18th International Conference on Space Operations, Montreal, Canada, 26 - 30 May 2025.

SpaceOps-2025, ID # 150

Towards Explainable and Secure AI for Space Mission Operations

Krzysztof Kotowski[a,*], Piotr Wilczyński[b], Dawid Płudowski[b], Agata Kaczmarek[b], Ramez Shendy[a], Jakub Nalepa[c,a], Przemysław Biecek[b], Evridiki Ntagiou[d,*]

[a] *KP Labs, Poland, {kkotowski, rshendy, jnalepa}@kplabs.pl*
[b] *Warsaw University of Technology, Poland,*
[c] *Silesian University of Technology, Poland, jnalepa@ieee.org*
[d] *European Space Operations Centre, European Space Agency, Germany, evridiki.ntagiou@esa.int*
* Corresponding Authors

Abstract

A multitude of AI systems are being developed for space applications for the ground, space, and user segments. However, their promising performance is not enough for the wide adoption of AI algorithms in practice – operational deployment requires addressing security challenges and enhancing the trust of users and stakeholders. This is why the topics of Security of AI (SAI) and Explainability of AI (XAI) are prioritized areas of the larger European Space Operations Centre initiative to use AI for the automation of space mission operations. Our project distills the state-of-the-art knowledge into catalogues of SAI and XAI for the space domain and implements actual software solutions to real SAI and XAI issues identified at the European Space Agency. The SAI catalogue describes common AI security risks with their mitigations and domain-relevant examples. The XAI catalogue lists methods to explain decisions of AI models for analyzing satellite telemetry, Earth observation, or mission documentation. The software solutions address use cases of XAI for anomaly detection in satellite telemetry, data poisoning detection in telemetry forecasting, XAI for ship detection in SAR data, and XAI/SAI for large language models. The project effectively establishes guidelines for both SAI and XAI in space applications. All materials are available at https://assurance-ai.space-codev.org.

Keywords: explainable AI, secure AI, assurance, trustworthiness

Acronyms/Abbreviations

AI – Artificial Intelligence
API – Application Programming Interface
ENISA – European Union Agency for Cybersecurity
ESOC – European Space Operations Centre
ML – Machine Learning
NIST – National Institute of Standards and Technology
OWASP – Open Worldwide Application Security Project
SAI – Security of AI
SAR – Synthetic Aperture Radar
SOE – Spacecraft Operations Engineer
SOM – Spacecraft Operations Manager
XAI – Explainability of AI

1. Introduction

The European Space Operations Centre (ESOC) has launched an innovative initiative to integrate data-driven Artificial Intelligence (AI) applications into the ground segment infrastructure in order to enhance the automation levels of space mission operations [1], [2], [3]. At the same time, the technology has a prominent place in the European Space Agency Strategy 2040 [4] as the key enabler for many space-related applications. Hence, a multitude of AI systems are being developed for space applications for the ground, space, and user segments. However, their promising performance is not enough for the wide adoption of AI algorithms in practice – operational deployment requires addressing security challenges and enhancing the trust of users and stakeholders. This is the main topic of the "Assurance for space domain AI applications" activity done at ESOC. The project effectively establishes best practices and guidelines for both the security of AI (SAI) and the explainability of AI (XAI) in space applications (see the definitions in sections 1.1 and 1.2). It distills the state-of-the-art knowledge into comprehensive catalogues of SAI and XAI for space applications and implements actual software solutions to solve relevant issues identified at the European

Space Agency. Importantly, instead of introducing yet another new AI applications, our work offers external methods to add a layer of assurance to existing AI models.

1.1. Explainability of AI

In the context of space operations, explainability of AI systems refers to the ability of an AI system to clearly and transparently communicate the reasoning behind its decisions or predictions, in a manner that is understandable and verifiable by human operators, engineers, or mission controllers. Given the critical and often autonomous nature of AI in space – where real-time human intervention may be limited or impossible – explainability ensures that decisions made by AI systems can be trusted, audited, and aligned with mission objectives and safety requirements. It is particularly important for identifying faults, ensuring accountability, supporting certification processes, and facilitating rapid troubleshooting during unexpected scenarios in space environments. The concept of explainability is often used interchangeably with interpretability or transparency. But for the purposes of space operations systems, we have adopted a distinction presented in Table 1.

Table 1. Explainability vs. Interpretability vs. Transparency of AI.

	Explainability of AI	**Interpretability of AI**	**Transparency of AI**
Definition	Ability to provide understandable reasons for AI decisions or behavior	The degree to which a human can directly understand how an AI model works	Openness about how an AI system is built, trained, and operates
Main focus	*Post hoc* understanding	Model structure and logic (i.e., use simpler models)	Development and data provenance
Relevance in space operations	Helps operators verify and trust autonomous system actions during missions	Crucial for engineering teams to assess risks and validate system logic	Supports compliance, traceability, and auditing of mission-critical systems

According to the study among spacecraft operators by Hulsmann & Forstner [5], the lack of trust and black-box characteristics of AI models are the key barriers to the wider adoption of AI applications in the space domain. The explainability of AI in space applications is also mentioned in both the Machine Learning Handbook of the European Cooperation for Space Standardization (ECSS) [6] and in the AI for Automation (A^2I) Roadmap of ESOC [1] as a crucial aspect of operationalization.

1.2. Security of AI

The *security of AI* aims to protect AI models from external threats and make them less vulnerable to manipulation or poisoning. Table 2 differentiates this topic from the topics of *AI for security* and the *safety of AI* which are not directly addressed in the paper.

Table 2. Security of AI vs. AI for security vs. safety of AI.

