FORGING DIGITAL TRUST ACROSS BORDERS – ASSESSING THE ROLE OF EXPLAINABLE AI (XAI) IN OVERCOMING GEOPOLITICAL BARRIERS TO INTERNATIONAL COLLABORATION

Dumitru-Cătălin VASILE [a1]

[a] *National University of Political Studies and Public Administration, Romania*

ABSTRACT

This paper assesses the critical role of Explainable AI (XAI) as a techno-diplomatic mechanism for overcoming geopolitical barriers to international collaboration. In an era defined by global challenges such as pandemics, climate change, and systemic cyber threats, AI-driven solutions are imperative. However, their development and deployment are increasingly crippled by a profound "digital trust deficit" between nations. This deficit is fueled by the "black box" nature of advanced AI, compounded by fears of embedded espionage, economic data theft, and algorithmic bias. We argue that XAI—comprising systems and methods that make AI decision-making transparent and auditable—is emerging as a foundational tool for forging this necessary trust. This paper analyzes how XAI provides a common, verifiable language that allows untrusting state actors to collaborate on joint AI projects. It examines XAI's application in three critical domains: 1) enabling federated learning for global health while protecting data sovereignty; 2) ensuring interoperability and trust in multinational military coalitions; and 3) creating verifiable "glass box" models for international regulatory and climate agreements. The expected conclusions are threefold. First, XAI is a necessary, though not sufficient, condition for high-stakes international AI collaboration. Second, its primary function is not purely technical but diplomatic—acting as a neutral mediator that allows for verification without full disclosure of proprietary code or data. Finally, we conclude that the effectiveness of XAI is contingent upon a parallel effort to establish global governance standards, creating an "IAEA for algorithms" to certify that explanations are themselves accurate and trustworthy.

KEYWORDS: *Explainable AI (XAI), Digital Trust, International Collaboration, Geopolitics, AI Governance, Data Sovereignty, Techno-Diplomacy*

DOI: 10.24818/IMC/2025/05.07

1. INTRODUCTION

The 21st century is defined by a deep and accelerating paradox. The world's most existential challenges—from pandemic response and climate modeling to critical infrastructure security and supply chain management—are now global, data-intensive problems that demand global, AI-driven solutions. Yet, the very nations that must collaborate to build these solutions are simultaneously retreating into postures of digital nationalism, walled off by competing regulations and deep-seated geopolitical mistrust. We are in the midst of a "balkanization of governance" (Ifri, 2025), engaged in a counter-productive "AI arms race" (Bryson, 2018) when the moment demands an "AI collaboration."

∗Corresponding author. E-mail address: *catalin.vasile@outlook.com*

This "Great Decoupling" is most evident in the US-China rivalry, but its effects are systemic. Trust, the fundamental lubricant of international relations, is being eroded at the digital level. The root of this paradox is the "black box" problem scaled to the geopolitical level. The complex, opaque, and often non-deterministic nature of modern machine learning models creates an intractable trust deficit. How can allied nations build a joint command-and-control system if one partner cannot verify the targeting algorithm of another? How can nations pool sensitive epidemiological data for medical research if they fear a partner's AI will steal valuable genetic information?

The National Security Commission on Artificial Intelligence (NSCAI, 2021) starkly warned that without a new framework for "AI-enabled partnerships," the U.S. and its allies would fail to achieve the interoperability needed to defend their interests, ceding the digital high ground to strategic competitors. The "black box" is no longer just a technical inconvenience for computer scientists; it is a fundamental barrier to 21st-century diplomacy.

This paper argues that Explainable AI (XAI) is the most critical technological and diplomatic tool for resolving this paradox. By rendering the decision-making processes of AI systems transparent, interpretable, and auditable, XAI functions as a techno-diplomatic bridge—a neutral, verifiable language that can forge digital trust between untrusting parties. It provides a mechanism to move the discussion from a weak, politically untenable "trust me" posture to a robust, verifiable "verify me" framework.

