SPATIAL

DESIGN PATTERNS DESIGN PRINCIPLES

TABLE OF CONTENTS

DESIGN PATTERNS

In this Section, we will explore the purpose and significance of software design patterns, understanding how they have traditionally served as the go-to solutions for software design challenges. We will then transition our focus to the world of SPATIAL design patterns, exploring how they have emerged as essential tools for addressing the intricate intricacies of AI-centric architectures, particularly in security domains.

By the end of this Section, you will have a clear understanding of the fundamental differences between these two pattern paradigms and how they contribute to the art and science of software engineering in distinct yet complementary ways.

4.1 THE PURPOSE OF SOFTWARE DESIGN PATTERNS

In software engineering, design patterns are typical solutions to commonly occurring problems in software design[13]. The pattern is not a specific piece of code, but a general concept for solving a particular problem. Patterns are often confused with algorithms because both concepts describe typical solutions to some known problems. While an algorithm always defines a clear set of actions that can achieve some goal, a pattern is a more high-level description of a solution. The code of the same pattern applied to two different programs may be different.

Most patterns are described very formally so people can reproduce them in many contexts and their description usually contains an intent, motivation, structure, and code example. The intent of a pattern briefly describes both the problem and the solution.

The motivation further explains the problem and the solution the pattern makes possible. The structure of classes shows each part of the pattern and how they are related. The code example in one of the popular programming languages makes it easier to grasp the idea behind the pattern.

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.2 SPATIAL DESIGN PATTERNS

SPATIAL design patterns are conceptual frameworks that address the unique challenges and requirements of creating secure, transparent, and accountable AI-driven systems. Unlike software engineering design patterns, these guidelines are particularly crafted keeping in mind the intricacies and nuances of AI-centric architectures in security domains.

Rather than serving as a direct code template, these patterns act as high-level blueprints, providing direction and best practices to developers for problem-solving within the realm of AI.

Table 5 provides a comprehensive overview of the key differences between these two paradigms, setting the stage for a deeper exploration of how these patterns shape the way we approach software and AI system design.

Aspect	Software Design Patterns	SPATIAL Design Patterns
Purpose	Software design patterns are time-tested and widely recognized solutions for addressing common challenges in software architecture and design. These patterns serve as a repository of best practices, helping software engineers create maintainable, scalable, and efficient systems. They offer a set of high-level templates that can be adapted to specific project requirements.	SPATIAL design patterns, on the other hand, are a specialized set of guidelines that are fit for the unique intricacies of AI-driven systems, with a strong focus on security, transparency, and accountability. These patterns provide high-level blueprints for designing AI-centric architectures in domains where security and compliance are paramount. They offer direction and best practices, allowing developers to tackle AI-specific challenges effectively.
Level of Abstraction	Software design patterns are abstract concepts that offer a generic approach to solving recurring design problems. They provide a framework for thinking about solutions, enabling adaptability across various software contexts.	SPATIAL design patterns maintain a high level of abstraction but are more focused on the intricacies of AI-driven systems in security domains. They offer a tailored blueprint for addressing challenges specific to AI technologies, ensuring that security, transparency, and accountability are central to the design process.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

Specificity	These patterns are intentionally designed to be general concepts, making them applicable to a wide range of software development scenarios. They are not tied to any one programming language or technology.	SPATIAL design patterns are specialized and tailored for AI systems, particularly in security domains. They are built using the specific requirements proposed in D1.3 and constraints of AI-driven applications and prioritize addressing issues related to data security, explainability, and compliance.
Code vs Blueprint	Software design patterns are not direct code templates but rather conceptual solutions. They guide developers on structuring their code and architecture but leave room for adaptation and implementation according to the unique needs of a project.	SPATIAL design patterns also do not provide specific code templates. Instead, they serve as high-level blueprints, offering guidance on how to approach the design of AI-driven systems within the context of security and transparency. Developers must adapt and implement these patterns to suit their specific AI projects.
Implementation Flexibility	These patterns are highly flexible and adaptable, allowing developers to apply them to a wide range of programming languages, platforms, and software domains.	SPATIAL design patterns, while offering flexibility, are primarily intended for AI-driven systems in security domains. This specialization ensures that the patterns align with the specific needs and challenges that arise in the context of AI technologies, emphasizing security and accountability in implementation.

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

Pattern Elements	Software design patterns typically comprise key elements, including intent (problem and solution description), motivation (context and rationale), structure (class relationships and interactions), and code examples (implementation hints). These elements make the patterns more accessible and understandable.	SPATIAL design patterns follow a similar structure, featuring 6 core elements that provide insights into the design approach. These elements include intent, applicability, structure, implementation plan, relationship with requirements and naming/classification.
Applicability and Context Awareness	Software design patterns are widely applicable across the software development landscape. They are versatile and can be employed in diverse contexts.	SPATIAL design patterns are context-aware and highly applicable in AI-driven systems, especially those where security and accountability are paramount. They offer guidance specifically tailored to this niche, ensuring that AI systems are designed with the necessary safeguards and transparency.

TABLE 5 Detailed comparison between Software Design patterns and SPATIAL Design Patterns. emphasizing their unique characteristics and applications

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

With the software design pattern structure in its foundation, each SPATIAL design pattern is defined using 6 main elements as depicted in Figure.

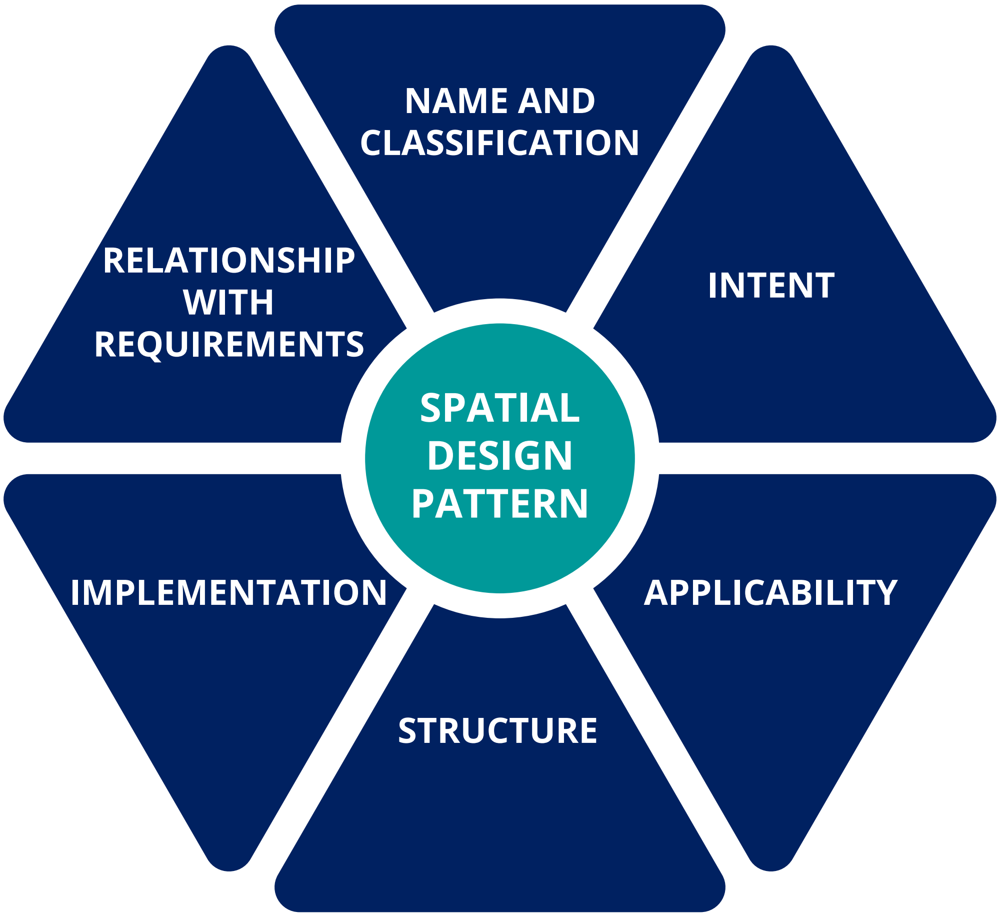


Figure 3 spatial design pattern structure including 6 core elements

The process of defining a SPATIAL pattern involves answering a set of predefined questions, beginning with establishing the **intent** of the proposed design pattern, through **applicability, structure, implementation plan, relationship with requirements, naming and classification.** Below, we present the SPATIAL design pattern questions along with sample answers to aid the explanation of the process followed.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

Intent

Q1. What is the problem that this pattern is addressing and what are the challenges or issues faced by the developers which necessitate the use of this design pattern?

Sample answer: When we speak of "leaking unnecessary data" in the context of ML predictions, we're referring to instances where the output (predictions) of the model reveals more information than intended. This could be data that was used during the training phase or other sensitive details that could be reverse-engineered from the predictions themselves. One of the primary concerns is the potential violation of user privacy. For instance, if a healthcare model leaks details about patients or their conditions, it can lead to severe privacy infringements. In sectors where data is a significant asset, like finance or business, leaking data can provide competitors or malicious actors with an unfair advantage. Many sectors have strict data protection and privacy regulations. Unintended data leaks can lead to hefty fines and legal consequences. If users or stakeholders discover that a model is leaking data, it can result in a loss of trust, which can be detrimental for businesses or institutions relying on ML models.

Q2. What are the objectives or benefits that can be realized by employing this design pattern to address the identified problems?

Sample answer: The sample objective of this pattern could be minimizing the exposure of properties of a model and the data used to train it is a critical objective in the field of machine learning and data science, especially when the model or data encompasses sensitive or proprietary information. The direct benefit is the assurance that sensitive, proprietary, or critical information about the model's inner workings and the data it was trained on is protected from unauthorized access, misuse, or reverse engineering.

Applicability

Three questions aid us in defining the applicability of the design pattern:

Q1. What are the general scenarios or conditions under which the proposed design pattern should be utilized?

Sample answer: This pattern can be very beneficial, especially when trust is paramount, such as in life-critical domains. In domains where human lives are at stake, systems need to be reliable, predictable and trustworthy. For example in medical software, where decisions can influence patient health or even life-and-death situations, using established design patterns can ensure that the software behaves as expected.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

Q2. What are the potential drawbacks of using this pattern? Where this pattern may not be applicable or should not be used?

Sample answer: The potential drawbacks are mainly associated with cost and latency. There may be an extra cost of implementation in the design and when implemented, the users can expect that a latency will be added to model prediction, so an evaluation has to be done before the implementation of the pattern.

Q3. How does this pattern fit into the SPATIAL use case?

The patterns presented in Section are relevant to the use cases presented in WP5. This correlation is shown for each pattern in this document. For example, it could be that a particular pattern is related to Use Case 3 because in that use case, high accountability is of utmost importance as decisions may affect people's lives.

Structure

The structure of each SPATIAL design pattern is presented using the UML (Unified Modeling Language) language, which highlights the significance of having a standardized and unified presentation method. UML is a visual language, allowing complex ideas, structures, and relationships to be communicated concisely and effectively through diagrams and it's an industry-standard language for modelling designs.

Having a unified presentation method, like UML, for SPATIAL design patterns ensures clarity, reduces ambiguities, and streamlines communication across various stages of the project lifecycle. The visual, standardized, and comprehensive nature of UML makes it an apt choice for representing complex design patterns, ensuring that they are understood, adopted, and implemented effectively across the board.