	Security of AI (addressed in the project)	**AI for security**	**Safety of AI**
Definition	Resistance of AI systems to external threats, cyberattacks, and data breaches	Using AI to increase cybersecurity and reliability of computer systems	Resistance to errors, biases, misuses, misunderstandings, ethical problems, and unintended or harmful consequences of using AI systems
Main focus	To make AI systems more reliable and less vulnerable to manipulation and poisoning, safeguarding the confidentiality, accessibility, and integrity of AI models.	To support humans in detecting cybersecurity threats.	To connect AI with human values and reduce the chance that AI systems have a negative impact on businesses and society.
Mitigations	To design new methods to make AI systems more resilient to external threats	To create AI-enhanced systems for cybersecurity	To increase transparency and prevent biases and harmful consequences of using AI systems

AI for security aims to use AI techniques to increase the cybersecurity and reliability of computer systems and the *safety of AI* aims to address inherent issues and harmful consequences of using AI systems. The *safety of AI* is thoroughly covered in the recent International AI Safety Report [7] and is also mentioned as a part of the European Union AI Act [8].

There are multiple recent white papers issued by renowned cybersecurity organizations describing common security risks of AI (see Table 3), but there is no comprehensive analysis of SAI for space applications. The initial analysis by Weber & Franke [9] mentioned SAI as one of the key aspects for the reliability of future AI applications in the space domain.

Table 3. White papers about the security of AI issued by renowned cybersecurity organizations.

Organization	Name of the resource
NIST	Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations [10]
Snowflake	AI Security Framework [11]
Google	Secure AI Framework (SAIF) [12]
ENISA	Multilayer Framework for Good Cybersecurity Practices for AI [13]
OWASP	OWASP Machine Learning Security Top Ten [14] OWASP Top 10 for LLM Applications [15]
MITRE	Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS) [16]

2. Catalogues of Explainable and Secure AI for space operations

AI is a rapidly evolving technology with enormous potential for scaling. However, numerous examples of successful solutions are accompanied by examples where, as a result of an error or deliberate adversarial action, AI systems do not work properly. For this reason, a growing number of regulations are emerging at the national as well as European level, which define what conditions secure AI solutions should meet. These regulations are accompanied by new standard specifications describing the processes and practices to build such secure solutions. The aim of catalogues is to critically distill these standards and select elements that are relevant to space mission operations.

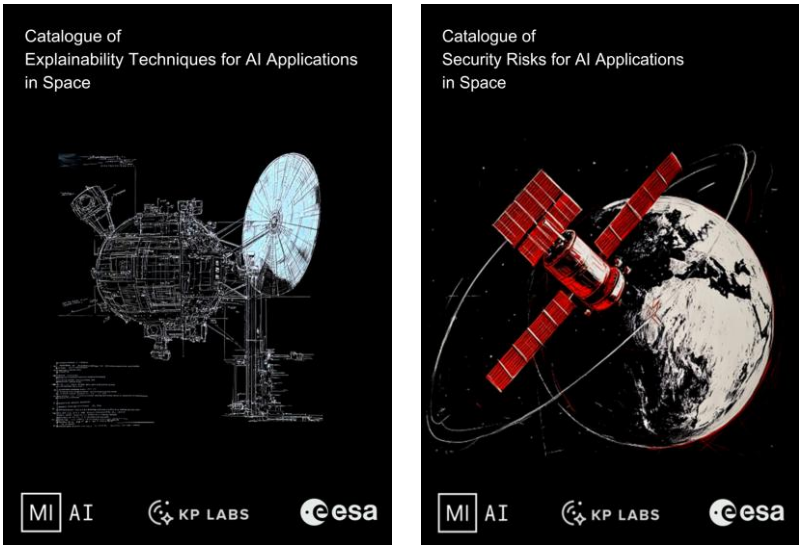


Figure 1. Covers of the XAI (left) and SAI (right) Catalogues.

2.1. XAI Catalogue

The overall structure of the XAI Catalogue is presented in Figure 2. It consists of four parts, each devoted to a different data modality – tabular, image, text, and time series data. This division is because explanations are often constructed depending on how the input data can be modified, and for each modality, the possible ways of input

modification are different. For each modality, there are subsections dedicated to AI end-users, AI developers and AI researchers.

The most common explainability strategies, i.e. the main concepts and goals behind constructing the model's explanatory method, are discussed for each modality. Based on the literature analysis, we have identified four main strategies described in more detail in the next section. In this XAI Catalogue, we focus on strategies for which methods have been proposed that can be used in applications typical to the space domain.

The next level is a list of specific XAI methods that implement a determined strategy for a particular modality. These lists cover only the most popular methods and are focused on the general concepts of how a particular method works. Since XAI is an area of very intensive analysis and development, we expect this list of methods for each modality to grow rapidly. The purpose of presenting these techniques is to characterise them briefly. Relevant references are provided to help a better understanding of each method. The catalog was designed as a living document, and with the development of XAI techniques, new components will appear in these catalogs.