This paper assesses the role of XAI in overcoming these geopolitical barriers. Section 2 will define the "black box" as a fundamental barrier to international collaboration, analyzing the three core fears that create the digital trust deficit: espionage, economic theft, and value misalignment. Section 3 will define XAI as a trust-building mechanism, moving beyond its technical definition (e.g., LIME, SHAP, interpretable models) to explore its role as a diplomatic mediator for verification, bias negotiation, and establishing a common technical language. Section 4 will assess XAI's practical application in three high-stakes case studies: enabling federated learning in global health, ensuring interoperability in military coalitions, and verifying compliance in international regulatory agreements. Section 5 will provide a critical analysis of XAI's significant limitations, arguing that it is not a silver bullet and can be deceptive. The paper will conclude that XAI is an essential foundation for 21st-century diplomacy, but one that requires a new, robust international governance framework—an "IAEA for algorithms"—to be truly effective.

2. THE GEOPOLITICS OF THE "BLACK BOX" – BARRIERS TO COLLABORATION

The "black box" is the core of the problem. In international relations, an unexplainable system is an untrustworthy one. This is not merely a technical inconvenience for engineers; it is a fundamental barrier to diplomacy. In a high-stakes, low-trust geopolitical environment, a "black box" system represents an unacceptable, unquantifiable risk. A nation's willingness to collaborate is built on its ability to verify a partner's actions and intentions. An opaque algorithm, whose decision-making logic is hidden, makes verification impossible.

This systemic lack of verifiability creates what we term the "digital trust deficit." It forces nations to assume worst-case scenarios about a partner's AI. Does a shared logistics AI contain a "kill switch"? Does a medical AI "steal" genetic data? Does a financial model embed a partner's economic bias? Without explainability, these questions cannot be answered. This geopolitical trust deficit, therefore, manifests in three primary, overlapping, and often mutually reinforcing barriers to collaboration, which this paper will now detail.

2.1 Barrier 1: National security and espionage (The "Trojan Horse" fear)

In the context of defense and intelligence, sharing an AI model with an ally is an unprecedented risk. An adversary—or even a "frenemy"—could theoretically embed hidden functionalities within a complex neural network. This "Trojan Horse" concern is twofold:

- **Data exfiltration**: A model shared for a joint task (e.g., logistics optimization) could be covertly designed to identify and exfiltrate sensitive data (e.g., troop locations, supply deficits, fuel-stock data) back to its home nation. Because the model's logic is opaque, the user nation has no way to verify *how* the model is processing its data, or what "hidden" computations it is performing.
- **Kill switches & hidden triggers:** An AI system for a joint weapons platform or a shared command-and-control (C2) system could contain a "backdoor" or a specific, non-obvious trigger. For example, a shared targeting AI could be secretly trained not to recognize a certain class of aircraft, or to fail when it detects a specific (and secret) electronic warfare signal. In a future conflict, an ally could find its "smart" systems disabled or even turned against it.

As Lim and Liu (2019) illustrates in *Army of None*, the speed and autonomy of AI systems mean that human oversight is already strained. In a coalition "sensor-to-shooter" link, where an AI sensor from one nation passes a target directly to an AI-guided weapon from another, trust must be absolute and instantaneous. If the *allied AI* itself cannot be trusted, the entire foundation of a coalition force collapses.

2.2 Barrier 2: Data sovereignty and economic competition (The "Data Theft" Fear)

The "data is the new oil" maxim has led to a new era of digital protectionism. Nations and blocs are erecting "data borders" to protect their economic and social interests, most notably the European Union's General Data Protection Regulation (GDPR) and China's Personal Information Protection Law (PIPL). These regulations strictly govern the transfer of citizen data, effectively halting many international AI research projects.

The "black box" is a direct threat to data sovereignty. A nation cannot share its sensitive data (e.g., proprietary economic data, citizen genetic records, or corporate IP) to train a joint AI model if it cannot get a verifiable guarantee of how that data is being used, stored, or if it is being "memorized" by the model. This is not a hypothetical risk. Advanced models have been shown to "memorize" and "regurgitate" specific pieces of their training data (Carlini et al., 2019). A corporation collaborating on a "black box" drug discovery model with a foreign state's research lab fears that the resulting model has "stolen" its proprietary chemical formulas. This fear of "data-mining-by-proxy" stifles collaboration in everything from green technology to pharmaceutical research.

