonomy:

- Parametric insurance for AI agent operations
- Coverage for algorithmic harm and autonomous decision consequences
- Risk-based premiums tied to manifesto compliance levels
- Claims processing based on decision trace verification

7.3 Standards Alignment

Current Standards Integration:

- ISA-95 for enterprise-control integration
- ISA-88 for batch process control
- IEC 61508 for functional safety
- IEC 62443 for industrial cybersecurity (with agentic AI extensions)
- NIST AI Risk Management Framework

Ontology and Knowledge Representation:

- Native support for industrial ontologies (IDO, DEXPI)
- Causal relationship modeling for industrial contexts
- IDTA Asset Administration Shell for digital twin integration
- ODRL (Open Digital Rights Language) for deontic tokens
- Foundation for future standardization efforts

8 The Path Forward

8.1 The Declaration

Industrial agents are not enhanced chatbots. They are a new category of operational technology requiring their own standards, evaluation criteria, and governance frameworks.

Organizations deploying agents without adherence to these laws accept risks that no responsible operator should bear. Organizations demanding compliance gain sustainable competitive advantage through operations that are simultaneously more capable and more trustworthy.

The Central Insight: Safety emerges not from the absence of autonomy but from its careful bounds. The symbolic layer creates trust. The separation of control ensures safety. The progressive autonomy framework respects organizational readiness. Continuous learning enhances capability without compromising protection.

The Promise: Industrial agents built to these ten laws enable bounded autonomy in safety-critical operations, with a clear path to bonded autonomy through independent verification and official validation.

9 References

[1] SAE International. “J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.” SAE International, 2021.

[2] Stanford University. “The Future of Work: Human-AI Collaboration Study.” Stanford Digital Economy Lab, 2024.

[3] Andersen, N.E. “Autonomous Operations: AI with Guardrails.” LNS Research, 2024.

[4] EEMUA. "Publication 191: Alarm Systems – A Guide to Design, Management and Procurement." 3rd Edition, 2013.

[5] Pearl, J. and Mackenzie, D. "The Book of Why: The New Science of Cause and Effect." Basic Books, 2018.

10 CONFORMANCE CHECKLISTS

In order to be compliant with "The 10 Laws of Trustworthy Autonomous Operations," organizations must meet the following conformance requirements.

10.1 FOR VENDOR EVALUATION

Law	Requirement	Evidence Required
1. Determinism	Deterministic validated actions for identical states	State replay tests with verification
2. Physics-Aware	Process constraints encoded in reasoning	Causal model documentation; conservation law validation
3. Symbolic Primacy	Symbolic layer bounds sub-symbolic AI	Architecture documentation; safety gate demonstrations
4. Separation of Control	Cognition isolated from execution; standardized protocols	Control flow diagrams; protocol support evidence
5. Emergency Stop	Immediate override; graceful degradation; HA architecture	Safety system demonstration; failover tests (<5 sec)
6. Industrial Protocols	Native/verified/partner OT connectivity	Integration catalog; conformance tests; latency benchmarks
7. Auditability	Complete decision trails with artifact specifications	Audit log examples; incident replay package
8. Progressive Autonomy	Graduated HAS levels supported	Configuration demonstration; deployment maturity path
9. Multi-Agent Orchestration	Specialized capabilities with safety hierarchy	Architecture documentation; conflict resolution demos
10. Continuous Learning	Offline learning with MOC-approved deployment	Shadow mode results; validation protocols; rollback tests

Table 10-1: Vendor Evaluation Checklist. Source: DTC Composability Framework Working Group, XMPro

10.2 Additional Verification Requirements

Category	Requirement	Evidence Required
SBOM Integrity	Three-component SBOM supporting time-based queries	SBOM examples; cryptographic verification tests; drift detection reports
Threat Model	Coverage of supply chain, RAG, prompt injection, drift, coordination threats	SBOM documentation; threat mitigation mapping; penetration test results
Identity & Authorization	Cryptographic identity; least privilege; approval workflows	Identity lifecycle documentation; authorization matrix; approval audit trails
Management of Change	MOC process for rules, models, autonomy, infrastructure	MOC procedure documentation; approval authority matrix; change audit trails
High Availability	Defined failover times and degraded mode behavior	HA architecture documentation; failover test results; network loss scenarios

Table 10-2: Additional Verification Requirements. Source: DTC Composability Framework Working Group, XMPro

11 Deployment Maturity Assessment

Organizations can assess their readiness for industrial agents using this framework:

Maturity Level	Characteristics	Recommended HAS Level	Typical Timeline
Level 1: Monitoring	Manual operations with basic alarming	HAS 1-2	0-6 months
Level 2: Advisory	AI recommendations with human decision-making	HAS 2-3	6-12 months
Level 3: Coordination	Workflow automation and resource optimization	HAS 3-4	12-18 months
Level 4: Autonomous	Closed-loop control within safety boundaries	HAS 4-5	18-24 months
Level 5: Bonded	Certified autonomous operations with insurance	HAS 5	24+ months

Table 11-1: Deployment Maturity Assessment. Source: DTC Composability Framework Working Group, XMPro

12 AUTHORS & LEGAL NOTICE

This document is a work product of the Digital Twin Consortium Composability Framework Working Group, chaired by Pieter van Schalkwyk (XMPro) and Sean Whiteley (Axomem).

Authors: Pieter van Schalkwyk (XMPro)

Contributors: The following persons contributed valuable ideas and feedback that significantly improved the content and quality of this document: Sean Whiteley (Axomem)

Technical Editors: Dan Isaacs (DTC CTO) and Will Thompson (DTC) oversaw the process of editing and organizing the contributions of the above Authors and Contributors into an integrated document.

Opinions expressed in this content are solely those of the authors and do not necessarily represent those of the Digital Twin Consortium, EDMA or any other related parties.