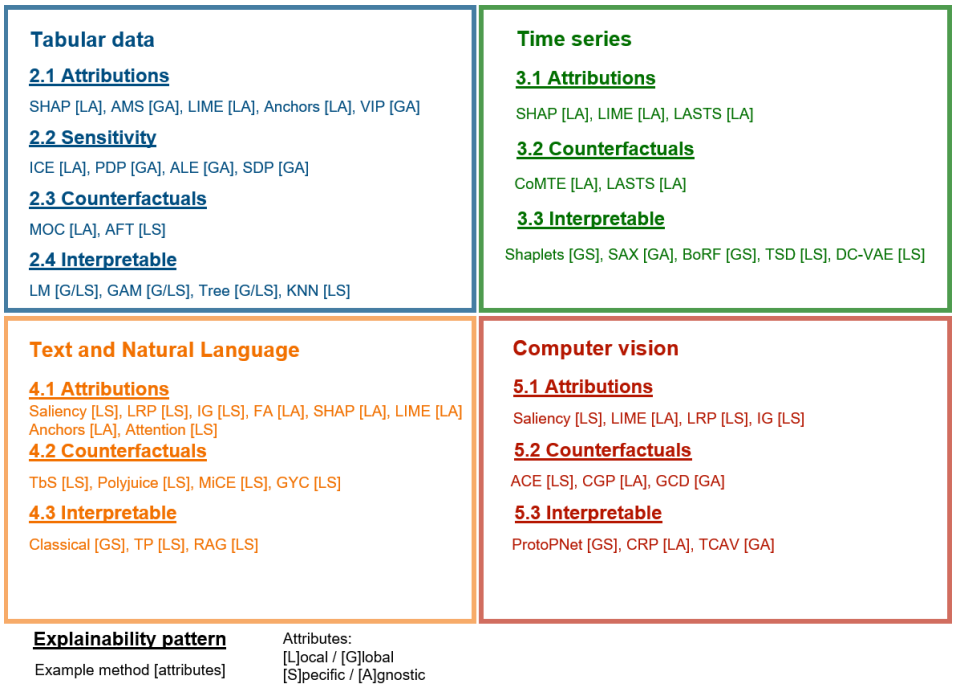


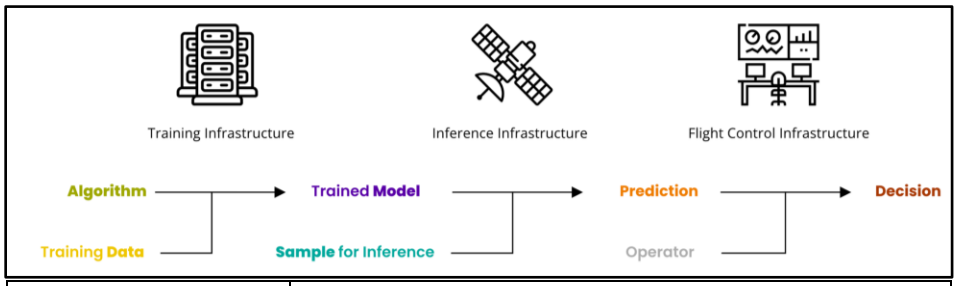
Figure 2. A summary of explainability methods described in the XAI Catalogue, divided into 4 data modalities and 4 explanation types.

2.2. SAI Catalogue

To create a baseline for the SAI Catalogue, various existing standards of AI security have been analysed (see Table 3) and unified for the space domain. Documents by the National Institute of Standards and Technology (NIST), Snowflake AI Security Framework, and the Open Worldwide Application Security Project (OWASP) have been chosen as the main point of reference, because they are kept up to date, are mainly created by practitioners, and cover most of the SAI areas. The summary of risks covered in the SAI Catalogue is given in Figure 3.

The vulnerabilities identified in these documents were then mapped to a set of nine high-level risks defined for the project. These risks were formulated with a focus on the type of access required for an attack and the potential consequences, rather than on detailed implementation specifics. This approach allowed the catalogue to capture the most impactful and context-relevant threats while maintaining clarity and focus.

Some of the risks in the catalogue encompass multiple vulnerabilities from the source documents, as those documents vary in how specifically or generally, they describe threats. In addition, certain vulnerabilities appear in more than one risk category within the catalogue. This reflects the overlapping nature of many AI security threats and the fact that clear-cut boundaries between risk categories are often difficult to define.



Security risk	Access to	Type of risk	At risk
Data Leakage	**Sample**	to extract information	**Data**
Data Poisoning	**Data**	to change behavior	**Model**
Adversarial Samples	**Sample**	to change behavior	**Prediction**
Sponge Samples	**Sample**	to affect availability	**Model**
Model Theft	**Sample**	to extract information	**Model**
Prompt Injection	**Sample**	to extract information	**Data**
Overreliance	**Prediction**	to change behavior	**Decision**
Supply Chain Attack	**Infrastructure**	to change behavior	**Algorithm**
Trojan Horse Attack	**Model**	to change behavior	**Prediction**

Figure 3. Nine security risks from the SAI Catalogue.

3. Explainability and Security as a Service (XSaaS)

XAI and SAI techniques work on top or as a part of existing AI applications. They are sometimes very application- or architecture-specific which makes them difficult to reuse in different scenarios. Our idea of *Explainability and Security as a Service* (XSaaS) aims to hide our specialized services behind the universal standardized interface. The goal is to hide the intricacies of each service from users, so it is enough for them to learn the universal interface to explain any model for space operations.

The user interacts with the universal XSaaS interface using simple and generic API, including 4 main parameters related to the explanation level of detail, model reference, sample to explain, and explanation context (reference data, mission name, user type). The interface obtains information about the model type (target task, architecture, and supported data modality), context (e.g., training data), and input/output specification (input size, input type, available output endpoints) from the model endpoint. Then, it identifies and runs proper specialized XSaaS for the provided sample. Finally, the interface formats the result according to the user's needs. The process is visualized in Figure 4.