2.3 Barrier 3: Ethical and value misalignment (The "Values" Fear)

AI systems are not neutral; they are "opinions embedded in code" (paraphrasing Doyle, 2017). An AI's objective function, trained on one nation's data and values, will necessarily reflect those values. This creates an "algorithmic culture clash" when these systems are shared.

- An AI model for social media content moderation trained in the U.S. (prioritizing free speech) would be functionally and legally useless in Germany (which has strict laws against hate speech).
- A "social credit" model from China, designed for state-sponsored surveillance and social cohesion, is antithetical to the liberal democratic values embedded in the EU's *AI Act*, which prioritizes individual rights.

The EU's *AI Act* (2024) is a prime example of this "value-based" governance. It places strict transparency and "right to explanation" requirements on "high-risk" AI systems. A U.S. or Asian tech firm cannot simply deploy its "black box" model in Europe. It must be able to *explain* its function to regulators. This value-driven legal barrier, in the absence of XAI, becomes a hard stop for international technology deployment and collaboration.

3. EXPLAINABLE AI (XAI) AS A TECHNO-DIPLOMATIC BRIDGE

Explainable AI (XAI) refers to a set of methods and technologies that produce "AI systems whose learned models and decisions can be understood and appropriately trusted by end users" (Gunning & Aha, 2019). The "black box" problem, as defined in the previous section, is characterized by models that only provide predictions (an answer), forcing partners to accept them on faith. XAI, in contrast, provides the reasoning behind the answer. It moves beyond just predicting (e.g., "this is a threat") to explaining (e.g., "this is a threat because it matches these three target signatures").

This paper argues that this technical capability has a profound and under-appreciated diplomatic function. We introduce the "techno-diplomatic bridge" framework to re-situate XAI not just as a tool for computer scientists or domestic regulators, but as a critical enabler of international relations in a low-trust environment.

Its function is diplomatic because it fundamentally alters the foundation of collaboration. It provides a mechanism for untrusting partners to bypass the need for political "trust" (e.g., "trust us, our model is safe") and replace it with technical, auditable "verification" (e.g., "verify our model for yourself"). XAI thus acts as a neutral third-party mediator—not a person, but a process—in which all sides can place their confidence. It provides a common, verifiable ground for untrusting partners, built on three core mechanisms which this section will now explore: verification, bias negotiation, and a common technical language.

3.1. Mechanism 1: Verification and auditability ("The Glass Box")

XAI's primary role in geopolitics is to enable verification. It allows a partner nation to "audit the algorithm" without needing access to the full, proprietary source code or training data—a critical compromise. This verification can take two forms:

- **Post-hoc explanations (Interrogating the Black Box):** These tools (e.g., LIME, SHAP, Counterfactuals) allow a partner to "interrogate" a black box model that is already trained. By feeding it hypothetical inputs, the partner can see *why* the model made specific decisions. For example, a military ally could test a targeting AI by feeding it 10,000 images of friendly, neutral, and hostile forces and *verifying* that the model's "reasons" (e.g., the key pixels it focused on) for classifying each are correct and not based on spurious or hidden factors.
- **Inherently interpretable models ("Glass Boxes"):** In many high-stakes cases, as Cynthia Rudin (2019) has argued, "Stop explaining black box models... and use interpretable models instead." Here, the diplomatic solution is to *forbid* black boxes for a specific collaborative task. Nations can collaboratively agree to use a simpler, *inherently interpretable model* (like a decision tree or a linear regression) where the logic is plain to all parties. XAI's role here is to *prove* that this simpler model is "good enough" for the task, trading a small amount of accuracy for 100% transparency.

This auditability directly counters the "Trojan Horse" fear. It allows a nation to build confidence that an AI partner is functioning *only* as specified.

3.2. Mechanism 2: Surfacing and negotiating algorithmic bias

When collaborating on AI for social good (e.g., health, economics), XAI is the only tool for detecting and negotiating bias *across borders*. A model deemed "fair" in one country may be deeply biased in another.