Implementation plan

An implementation plan for a design pattern in an AI system is vital for several reasons. Design patterns presented in Section are tested solutions to recurring problems identified during the development of a SPATIAL platform and AI systems in general. However, how these patterns are applied can vary based on the specifics of the problem domain and the system's architecture. An implementation plan gives developers a clear roadmap on how to adapt the generic pattern to the system's unique requirements. A well-thought-out plan can highlight these potential issues, ensuring that developers are aware and can take preventive measures. Another reason for a good implementation plan is that stakeholders (like architects, business analysts, or product managers) need to be aware of significant architectural or design decisions.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

An implementation plan can be a communication tool to ensure everyone is aligned. Finally, design patterns don't exist in isolation. They often need to integrate with other parts of the system. A plan can highlight these integration points and any potential challenges, ensuring smooth integration.

The basic implementation plan that can serve as a foundation for most design patterns is depicted in Figure 4, however, as presented in Section, implementation plans are unique to each pattern and should be thoroughly designed.

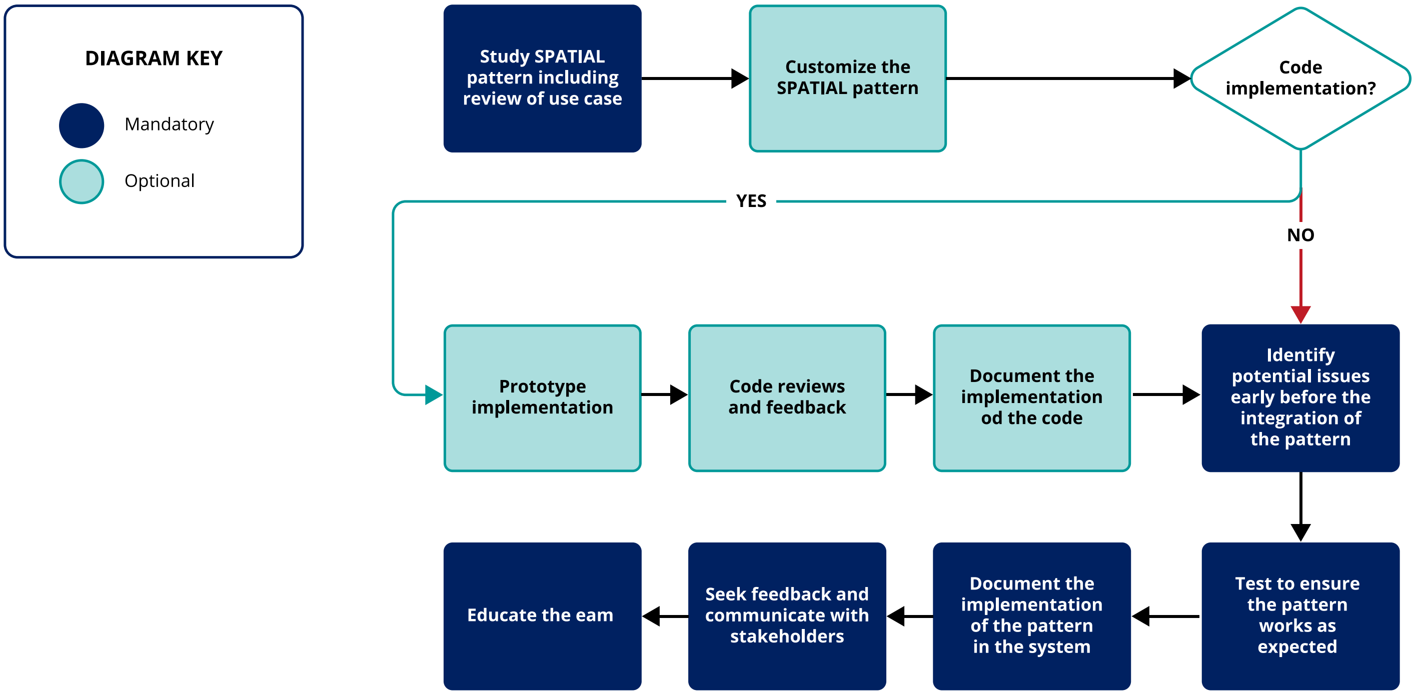


DIAGRAM KEY
Mandatory
Optional
Study SPATIAL pattern including review of use case
Customize the SPATIAL pattern
Code implementation?
YES
NO
Prototype implementation
Code reviews and feedback
Document the implementation od the code
Identify potential issues early before the integration of the pattern
Test to ensure the pattern works as expected
Document the implementation of the pattern in the system
Seek feedback and communicate with stakeholders
Educate the eam

Figure 4 suggested implementation plan of a spatial design pattern

Relationship with requirements

Within the scope of document D1.3, a thorough requirements analysis was conducted. This detailed examination aimed to identify 85 key recommendations for AI-based systems, and then capture them as precise requirements. Using these initial findings, we further proposed 255 system requirements, specifically for the SPATIAL use cases and the platform. A sample requirement is presented in, along with a priority-level explanation in.

A primary foundation for these requirements is the deep industrial domain knowledge of our consortium partners. Additionally, the four SPATIAL use cases, which include Mobile Edge Systems, Cybersecurity Applications and Analytics, IoT, and eHealth, have been vital.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

They provided key insights and basic design principles essential for effectively integrating and using AI algorithms and frameworks. From an extensive literature review, along with insights from the project, we discovered more requirements and recommendations. These are especially relevant in the context of SPATIAL.

Importantly, to enhance the value of this document, we are now linking these requirements with each proposed design pattern. This important step ensures that for every design pattern presented, its connection with specific requirements is clear. This linkage aims to offer better guidance, aiding a smoother and more informed implementation of patterns within AI systems.

Naming and classification

Naming, classifying, and tagging design patterns are crucial practices in software engineering and system design, offering significant benefits in terms of clarity, communication, and organization. For instance, a pattern named “POST-HOC AI-INSIGHT PATTERN” that is presented in Section 4.3.1 immediately provides a reference point. This name, while concise, encapsulates a complex set of ideas, making it easier for developers, designers, and stakeholders to discuss and apply it in different contexts. Classifying this pattern under ‘Explainability’ further organizes our understanding, categorizing it among patterns that enhance understanding and clarity of AI systems. This classification aids in quickly identifying the type of solutions the pattern offers and its relevance to specific scenarios. Similarly, the “TRUSTED EXECUTION ENVIRONMENT COMPUTING PATTERN” presented in Section 4.3.7 falls under the ‘Security’ class, immediately signalling its focus on safeguarding systems.

Tags such as ‘Modular’, ‘Secure Integration’, and ‘Confidentiality’ for the provide additional context. These tags highlight the pattern’s key aspects and primary benefits, serving as a quick summary of its features and applications.

In conclusion, such a structured approach is valuable in facilitating the selection of appropriate patterns for specific problems. It helps in filtering and comparing different patterns, guiding designers and developers in choosing the most suitable pattern based on the requirements and constraints of their projects. For both newcomers and experienced professionals, well-named, classified, and tagged patterns serve as essential learning resources. They allow quick grasping of a pattern’s essence and understand where and how it can be applied. This approach not only streamlines the learning process but also fosters consistency in how design solutions are implemented and discussed across different projects and teams. Finally, this structured approach fosters best practices and standardization in the field, streamlining the learning process and adoption of these patterns across the industry.

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.3 PROPOSED PATTERNS

This chapter introduces eleven innovative design patterns, each addressing specific challenges and needs in AI system development. These patterns encompass various aspects of AI system design, including explainability, privacy, integration, scalability, data integrity, and security. The proposed patterns aim to streamline processes, enhance system robustness, and ensure that AI applications not only meet technical requirements but also adhere to ethical and privacy standards. The summary of proposed patterns is presented in Table 6.

Pattern Number (subsection)	Pattern Name	Class
4.3.1	POST-HOC AI-INSIGHT PATTERN	Explainability
4.3.2	PROPERTY EXPOSURE MINIMISATION PATTERN	Privacy
4.3.3	API GATEWAY PATTERN	Integration
4.3.4	LOAD BALANCING PATTERN	Scalability
4.3.5	UNIFIED AND USER-ORIENTED EXPLANATIONS PATTERN	Explainability

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.3.6	PRIVACY-FOCUSED ML TRAINING PATTERN	Privacy
4.3.7	TRUSTED EXECUTION ENVIRONMENT COMPUTING PATTERN	Security
4.3.8	ADVERSARIAL TRAINING PATTERN	Security
4.3.9	LOAD BALANCING PATTERN	Data Integrity
4.3.10	UNIFIED AND USER-ORIENTED EXPLANATIONS PATTERN	Explainability
4.3.11	UNIFIED AND USER-ORIENTED EXPLANATIONS PATTERN	Security

Table 6 summary of the proposed design patterns

4.3.1 POST-HOC AI-INSIGHT PATTERN

INTENT

To create a robust, modular, and secure framework that will facilitate the easy integration and application of xAI methods across a range of different models while ensuring the protection of sensitive information.

Problems that this pattern addresses

- Model-specific XAI method would limit the selection of compatible AI models to test during the developmental stage. Decoupled post-hoc type design would remove this limitation as they only require query level access only.
- Extreme coupling of the xAI methods with AI methods (e.g.: In-model xAI) can complicate the development process
- Other xAI implementation methods (e.g.: In-Model) can expose the internal structural details of AI methods.

Aims that this pattern achieves

- Broadens AI model compatibility with xAI methods for effective understanding and interpretation, regardless of model complexity.
- Implements decoupled architecture to simplify development, improve efficiency, and enhance component maintenance, upgrades, and scalability.
- Focuses on separating development components for streamlined testing, deployment, and scalable solutions.
- Strengthens security protocols to secure proprietary information, ensuring privacy, integrity, and compliance with privacy standards.

APPLICABILITY

Usage scenarios

- When the accuracy of the explanations is only expected to be of a moderate level
- When the explanations are not required for real-time computations (e.g. SHAP)
- When the model only exposes query level access to the XAI methods.

Relationship with SPATIAL use case

- Mal-DoC use case (WithSecure): the model and the explanations are decoupled and the proprietary model information wouldn’t be required for generating explanations.
- For user activity classifications (Montimage) the post-hoc methods were more suitable due to compatibility with metric calculations that followed