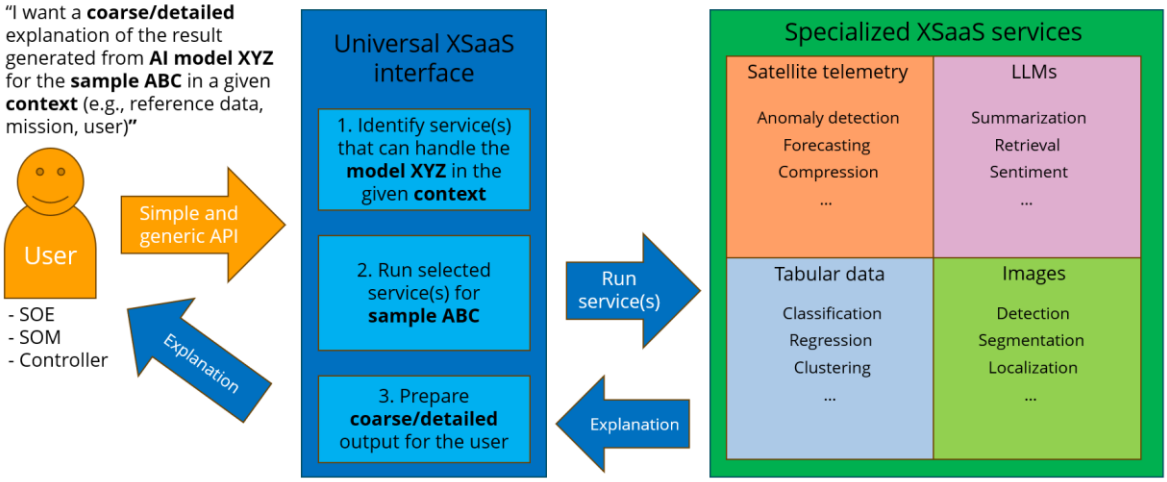


Figure 4. The idea of the universal XSaaS interface.

Our software solutions are designed to be as easy as possible to integrate with the proposed interface using REST API. Our idea is meant to be implemented in the future as a part of the AInabler platform of ESOC (https://ainabler.space-codev.org/), so our software uses the same data storage (MinIO), event streaming (Kafka), and containerization (Docker) infrastructure.

4. Software solutions

In the scope of the project, we have implemented 4 different XAI and SAI software solutions addressing different AI models, tasks, and data modalities. All materials have been created in cooperation with end users, contractors, and stakeholders from the Agency. There were several workshops to collect ideas for scenarios and feedback about proposed solutions. The ultimate goal of the implementation phase was to generate tangible benefits for end users of existing and future AI systems. Mainly, to add a layer of assurance in specific AI applications. The demonstrations of actual software solutions have been a powerful way to explain the purpose and importance of the project to all parties. Thanks to these presentations, the project met with great interest from users and contractors struggling with a lack of explainability, e.g., in responses from the LLM-based version Operation CompAnIon (OCAI-NG) [17] or in automation of satellite health and ground operations (AISHGO) [18]. Also, the results gained attention of users debugging ship detection models at European Space Research and Technology Centre (ESTEC), improving anomaly investigation process at ESOC, qualifying AI in space, and trying to improve the security of AI for space operations at ESOC's Cyber Security Operations Center (CSOC).

4.1. XAI for anomaly detection in satellite telemetry

This XSaaS architecture delivers on-demand, interpretable explanations for anomaly detection (AD) models used in satellite telemetry time series data. Compatible with models that follows a predefined interface, it produces multi-level explanation reports—including parameter importance via LIME [19] and counterfactuals via CoMTE [20]—in both interactive HTML and raw JSON formats. The interactive HTML report was developed in consultation with expert end users during dedicated workshops, these reports are tailored to offer different levels of support, enhancing the transparency and accountability in model decision-making. Figure 4 shows the different levels of explainability detail provided by the service, highlighting how explanations can be tailored to varying user needs. The XSaaS for AD is designed to support four primary use cases: building trust by increasing user confidence in model predictions, ensuring regulatory compliance through transparent and explainable outputs, aiding in debugging by uncovering biases or errors in model behavior, and supporting model auditing to verify that predictions are based on relevant and fair factors. As a model-agnostic solution, XSaaS does not require direct integration with any ML model but instead interfaces with a model server via well-defined APIs, reducing dependency and integration complexity.

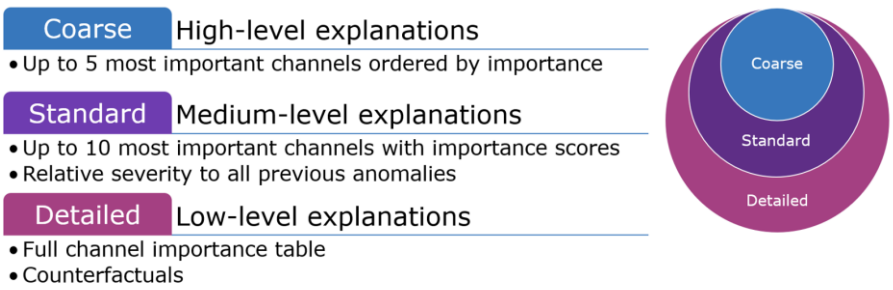


Figure 5. Explainability levels of detail for the anomaly detection in satellite telemetry service.

The XSaaS for AD supports a range of user needs across the identified target groups—spacecraft operations managers (SOMs), spacecraft operations engineers (SOEs), and AI engineers. For SOMs, the service provides concise explanations of anomalies and their root causes to enable quick, high-level decision-making. SOEs can also benefit from detailed reports on the AI model's reasoning, prioritized list of impacted telemetry parameters or influential factors which can be provided as shown in Figure 6. They can also see comparisons with expected nominal signals to better diagnose and respond to anomalies which can provided as shown in Figure 7. AI engineers can use the service

to investigate false alarams and refine model performance, ensuring more reliable time series anomaly detection over time.

