Agentic AI-Driven Interoperability for Safe and Secure Multi-Domain Operations

Bruno Carron, Anil Ranjitbhai Patel – Airbus Defence and Space, Germany

1. Introduction

Multi-Domain Operations (MDO) require speed, adaptability, and coordination across the land, air, space, and cyber domains. Today's command-and-control (C2) architectures cannot deliver this: built on static, protocol-defined interfaces, they need weeks of integration to add a sensor or re-route a mission, and offline safety checks cannot react within the sub-second window imposed by fleeting targets. Agentic AI, that is, autonomous systems with persistent goals, memory, and tool use [1], offers a way forward, but it also brings serious risks: untrusted coordination, collusion, semantic misalignment, and loss of human oversight [2]. Existing frameworks treat trust as a policy layer, not as an enforced property [3], and current military agentic architectures still lack semantic interoperability and runtime assurance [4]. **Our position** is therefore simple: interoperability must be redefined, not as connectivity, but as *safe, verifiable, and governable coordination*, and this property must be engineered into the communication and contract layer itself, not certified afterwards.

2. Identification of Use Cases

Joint fire support that combines a field artillery system with an autonomous loitering-drone swarm. Drones detect and track fleeting targets, an AI-fusion node builds a unified tactical picture, and a fire solution is negotiated and executed across the coalition, covering detection, planning, and effects. Potential users are joint task forces, artillery units, and swarm operators. The problem is to close the sensor-to-shooter loop within operational time bounds ($<$1.5 s) while preserving auditability, Rules-of-Engagement (RoE) compliance, and human override in a lethal context.

A forest-fire suppression scenario uses the same architecture (Fig. 1). A water-drop helicopter, a water-bomber, a reconnaissance aircraft, and autonomous UAVs streaming optical and infrared imagery publish their functions as discoverable services, for example `TargetTrack` for live fire maps. A command vehicle fuses this with satellite, terrain, and wind data. Suppression contracts with explicit safety limits (collateral probability, wind speed, exclusion zones) are negotiated over a temporary 4G/5G ad-hoc network that acts as the AI Service Bus; any mission can be aborted within 50 ms, and the design target is under 1.5 s from detection to water drop. Potential users include civil-protection agencies and firefighting coalitions.

Dual-use by design. Both scenarios share one coordination pattern (sense, fuse, decide, effect) under the same constraints: hard latency bounds, collateral limits, exclusion zones, and human override. An architecture proven on the civilian case therefore transfers to defence with different effectors and RoE, but the same structure. The civilian scenario is also a *legally deployable testbed*. Real flight campaigns can generate the operational data that trust models require, and build the certification evidence base before any defence deployment. This de-risks precisely the aspects that block the adoption of autonomy in lethal contexts: runtime assurance and human oversight.

3. Proposed Technical Approach

We propose the *Agentic AI Interoperability Framework (AIF)*. AIF is a safety-and-security-first architecture: trust, semantics, and governance are built into the communication and contract layer, instead of being added on top.

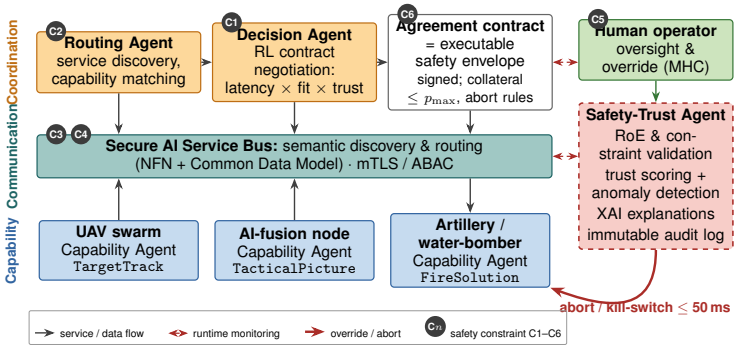

Figure 1: The AIF architecture: a functional plane (Capability, Routing, and Decision Agents over the Secure AI Service Bus) delivers plug-and-play coordination; an assurance plane (Safety-Trust Agent under human oversight) enforces C1–C6 at runtime. Line styles distinguish data, monitoring, and abort paths for black-and-white legibility.

Methodological choices. AIF is co-designed around six non-negotiable safety constraints [1, 3]: **C1** bounded real-time autonomy ($<$1.5 s loops, auditable post-hoc); **C2** provable trustworthiness, with misbehaving agents detected and isolated; **C3** resilient communication under Byzantine faults, jamming, and spoofing; **C4** semantic safety, so that misinterpretation degrades gracefully and never causes a kinetic misfire; **C5** scalable Meaningful Human Control (MHC) with override authority; **C6** full provenance of every negotiation and execution.