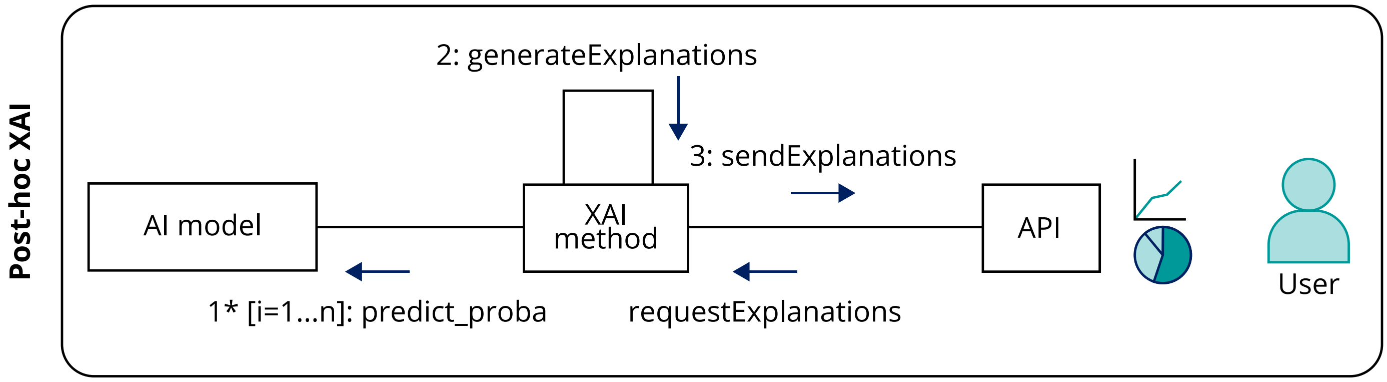
Drawbacks and limitations

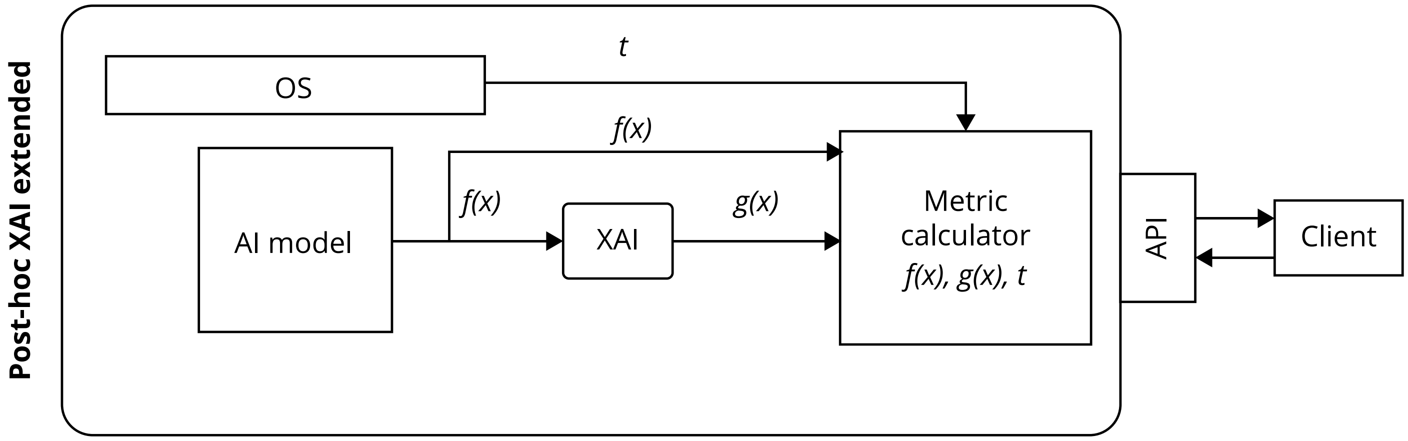
- Third parties can manipulate the explanations of post-hoc explainers through the AI models.
- The post-hoc method can it-self become an attack vector into the system.
- Limited access to the model limits the insights of the model that can be obtained for the stakeholders

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

STRUCTURE

The post-hoc xAI pattern is a versatile method for obtaining explanations from black-box AI models. Here, the user's request for explanations around a specific data point or a model can trigger the API to request explanations from the post-hoc xAI method. Then, the xAI method can start generating the explanations by querying the predict_proba endpoint of the AI model. This call can be computationally expensive, depending on the xAI method utilized. Nevertheless, it will be an iterative call to the AI model. Finally, the xAI method will generate explanations from the output of the AI model received and send the explanations to the API for the user's visualization purposes. Furthermore, this can be extended to include intermediary modules such as metric components to enhance the explainability by providing additional measurements on the model's accountability, resilience and privacy from the explanations.





t System performance stats

f(x) Predictions

g(x) Explanations

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

SUGGESTED IMPLEMENTATION PLAN

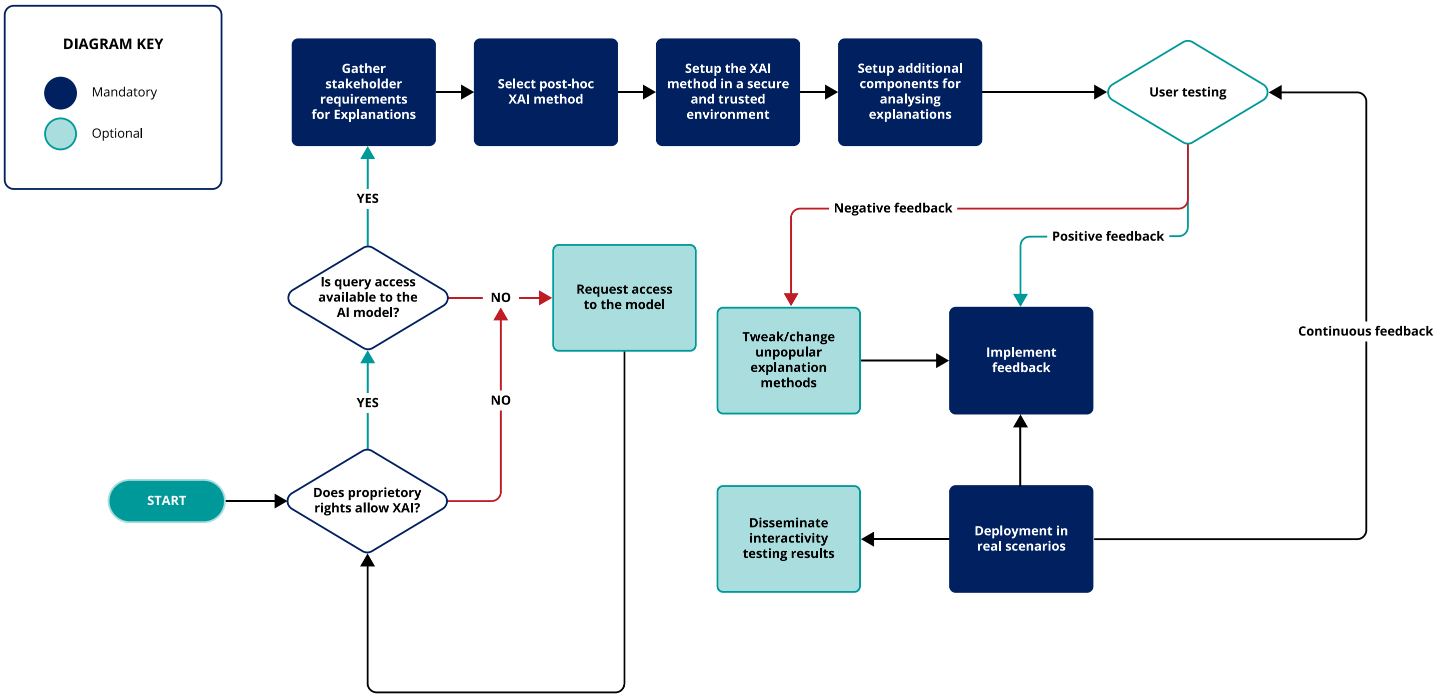


DIAGRAM KEY
● Mandatory
● Optional
START
Does proprietory rights allow XAI?
YES
NO
Is query access available to the AI model?
YES
NO
Request access to the model
Gather stakeholder requirements for Explanations
Select post-hoc XAI method
Setup the XAI method in a secure and trusted environment
Setup additional components for analysing explanations
User testing
Negative feedback
Positive feedback
Tweak/change unpopular explanation methods
Implement feedback
Continuous feedback
Disseminate interactivity testing results
Deployment in real scenarios

RELATIONSHIP WITH SPATIAL REQUIREMENTS

DAT.RQ.10	Data quality SHOULD be measured by quantifying the performance of the AI model
DAT.RQ.11	Data quality for AI training SHOULD be also explicitly defined by data dimensions (e.g. accuracy, currency, and consistency).
DAT.RQ.13	Pre-processed input data SHOULD be linked with prediction outputs of AI models to derive quantifiable explanations to users.
DAT.RQ.14	AI-models can be continually trained with aggregated data, but consistency and integrity of data MUST be preserved through quantifiable estimations.
MOD.RQ.1	The ML model MUST have a high accuracy.
MOD.RQ.8	ML models SHOULD be testable to verify they fulfil expectations on their outputs.
MOD.RQ.10	ML models MUST provide objective evidence that requirements and a specific intended use have been fulfilled.
PRV.RQ.6	There SHOULD be proper metrics defined for privacy to support privacy protection measures.
LEG.RQ.6	According to the AI Act, there MAY need to be a testing process to identify risks and determine appropriate mitigation measures, and to validate that the system runs consistently for the intended purpose, with tests made against prior metrics and validated against probabilistic thresholds.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

USB.RQ.9	Users of AI-based systems SHOULD be able to identify, report, and correct mistakes in the decision-making of AI models.
ACC.RQ.3	Non-Functional requirements relating to the ML process (e.g. accuracy and generalizability) SHOULD be documented in a way that is accessible to lay users.

CLASSIFICATION

Class: Explainability

Tags: Decoupling, Flexibility, AI Model Adaptability, Secure Integration, Information Protection

4.3.2 PROPERTY EXPOSURE MINIMISATION PATTERN

INTENT

To develop a secure and privacy-preserving framework for ML models that minimises the exposure of sensitive properties of model and training data. Most of the membership and property inference attacks require learning more information about the target model. For this, the adversaries tend to query the model for large number of inputs, especially if the model predictions are offered as a service. Having more information like higher floating points on the prediction vector from the model can potentially expose details of the underlying decision-making process of the models, which the adversaries can exploit via these attacks.

Yet the model/data owners or developers may require access to the full model output predictions for requirements such as evaluating explainability of the models. Therefore, this pattern aims to mitigate privacy leakage via model parameters meanwhile providing access to authorised parties.

Problems that this pattern addresses

- Leakage of unintended data from ML model predictions.
- The possibility of privacy related attack scenarios like inference attacks.
- Loss of fine-grained information from model predictions for development and post-hoc explanations with privacy enhancements.

Aims that this pattern achieves

- Implementing strategies to limit the disclosure of the intrinsic characteristics of ML models and the data utilized in their training process.
- Strengthening security measures to mitigate vulnerabilities and risks posed by privacy attacks exploiting model outputs.
- Establishing sophisticated access control mechanisms that grant authorized users precise and comprehensive insights into model predictions.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

APPLICABILITY

Usage scenarios

- When exposing the model outputs to third parties as APIs/interfaces.
- For collaborative learning scenarios like Federated Learning.
- When designing and developing ML model architectures.

Relationship with SPATIAL use case

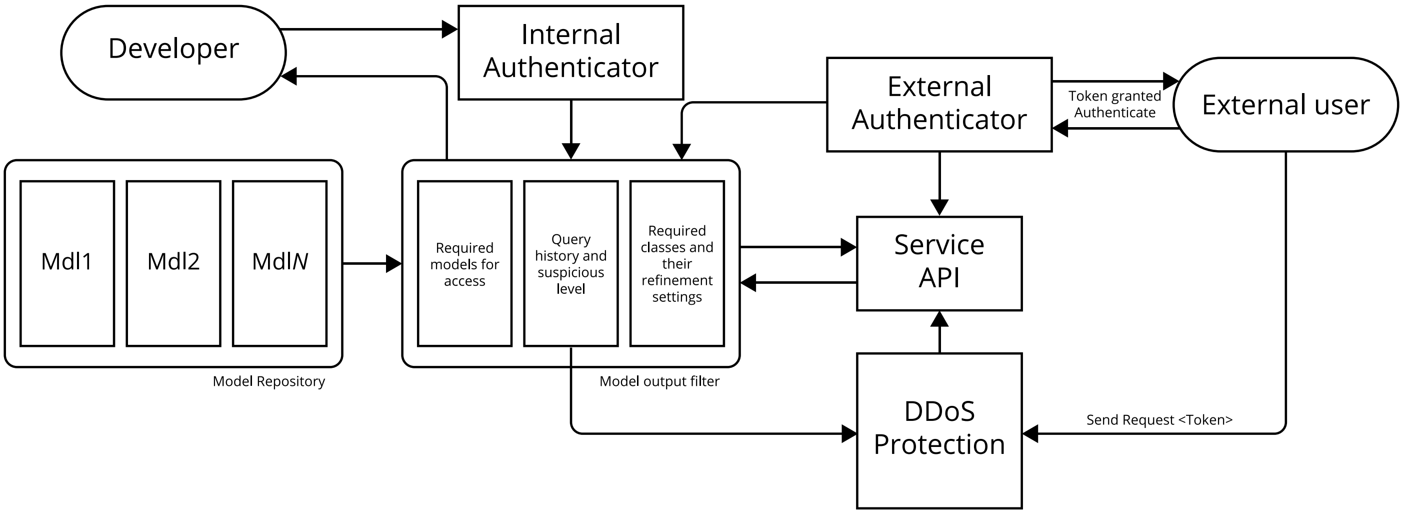
- For the activity classification use case (MI), the predictions can be exposed only with the required class, instead the prediction vector which can leak details on model confidence on predicting certain inputs.
- For Federated Learning (TID), implementing techniques for minimising model outputs before sending the models to an aggregator by clients or limiting the predictions from the global model to a third-party can mitigate the risk of inference attacks.