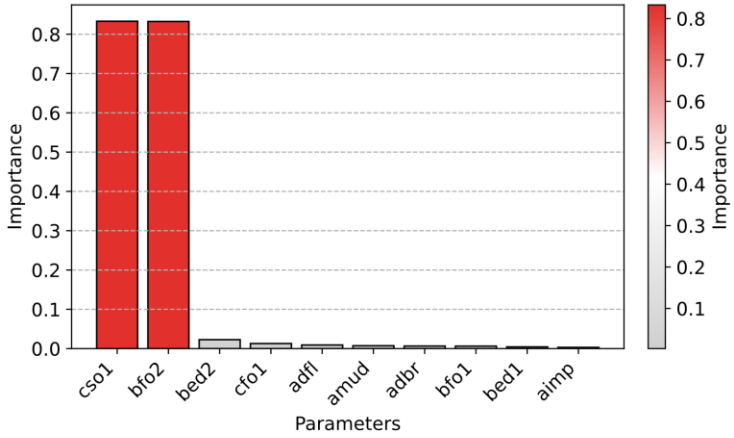


Figure 6. Example list of impacted telemetry parameters during an anomaly.

The XSaaS for AD is designed with extensibility in mind, allowing the integration of additional mode-agnostic explaianability techniques beyond the adopted ones. This flexibility empowers AI researchers and developers to add more methods according to their needs including those focused on feature importance or counterfactual reasoning. The service has been tested on anomalies from the CATS [21] dataset using a baseline AD model DC-VAE [22], with results reviewed by SOEs.

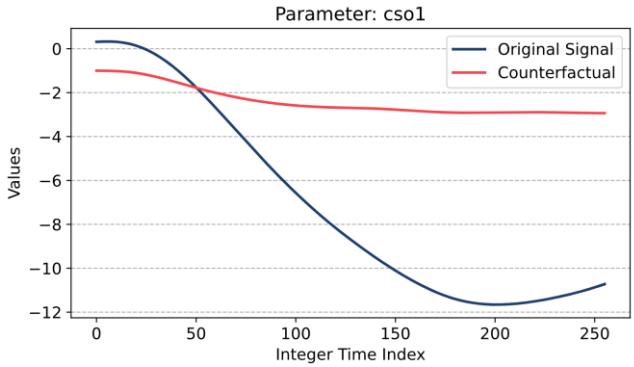


Figure 7. Example comparison between anomalous and nominal (counterfactual) signals during an anomaly.

4.2. XAI and SAI for satellite telemetry forecasting

The XSaaS for telemetry forecasting (TF) is a solution designed to provide explanations for predictions made by ML and AI models used for forecasting time series data, such as satellite telemetry. In addition to standard explanations, it includes a security feature to detect data poisoning or drift by comparing model explanations before and after retraining with new data. The service can be called on-demand by external systems or operators whenever a forecasting model is retrained. It is compatible with any forecasting model that follows a predefined interface and outputs importance scores for context time windows, along with comparisons of model attributions to monitor changes in reasoning. Figure 8 shows attributions difference between the reference model and the retrained (target) model. The comparison reveals a significant shift in model reasoning: while the reference model focuses on the nominal signal (highlighted in pink), the retrained model attends primarily to abnormalities or peaks in the signal (highlighted in blue). This change in behavior suggests that retrained model may have been exposed to poisoned data while retraining indicating a data poisoning event.

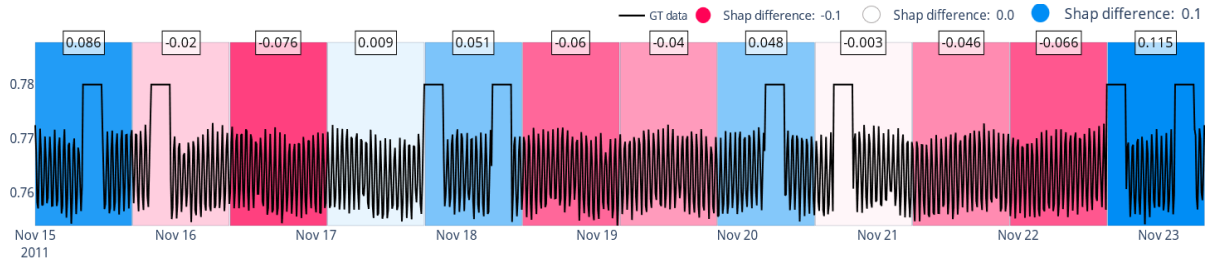


Figure 8. Attributions difference between target and reference models.

The XSaaS for TF supports four key use cases: enhancing trust in AI model forecasts, aiding AI qualification and validation through explainability, enabling debugging by identifying biases or errors, and mitigating security risks by detecting data poisoning between model versions. The service is designed for various end users, including SOEs, spacecraft controllers (SpaceCons), SOMs, and AI engineers. User stories highlight the need for AI engineers to be alerted to significant changes in feature importance and attribution shifts to guide model deployment decisions and root cause analysis. Similarly, SOEs require insights into how model behavior evolves across iterations and which parts on the input data the model focuses on during forecasting, supporting mission planning, and system monitoring.

The service leverages our novel explanation method, Shapley Values for Time Series (ShapTS), which adapts the well-established Shapley Values [23] approach from cooperative game theory to time series data. In this context, the time series is divided into smaller time windows—which serve as individual features to capture temporal dependencies. ShapTS highlights the importance of these time windows in influencing the model's forecast, with darker colors in the visualizations indicating greater significance attributed to a particular time window by the model.

The service has been tested on the data from the ESA-ADB dataset [24] and the N-HiTS forecasting model [25].

