RESEARCH ARTICLE OPEN ACCESS

Enhancing the Security of AI-Driven Systems in Space Exploration and Research

Audrey Tobesman[1], Dr. Alex Mathew [2]
Bethany College, West Virginia, USA

This research was made possible by NASA West Virginia Space Grant Consortium, NASA Agreement #80NSSC25M7079

ABSTRACT
AI's increasing presence in space exploration platforms has allowed mission opportunities from autonomous navigation to real-time anomaly detection. However, this progress brings with it a new category of cybersecurity problems that traditional cybersecurity models do not expect. This paper provides a comprehensive review of the adversarial attack vectors for AI systems in space environments, such as input manipulation, poisoning of training data, degradation of AI models, and exploitation of the supply chain. Based on the 2022 incident with ViaSat KA-SAT, which knocked out approximately 45,000 modems in Europe, this research highlights major shortcomings of today's space security stances. It is proposed to implement a five-layer defense-in-depth approach, based on guidance from NIST CSF 2.0 [5], NIST AI RMF 1.0 [6], MITRE ATLAS [7], and CCSDS standards [9]. Architecture is designed to address data integrity, model robustness, infrastructure hardening, encrypted communications, and autonomous operational resilience, all of which are specific to space deployment restrictions. The framework is validated as it is practically effective in the case of the KA-SAT attack chain.

Keywords:- space cybersecurity, adversarial machine learning, satellite security, zero trust architecture, NIST AI RMF, MITRE ATLAS, defense-in-depth

I. INTRODUCTION

A. Background and Motivation

The traditional software pipelines, which are deterministic, are no longer used for modern space missions. AI is now a component of several primary mission capabilities, including terrain recognition, detecting anomalies in orbit, and developing an autonomous path for deep space rovers. For programs such as NASA's Artemis mission and interplanetary robotic missions, where the ground controller cannot monitor everything in real-time, machine learning models are used to process and analyze the environmental data and make quick decisions [8].

However, this dependency brings a new threat scenario to mission architectures. The traditional satellite security techniques involved link encryption, software integrity checks, and handling of access credentials. An AI system can be vulnerable, just like any system, at the mathematical model level: Adversarial perturbations are tiny changes in the state space that are undetectable by the human eye, but sufficiently large to fool the neural network into thinking that a landing site is safe when it actually is not [6, 7]. The training data might be dormant for many years, with only a specific set of operating conditions to take effect from the use of the training data [12]. Unlike conventional software vulnerabilities, adversarial ML weaknesses are often rooted in model behavior and may require retraining, architectural modification, or operational mitigation rather than traditional patching.

The distinctive operational complexities in space only compound these risks when the same is paired with AI fragility. The time for a round-trip in deep-space communications may be more than 40 minutes. After the launch, no technician can get on board a spacecraft. Radiation-

hardened processors have a much lower speed than processors on Earth [16]. These realities render traditional cyber-defense approaches such as patching, validating with humans, and changing hardware impractical.

B. Problem Statement

Space environments were not a factor in the development of existing cybersecurity frameworks. The NIST AI Risk Management Framework offers comprehensive guidance for managing AI-related risk. However, it does not consider extreme signal latency, hardware degradation due to radiation, or long-term deployments with only 10-year access of space-based systems [6]. MITRE ATLAS enumerates adversarial ML tactics but assumes ground-based infrastructure with the ability to intervene rapidly by humans. MITRE ATLAS lists adversarial ML tactics but assumes a ground-based infrastructure with rapid human intervention. The CCSDS security standards are for space communication protocols, but these standards were written before the widespread use of machine learning in operation [9]. The NIST CSF 2.0 provides enterprise cybersecurity controls, but it does not include any specific controls for AI [5]. With no common, space-appropriate AI security standard, mission architects are left with no clear guidance, just at the time when spacecraft autonomy is increasing [1] [4].