Drawbacks and limitations

- The implementation may introduce additional response time during the prediction phase.
- Adopting this process may incur additional financial resources in the design stage.
- In domains or applications with lower privacy concerns, the added value of this implementation may be minimal.
- The necessity of this method is more pronounced in models characterized by a higher number of classes and extensive floating-point predictions.

STRUCTURE

In an API-based service, an external user may contact an authenticator service to get authenticated. Based on the level of authorisation, the user will get a token, which can be used to obtain information that the user is allowed to perform. If the user has authorisation to the actual model outputs, the prediction vector will be provided via a secure endpoint. Otherwise, the public endpoint will provide only the class without revealing the actual prediction vector. Furthermore, the API can maintain DDoS protection system such that attackers may not attempt to repetitively query the model to identify critical decision boundary changes of the model via the predictions.



SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

SUGGESTED IMPLEMENTATION PLAN

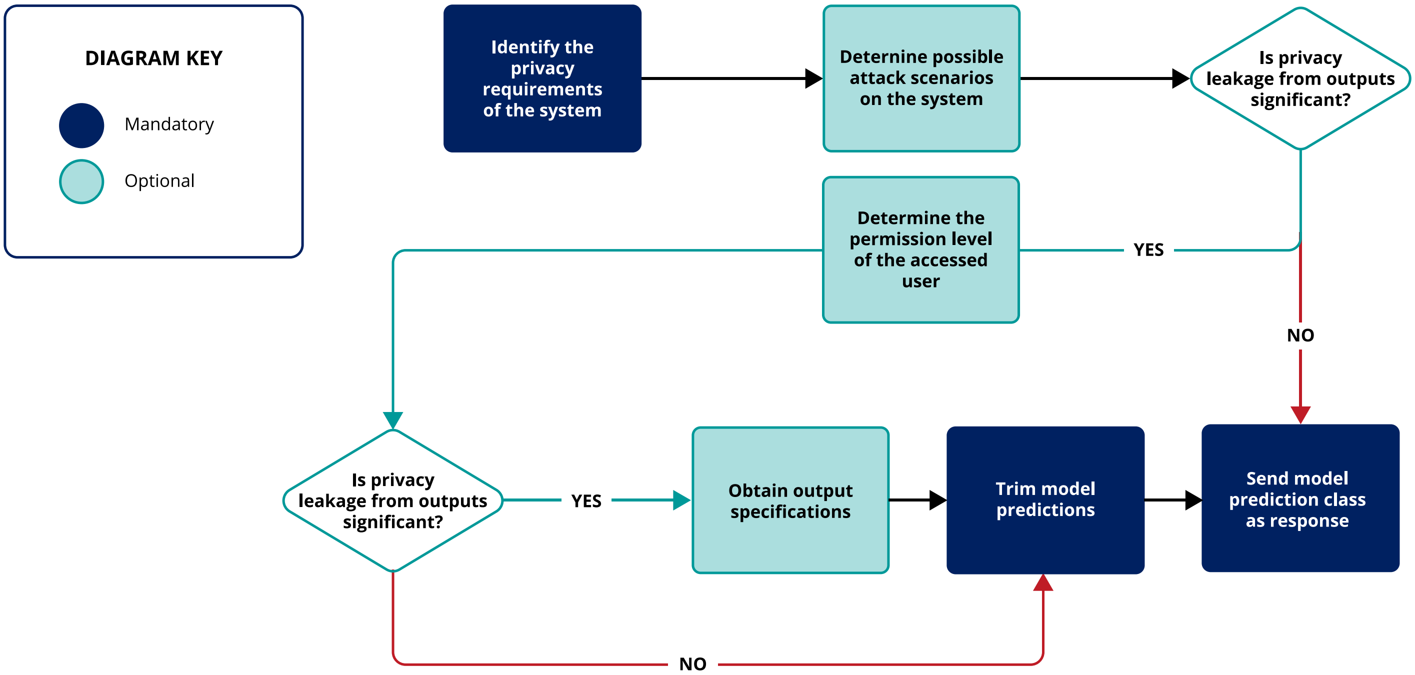


DIAGRAM KEY
● Mandatory
○ Optional
Identify the privacy requirements of the system
Deternine possible attack scenarios on the system
Is privacy leakage from outputs significant?
YES
Determine the permission level of the accessed user
NO
Is privacy leakage from outputs significant?
YES
Obtain output specifications
Trim model predictions
Send model prediction class as response
NO

RELATIONSHIP WITH SPATIAL REQUIREMENTS

PRV.RQ.1	The category of collected data SHOULD be identified for maintaining privacy measures based on the category.
PRV.RQ.2	Privacy measures for data SHOULD be considered in each stage of data generation, processing, and storage.
PRV.RQ.5	The possibility of privacy-related attacks on AI, system and data MUST be assessed and protection or mitigation processes MUST be made.
PRV.RQ.7	The privacy by design approaches SHOULD be included during the system design process.
PRV.RQ.8	Trade-offs between model performance and privacy MAY be considered when implementing privacy.
SEC.RQ.7	AI-based systems MUST be resilient against property interference attacks.
SEC.RQ.8	AI-based systems MUST be resilient against membership interference attacks.
SEC.RQ.11	AI-based systems, dealing with sensitive or confidential data, MUST preserve the confidentiality of the data during the operational phase.

CLASSIFICATION

Class: Privacy

Tags: Data Leak Prevention, Inference Attack Resilience, Confidentiality

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.3.3 API GATEWAY PATTERN

INTENT

To seamlessly integrate and orchestrate microservices dedicated to the rigorous verification and validation of accountability, resilience, and privacy within the framework of AI-based systems with the view to establish trustworthiness for AI-based system.

Problems that this pattern addresses

- Addressing the complexity of multiple access points by establishing a unified entry point for client requests, enhancing efficiency and coherence in microservice interactions.
- Mitigating the performance impact of numerous round trips by consolidating responses from diverse microservices, thereby streamlining data flow and reducing latency.
- Ensuring that robust and consistent security protocols are uniformly enforced across all API interfaces, safeguarding the integrity of the system and maintaining high-security standards.

Aims that this pattern achieves

- Simplifying client access and reducing the need for clients to be aware of individual service endpoints.
- Maintain a balanced and responsive system.
- Enables the aggregation of responses from multiple microservices into a single, coherent response.
- Protecting sensitive data and ensuring compliance.

APPLICABILITY

Usage scenarios

- When system involves collection of services that interact to offer some defined services
- To simplify the experience for clients by offering a single, well-defined API endpoint.
- When enhancing the security of the microservices is a priority, the API Gateway can serve as a centralized security layer.

Relationship with SPATIAL use case

- In SPATIAL project, the API Gateway acts as a central hub for accessing spatial data from diverse stakeholders. It simplifies the process by abstracting the complexities of interacting with multiple data sources, standardizing data formats, and ensuring that spatial data is presented consistently and ready for use by clients.
- In SPATIAL applications with resource-intensive operations and extensive data, the API Gateway's response aggregation minimizes network round trips, enhancing performance and delivering a seamless user experience.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

APPLICABILITY

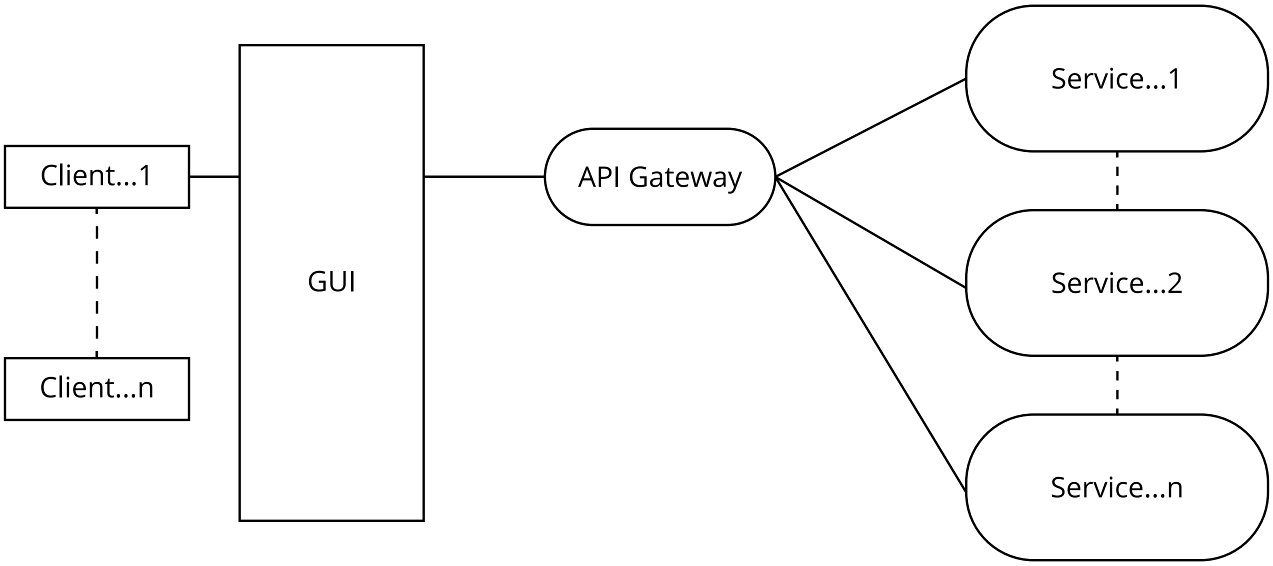
⊖ **Drawbacks and limitations**

- Third parties can manipulate the explanations of post-hoc explainers through the AI models.
- The post-hoc method can it-self become an attack vector into the system.
- Limited access to the model limits the insights of the model that can be obtained for the stakeholders