4.3. XAI for ship detection in SAR data

The SAR Ship Detection explainability solution is a web-based tool designed to support AI developers—particularly within ESA's Radio Frequency Payloads and Technology team (ESTEC) – in understanding, debugging, and improving AI models for ship detection using SAR data [26]. It addresses key use cases including gaining expertise by comparing model behavior on raw versus processed data, identifying and correcting model biases, and building trust in AI applications. The primary end users are AI engineers working with SAR imagery who require deeper insight into how their models make decisions.

User stories for this use case focus on understanding which parts of the SAR image contribute most to a model's decision (e.g., for detecting ships), identifying causes of false positives like ghost ships, and comparing the reasoning of models trained on different data types. Figure 9 shows the architecture of the explainability pipeline designed to support AI-based ship detection using SAR data. It showcases two parallel processing pipelines: one for raw SAR images and one for processed Single Look Complex (SLC) SAR images.

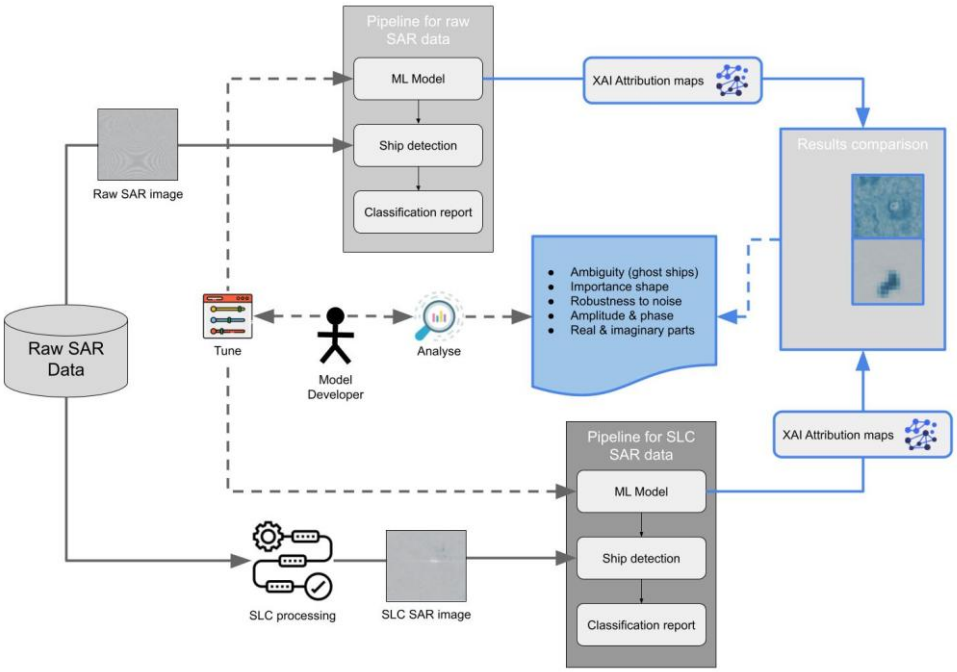


Figure 9. Use case diagram of the application.

The software allows users to select among six attribution methods [27-32] and target classes (e.g., ships or oceans). It visualizes inputs, predictions, and attribution maps for models trained on both raw and processed SAR data, it also distinguishes between the real and imaginary parts of the input. Input to the system include: raw SAR images, SLC images, and ground truth labels.

The attribution maps generated by the system indicate the importance of each pixel in the model's decision-making process, helping developers reasoning inconsistencies or model blind spots. The system was tested in collaboration with ESA, where it successfully provided actionable insights for improving AI-based ship detection performance. Figure 10 compares model predictions and explainability results for raw SAR data and processed SLC data in the context of ship detection. Top row displays the binary predictions of two models—one trained on raw SAR data (left) and the other on SLC data (right). Bottom row displays the corresponding attribution maps, highlighting which parts of the image were most influential in the model's decision-making.

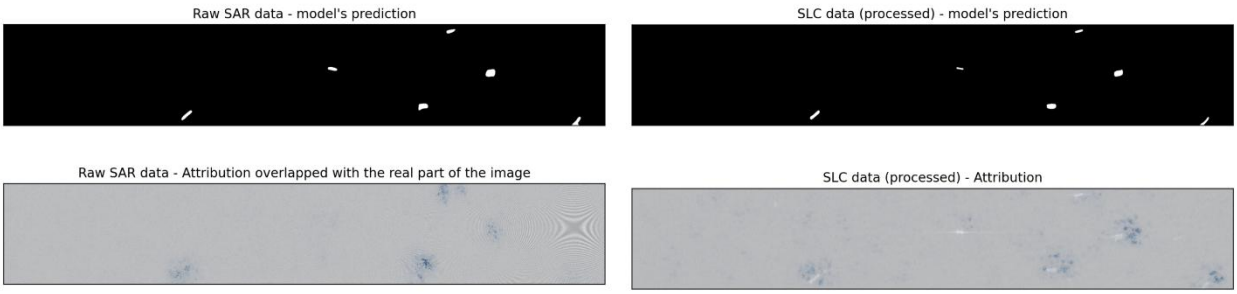


Figure 10. Attribution maps of both raw and processed SAR images.

4.4. XAI and SAI for LLM-based summarization

Large Language Models (LLMs) are powerful tools to increase the effectiveness of information retrieval from different documents, but they come with multiple security and explainability issues, i.e., they hallucinate and are prone to malicious instructions in source documents. Moreover, users tend to rely too much on LLM outputs. Our goal was to address both XAI and SAI issues when generating summaries and answers about space mission reports generated at

ESOC. The proposed software solution is a secure and explainable interface for LLM-based summarization and question-answering based on source documents.

