DOI: 10.53477/3045-2309-26-14

SECURING THE BLACK BOX: A TECHNO-DIPLOMATIC FRAMEWORK FOR AI INTEGRATION IN MODERN DEFENCE ALLIANCES

Dumitru-Cătălin VASILE, Dipl. Eng.,
PhD Candidate, National School of Political and Administrative Studies, Bucharest, Romania
Master's Student, "Carol I" National Defence University, Bucharest, Romania
E-mail address: catalin.vasile@outlook.com

Abstract*: Integrating Artificial Intelligence (AI) into modern defence alliances creates a serious trust issue: the "black box" problem. When algorithms are opaque, they undermine the inter-state trust that coalitions need to operate effectively. This paper examines the conflict between strict national data sovereignty and operational demands for shared intelligence. Using a qualitative approach that draws on recent developments from NATO's DIANA initiative and AUKUS Pillar 2, I propose a "Techno-Diplomatic Framework". The study argues that traditional hardware standards are insufficient for managing probabilistic machine learning. Instead, it suggests a governance model that utilises Privacy-Enhancing Technologies, specifically Federated Learning and Confidential Computing, as diplomatic enablers. By harmonising technical tools with normative alignment, alliances can achieve "cognitive interoperability". This approach allows nations to collaborate on sensitive training without exposing sovereign data, effectively transforming the black box from a vulnerability into a secured asset for collective defence.*

Keywords: *Techno-diplomacy; Artificial Intelligence; Interoperability; Black Box Problem; Federated Learning; NATO.*

Introduction

The contemporary global security environment is defined by the collision of kinetic threats and digital disruptions, forcing defence alliances to fundamentally reimagine the nature of collective security (Hadean, 2025). Central to this transformation is the integration of Artificial Intelligence (AI) into the military instrument of power. Unlike previous revolutions in military affairs, driven by gunpowder, the internal combustion engine, or nuclear technology, the AI revolution is cognitive (Jensen & Mishra, 2025). It promises to compress the OODA loop (Observe, Orient, Decide, Act) from minutes to milliseconds, analysing vast sensor arrays to predict adversary behaviour and optimise logistics with superhuman precision (Bozkurt et al., 2025). For alliances such as the North Atlantic Treaty Organization (NATO), the AUKUS partnership, and the Five Eyes intelligence consortium, the adoption of AI is not merely an option for modernisation; it is an imperative for survival in an era of "techno-strategic" competition with revisionist powers (European Parliamentary Research Service [EPRS], 2025).

The transition to AI-enabled warfare introduces a structural vulnerability that is as much diplomatic as it is technical: the "black box" problem (Sullivan & Ricket, 2024). The most potent AI systems, particularly those utilising deep reinforcement learning and large language models (LLMs), operate through internal logics that are often unintelligible to their human operators. In a sovereign defence context, this opacity complicates testing and evaluation. In a multinational coalition, it creates a crisis of trust. When a commander from Nation A is asked to authorize a strike based on a target identified by Nation B's algorithm, the inability to interrogate the system's reasoning to "open the black box" becomes a political and legal liability.

How can an alliance maintain unity of command when its constituent digital systems rely on opaque algorithmic processes whose outputs are inherently inexplicable? The data fuelling these systems is subject to increasingly stringent sovereignty regimes (NATO, 2025). The traditional model of intelligence sharing, predicated on the exchange of finalised reports or raw signals, is ill-suited for the era of machine learning, which requires massive, continuous datasets to train and refine models. Nations are rightfully wary of pooling sensitive data into centralized repositories due to security risks and the potential for "model inversion" attacks that could expose national secrets. This creates 'data silos'. These silos fragment the alliance's view, making us weaker together than we are apart.

We argue that resolving these challenges requires a new form of statecraft: "techno-diplomacy" (World Economic Forum [WEF], 2023). It argues that the technical architectures of AI systems must be elevated to the level of diplomatic protocols. By moving beyond the static "hardware interoperability" of the 20th century (standardised ammunition and fuel) to a dynamic "cognitive interoperability", alliances can secure the algorithmic bond (Belfer Center for Science and International Affairs, 2025). This research creates a unified Techno-Diplomatic Framework that leverages emerging privacy-enhancing technologies (PETs), specifically Federated Learning (FL) and Confidential Computing, to reconcile the competing demands of data sovereignty and collective intelligence.