STRUCTURE

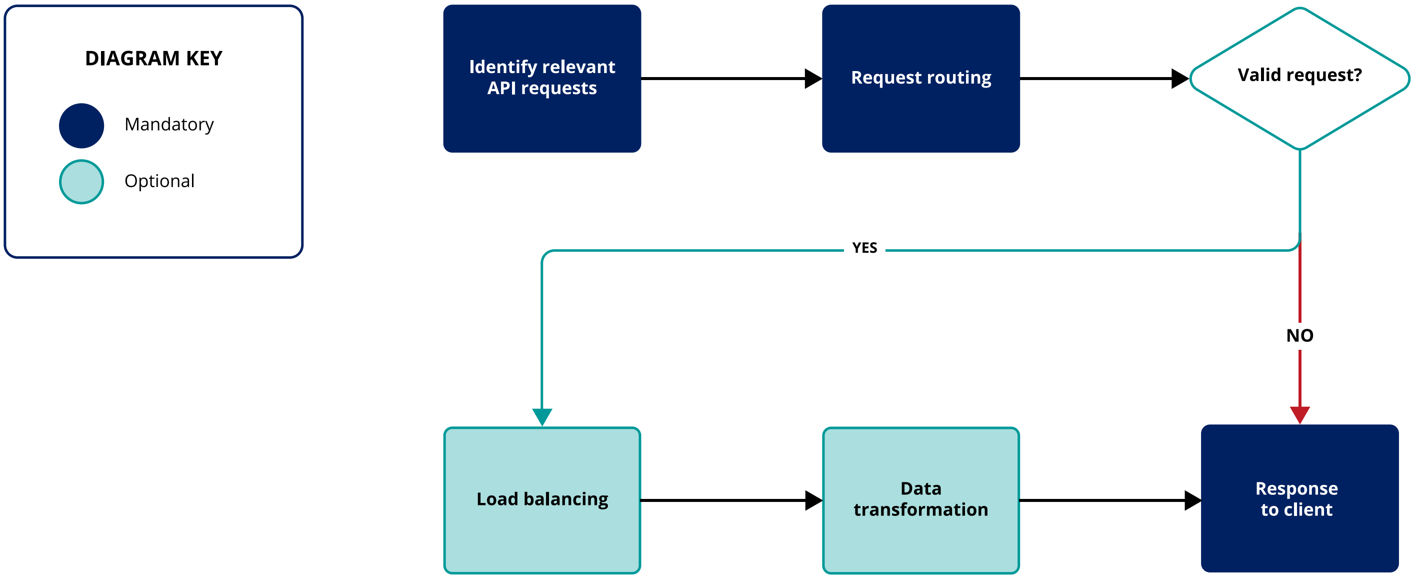
The API Gateway Pattern strategically manages communication between clients and backend services by serving as a centralized entry point, directing requests to relevant microservices. As a microservice pattern itself, the API Gateway orchestrates communication within distributed systems. Adapted to a SPATIAL pattern, it not only facilitates disaggregation between resource-intensive and less intensive services but also enhances user-friendliness and optimizes response times through intelligent backend processing. This spatial partitioning and temporal aggregation contribute to a more scalable and responsive system.

The process begins when the API Gateway receives an incoming client request. API Gateway identifies the appropriate microservice or backend service to handle the request. If multiple instances of the same service exist, the API Gateway implements load balancing. Data transformation and response formatting are carried out if necessary. The API Gateway ensures that the response is presented in a format suitable for the client. API Gateway sends the processed response back to the client. This response is typically in a format that the client can understand and use.



SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

SUGGESTED IMPLEMENTATION PLAN



RELATIONSHIP WITH SPATIAL REQUIREMENTS

PLAT.RQ.36 The category of collected data SHOULD be identified for maintaining privacy measures based on the category.

PLAT.RQ.37 Privacy measures for data SHOULD be considered in each stage of data generation, processing, and storage.

PLAT.RQ.38 The possibility of privacy-related attacks on AI, system and data MUST be assessed and protection or mitigation processes MUST be made.

PLAT.RQ.39 The privacy by design approaches SHOULD be included during the system design process.

CLASSIFICATION

Class: Integration

Tags: Microservice Management, Security, Simplification, Scalability, Flexibility, Interoperability, Performance Optimization

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.3.4 LOAD BALANCING PATTERN

INTENT

To effectively managed traffic and ensure high-performance, scalable and fault-tolerant microservices in AI-based system.

Problems that this pattern addresses

- As SPATIAL's workload grows with XAI and metrics, load balancing ensures efficient distribution of tasks across multiple servers. Utilizing a micro-services architecture, dedicated metrics can be easily added or removed, allowing SPATIAL to adapt its diagnosis profile for trustworthiness in compliance with evolving regulations.
- In a distributed system, server failures or maintenance can occur. It will affect for the High Availability of the system.

Aims that this pattern achieves

- Facilitates horizontal scaling by distributing incoming requests across multiple servers, enabling the system to efficiently handle increased load by adding additional server resources.
- Achieves superior response speeds, contributing to a more agile and responsive user interaction, thereby elevating the overall user experience.
- Strengthens the system's resilience to failures, ensuring that client requests are consistently directed away from malfunctioning servers, thus maintaining uninterrupted service delivery.

APPLICABILITY

Usage scenarios

- When an application experiences high levels of incoming traffic that cannot be effectively handled by a single server or service instance.
- To maintain the availability of the application in the face of server failures or maintenance.

Relationship with SPATIAL use case

- Load balancing ensures high availability for the SPATIAL application by evenly distributing incoming requests across multiple service instances.
- • Load balancing allows to add more instances of stakeholders' services as needed, ensuring that the application can handle increased traffic during peak times or as user base grows.

Drawbacks and limitations

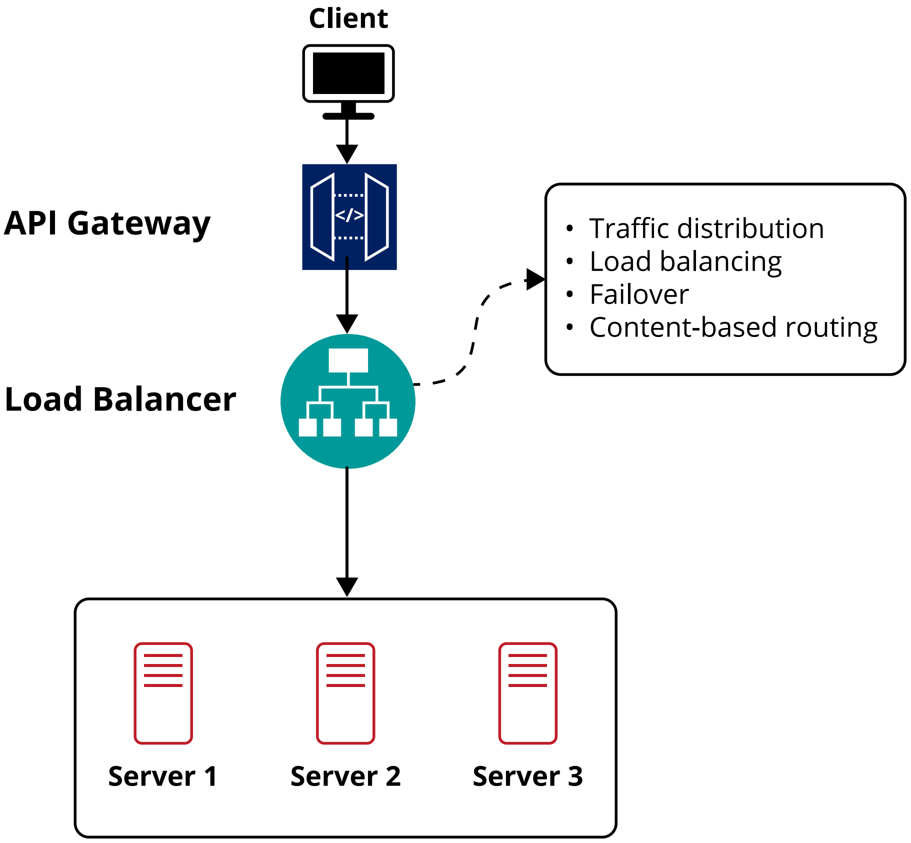
- Implementing and managing load balancing can introduce complexity into the system. Configuration, monitoring, and maintenance of load balancers require expertise and ongoing attention.
- In instances where the load balancer encounters operational difficulties or malfunctions, it can significantly impede the effective distribution of traffic to backend servers, potentially leading to service interruptions or degraded performance.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

STRUCTURE

Clients initiate requests to the system, sending them to the load balancer, which serves as the entry point. The load balancer employs a distribution algorithm to allocate incoming requests among the available server instances in the pool. This ensures an even distribution of the workload. Individual server instances process the requests, executing the application logic and generating responses.

The load balancer continuously monitors the health and performance of servers, adjusting the distribution of requests or triggering auto-scaling mechanisms based on predefined metrics. Once a server processes a request, the load balancer forwards the response back to the client. This dynamic and adaptive flow optimizes resource utilization, enhances system performance, and ensures fault tolerance in the face of changing workloads.



SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

SUGGESTED IMPLEMENTATION PLAN

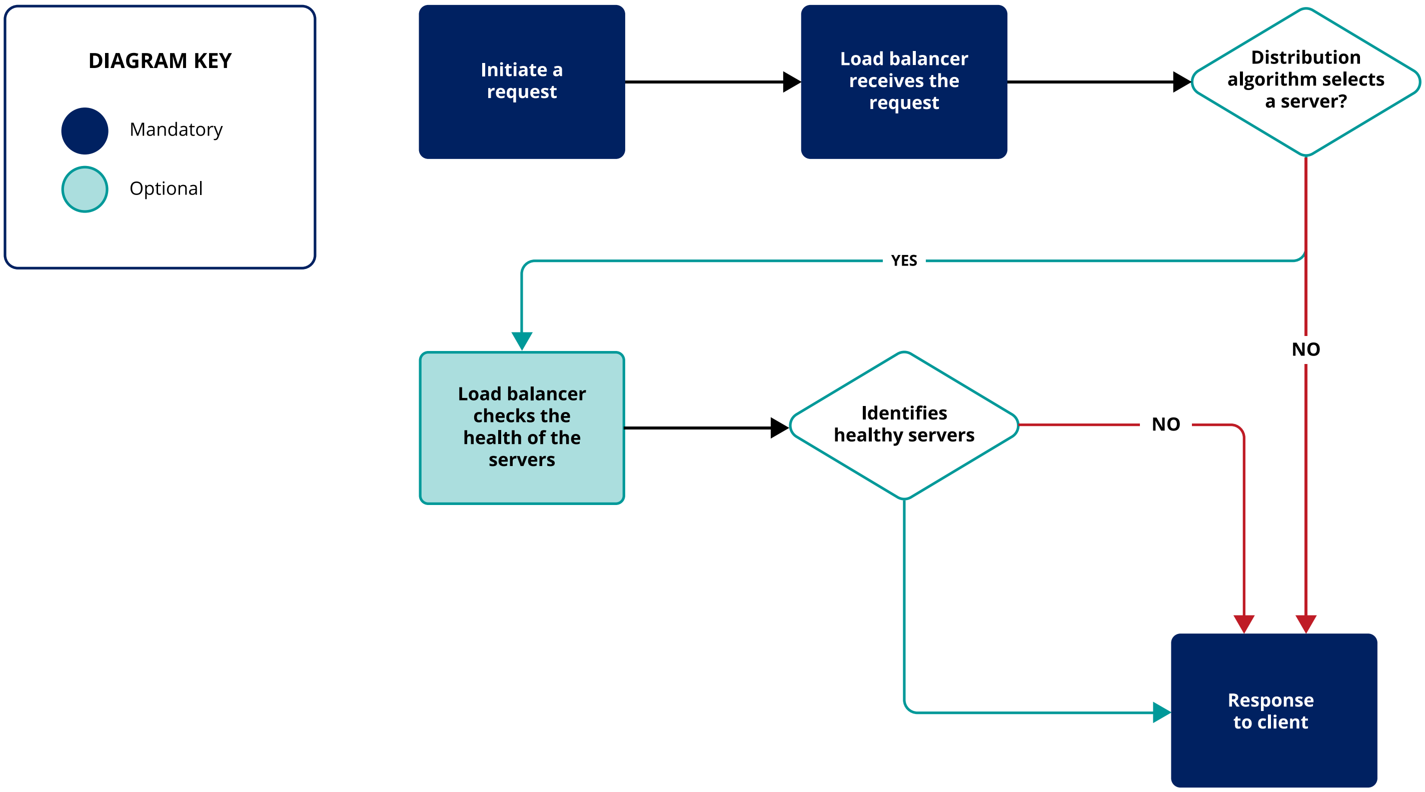


DIAGRAM KEY
● Mandatory
○ Optional
Initiate a request
Load balancer receives the request
Distribution algorithm selects a server?
YES
NO
Load balancer checks the health of the servers
Identifies healthy servers
NO
Response to client

RELATIONSHIP WITH SPATIAL REQUIREMENTS

PLAT.RQ.37 The “API gateway” MUST be able to connect to individual platform components (realized as microservices) and forward client requests to them. Hence, the API gateway MUST provide a solution for connecting the loosely coupled SPATIAL services.