C. Research Contributions

This paper delivers five original contributions:

1. A space-adapted taxonomy of AI-specific cyber threats organized by attack mechanism, target layer, and mission consequence.
2. Root cause identification of five vulnerability classes unique to AI-enabled spacecraft.
3. A five-layer, defense-in-depth security framework reconciling NIST CSF 2.0 [5], NIST AI RMF [6], MITRE ATLAS [7], and CCSDS [9] into a unified space architecture.
4. Retrospective framework validation against the 2022 ViaSat KA-SAT cyberattack [2], [3].
5. Constraint-aware implementation guidance for each security layer.

II. RESEARCH METHODOLOGY

The research was divided into two complementary phases. Phase I reviewed relevant literature from 2018 to 2026, incident reports from documented space cyberattacks [2, 3], and the NIST CSF 2.0 [5], NIST AI RMF [6], MITRE ATLAS [7], and CCSDS [9] standards. Phase II synthesized findings and came up with the proposed security framework.

The MITRE ATLAS matrix structure [7] was adapted for threat identification, taking into consideration amplification factors associated with the space. These threats were categorized along four dimensions: attack vector, targeted architectural layer, mission impact category, and space-specific risk multiplier. The gap analysis, extraction of applicable cross-standard controls, constraint mapping, and retrospective validation against the KA-SAT incident [2] and [3] were the steps followed in the design of the framework.

III. THREAT LANDSCAPE FOR AI-DRIVEN SPACE SYSTEMS

A. Threat Taxonomy

Table 1 provides a structured classification of threats specific to AI in space mission environments along with space-relevant examples and mission impacts [7, 11, 17].

Table 1: AI-Specific Threat Taxonomy for Space Environments

Threat Class	Attack Mechanism	Space Example	Mission Impact
Adversarial Evasion	Imperceptible input perturbation	Terrain misclassification from altered imagery	Navigation or landing errors

Data Poisoning	Malicious injection into training sets	Corrupted Earth observation datasets	Backdoored model behavior
Model Drift	Distribution shift over deployment lifetime	Martian dust altering sensor baselines	Undetected accuracy degradation
Backdoor Attack	Trigger-conditioned malicious response	Sensor pattern activating unsafe command	Latent mission-critical failure
Supply Chain Attack	Compromised third-party model components	Embedded backdoor in pretrained vision module	Multi-mission systemic vulnerability
Model Extraction	Query-based parameter reconstruction	API probing of space-ground ML service	Intellectual property loss; attack replication

B. Space-Specific Vulnerability Amplifiers

AI security risks are exacerbated in space operations by six characteristics of the structure of space operations [4], [14]:

Communication Latency: The time for a round trip to Mars is more than 48 minutes, which makes it impossible to verify in real time the decisions made by AI systems [16].

Patch Infrequency: Radiation-hardened storage is limited in write endurance, and orbit constraints limit update windows. Any vulnerabilities found after launch could last the entire mission [4].

Hardware Constraints: The space-grade processors, like the RAD750, run at ~200MHz, and are not equipped with the trusted execution environments and cryptographic accelerators common in terrestrial AI processors [9].

Extended Mission Lifecycles: The models deployed today should continue to perform well for 10-20 years in the changing environment and against the changing capabilities of adversaries [1], [16].

Dual-Use Tensions: Military payloads are regularly flown on commercial satellites. Commercial cost pressures lead to security gaps, which are exploited by adversaries as in the case of KA-SAT [2,3].

Physical Inaccessibility: Post-launch hardware inspection, replacement, and forensic analysis are not possible. Only software-only remediation can be done for recovery [4].

C. Existing Standards Gap Analysis

Table 2 compares the existing frameworks with the space-AI security requirements. No single current standard covers all aspects of AI-specific threats in space-operational environments, as illustrated [5], [6], [7], and [9].