The analysis proceeds in four parts: Chapter 1 dissects the "Strategic Black Box", analysing the operational and legal risks posed by opaque AI in coalition settings. Chapter 2 examines the "Data Sovereignty Crisis", exploring the friction between national classification regimes and the technical needs of AI training. Chapter 3 defines the emerging field of "Techno-Diplomacy" and evaluates current institutional efforts, such as NATO's DIANA and the EU-US Trade and Technology Council. Chapter 4 provides a technical analysis of the proposed solutions, Federated Learning and Trusted Execution Environments, demonstrating their utility as diplomatic tools. Finally, the paper synthesises these elements into the proposed framework, offering concrete recommendations for the "complex and dynamic" security environment.

This research employs a qualitative methodology to analyse the intersection of artificial intelligence technologies and coalition diplomacy, grounded in a comprehensive review of strategic literature and primary policy documents. The study focuses specifically on modern defence alliances, including the North Atlantic Treaty Organization (NATO), the AUKUS partnership, and the Five Eyes intelligence consortium. The research scope is defined by the "black box" paradox, addressing the fundamental tension between the operational necessity of collective algorithmic intelligence and the sovereign constraints of data security.

To resolve this paradox, the study adopts a dual-track analytical approach that bridges operational theory with technical architecture. First, it conducts a techno-strategic analysis to examine the operational and legal risks posed by opaque AI systems such as Deep Learning and Large Language Models within multinational coalitions, with a particular emphasis on the challenges of "cognitive interoperability" and compliance with International Humanitarian Law (IHL). Second, the research performs an architectural synthesis to evaluate the utility of emerging Privacy-Enhancing Technologies (PETs), specifically Federated Learning (FL) and Confidential Computing. These technologies are analysed not merely as technical tools, but as diplomatic enablers capable of resolving the "Third Party Rule" dilemma and mitigating data sovereignty conflicts.

Data sources for this analysis include official alliance publications, such as those on NATO's DIANA initiative and AUKUS Pillar 2 developments, as well as technical specifications for secure computing mechanisms, such as Trusted Execution Environments (TEEs) and Secure Multi-Party Computation (SMPC). By synthesising these diverse elements, the paper proposes a unified "Techno-Diplomatic Framework" that demonstrates how specific technical architectures can serve as the foundational bedrock for new diplomatic protocols in the era of algorithmic warfare.

1. The strategic black box: operational and trust challenges in coalition AI

The integration of Artificial Intelligence into the defence sector marks a definitive shift from platform-centric warfare, defined by the capabilities of tanks, ships, and aircraft, to data-centric warfare, defined by the superiority of algorithms and information flows (NATO, 2024). In this new paradigm, the "black box" problem is not merely a technical glitch to be debugged; it is a structural impediment to alliance cohesion that creates operational, legal, and strategic problems.

1.1. The opacity of deep learning and the crisis of explainability

The fundamental challenge of modern AI lies in the trade-off between performance and explainability. The most advanced systems, deep neural networks (DNNs) used for computer vision, signal processing, and predictive analytics, operate by processing inputs through millions or billions of parameters (weights) across multiple hidden layers (Sullivan & Ricket, 2024). While these systems can achieve superhuman accuracy in tasks such as target recognition or cyber anomaly detection, their internal architecture presents a unique challenge for coalitions. The pathway from input to output is not merely nonlinear in a mathematical sense, but relies on deep neural networks with millions of parameters in 'hidden layers' that defy reverse-engineering. Unlike a rule-based system where logic is transparent (if X, then Y), a deep learning model generates outputs based on high-dimensional feature correlations that are inaccessible to human operators. This creates a 'black box' phenomenon where the system provides a conclusion without an auditable chain of evidence, leaving the user with no mechanism to understand or reconstruct how the result was derived.