PLAT.RQ.38 The “API gateway” MUST be able to operate with multiple backends, meaning that components can be deployed in different networks.

CLASSIFICATION

Class: Scalability

Tags: Traffic Management, High-Performance, Fault Tolerance, Scalability, Redundancy, High-Availability

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.3.5 UNIFIED AND USER-ORIENTED EXPLANATIONS PATTERN

INTENT

This pattern aims toward standardized and unified outputs of different explanation methods in an explainable AI system. This enables the easy interpretation, comparison, and automated analysis of explanations. By involving stakeholders into the process, specific needs and requirements can be incorporated into the explanations' design.

Problems that this pattern addresses

- Inconsistent outputs across different xAI methods: Different xAI methods may produce varying output formats, e.g., different dimensions or different scales of the explanations. This makes it difficult to compare different explanations for the same input.
- Lack of interpretability: Default explanations generated by xAI methods may not be easily interpretable by end-users. For instance, complex visualizations like heatmaps can be challenging to understand for non-technical end-users. Therefore, default explanations should be transformed into a format that is easy interpretable by the end-users.
- Different needs for different users: There can exist different end-users that have different needs and requirements for the explanation. Therefore, different explanation formats for different users should be favoured over one single format.

Aims that this pattern achieves

- For each group of end-users, one or more standardized explanation formats are defined, enabling an easy interpretation and comparison of explanations.
- Due to the consistent and standardized output formats, an automated accountability analysis can be integrated into the xAI system.
- By involving stakeholders such as developers, analysts and end-users into the design of explanation formats, the explanations are tailored to the users' background and needs.

APPLICABILITY

Usage scenarios

- The pattern can be applied to each AI system that requires explanations, ensuring interpretability and comparability of the explanations. Especially when explanations from different xAI methods are deployed, this guarantees the consistency of explanations across these methods.
- The pattern should be considered when different groups of users with different technical knowledge should be able to interpret the explanations.
- It should also be applied when the model is aimed to be accountable, and therefore, an accountability analysis should be part of the system.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

Relationship with SPATIAL use case

- In SPATIAL UC3, there are different stakeholders such as developers, cardiologists and users with less medical background that all should be provided with explanations. Due to the different technical and medical background of these user groups, the explanations must be tailored for each (e.g., fine-grained heatmaps for developers with technical background, more coarse-grained heatmaps for cardiologists and coarse-grained heatmaps along with clear text explanations for users without any technical nor medical background).
- In UC3, an understanding of the model's decision is of utmost importance as these decisions affect people's life. Performing an accountability analysis helps to build trust into the model and its explanations.

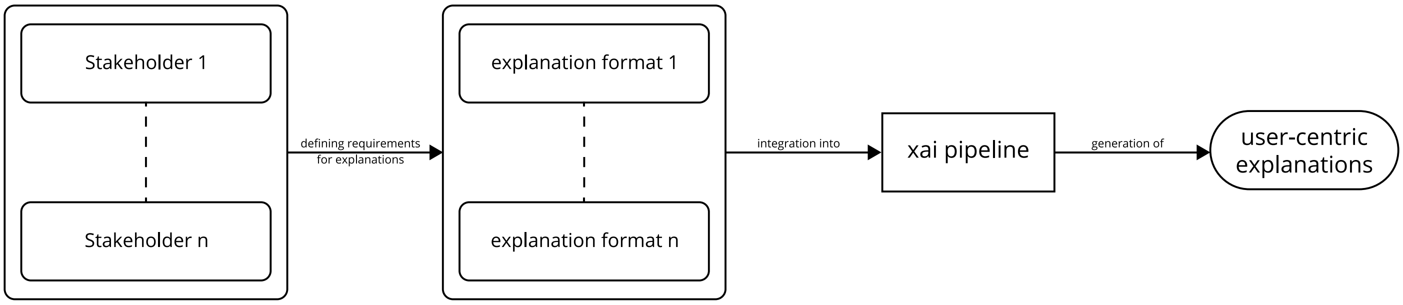
Drawbacks and limitations

- To involve stakeholders into the process of defining explanation formats can be time-consuming. It can involve iterative adaptions, as well as compromises due to different preferences by different users.
- Emphasizing user-friendly explanations may involve the simplification of explanations, potentially impacting the accuracy of the explanations. For example, averaging relevance scores of one segment may result into interpreting the segment to be not relevant, even though part of it may have been encountered to be highly relevant.
- Incorporating the transformation into the defined explanation formats adds computational cost to the system.

STRUCTURE

The first step is to identify relevant stakeholders who will use the explainable AI system. For each of the identified stakeholders, interviews should then be conducted to define requirements for the explanations to be interpretable. Once the requirements are defined, specific explanation formats for each of the stakeholders can be determined, such as the type, shape and scaling of the explanations.

Based on the data structure of the training set, the type of model that is chosen for the underlying task and the defined explanation formats, appropriate xAI methods can be selected. Finally, the default explanations from the selected xAI methods need to be transformed into the defined formats. As a result, explanations can be generated by the developed xAI system, that align with the specific needs of the different user groups.



SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

SUGGESTED IMPLEMENTATION PLAN

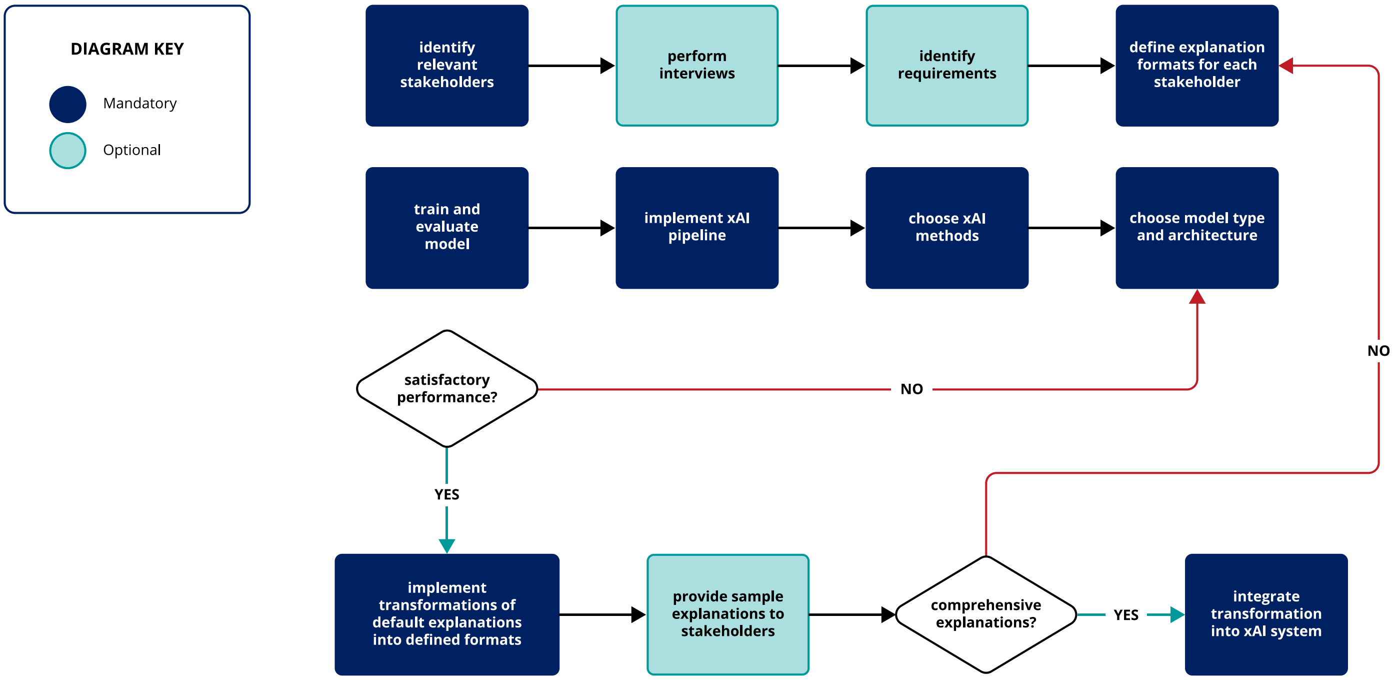


DIAGRAM KEY
● Mandatory
○ Optional
identify relevant stakeholders
perform interviews
identify requirements
define explanation formats for each stakeholder
train and evaluate model
implement xAI pipeline
choose xAI methods
choose model type and architecture
satisfactory performance?
NO
NO
YES
implement transformations of default explanations into defined formats
provide sample explanations to stakeholders
comprehensive explanations?
YES
integrate transformation into xAI system

RELATIONSHIP WITH SPATIAL REQUIREMENTS

MOD.RQ.6	ML models' predictions SHOULD provide high-level of explainability and should be understandable by humans.
USB.RQ.4	All decisions and outputs of AI-based systems SHOULD be as consistent as possible and follow pre-specified and interpretable formats.
USB.RQ.1	AI-based systems MUST provide comprehensible, uniform, and easy-to-use interfaces
USB.RQ.6	AI-based systems MUST provide explanations for individual decisions of the deployed AI models.

CLASSIFICATION

Class: Explainability

Tags: Standardization, User-Centric, Automated Analysis, Interpretability

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.3.6 PRIVACY-FOCUSED ML TRAINING PATTERN

INTENT

Enabling the training of stochastic-gradient-based ML models in a distributed way without the need of transferring raw training data, addressing critical issues as data privacy, data security, data access rights and access to heterogeneous data.

Problems that this pattern addresses

- Limited availability of real, centralized and heterogeneous datasets that could be useful for training ML models
- Privacy concerns might limit data sharing, what constraints the amount of information that can be used for training ML models
- Availability of limited communication bandwidth in certain deployments might constrain data sharing capabilities, what has a direct effect on the training feasibility.

Aims that this pattern achieves

- Enabling ML training without data acess, what prevents from incurring in data-handling and data-governance issues.
- Enabling collaborative ML model building across organizations with similar goals without the need of exchanging protected data.This is often referred as the cross-silo setting.

APPLICABILITY

Usage scenarios

- In privacy-sensitive sensitive data applications where data is distributed among multiple clients/devices. This is often referred as the cross-devices setting.
- In privacy-sensitive sensitive scenarios where different organizations aim at collaboratively building ML models without exchanging information between them. This corresponds to the cross-silos setting.

Relationship with SPATIAL use case

- Use Case 1 (Privacy Preserving AI on the edge and beyond) employs Federated Learning framework to build machine learning models.