- **Example:** A medical AI trained on predominantly Western European data to detect skin cancer may be 99% accurate for that demographic but fail dangerously on individuals with darker skin tones (Adadi & Berrada, 2018).

- **XAI's role:** Before deploying this model, a partner nation in Africa or South Asia could use XAI tools to *diagnose* this bias. The XAI would explicitly show *why* the model is failing—e.g., "it has learned to associate 'melanoma' with 'low-melanin-pixel-regions'." This technical, verifiable proof of bias allows the international team to move beyond political accusations of "algorithmic racism" and into a technical solution, such as targeted data augmentation or re-weighting the model.

3.3. Mechanism 3: A common, neutral language for diplomacy

In a politically charged negotiation, "trust" is a flimsy word. XAI provides a common, neutral, and *technical* language for discussion. Instead of a diplomat saying "We promise our AI is fair," a data scientist from their delegation can say, "Here is the SHAP plot showing the feature-importance for this decision; let's analyze it together." This moves the conversation from the political domain to the technical, allowing experts from untrusting nations to find common ground based on verifiable, reproducible evidence (Miller, 2017). This "techno-diplomatic" dialogue is essential for building the confidence needed for high-stakes collaboration.

4. ASSESSING XAI'S ROLE - CASE STUDIES IN INTERNATIONAL COLLABORATION

The true test of XAI is not in theory but in practice. Its role can be assessed in three critical domains where collaboration is currently stalled.

Case 1: Global health and Federated Learning (FL)

- **Problem:** To build a powerful AI to predict the next pandemic, researchers need diverse, global health data. However, no nation will (or legally can) share its sensitive citizen health records (HIPAA, GDPR) to a central server.
- **The FL solution:** Federated Learning (FL) is a technical solution where the private data never leaves its home country. Instead, the global AI *model* is sent to each nation, trained locally on their private data, and then only the *model updates* (mathematical parameters) are sent back to a central server to be aggregated.
- **The Trust gap:** This still requires trust. How does Germany know the U.S. model isn't a "Trojan Horse" designed to "leak" or "memorize" its data in the model updates? (Abdar et al., 2024).
- **XAI's role as bridge:** XAI is the *auditing layer* for FL. A partner nation's data scientists can use XAI tools to interrogate the model *before* and *after* it trains on their local data. They can use counterfactual explanations ("If I changed Patient X's record, how much does the model's update change?") to *prove* that the model is only learning high-level statistical patterns, not specific, private individual data. This provides the verifiable, technical guarantee of privacy that allows a nation to participate.

Case 2: Military coalitions and AI-Speed interoperability

- **Problem:** A core goal for NATO and other alliances is "interoperability." With AI, this is a crisis. A U.S. autonomous drone swarm cannot effectively partner with a German air defense system if the two systems' "black box" AIs cannot trust or understand each other's recommendations. This creates "a risk of fatal hesitation or error" (NSCAI, 2021). The "OODA loop" (Observe-Orient-Decide-Act) is now moving at machine speed.
- **XAI's role:** XAI provides **"explainability-for-interoperability."** When an AI system (e.g., a U.S. drone) passes a recommendation to a human (e.g., a German commander), it must be accompanied by an XAI-generated "explanation."
- **Assessment:** The recommendation "Target T-72 at 10:00" will be rejected. The recommendation "Target T-72 at 10:00; *Reason: [99% visual match], [95% thermal signature], [88% signals intercept], [0% friendly signals detected]*" can be trusted and acted upon. Critically, this explanation must be both **human-readable** (for the commander) and

machine-readable (so the German AI can *ingest* the explanation and fuse it with its own data). XAI thus becomes the high-speed protocol for "meaningful human control" *across* coalition forces.