In a coalition environment, this opacity creates a severe "trust deficit" (Reynolds & Atalan, 2024). Interoperability has traditionally been built on deterministic standards: a NATO commander knows that a 5.56 mm round will fire from a compliant rifle because the physical specifications are standardised and verifiable. "Cognitive interoperability", the ability of systems to share understanding, lacks these physical guarantees. This technical opacity has immediate strategic implications. If a US-developed AI system deployed in a joint operations center identifies a potential threat, a German or French commander receiving that data has no mechanism to verify the assessment if the system cannot explain its reasoning. Unlike traditional intelligence sharing, where analysts can scrutinise source credibility and raw data to validate a conclusion, an opaque AI output demands an operational leap of faith. The allied commander cannot discern whether the target identification was driven by robust tactical indicators or a statistical artifact within the model's hidden layers. Consequently, the inability to interrogate the machine's logic creates a fracture in collective decision-making: the ally must either accept a high-risk automated judgment blindly, potentially violating their specific Rules of Engagement or hesitate, thereby negating the speed advantage the AI was meant to provide.

This crisis of explainability introduces severe operational risks, primarily manifesting in operator psychology and alliance cohesion. Operators may vacillate between "automation bias", in which they uncritically accept the AI's output under time pressure, and "algorithm aversion", in which they reject valid AI insights due to a lack of understanding. In a high-tempo coalition environment, this inconsistency can lead to disjointed decision-making and a breakdown in the OODA loop (Bozkurt et al., 2025).

A "lowest common denominator" effect may emerge, as alliances are composed of sovereign states with varying risk appetites. If an AI system serves as a black box, the alliance may be forced to default to the Rules of Engagement (ROE) of the most risk-averse member, effectively vetoing the use of advanced capabilities. As a result, a system that cannot demonstrate its adherence to specific national legal interpretations of "necessity" may be sidelined, negating the investment in the technology.

1.2. The legal quagmire: IHL and the black box

The application of International Humanitarian Law (IHL) to AI-enabled operations is perhaps the most contentious aspect of the black box problem, primarily because IHL principles, specifically distinction, proportionality, and precaution, are inherently subjective and context-dependent (Pollard, 2024). A critical challenge arises regarding the principle of distinction, which requires a system to distinguish between combatants and civilians. A black box model might identify a target based on opaque correlations – such as a specific radio frequency combined with movement patterns – that are statistically valid but legally insufficient. This creates a critical accountability asymmetry. When a human soldier makes a catastrophic error based on flawed intuition (as seen in the Daunte Wright case in Minnesota), the legal system can interrogate their intent, negligence, and adherence to training to assign liability. In contrast, an AI 'hallucination' offers no such recourse. If a commander authorises a strike based on a machine's opaque recommendation that turns out to be a school bus, the chain of responsibility ruptures: the commander can claim reasonable reliance on a certified system, while the developer can claim the system functioned within its statistical error rate, leaving the violation of International Humanitarian Law without a punishable perpetrator.

The principle of proportionality requires weighing the anticipated military advantage against the expected collateral damage, a task that remains a value judgment rather than a mathematical calculation. If an AI system recommends a strike, a human commander acts as the moral agent responsible for that judgment; if the commander cannot understand the parameters the AI used to estimate collateral damage, their ability to exercise "meaningful human control" is illusory. For NATO and its partners, this creates a diplomatic rift, as some allies may interpret "meaningful human control" as requiring an understanding of the system's internal logic. In contrast, others may accept statistical reliability as a proxy for control (Department of Defense [DoD], 2024). Without a framework to bridge these interpretations, the black box becomes a wedge issue that could fracture the alliance during joint operations.

1.3. Adversarial vulnerabilities: the poisoned box

The opacity of AI systems also expands the attack surface for adversaries. Because the decision-making logic of a black box is hidden, it is uniquely susceptible to "adversarial attacks", subtle manipulations of input data that cause the model to fail with high confidence. One primary method is data poisoning, in which an adversary can subtly corrupt open-source datasets often used to pre-train military models. A "Trojan" behaviour could be embedded in the model, triggered only by a specific visual or digital signal (Scaleout Systems, 2025). In a coalition, if Ally A's model is poisoned and shares erroneous targeting data with Ally B, the corruption spreads across the network. Because the model is opaque, Ally B has no easy way to audit the incoming data for integrity.