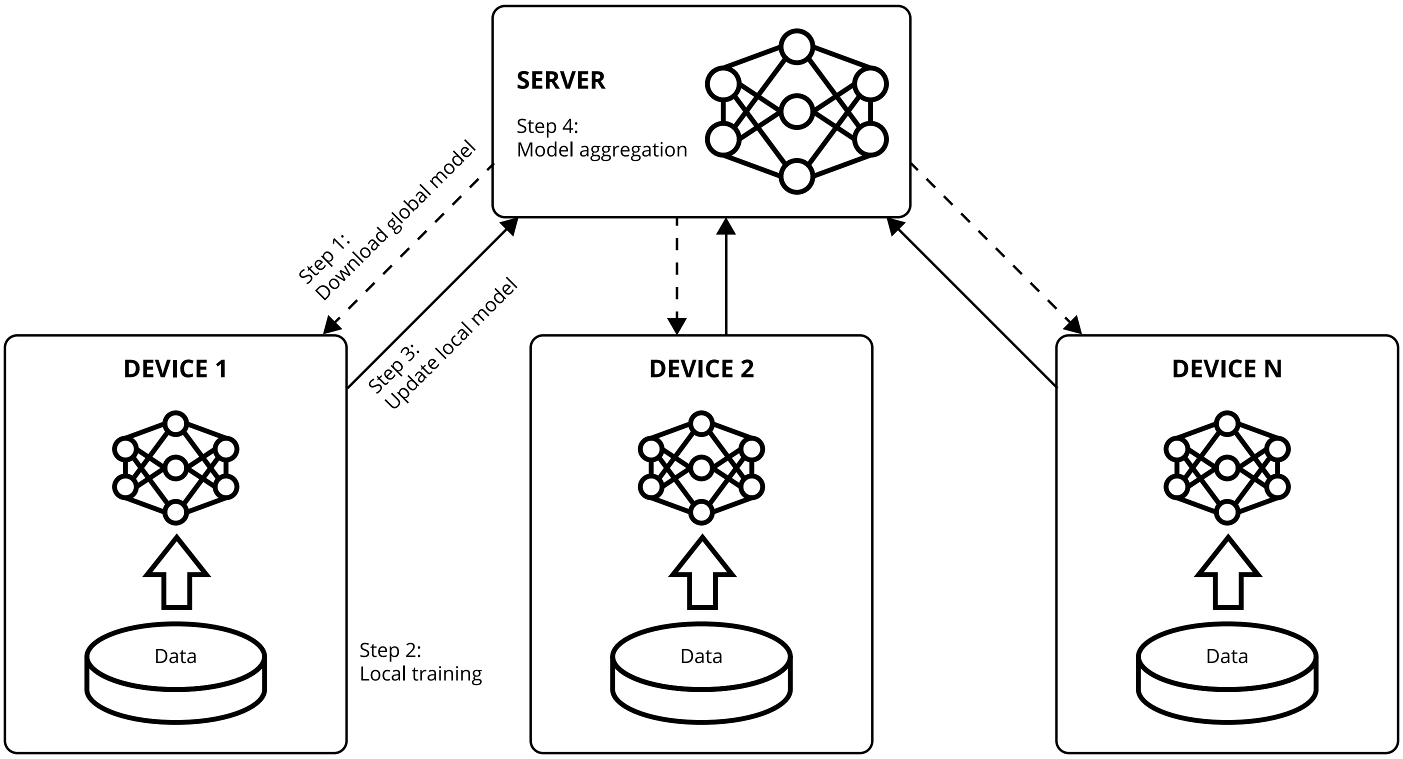
Drawbacks and limitations

- There may be limitations in achieving effective convergence in scenarios involving small, non-Independently and Identically Distributed (non-IID) data sets, impacting the model's learning efficiency.
- In situations where client devices possess constrained capabilities, the requirement to conduct local model training can be impractical, rendering the approach less effective.
- The efficiency of the FL process can be significantly reduced in scenarios where client devices consistently choose not to participate, thereby hindering the collaborative aspect of the learning process.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

STRUCTURE

In a federated learning setting, there is a central server coordinating the process. Local devices, such as smartphones or IoT devices, have their own local models and contribute to the global model's training without sharing raw data. The central server sends the global model to devices, which update it using their local data and send back these updates. These updates are aggregated to improve the global model, and this iterative process continues until the model reaches a satisfactory level of accuracy. This architecture ensures collaborative model training while preserving data privacy and security.



SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

SUGGESTED IMPLEMENTATION PLAN

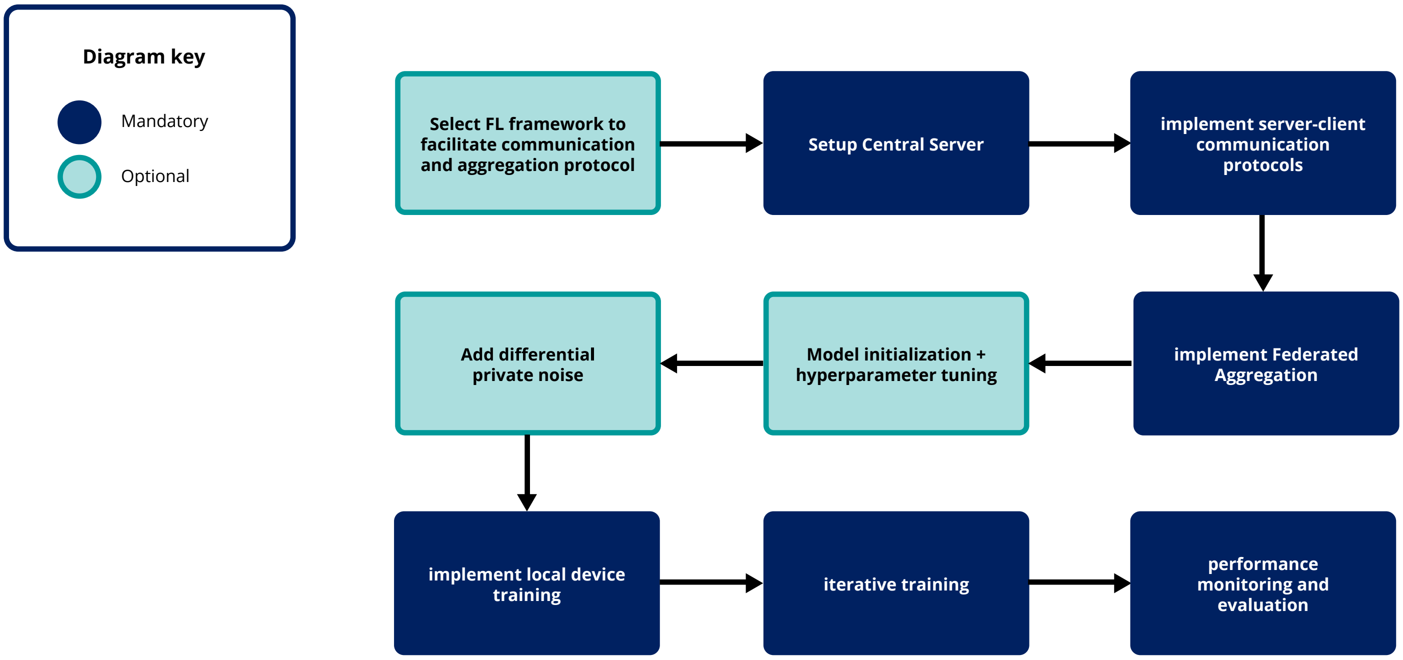


Diagram key
● Mandatory
○ Optional
Select FL framework to facilitate communication and aggregation protocol
Setup Central Server
implement server-client communication protocols
implement Federated Aggregation
Model initialization + hyperparameter tuning
Add differential private noise
implement local device training
iterative training
performance monitoring and evaluation

RELATIONSHIP WITH SPATIAL REQUIREMENTS

DAT.RQ.6	AI-based systems MUST be resilient against data reconstruction attacks
PRV.RQ.5	The possibility of privacy-related attacks on AI, system and data MUST be assessed and protection or mitigation processes MUST be made.
PRV.RQ.7	The privacy by design approaches SHOULD be included during the system design process.

CLASSIFICATION

Class: Privacy

Tags: Confidentiality, Distributed computing, Distributed Computing, Data Privacy, Security

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.3.7 TRUSTED EXECUTION ENVIRONMENT COMPUTING PATTERN

INTENT

Enabling Confidential Computing support without need to update the code, or significantly update the flow. Overhead added by setting up and enabling Confidential Computing layer and remote attestations is minimal for the end user.

Problems that this pattern addresses

- Complexity of setting up and monitoring trusted execution environments.
- Verification (attestation) of AI computations running in TEEs.
- Secure upload of AI algorithm and datasets and result download in multi-party computation scenarios.

Aims that this pattern achieves

- Executing computations in trusted execution environments is as close as possible to transparent from the end-user perspective.
- Setting up, monitoring, and management of the computations running in the TEE is safe and simple.
- Remote attestations procedures for computation running in TEE exist.
- Enable seamless integration of confidential computing using trusted execution environments.
- Executing computations in trusted execution environments is as close as possible to transparent from the end-user perspective.
- Setting up, monitoring, and management of the computations running in the TEE is safe and simple.
- Remote attestations procedures for computation running in TEE exist.
- Enable seamless integration of confidential computing using trusted execution environments.
- Simplify setup, monitoring, and management of TEEs with support for remote attestations to verify computations executed in TEE.

APPLICABILITY

Usage scenarios

- The pattern can be applied to each AI system that requires explanations, ensuring interpretability and comparability of the explanations. Especially when explanations from different xAI methods are deployed, this guarantees the consistency of explanations across these methods.
- The pattern should be considered when different groups of users with different technical knowledge should be able to interpret the explanations.
- It should also be applied when the model is aimed to be accountable, and therefore, an accountability analysis should be part of the system.

Relationship with SPATIAL use case

- UC1 to set up and use privacy-preserving environment.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

APPLICABILITY

⊖

Drawbacks and limitations

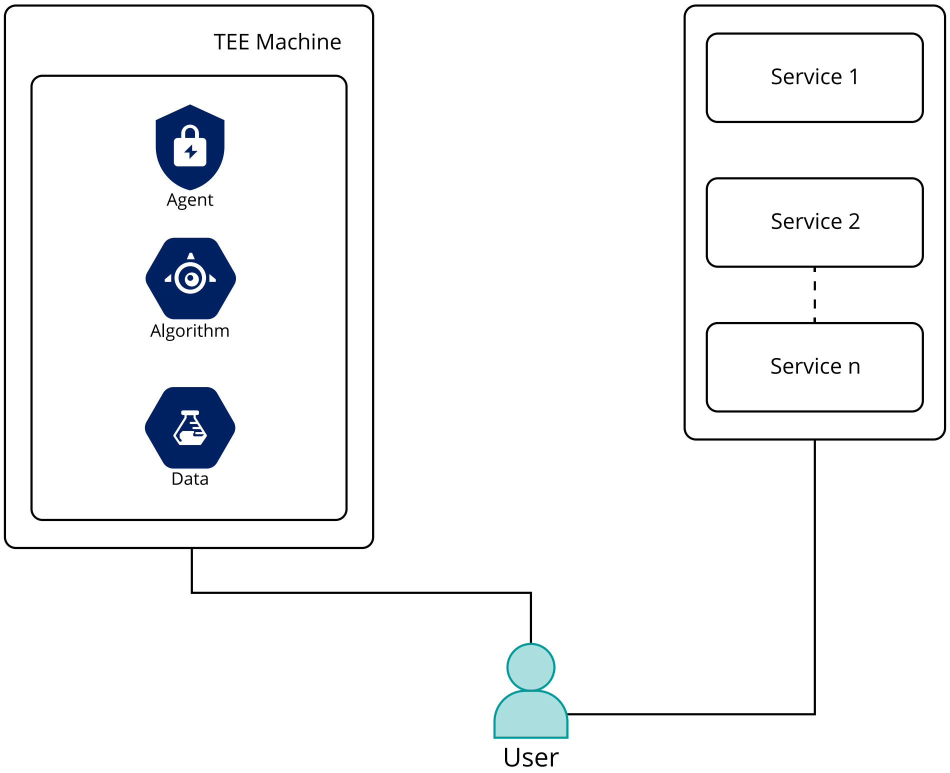
- The implementation may lead to increased consumption of system resources, necessitating careful consideration of resource allocation and management.
- Managing TEEs involves intricate orchestration to ensure secure and efficient operation, presenting a significant challenge in terms of technical complexity.
- The necessity for networking in the system architecture introduces potential security risks, as it opens an avenue for cyber attacks, requiring robust protective measures.