The purpose of the XAI part of the application is to highlight the most relevant (according to the LLM) input file fragments, so the user can evaluate if they are consistent with the model answer. To achieve this, we used the attention-based XAI method from the Inseq library (inseq.org) and the feature ablation method from the Captum library [33]. The user must provide an LLM model to use (either as a path to the local folder or the HuggingFace repository) and input files to use. For the summarization task, the prompt is created internally in the application based on the number of N output bullet points provided by the user (i.e., "Summarize the above text in N bullet points"). For each bullet point, our solution generates M (3 in default) most relevant input fragments with their relevance scores between 0 and 1, as presented in Figure 11.

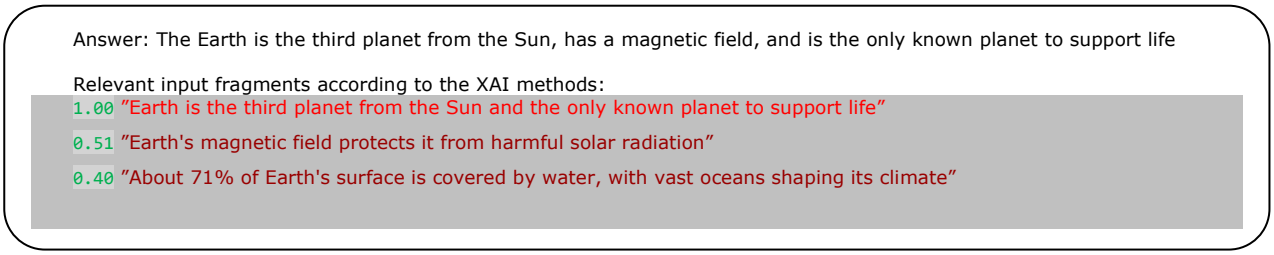


Figure 11. Example explanation for a single output bullet point from an LLM.

The purpose of the SAI part of the application is to implement mitigation measures for addressing potential prompt injection and adversarial modifications in an LLM system by employing monitoring, structured input formats, prompt adjustments, and output filtering techniques. This is achieved by implementing an indirect prompt injection mitigation pipeline that replaces the existing prompting pipeline, enhancing the LLM's resistance to malicious threats by using specially crafted prompts.

6. Conclusions

The topics of Security of AI (SAI) and Explainability of AI (XAI) are rapidly gaining importance in all domains of the high-tech industry. This is reflected in a growing number of standards, regulations, and incidents related to the use of AI in different applications. Our project effectively establishes best practices and guidelines for both SAI and XAI in space applications. It not only distills the state-of-the-art knowledge into catalogues of SAI and XAI for space domain AI applications but also implements actual software solutions to real SAI and XAI issues identified at the Agency.

The Catalogues created in the project are a valuable resource for all parties working on AI applications for space and a strong baseline for all further activities related to security, explainability, trustworthiness, assurance, qualification, validation, and verification of AI systems in the space domain. The Catalogues are going to be used to update the ECSS Machine Learning Handbook [6] and to prepare checklists for all projects providing AI solutions at the Agency.

AI applications in space, especially in space operations, are still at relatively low Technology Readiness Level (TRL) and not widely adopted, so it was difficult to work on specific real-life examples of explainability or security issues in some use cases. However, one of the main purposes of the project was to prepare a ground for wider adoption of such applications in real operations and to upskill users and developers in the topic already before deploying AI solutions in operations. This challenge was addressed by proposing realistic mitigation scenarios of risks and explainability issues and running multiple review iterations with end users. This way, we addressed the actual needs of end users and, at the same time, educating them on the topic of the project.

The proposed software solutions' scope and effectiveness are limited due to the relative novelty and immaturity of XAI/SAI methods, such as long computational times for LLM XAI, limited choice of model-agnostic methods to implement as XSaaS, or a lack of real-life scenarios to calibrate the data poisoning detection in telemetry forecasting. Currently, the software does not attempt to remove the human operator from the process but just offers additional information that should help users to better understand the predictions from AI models and make them aware of security risks. The solutions have been tested in an ESOC computational environment and there is an active effort to integrate

them with the AInabler platform. Nevertheless, the project is a pioneering work as the first comprehensive attempt to structure the knowledge about SAI/XAI in the space domain and implement actual software solutions to address the assurance for space domain AI applications. It has a real potential to generate a significant impact in the community and will be actively promoted as such.

Acknowledgments

This work is supported by the European Space Agency under contract number 4000144194/23/D/BL "Assurance for Space Domain AI Applications". The authors thank the ESA workforce involved in the project.

References

[1] G. De Canio, J. Eggleston, J. Fauste, A. M. Palowski, and M. Spada, "Development of an actionable AI roadmap for automating mission operations," in *2023 SpaceOps Conference*, Dubai, United Arab Emirates: American Institute of Aeronautics and Astronautics, Mar. 2023. Accessed: Feb. 12, 2024. [Online]. Available: https://star.spaceops.org/user_manudownload.php?doc=303__bm05ydei.pdf

[2] G. De Canio, E. Ntagiou, F. Antonello, and J. Eggleston, "Artificial Intelligence for mission operations automation roadmap: the European Space Operations Centre updates and vision," in *2025 SpaceOps Conference*, Montreal, Canada: Canada Space Agency, 2025.

[3] E. Ntagiou, J. Eggleston, P. Collins, and K. Cichecka, "DataX: Pioneering data strategies for enhanced mission operations," in *2025 SpaceOps Conference*, Montreal, Canada: Canada Space Agency, 2025.