Additionally, adversaries can employ model evasion to exploit the "blind spots" of a black box model. For instance, applying a specifically crafted pattern of tape to a tank might cause an image recognition algorithm to classify it as a civilian truck. Without transparency into which features the model prioritises, coalition forces remain vulnerable to these cognitive exploits. A convergence of operational opacity, legal uncertainty, and adversarial vulnerability thus defines the strategic black box. Solving this requires more than better engineering; it requires a diplomatic architecture that can manage risk without demanding impossible levels of transparency.

2. The data sovereignty crisis: fragmentation in the age of fusion

If algorithms are the engines of modern warfare, data is the fuel. The efficacy of AI systems is strictly limited by the diversity and volume of the data upon which they are trained (DefenseScoop, 2025). For a defence alliance, the theoretical advantage is immense: a coalition of 32 NATO nations should theoretically possess a dataset 32 times richer than any single member. In practice, the reality is a landscape of fragmented "data silos" enforced by rigid sovereignty concerns and legacy classification regimes.

2.1. The "third-party rule" and intelligence barriers

The primary diplomatic barrier to AI integration is the "Third Party Rule", a foundational principle of intelligence sharing that dictates that information received from a partner cannot be shared with a third party without the originator's explicit consent. In the context of static documents, this rule is manageable. In the context of AI training, it is a bottleneck. Training a robust coalition model, for example, an acoustic detection model for submarines, requires aggregating raw sonar data from the US, UK, France, and Norway. Under current rules, moving this raw, classified data into a central "training lake" is legally fraught. It would require complex multi-lateral agreements that are slow to negotiate and difficult to enforce.

This means nations default to training their own models on their sovereign data (NATO, 2025). The result is that the US Navy trains its models on Pacific and Atlantic data, while the Norwegian Navy trains on Arctic data. When a US ship operates in the Arctic, its AI is effectively "blind" to the local environmental nuances that the Norwegian model understands perfectly. This lack of data interoperability creates operational seams that adversaries can exploit.

2.2. Techno-nationalism and the fear of exposure

Beyond intelligence handling rules, the rise of "techno-nationalism" has made states increasingly protective of their national data assets. Data is now viewed as a strategic economic resource. There is a palpable fear among defence ministries that sharing high-fidelity training data might inadvertently reveal critical vulnerabilities. First, high-quality training data reveals not only what a military can detect, but also what it cannot, exposing gaps in coverage. Second, regarding sensitive sources and methods, even anonymised data can sometimes be "deanonymised" or reverse-engineered to reveal the location or nature of the sensor that collected it, compromising the source. Finally, an industrial advantage concern exists; as AI becomes a driver of the defence industrial base, nations are incentivised to hoard data to give their domestic defence primes, such as Thales, Leonardo, or Lockheed Martin, a competitive edge in developing superior algorithms.

2.3. The legacy trap: case study of the eastern flank

The data sovereignty crisis is further complicated by the disparity in digital maturity across the alliance. As highlighted in the context of the Romanian administration, many allies rely on "legacy systems" and "cloud private" architectures segregated, on-premise infrastructure that are mandated by national laws for the protection of classified information, such as Romania's strict interpretation of OUG 89/2022 regarding government cloud interoperability (Curtea de Conturi a României, 2023).

This friction was palpably demonstrated during the recent NATO Coalition Warrior Interoperability Exercise (CWIX 25) in Bydgoszcz. While digitally mature allies like the US and UK demonstrated real-time data fusion using cloud-native APIs, detachments from the Eastern Flank often faced a "digital hard stop". Personal observations from the exercise floor revealed that operators were frequently forced to manually bridge the gap between national secure networks and the NATO Mission Secret network. Instead of automated data streams, critical intelligence often had to be moved via "air-gapped" procedures, physically transferring data on secure hard drives, to comply with sovereign data laws.

These legacy systems are often physically incapable of the high-speed data transfer required for centralised AI training. The lack of standardised data labeling and metadata (as required by STANAG 5636) means that even if the political will to share data existed, the technical capability to ingest it is effectively absent (NATO Allied Command Transformation, 2025). This reality creates a "two-speed alliance" where data-rich, digitally mature nations accelerate away from data-poor or digitally distinct allies, fracturing the very interoperability that is the alliance's center of gravity.