Table 2: Standards Gap Analysis for Space-AI Cybersecurity

Standard	Primary Scope	Space-AI Coverage	Critical Gap
NIST CSF 2.0 [5]	Enterprise cybersecurity	None	No AI controls; no space operational constraints
NIST AI RMF [6]	AI risk management	Indirect	No latency, radiation, or patch-constraint guidance
MITRE ATLAS [7]	Adversarial ML tactics	None	Assumes terrestrial, human-accessible environments
CCSDS [9]	Space data communication	Pre-AI only	No adversarial ML guidance whatsoever

IV. CASE STUDY: VIASAT KA-SAT ATTACK (2022)

A. Incident Summary

On February 24, 2022, the first day of Russia's invasion of Ukraine, a coordinated cyber-attack disrupted ViaSat's KA-SAT satellite communications network. Some 40,000 to 50,000 modems were disabled throughout Ukraine and European other countries, effectively shutting down the military's communications at its most vulnerable time. Collateral disruption spread to 5,800 wind turbines in Germany, and civilian internet services in several countries [2] [3].

B. Attack Chain

Phase 1 – Credential Exfiltration (2021): The attackers exploited a publicly disclosed FortiGate VPN vulnerability, CVE-2018-13379, to steal credentials from almost 500,000 devices worldwide, including those protecting the network infrastructure of satellite communications provider ViaSat [2].

Phase 2 – Initial Intrusion (February 24, 2022): The attackers used stolen credentials to log in to the KA-SAT ground segment VPN. No multi-factor authentication was enforced, and the VPN gave unrestricted access to the internal network management system [3].

Phase 3 – Lateral Movement and Payload Delivery: The attackers used the compromised VPN appliance to propagate laterally to the network management system and then installed AcidRain, a destructive wiper specifically designed to overwrite the modem firmware, to about 45,000 endpoints [3].

Phase 4 – Sustained Denial of Service: Wiped modems would need to be replaced or manually reflashed. At the scale of the conflict, it was impossible to recover from the attack during the attack window [2, [3].

C. Lessons for AI-Enabled Space Systems

The AI-specific techniques used in the KA-SAT attack were not present, but the incident highlights a critical future risk: an attacker who can disable 45,000 satellite modems today by using ground segments could do the same tomorrow in an AI-controlled spacecraft, causing collision trajectories, corrupting navigation models, and triggering dangerous autonomous actions [4] [17]. Four lessons can be directly transferred:

- Ground infrastructure needs to be considered as a space asset: model repositories, training pipelines, and validation systems are as crucial as the spacecraft bus itself [2].
- Known vulnerabilities are the main attack vector, as is the case with adversarial ML, which uses well-known model architectures to do so, similarly to how CVE-2018-13379 exploited a known vulnerability in a VPN [6] [7].
- No segmentation means cascade failures – one compromised trust boundary can be passed to 45,000 endpoints [3].
- Autonomous defense is not an option: space systems outside of cislunar distance cannot wait for human intervention in case of active attack [16].

V. PROPOSED FIVE-LAYER SECURITY FRAMEWORK

The framework follows a defense-in-depth approach [5] that assumes no single layer of the system bears the entire security responsibility. When one layer fails, other layers prevent it from affecting the whole system and alert the rest of the system of a failure. All controls are developed under space operational conditions: limited computing power, infrequent contact with the ground, radiation exposure, multi-decades lifetime [4], [9].

Layer 1 – Data Security

Objective: Maintain the integrity, authenticity, and confidentiality of all data that is ingested or generated by AI systems during the mission.

All telemetry, command, and scientific data streams are transmitted in a manner that provides confidentiality (AES-256-GCM) and integrity verification (HMAC-SHA256), as recommended by CCSDS [9]. The long-term missions need to use post-quantum algorithms, namely CRYSTALS-Kyber for key encapsulation and CRYSTALS-Dilithium for digital signatures [4]. Pre-ingestion pipelines use statistical outlier-detection algorithms, namely isolation forests and autoencoder anomaly detectors [6]. The sensor data is checked for spoofing or injection before reaching the model inference through cross-sensor fusion and physics-based consistency checks [14].