[4] European Space Agency, "ESA Strategy 2040." European Space Agency, 2025. Accessed: Apr. 03, 2025. [Online]. Available: https://esamultimedia.esa.int/docs/corporate/ESA_Strategy_2040_InDepth.pdf

[5] M. Hulsmann and R. Forstner, "AI-based Spacecraft Operations and the Issue of Lacking Trust - First Results of an AI Trustability Survey in the Space Domain," in *74th International Astronautical Congress*, Baku, 2023.

[6] ESA, "Machine Learning Handbook," ESA, ECSS-E-HB-40-02A, 2024. [Online]. Available: https://ecss.nl/home/ecss-e-hb-40-02a-15-november-2024/

[7] Y. Bengio *et al.*, "International AI Safety Report," Jan. 29, 2025, *arXiv*: arXiv:2501.17805. doi: 10.48550/arXiv.2501.17805.

[8] European Union, *European Union Artificial Intelligence Act*, vol. PE/24/2024/REV/1. 2024. Accessed: Apr. 07, 2025. [Online]. Available: https://artificialintelligenceact.eu/the-act/

[9] A. Weber and P. Franke, "Space-Domain AI Applications need Rigorous Security Risk Analysis," in *Proceedings 2024 Workshop on Security of Space and Satellite Systems*, San Diego, CA, USA: Internet Society, Mar. 2024. doi: 10.14722/spacesec.2024.23008.

[10] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," National Institute of Standards and Technology, NIST Artificial Intelligence (AI) 100-2 E2023, Jan. 2024. doi: 10.6028/NIST.AI.100-2e2023.

[11] Snowflake, "AI Security Framework," 2024. [Online]. Available: https://www.snowflake.com/en/resources/white-paper/snowflake-ai-security-framework/

[12] Google Safety Center, "Secure AI Framework," 2024. Accessed: Apr. 07, 2025. [Online]. Available: https://safety.google/cybersecurity-advancements/saif/

[13] European Union Agency for Cybersecurity, "Multilayer Framework for Good Cybersecurity Practices for AI," Feb. 2024. Accessed: Apr. 07, 2025. [Online]. Available: https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai

[14] OWASP, "OWASP Machine Learning Security Top Ten," 2023. Accessed: Apr. 07, 2025. [Online]. Available: https://owasp.org/www-project-machine-learning-security-top-10/

[15] OWASP, "OWASP Top 10 for Large Language Model Applications." Accessed: Apr. 07, 2025. [Online]. Available: https://owasp.org/www-project-top-10-for-large-language-model-applications/

[16] MITRE, "MITRE ATLAS™," 2021. Accessed: Apr. 07, 2025. [Online]. Available: https://atlas.mitre.org/matrices/ATLAS

[17] E. Ntagiou et al., Smart Space Operations: OCAI's Contribution to Operational Excellence, IAF Space Operations Symposium, Milan, Italy, 14-18 October 2024.

[18] N. Policella *et al.*, "Innovating Space Operations with AI: The AISHGO Project," in *2025 SpaceOps Conference*, Montreal, Canada: Canada Space Agency, 2025.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining, in KDD '16. New York, NY, USA: Association for Computing Machinery, Sierpie 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[20] E. Ates, B. Aksar, V. J. Leung, and A. K. Coskun, "Counterfactual Explanations for Multivariate Time Series," in *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, May 2021, pp. 1–8. doi: 10.1109/ICAPAI49758.2021.9462056.

[21] P. Fleith, "Controlled Anomalies Time Series (CATS) Dataset." Feb. 2023. doi: 10.5281/zenodo.7646897.

[22] G. G. González, S. M. Tagliafico, A. Fernández, G. Gómez, J. Acuna, and P. Casas, "One Model to Find Them All Deep Learning for Multivariate Time-Series Anomaly Detection in Mobile Network Data," *IEEE Transactions on Network and Service Management*, 2023, doi: 10.1109/TNSM.2023.3340146.

[23] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Apr. 07, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[24] K. Kotowski *et al.*, "European Space Agency Benchmark for Anomaly Detection in Satellite Telemetry," Jun. 06, 2024, *arXiv*.

[25] C. Challu, K. G. Olivares, B. N. Oreshkin, F. G. Ramirez, M. M. Canseco, and A. Dubrawski, "NHITS: Neural Hierarchical Interpolation for Time Series Forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, Art. no. 6, Jun. 2023, doi: 10.1609/aaai.v37i6.25854.

[26] K. De Sousa, G. Pilikos, M. Azcueta, and N. Floury, "Ship Detection From Raw SAR Echoes Using Convolutional Neural Networks," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 17, pp. 9936–9944, 2024, doi: 10.1109/JSTARS.2024.3399021.

[27] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," Apr. 19, 2014, *arXiv*: arXiv:1312.6034. Accessed: Jul. 10, 2024. [Online]. Available: http://arxiv.org/abs/1312.6034

[28] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, Jul. 2017, pp. 3319–3328. Accessed: Sep. 27, 2024. [Online]. Available: https://proceedings.mlr.press/v70/sundararajan17a.html

[29] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," Nov. 28, 2013, *arXiv*: arXiv:1311.2901. doi: 10.48550/arXiv.1311.2901.

[30] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," Apr. 13, 2015, *arXiv*: arXiv:1412.6806. doi: 10.48550/arXiv.1412.6806.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[32] L. Merrick, "Randomized Ablation Feature Importance," Oct. 01, 2019, *arXiv*: arXiv:1910.00174. doi: 10.48550/arXiv.1910.00174.

[33] V. Miglani, A. Yang, A. Markosyan, D. Garcia-Olano, and N. Kokhlikyan, "Using Captum to Explain Generative Language Models," in *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, L. Tan, D. Milajevs, G. Chauhan, J. Gwinnup, and E. Rippeth, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 165–173. doi: 10.18653/v1/2023.nlposs-1.19.