3. Techno-diplomacy: the new statecraft of alliance management

To close the gap between the technical reality of the black box and the political reality of data sovereignty, a new form of statecraft is emerging: "techno-diplomacy" (Bano et al., 2024). This practice integrates the technical governance of digital infrastructure with the strategic objectives of foreign policy, acknowledging that in the 21st century, technology standards are the new treaties.

3.1. Defining techno-diplomacy in a defence context

Techno-diplomacy differs from "science diplomacy" (which focuses on scientific cooperation for peace) and "digital diplomacy" (using digital tools for public messaging). It is the strategic negotiation of the rules, norms, and architectures that govern critical technologies. For defence alliances, techno-diplomacy serves three critical functions. The first is normative harmonisation, which involves aligning the ethical and legal frameworks that govern AI use to ensure that allies share a common understanding of "responsible use" (DoD, 2024). The second function is regulatory alignment, which aims to coordinate export controls, investment screening, and certification standards to create a trusted "technology zone" where innovation can flow freely. The third function is architectural consensus, which requires agreeing on the technical designs, such as zero-trust architectures or federated networks, that will underpin shared systems. This shift is visible in the appointment of "Tech Ambassadors" and the restructuring of foreign ministries, such as the US State Department's Bureau of Cyberspace and Digital Policy, to explicitly address the geopolitics of technology (WEF, 2023).

3.2. Institutional vehicles for AI governance

Several key institutions are currently pioneering techno-diplomacy, serving as laboratories for the framework proposed in this paper. NATO's DIANA and Innovation Fund serve as a prime example. The Defense Innovation Accelerator for the North Atlantic (DIANA) establishes a network of test centers and accelerators across the alliance, creating a mechanism for "interoperability by design" (NATO, 2021). It allows a startup in Estonia to validate its AI model on a test range in Portugal, ensuring that the technology is compatible with allied standards from inception. The €1 billion Innovation Fund complements this by providing "sovereign capital", reducing reliance on non-aligned investment and fostering a shared industrial base.

Additionally, the AUKUS partnership (Australia, UK, US) represents a "minilateral" acceleration of techno-diplomacy, particularly through Pillar 2, which focuses specifically on "advanced capabilities" including AI and autonomy (U.S. Department of War, 2026). By negotiating exemptions to strict export controls (like the US ITAR regime), AUKUS partners are creating a "free trade zone" for algorithms and military IP. This demonstrates that smaller, high-trust groups can achieve deeper integration than larger alliances, potentially serving as a pathfinder for broader NATO efforts. The EU-US Trade and Technology Council (TTC) serves as the primary forum for resolving the trans-Atlantic "regulatory gap". With the EU moving towards comprehensive regulation through the AI Act and the US favoring a risk-based, sectoral approach, the TTC works to align on definitions of "trustworthy AI" and coordinate on semiconductor security. This alignment is a prerequisite for military interoperability; without it, US-made military AI might fail to meet European legal certification standards, blocking its deployment in European theaters.

3.3. Project Maven: a model for coalition integration

Operationalising techno-diplomacy is best exemplified by the evolution of Project Maven. Originally a US initiative to automate the analysis of drone video feeds, Maven has evolved into the "Maven Smart System" (MSS), a platform now being extended to coalition partners (Atlantic Council, 2024). MSS acts as a techno-diplomatic bridge: it ingests data from diverse allied sensors, processes it through US-trained algorithms, and shares the insights (the "cognitive output") back to

the allies. This allows partners to benefit from advanced AI without needing to access the black-box algorithm itself or expose their raw data to a central US repository. It is a working prototype of the "Techno-Diplomatic Framework" in action.

4. Technical architectures for sovereign interoperability

While techno-diplomacy provides the political will, "Privacy-Enhancing Technologies" (PETs) provide the technical way. To secure the black box and resolve the data sovereignty crisis, alliances must adopt architectures that enable collaboration without mutual exposure. The framework relies on three specific technologies: Federated Learning, Confidential Computing, and Secure Multi-Party Computation.