Architecture (Fig. 1). AIF separates a *functional plane* from an *assurance plane*. In the functional plane, *Capability Agents* expose platform sensors and effectors as discoverable services, *Routing Agents* match requests to capable providers (C2), and *Decision Agents* negotiate contracts with reinforcement-learning policies that balance latency, capability fit, and trust under hard time bounds (C1). The assurance plane is anchored by the novel *Safety-Trust Agent (STA)*. The STA validates every contract against the RoE, computes dynamic trust scores by anomaly detection on communication patterns, latency, and semantic consistency, and holds a $\leq$50 ms circuit-breaker channel to the human operator (C3, C5).

Technologies mobilised. The *AI Service Bus* is a secure, message-oriented edge middleware enabling service discovery and contract negotiation without pre-wired integration; it extends Named Function Networking principles with semantic routing over a Common Data Model aligned with NATO ISR data standards [5], so requests are matched by intent and context rather than syntax (C4). Agreement contracts are lightweight, signed JSON *safety envelopes*. Their embedded constraints (Fig. 1) bound every action: collateral probability, authorisation mode, abort conditions, and kill-switch latency. Any violation triggers an abort and human escalation (C5), and blockchain-anchored audit trails ensure non-repudiation (C6). Mutual TLS with attribute-based access control secures authentication, and explainable AI makes contract choices transparent to operators (MHC).

Why the approach can work. Each claimed benefit rests on a clear mechanism. *Latency:* in conventional C2, the dominant delay is organisational, coming from staffing chains and pre-mission integration. AIF replaces both with machine-speed discovery and negotiation, and contracts pre-encode authorisation, which turns approval into a millisecond-scale check. Reviews of defence C2 already identify agentic architectures at the tactical edge [4], and networked-control designs are established for drone swarms [6]. *Safety:* runtime enforcement checks constraints against the real situation at execution time, and the abort path (STA $\rightarrow$ effector, Fig. 1) is independent of the mission path, so a coordination failure cannot disable the ability to stop it. *Trust:* anomaly detection over communication patterns, latency, and semantic consistency targets exactly the known multi-agent failure modes, namely collusion and memory poisoning [2].

Open challenges and evaluation plan. Open questions remain: early detection of subtle sensor faults, trust-model reliability in new environments, Service-Bus scalability at swarm scale, operator bandwidth under stress, and formal verification of the protocol. We propose to evaluate AIF first in simulation on the forest-fire scenario, then in hardware-in-the-loop flight campaigns, scaling from a few agents to swarm size via hierarchical routing and edge-offloading. The criteria are measurable: sensor-to-effect latency against a conventional C2 baseline; negotiation and abort latencies against the 1.5 s and 50 ms budgets; trust-model precision and recall under injected faults, jamming, and spoofing; and operator workload at scale (C5).

Maturity level. This work is a *concept-stage position* (TRL 1–2). The architecture, agent taxonomy, contract schema, and governance mechanisms are specified, but no experimental validation has been performed yet: the evaluation plan above is exactly the experimentation we propose, starting with the civilian testbed. Later steps toward a pre-operational solution are formal verification of the contract protocol by model checking, and quantum-resistant trust models.

References

[1] S. Raza, R. Sapkota, M. Karkee, and C. Emmanouilidis, "TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems," *arXiv:2506.04133*, 2025.

[2] S. J. Lazer, K. Aryal, M. Gupta, and E. Bertino, "A Survey of Agentic AI and Cybersecurity: Challenges, Opportunities and Use-case Prototypes," *arXiv:2601.05293*, 2026.

[3] A. Habbal, M. K. Ali, and M. A. Abuzaraida, "Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions," *Expert Systems with Applications*, vol. 240, art. 122442, 2024.

[4] S. Joshi, "Agentic AI for National Defense: Review of Architectural Patterns and Recent Developments," *TechRxiv*, 2026.

[5] NATO Standardization Office, "Air Reconnaissance Primary Imagery Data Standard," *STANAG 7023*, Ed. 5, 2022.

[6] G. Asaamoning, P. Mendes, D. Rosário, and E. Cerqueira, "Drone Swarms as Networked Control Systems by Integration of Networking and Computing," *Sensors*, vol. 21, no. 8, art. 2642, 2021.