Layer 2 – Model Security

Objective: Make AI models more resistant to adversarial manipulation, unauthorized extraction, and performance degradation throughout the mission.

Adversarial examples are generated using Projected Gradient Descent and Carlini and Wagner methods and then used to train models with a mixture of the two training examples [7]. All models are evaluated in a red-team scenario prior to deployment, against the MITRE ATLAS tactic matrix [7], which includes scenarios for evasion (TTA0001), poisoning (TTA0002), and extraction (TTA0003). Ed25519 signatures are used to sign and cryptographically verify model updates that are sent to spacecraft before loading [5]. In continuous drift monitoring, the maximum mean discrepancy between training and deployment distributions is used. If the drift is above a certain threshold, the system switches back to conventional deterministic algorithms [6, 12].

Layer 3 – System Security (Zero Trust Architecture)

Objective: Eliminate implicit trust at each level of the system stack to protect spacecraft and ground computing infrastructure.

Any access, either within or outside the network, is authenticated and authorized before being executed, following the basics of Zero Trust Architecture [5]. With micro-segmentation, AI inference, training, and data preparation are all restricted to their own designated trust zones and only permitted to communicate with each other via secure, encrypted channels that are mutually authenticated. The mode that uses hard-wired ID based on physically unclonable functions (PUFs) is used to combat the spoofed authentication in radiation-hardened field-programmable gate arrays (FPGAs) [9]. A validated Software Bill of Materials cryptographically secures all third-party libraries and pre-trained models [13]. VPN concentrators remain in remote DMZ locations that do not have direct interfaces to network management systems (as in the case of the KA-SAT failure).

Layer 4 – Communication Security

Objective: Protect all data connections from the space-ground-relay network from interception, injection, replay, and jamming.

All Telemetry and Telecommand channel encryptions are under the control of the CCSDS Space Data Link Security Protocol [9]. AES-256-GCM using different nonces per packet protects against replay attacks. To jam-proof critical command links, frequency-hopping spread spectrum is used. The communication paths are logically partitioned according to the level of trust: emergency command, operational telecommand, payload downlink, and inter-satellite links have different levels of trust and bandwidth requirements [9, 11]. A command rate and pattern sequence anomaly detection detects injection attempts, before

they reach mission-critical systems, on command [15].

Layer 5 – Operational Resilience

Objective: Enable continuous security visibility, autonomous threat detection, and self-contained response when there is no way to intervene from the ground.

Real-time monitoring is provided for all four lower layers and passed to an onboard Security Information and Event Management (SIEM) system with anomaly detection using an autoencoder algorithm trained on normal mission telemetry [6]. Predefined response playbooks are automatically executed: if a poisoning event is detected, rollback is performed to the last cryptographically verified model version, and if an unauthorized access is detected, the corresponding trust zone is isolated, and the hardened emergency command channel is activated [5, 7]. Byzantine fault tolerance [16] is achieved by using N-version redundancy, where three or more independently developed AI models are used for each safety-critical function. Adversaries can try to remove evidence after an incident, but tamper-evident, hash-chained logs will allow forensic capability to be retained after the incident [18].

VI. FRAMEWORK VALIDATION AGAINST KA-SAT

Table 3 shows a correlation between each failure documented in KA-SAT [2] and [3] and the corresponding framework of control that could have acted.

Table 3: Retrospective Validation of Framework Against KA-SAT Attack Chain

KA-SAT Failure	Framework Layer	Applicable Control	Projected Outcome
Known CVE unpatched for 3 years [2]	Layer 3: System Security	Automated vulnerability scanning; patching at ground contact windows	An attack requires a zero-day; it substantially raises the attacker's cost
No MFA on VPN access [3]	Layer 3: System Security	Mandatory multi-factor authentication for all privileged access	Stolen credentials alone are insufficient—the attack was prevented
VPN directly accessed NMS [3]	Layer 3: System Security	VPN in DMZ; micro-segmentation blocks NMS path [5]	Lateral movement blocked; attacker contained to the perimeter
45,000 modems infected before detection [2]	Layer 5: Resilience	Behavioral baselining; rate anomaly detection on command patterns [6]	Wiper propagation detected within the first deployment cluster
No segment isolation capability [3]	Layer 5: Resilience	Automated containment playbook; safe-mode failover	Infection contained; unaffected modems preserved
No forensic evidence retained [2]	Layer 5: Resilience	Tamper-evident hash-chained logs; selective downlink on anomaly detection	Post-incident attribution and analysis accelerated