4.1. Federated Learning (FL): bringing the code to the data

Federated Learning (FL) fundamentally reverses the traditional paradigm of machine learning. Instead of moving sensitive data to a central server for training an action that often violates sovereignty and classification protocols FL moves the *model* to the data (Gradiant, 2021). In this architecture, a central "aggregator" (e.g., managed by Allied Command Transformation) distributes a baseline AI model to the secure local clouds of participating nations. Each nation trains the model locally on its own classified data, computing only the mathematical updates (gradients) which are then sent back to the aggregator to update the global model.

To operationalise this, consider an Anti-Submarine Warfare (ASW) scenario in the Baltic Sea. A US Navy P-8 Poseidon deployment may arrive with acoustic detection models trained primarily on deep-water Atlantic or Pacific datasets, making them less effective in the shallow, brackish, and acoustically cluttered environment of the Baltic. Under an FL framework, the US model could be sent to a secure Norwegian or German naval cloud. There, it would "learn" from local, high-fidelity sonar logs capturing specific thermal layers and salinity profiles without those raw, highly classified logs ever leaving the host nation's custody.

This mechanism serves as a potent diplomatic enabler. The US P-8 flies with a smarter model that understands the Baltic environment, while the Norwegian Navy contributes to alliance security without violating the "Third Party Rule" or exposing its sensitive acoustic libraries. It resolves the dilemma of collective intelligence by sharing the *mathematical learnings* (the "cognitive output") rather than the raw intelligence itself. Since the central model is exposed to external inputs, FL needs strong safeguards against "model poisoning", which calls for incorporating the next technology: Confidential Computing.

4.2. Confidential computing and trusted execution environments (TEEs)

Confidential Computing protects data in use (Confidential Computing Consortium, 2025). Standard encryption protects data at rest (on disk) and in transit (over the wire), but data must typically be decrypted in memory before a CPU can process it. This "clear text" phase is a vulnerability. TEEs (hardware enclaves such as Intel SGX or AMD TDX) encrypt memory at the hardware level. The AI model and the data are loaded into this enclave and are invisible to both the host operating system and the cloud provider.

For techno-diplomacy, TEEs enable "Zero Trust" collaboration (Anjuna, 2025). A US-developed targeting algorithm can be sent to a Polish server to process local data. The algorithm runs inside a TEE; the Polish operators cannot see the US proprietary code, and the US developers cannot see the Polish raw data. The hardware guarantees the isolation. This allows for the deployment of "black box" systems on allied infrastructure with mathematical guarantees of IP and data security.

4.3. Secure multi-party computation (SMPC)

SMPC allows multiple parties to jointly compute a function over their inputs while keeping those inputs private. Data is split into "secret shares" and distributed among the parties. No single party ever holds the complete data. They perform computations on these shares, and the results are combined to reveal only the final output. This is ideal for "Private Set Intersection" (PSI), allowing intelligence agencies to check if a suspect appears in each other's databases without revealing the broader dataset (Gradiant, 2021). This minimises the "blast radius" of intelligence sharing to the strict "need to know".

5. The techno-diplomatic framework: a proposed architecture

To secure the black box, alliances must integrate these technical enablers into a cohesive governance structure. This paper proposes a three-layered Techno-Diplomatic Framework.

5.1. The normative layer: defining the "Rules of the Road"

This layer establishes the political and ethical boundaries for AI use, building on NATO's Principles of Responsible Use. It begins with a harmonised risk taxonomy in which allies agree on a standard classification of AI risk. The EU AI Act's tiered model serves as a baseline but must be adapted for military contexts; a "NATO AI Risk Standard" should categorize systems based on the consequence of error (e.g., Kinetic-Lethal vs. Logistics-Non-Lethal) to determine the necessary level of TEV&V (Bozkurt et al., 2025). The framework necessitates context-appropriate Meaningful Human Control (MHC). To bridge the gap between US pragmatism and European caution, MHC is defined not as a static "man-in-the-loop" requirement, but as a dynamic variable that scales with risk: high-risk systems, such as Lethal Autonomous Weapons, require strict oversight, while low-risk systems, such as logistics optimisation, allow greater autonomy. Finally, to mitigate the black box trust deficit, allies must agree to share AI Bills of Materials (AI-BOMs). These documents certify the provenance of the training data (e.g., "trained on validated NATO-standard imagery") and the model architecture, providing a "nutrition label" for the algorithm without revealing the proprietary "recipe".