Case 3: International regulation (Climate and Finance)

- **Problem:** To enforce international agreements (like the Paris Agreement or anti-money laundering regulations), nations need to use AI to monitor and verify compliance. How can the world trust a nation's "black box" AI model that *claims* it has met its carbon emissions targets or that its banks are not laundering money? The risk of "algorithmic greenwashing" or financial deception is enormous.
- **XAI's role:** Here, XAI's role is to *advocate for simplicity and transparency*. As argued by Rudin (2019), for high-stakes public policy, "black boxes" should be rejected entirely.
- **Assessment:** XAI's role is to *prove* that a simpler, "glass box" model is "good enough." International bodies (like the UN, IMF, or the IPCC) can mandate that all member states use an *inherently interpretable model* for their reporting. For example, a "carbon compliance" model could be a simple, open-source decision tree that all nations can audit. Its logic would be transparent: "Emissions are marked 'high risk' *if* [Satellite-Data-Source-X = 'High Methane'] *and* [Shipping-Manifest-Data = 'False']." The model's shared, open-source, and explainable nature serves as the trust mechanism.

Case 4: Transnational Counter-Terrorism and intelligence sharing

- **Problem:** Combating global terrorist networks or organized crime syndicates requires multiple nations to share highly sensitive, "sources-and-methods" intelligence data to build a complete picture. This is arguably the lowest-trust environment in international relations. If Agency A (from Country A) shares its data with an AI analysis platform built by Country B, it has two core fears: 1) Country B's AI will "steal" its sources (e.g., memorize the identity of a human asset) or 2) The AI is a "Trojan Horse" that will feed Agency A's analysts bad, manipulated intelligence.
- **XAI's Role:** XAI provides **"sources and methods verification."** Before sharing any real data, Country A's technical experts can "red team" Country B's AI in a secure sandbox.
- **Assessment:** Using post-hoc explanation tools, Country A can feed the AI thousands of pieces of test data. They can then verify the AI's reasoning: "Did the AI flag this person as a 'high-value target' *because* of the correct data patterns, or because they were associated with a 'honeypot' keyword that a malicious actor embedded?" Furthermore, they can use XAI to test the model for "memorization" (Carlini et al., 2019), ensuring that when the model is trained, it does not store and cannot "regurgitate" the specific names, locations, or identifiers of their sensitive assets. This technical audit provides the *only* possible basis for collaboration in such a high-stakes domain.

5. THE LIMITS OF EXPLANATION - "WHO EXPLAINS THE EXPLAINER?"

Assessing XAI's role requires a critical examination of its significant limitations. Believing that a technical tool can, by itself, solve a deeply political problem like mistrust is a dangerous oversimplification. In a geopolitical context, any proposed solution will be immediately and rigorously "red-teamed" by all parties. Nations will not just ask "does it work?" but "how can it be cheated?" and "what are its hidden risks?".

This critical analysis is therefore essential to the paper's central argument. If XAI is to be a viable techno-diplomatic bridge, we must first understand its weaknesses and the new forms of risk it introduces. The following limitations are not just technical inconveniences; they are fundamental barriers to diplomatic acceptance. The question "Who explains the explainer?" is not just philosophical; it is the central governance challenge. If the explanation itself cannot be trusted, the entire framework collapses.

5.1. The Accuracy vs. Explainability trade-off

The most powerful AI models, deep neural networks, are often the least transparent. The most transparent models, like decision trees, are often less accurate. This forces a difficult choice for international partners: Do they use a "glass box" model that is 100% transparent but only 90% accurate, or a "black box" that is 99% accurate but only 30% explainable? In a high-stakes domain like military defense or medical diagnosis, that 9% gap can mean lives. This trade-off must be explicitly negotiated at the diplomatic level for each specific collaborative project.

5.2. Explanations can be misleading or gamed (Adversarial XAI)

This is the most dangerous limitation. Post-hoc explanation methods (like LIME or SHAP) are not perfect "ground truth." They are *approximations* of the model's behavior. A sophisticated nation-state could, in theory, engage in **"adversarial XAI"**—designing a malicious AI that *appears* to be making a fair decision and can even *generate a plausible-sounding fake explanation* for it, all while its true, hidden logic is malicious. The explanation itself becomes a new attack surface. This means that a nation cannot just trust an explanation; it must have the deep technical expertise to *verify the explanation method itself*.