The retrospective analysis shows that none of the controls alone are effective in preventing the attack. But when multiple layers of redundancy exist, the event is a top-to-tail failure handled as an itemized event that can be discovered at various stages. The attack chain—VPN compromise → NMS access → 45,000 modem wipeout—required every security layer to fail simultaneously [2], [3]. The suggested framework separates this dependence at three different points.

VII. LIMITATIONS AND FUTURE RESEARCH

A. Limitations

The framework has been tested in a single well-documented event and validated by retrospective testing. The future is empirical validation through the implementation of controls in representative space hardware or high-fidelity simulation. The performance thresholds, such as acceptable inference latency, the minimum anomaly detection rates, are mission-specific and must be analyzed mission-specifically [9]. The adversary model assumes an advanced, nation-state-level adversary similar to the KA-SAT attribution [4]. Side channel model inversion and electromagnetic fault injection are new attack methods that could necessitate framework extensions [7], [17].

B. Future Research Directions

- Test and evaluate framework controls on CubeSat-class platforms, assessing security benefits in terms of compute and power consumption [9, 16].
- Create automated red-teaming toolchains to create mission-realistic adversarial examples for spacecraft AI systems [7].
- Generalize formal verification techniques for neural networks to space-relevant properties, e.g., attitude control stability bounds [6].
- Create satellite constellation-based federated threat intelligence protocols for anonymously sharing attack indicators [15].
- Compare and contrast the performance of CRYSTALS-Kyber and CRYSTALS-Dilithium on radiation-hardened FPGAs for deep-space quantum-resistant communications [4].

VIII. CONCLUSION

The use of artificial intelligence and autonomous spacecraft systems provides mission capabilities that were not previously available, and at the same time presents new cyber threats to those missions that are not yet addressed in a unified guidance [1], [4]. By systematically categorizing threats, analyzing the vulnerabilities of AI architectures [6, 7], and examining the 2022 attack on ViaSat KA-SAT [2, 3], this paper pinpoints the exact vulnerabilities that make AI-based space systems vulnerable.

The proposed five-layer defense-in-depth architecture (data security, model robustness, zero trust system architecture [5], encrypted communications [9], and autonomous operational resilience) offers an actionable and constraint-aware system architecture for existing and future space missions. This research combines the NIST CSF 2.0 [5] with the NIST AI RMF 1.0 [6] and the MITRE ATLAS [7] and the CCSDS [9] standards into a unified space-adapted model to address the inconsistencies across the standards and provide clear implementation guidance.

The time delay of Earth-based supervision is too great to make a crewed presence to the moon, Mars, and beyond practical [16]. Space systems need to be designed to autonomously, continuously, and reliably defend themselves for decades of mission life. The framework outlined here is a starting point towards AI systems that are not only intelligent but also trusted stewards of human expansion into the solar system [8].

REFERENCES

[1] V. R. Glick, A. H. Bruder, A. Futerman, K. Steval, H. G. Diament, and Z. Wawrzyniak, "Securing the final frontier: Cybersecurity risk, regulation, and compliance trends in space and satellite operations," Mayer Brown, Dec. 2025. [Online].

Available: https://www.mayerbrown.com/en/insights/publications/2025/12/securing-the-final-frontier-cybersecurity-risk-regulation-and-compliance-trends-in-space-and-satellite-operations

[2] N. Boschetti, N. Gordon, and G. Falco, "Space cybersecurity lessons learned from the ViaSat cyberattack," in *Proc. AIAA ASCEND Conf.*, 2022. [Online]. Available: https://www.researchgate.net/publication/363558808

[3] A. Kazi, S. Kazi, and S. Bhosale, "Invisible battlefields: Analyzing the Viasat attack and its broader implications," *Scientific Bulletin*, 2025, doi: 10.2478/bsaft-2025-0007.