5.2. The technical layer: the federated interoperability backbone

This layer builds the infrastructure for the "Alliance Cloud". The Federated Interoperability Backbone (FIB) is a standing digital infrastructure that connects national AI centers via a secure, federated network. The FIB facilitates the exchange of model updates (via FL) and supports "containerised" deployment of AI capabilities. Critical to this is the adoption of standardised APIs; open, standardised Application Programming Interfaces ensure that an AI module developed by Nation A can plug into Nation B's Command and Control (C2) system, regardless of the underlying legacy hardware. Additionally, metadata standardisation via the full implementation of STANAG 5636 (NATO Core Metadata Specification) is required to ensure data is "AI-ready" and discoverable across the federation (NATO Allied Command Transformation, 2025).

5.3. The operational layer: continuous assurance

Trust must be continuously validated through rigorous testing. First, the alliance must expand on DIANA to create Joint AI Proving Grounds, acting as "sandboxes" where allies can red-team each other's models. These environments enable adversarial testing (e.g., attempts to spoof a vision system) to verify robustness before deployment (Atlantic Council, 2024). Second, to address the explainability gap in real time, the framework proposes deploying "Observer" modules alongside black-box systems. These are simple, rule-based AI agents that monitor the inputs and outputs of the complex model. If the black box suggests an action that violates a defined Rule of Engagement (e.g., targeting a protected site), the Observer acts as a "circuit breaker", flagging the decision for human review.

5.4. Limitations and technical challenges

While the Techno-Diplomatic Framework offers a path toward cognitive interoperability, its implementation faces significant technical and environmental hurdles. A primary limitation is the communication overhead associated with Federated Learning. In contested or degraded tactical environments, the high bandwidth required to continuously transmit model gradients between national clouds and an alliance aggregator can lead to significant latency, potentially desynchronising the collective intelligence. While Trusted Execution Environments (TEEs) provide hardware-level isolation, they are not immune to sophisticated side-channel attacks that may attempt to infer proprietary code or data through power consumption or timing analysis.

Additionally, the framework assumes digital maturity across all member states. The "two-speed alliance" problem remains a critical risk; nations relying on disconnected legacy systems may find themselves unable to participate in the Federated Interoperability Backbone, regardless of the diplomatic will to do so. Finally, the move toward "Zero Trust" collaboration via encryption and TEEs can complicate digital forensics and post-incident auditing. If an AI-driven decision leads to an unintended kinetic outcome, the very technologies used to protect sovereign data may make it more difficult for an alliance to conduct a transparent investigation into the algorithmic failure.

Conclusions

The integration of Artificial Intelligence into defence alliances is not a distant future; it is a current reality that demands a sophisticated response. The "black box" problem represents a critical vulnerability in this transition, threatening to fracture alliance cohesion through mistrust and operational paralysis. As this paper has demonstrated, the challenge is surmountable through the application of a Techno-Diplomatic Framework. By fusing the normative power of diplomacy with the architectural guarantees of Federated Learning and Confidential Computing, alliances can shift from a "need to know" to a "need to share" posture without compromising sovereignty. The shift from "hardware interoperability" (STANAGs) to "cognitive interoperability" (shared models and norms) is the defining task for the next decade of collective defence. The success of this framework depends on alliance leaders' willingness to engage with the technical nuances of AI and on technologists' willingness to design for diplomatic constraints. If successful, this approach will not only secure the black box, but also ensure the black box remains secure. But in the end, it will become the foundation of a stronger, smarter, and more united alliance that is well-equipped to face the complex challenges of the 21st century.