5.3. Cognitive overload and the "Useless" explanation

An explanation is only useful if it is *understandable* to its human user. A 1000-page SHAP plot (a common XAI output) delivered to a general or a diplomat is not a useful explanation; it is cognitive "noise." A key challenge is "human-in-the-loop" design: creating explanations that are tailored to the *user* (expert vs. novice) and the *time-frame* (a millisecond explanation for a soldier vs. a one-week explanation for a regulator). An "explanation" that cannot be understood is functionally identical to no explanation at all (Lipton, 2016).

5.4. The Governance chasm, a "Tower of Babel"

This is the central geopolitical barrier. XAI is useless as a trust mechanism if there is no shared, international standard for *what constitutes a good, clear, or sufficient explanation*.

- The EU's *AI Act* may demand one level of explanation (e.g., "right to explanation").
- The U.S. NIST *AI Risk Management Framework* may suggest another (e.g., risk-based "explainability tiers").
- China's governance rules will demand a third.

Without a common, enforceable set of XAI standards, we create a "Tower of Babel" where everyone is "explaining" in a different language. This is where the technical problem becomes a purely diplomatic one.

6. CONCLUSION

The "black box" of AI, when scaled to the geopolitical stage, is a powerful engine of mistrust, fragmentation, and strategic paralysis. It is a critical barrier that prevents humanity from collaborating on its most pressing shared challenges.

This paper has argued that Explainable AI (XAI) is the most viable techno-diplomatic tool for bridging this digital trust deficit. It is not merely a technical feature but a political mediator. By providing a common, verifiable language, XAI allows untrusting nations to engage in a "verify, then trust" framework. It enables collaboration in high-stakes domains—from global health and climate change to the very interoperability of military alliances—by allowing partners to audit algorithms, detect biases, and verify behavior without demanding the full, compromising disclosure of proprietary data or code.

However, this assessment concludes that XAI is not a silver bullet. Its technical limitations—the trade-off with accuracy, the potential for deceptive explanations, and the risk of cognitive overload—

mean that XAI itself must be governed. The ultimate challenge is not technical but diplomatic. We must not only *build* explainable systems; we must build the *international governance*—the shared standards, certification bodies, and treaties—that make those explanations politically meaningful.
There is a clear and urgent need for a new international body, an **"IAEA for Algorithms,"** that can serve as a neutral, third-party auditor to *certify* AI explanations and build a common, technical language for digital trust. Forging this "glass box" governance is the great, hybrid diplomatic-technical project of the 21st century.

REFERENCES

Abdar, M., Zomorodi-Moghadam, M., Kakhani, M., et al. (2024). SemFedXAI: A Semantic Framework for Explainable Federated Learning in Healthcare. *Electronics, 13*(6), 435. https://doi.org/10.3390/info16060435

Adadi, A. & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/access.2018.2870052

Bryson, J. J. (2018). *The 'AI arms race' and international stability.* Center for a New American Security (CNAS).

Carlini, N., Farid, U., Pagnan, C., et al. (2019). *The Secret Sharer: Measuring Unintended Data Leakage in ML*. 28th USENIX Security Symposium.

Gunning, D. & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine, 40*(2), 44–58. https://doi.org/10.1145/3301275.3308446

Institut Français des Relations Internationales (Ifri). (2025). *Artificial Promises or Real Regulation? Inventing Global AI Governance*.

Lipton, Z. (2016). The Mythos of Model Interpretability. *Communications of the ACM. 61*. https://doi.org/10.1145/3233231

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*. https://doi.org/267. 10.1016/j.artint.2018.07.007

National Security Commission on Artificial Intelligence (NSCAI). (2021). *Final Report*. https://www.google.com/search?q=https://www.nscai.gov/reports/

Doyle, T. (2017). Weapons of Math Destruction: How big data increases inequality and threatens democracy. *The Information Society, 33*(5), 301–302. https://doi.org/10.1080/01972243.2017.1354593

Rudin, C. (2019). Stop explaining black box models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Lim, D., & Liu, R. (2019). Army of None: Autonomous weapons and the future of war. *Journal of Military Ethics, 18*(2), 165–167. https://doi.org/10.1080/15027570.2019.1637555