[4] C. Poirier, "Hacking the cosmos: Cyber operations against the space sector," Center for Security Studies, ETH Zürich, Zürich, Switzerland, Tech. Rep., 2024. [Online]. Available: https://www.research-collection.ethz.ch/handle/20.500.11850/697348

[5] National Institute of Standards and Technology, "The NIST cybersecurity framework (CSF) 2.0," NIST, Gaithersburg, MD, USA, Tech. Rep. NIST CSWP 29, Feb. 2024. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf

[6] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST, Gaithersburg, MD, USA, Tech. Rep. NIST AI 100-1, Jan. 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[7] MITRE Corporation, "MITRE ATLAS: Adversarial threat landscape for artificial intelligence systems," MITRE, McLean, VA, USA, 2023. [Online]. Available: https://atlas.mitre.org/matrices/ATLAS

[8] National Aeronautics and Space Administration, "Artificial intelligence (AI) at NASA," NASA, Washington, DC, USA, 2023. [Online]. Available: https://www.nasa.gov/artificial-intelligence

[9] Consultative Committee for Space Data Systems, "Security guide for space data systems," CCSDS, Tech. Rep. CCSDS 350.1-G-3, 2022. [Online]. Available: https://public.ccsds.org/Pubs/350x1g3.pdf

[10] F. Cremer et al., "Cyber risk and cybersecurity: A systematic review of data availability," *The Geneva Papers on Risk and Insurance*, vol. 47, no. 3, pp. 698–736, 2022, doi: 10.1057/s41288-022-00266-6.

[11] X. Wu et al., "Threat analysis for a space information network based on network security attributes: A review," *Complex Intell. Syst.*, vol. 8, pp. 3921–3936, 2022, doi: 10.1007/s40747-022-00679-7.

[12] Z. Yang, J. Zhang, W. Wang, and H. Li, "Invisible threats in the data: A study on data poisoning attacks in deep generative models," *Appl. Sci.*, vol. 14, no. 19, p. 8742, Sep. 2024, doi: 10.3390/app14198742.

[13] F. Marulli, S. Marrone, and L. Verde, "Sensitivity of machine learning approaches to fake and untrusted data in the healthcare domain," *J. Sens. Actuator Netw.*, vol. 11, no. 2, p. 21, Apr. 2022, doi: 10.3390/jsan11020021.

[14] I. Jada and T. O. Mayayise, "The impact of artificial intelligence on

organizational cyber security: A systematic literature review," *Data Inf. Manage.*, vol. 8, no. 2, p. 100063, 2023, doi: 10.1016/j.dim.2023.100063.

[15] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial intelligence for cybersecurity: Literature review and future research directions," *Inf. Fusion*, vol. 97, p. 101804, Sep. 2023, doi: 10.1016/j.inffus.2023.101804.

[16] I. A. D. Nesnas, L. M. Fesq, and R. A. Volpe, "Autonomy for space robots: Past, present, and future," *Current Robot. Rep.*, vol. 2, no. 3, pp. 251–263, Sep. 2021, doi: 10.1007/s43154-021-00057-2.

[17] M. M. Yamin, M. Ullah, H. Ullah, and B. Katt, "Weaponized AI for cyber-attacks," *J. Inf. Secur. Appl.*, vol. 57, p. 102722, Mar. 2021, doi: 10.1016/j.jisa.2020.102722.

[18] M. Pooyandeh, K.-J. Han, and I. Sohn, "Cybersecurity in the AI-based metaverse: A survey," *Appl. Sci.*, vol. 12, no. 24, p. 12993, Dec. 2022, doi: 10.3390/app122412993.