BIBLIOGRAPHY:

Anjuna. 2025. U.S. Navy charters secure AI course with confidential computing. https://www.anjuna.io/case-studies/united-states-navy

Atlantic Council. 2024. *A marketplace for mission-ready AI: Accelerating capability delivery to the Pentagon* (Strategic Insights Memo). https://www.atlanticcouncil.org/content-series/strate gic-insights-memos/a-marketplace-for-mission-ready-ai-accelerating-capability-delivery-to-the-pentagon/

Bano, M., Chaudhri, I., & Zowghi, D. 2024. *Diplomacy in the age of generative AI* (arXiv:2401.05415). arXiv. https://doi.org/10.48550/arXiv.2401.05415

Belfer Center for Science and International Affairs. (2025, March 21). *Boosting interoperability of joint forces with AI: A unified language for joint warfighting*. https://www.belfercenter.org/research-analysis/boosting-interoperability-joint-forces-ai-unified-language-joint-warfighting

Bozkurt, M., Saylam, S., Saylam, R., & Gündoğdu, F. K. (2025, August 6). *Risk assessment of artificial intelligence support in the command and control cycle using spherical fuzzy z-number best-worst decision-making method*. NATO C2COE. https://c2coe.org/risk-assessment-of-ai-in-c2/

Confidential Computing Consortium. (2025). *Confidential computing: The future of data security*. https://confidentialcomputing.io/wp-content/uploads/sites/10/2025/11/US53866125.pdf

Curtea de Conturi a României. (2023). *Raport privind digitalizarea administrației publice* [Report on the digitization of public administration].

DefenseScoop. 2025, November 12. *Too much data, too few analysts: How AI offers a 'force multiplier' for intelligence analysts*. https://defensescoop.com/2025/11/12/too-much-data-too-few-analysts-how-ai-offers-a-force-multiplier-for-intelligence-analysts/

Department of Defense. (2024). *Responsible AI strategy and implementation pathway*. https://media.defense.gov/2024/Oct/26/2003571790/-1/-1/0/2024-06-RAI-STRATEGY-IMPLEMENTATION-PATHWAY.PDF

European Parliamentary Research Service. 2025. *Artificial intelligence in the military: Opportunities and challenges* (Briefing No. 769580). https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/769580/EPRS_BRI(2025)769580_EN.pdf

Gradiant. 2021. *Federated learning and secure multi-party computation: A powerful alliance for pri vacy-preserving AI*. https://gradiant.org/en/blog/trumpet-federated-learning-computation-secure/

Hadean. (2025). *Interoperability at the edge: The strategic imperative for NATO in an era of complex threats*. https://hadean.com/blog/interoperability-at-the-edge-the-strategic-imperative-for-nato-in-an-era-of-complex-threats/

Jensen, B., & Mishra, B. 2025. *Code, command, and conflict: Charting the future of military AI*. Belfer Center for Science and International Affairs. https://www.belfercenter.org/research-analysis/code-command-and-conflict-charting-future-military-ai

NATO. 2021. *Summary of the NATO artificial intelligence strategy*. https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2021/10/22/summary-of-the-nato-artificial-intelligence-strategy

NATO. 2024. *Summary of NATO's revised artificial intelligence (AI) strategy*. https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2024/07/10/summary-of-natos-revised-artificial-intelligence-ai-strategy

NATO. 2025. *Data strategy for the alliance*. https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2025/05/05/data-strategy-for-the-alliance

NATO Allied Command Transformation. (2025, June 23). *CWIX 25 concludes*. https://www.act.nato.int/article/cwix25-concludes/

Pollard, M. 2024. *Autonomous weapons systems and the inherent right of self-defense* [Doctoral dissertation, University of Buckingham]. http://bear.buckingham.ac.uk/624/1/1404758%20Michael%20Pollard%20Final%20Thesis.pdf

Reynolds, I., & Atalan, Y. 2024, July 8. *Calibrating NATO's vision of AI-enabled decision support*. Center for Strategic and International Studies.
https://www.csis.org/analysis/calibrating-natos-vision-ai-enabled-decision-support

Scaleout Systems. 2025. *Defense and security: Secure intelligence for mission-critical applications*. https://www.scaleoutsystems.com/defense-and-security

Sullivan, A., & Ricket, T. 2024. The black-box problem in AI-based weapon systems. In *CyCon 2024*. NATO CCDCOE. https://ccdcoe.org/uploads/2024/05/CyCon_2024_Sullivan_Ricket-1.pdf

U.S. Department of War. (2026, January 12). *Artificial intelligence acceleration strategy*. https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF

World Economic Forum. (2023, February). What is tech diplomacy? Experts explain. https://www.weforum.org/stories/2023/02/what-is-tech-diplomacy-experts-explain/