STRUCTURE

There are two types of users in multiparty computations scenario:

- data providers
- algorithm providers

The first step in preparation is to set execution environment that will be running inside TEE. After that, the user creates computation manifest and uploads metadata, such as keys used to open a secure connection once TEE is established. Once all the conditions are met, the software agent is started in the TEE that will be used for monitoring and managing the computation. The end-users will use agent to upload executable algorithm(s) and encrypted datasets, as well as to receive computation result (AI model or any other) and to receive attestation to very data consistency and privacy.



SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

SUGGESTED IMPLEMENTATION PLAN

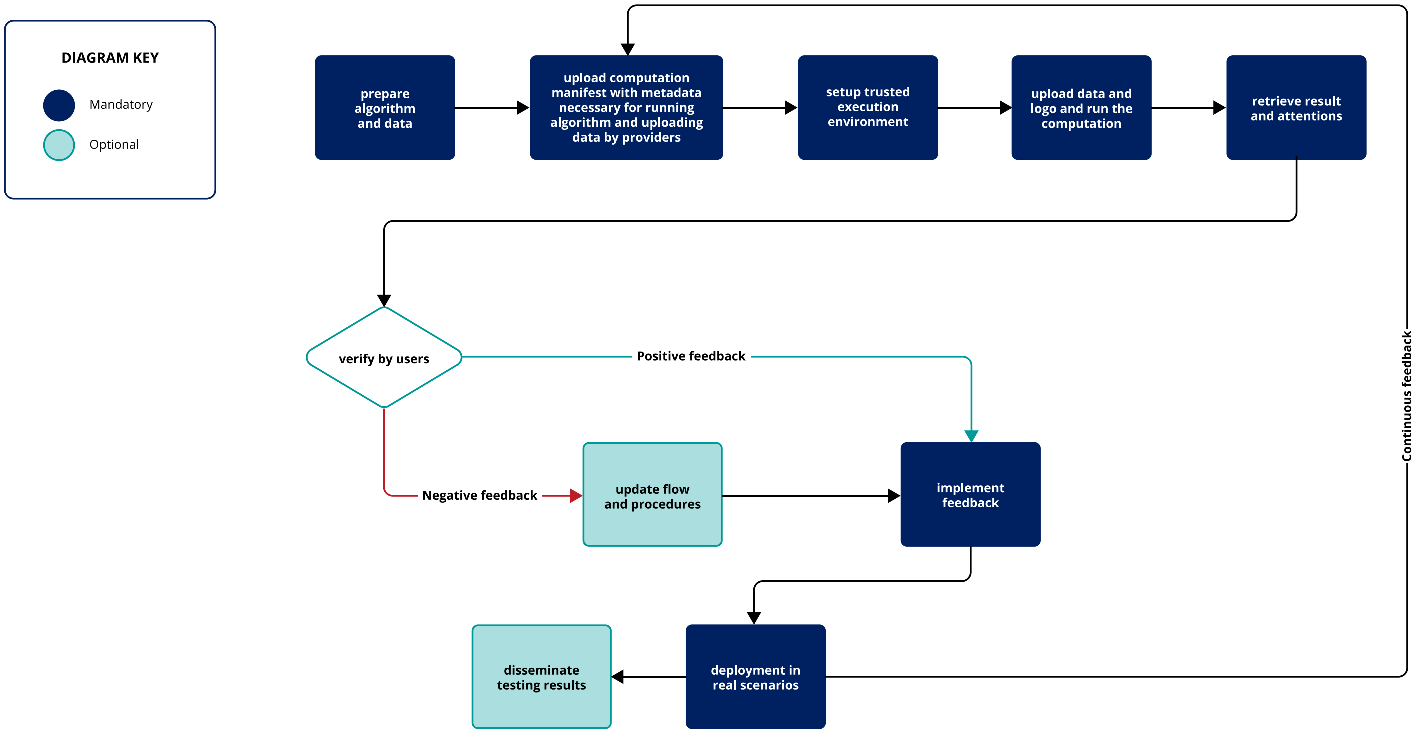


DIAGRAM KEY
● Mandatory
○ Optional
prepare algorithm and data
upload computation manifest with metadata necessary for running algorithm and uploading data by providers
setup trusted execution environment
upload data and logo and run the computation
retrieve result and attentions
verify by users
Positive feedback
Negative feedback
update flow and procedures
implement feedback
deployment in real scenarios
disseminate testing results
Continuous feedback

RELATIONSHIP WITH SPATIAL REQUIREMENTS

SEC.RQ.21 AI-based systems, dealing with sensitive or confidential data, MUST preserve the confidentiality of the data during the operational phase.

SEC.RQ.22 The integrity of the training data MUST be guaranteed between the data sources and the training platform.

SEC.RQ.23 The integrity of the machine learning model MUST be guaranteed between the training platform and the inference platform.

SEC.RQ.24 The number of parties involved in machine learning model training and inference SHOULD be restricted to the required minimum.

SEC.RQ.25 The machine learning model SHOULD be as isolated as possible from its clients and every interaction must be monitored to detect potential abuse.

SEC.RQ.26 The training platform for the machine learning model MUST be properly secured to prevent any compromise.

SEC.RQ.28 The provenance and integrity of inputs provided or computed by external parties SHOULD be verified.

DAT.RQ.14 AI-models can be continually trained with aggregated data, but consistency and integrity of data MUST be preserved through quantifiable estimations.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

CLASSIFICATION

Class: Security

Tags: Confidential Computing, TEE, Multiparty Computation

4.3.8 ADVERSARIAL TRAINING PATTERN

INTENT

Improving the robustness of AI models against evasions attacks by training with both clean (regular) data and adversarial examples that have been specifically crafted to deceive the AI systems.

Problems that this pattern addresses

- Susceptibility of machine learning models, particularly deep neural networks, to evasion attacks in which intentionally crafted perturbations are injected into input data with the goal of misleading the model
- AI models' sensitivity to small changes in input data, which humans may not even notice.
- Adversaries extract sensitive information from models through adversarial means

Aims that this pattern achieves

- Adversarial training seeks to make models more robust and resistant to such evasion attacks, thereby improving their trustworthiness.
- Adversarial training can contribute to preserving privacy by making it harder for adversaries to extract sensitive information from models through adversarial means.
- Adversarial training is an ongoing process, recognizing that adversarial attacks are continually evolving. It addresses the dynamic nature of adversarial threats and the need for continuous model improvement to stay ahead of attackers

APPLICABILITY

Usage scenarios

- Adversarial training should be considered in all general AI-based applications. Especially in applications where AI systems are used for critical tasks like autonomous vehicles, security systems, and fraud detection, it is essential that the models are secure and trustworthy.
- Adversarial training enhances the security of these systems by reducing their susceptibility to adversarial manipulation, thereby improving their trustworthiness. It should also be applied when the model is aimed to be accountable, and therefore, an accountability analysis should be part of the system.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

Usage scenarios

- Adversarial training should be considered in all general AI-based applications. Especially in applications where AI systems are used for critical tasks like autonomous vehicles, security systems, and fraud detection, it is essential that the models are secure and trustworthy.
- Adversarial training enhances the security of these systems by reducing their susceptibility to adversarial manipulation, thereby improving their trustworthiness.
- Adversarial training is also important in scenarios where privacy is a concern, such as in healthcare and financial applications.

Relationship with SPATIAL use case

- Adversarial training has been assessed in the context of Use Case 2 - Network Traffic Analysis for Anomaly Detection to make the application resilient against evasion attacks.

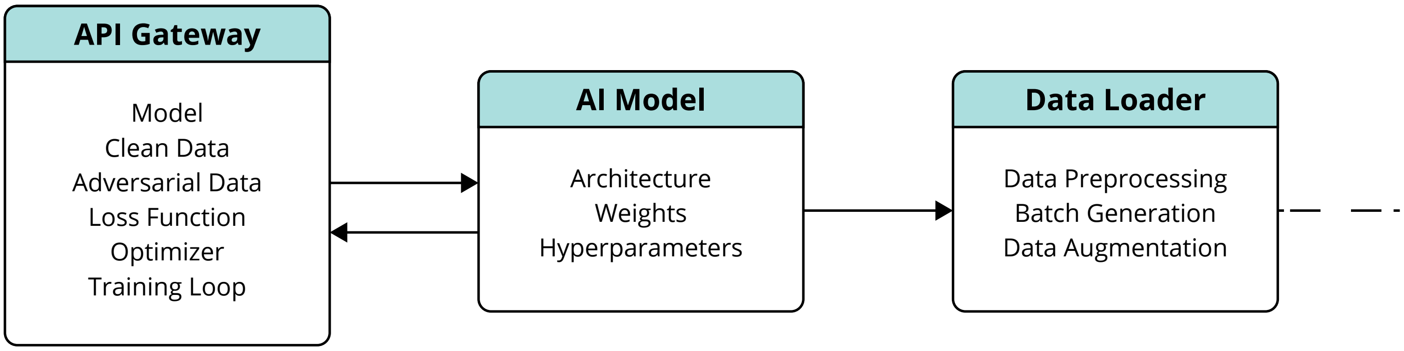
Drawbacks and limitations

- Adversarial training requires **additional iterations** during model training, where both clean and adversarial examples are used. This can significantly increase the **computational resources and time** needed for training.
- In some cases, adversarial training might trade off model accuracy on clean data for improved robustness against adversarial attacks. Striking the right balance between accuracy and robustness can be challenging.
- Effective adversarial training relies on having a representative set of adversarial examples that mirror real-world attack scenarios. Creating these examples can be challenging, and the model's robustness is limited to the types of adversarial examples used during training. Intensive adversarial training can lead to overfitting, where the model becomes overly specialized in recognizing specific adversarial examples. This can reduce the model's ability to generalize to new, unseen data.

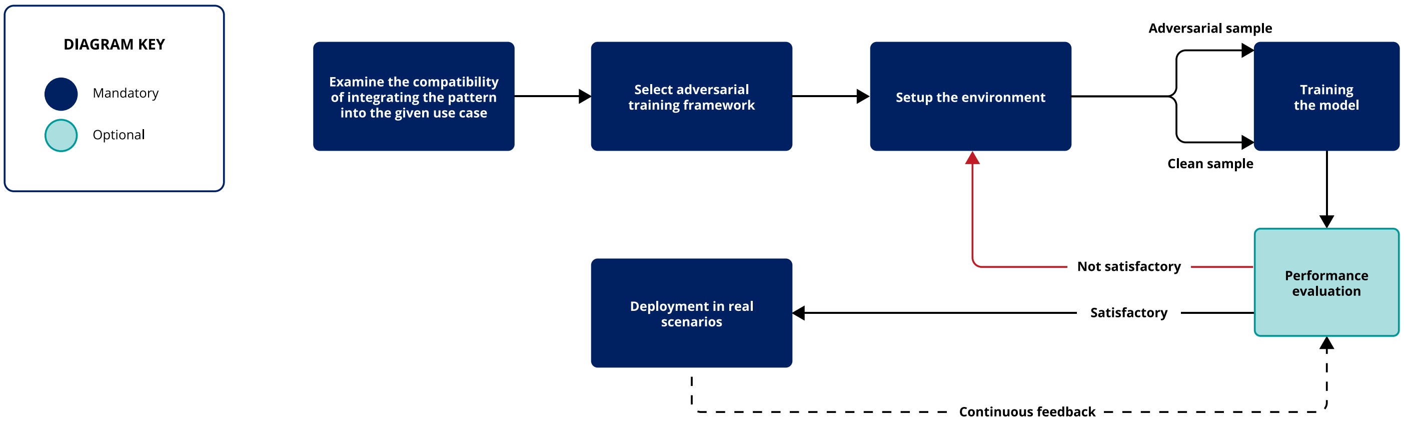
STRUCTURE

Here below a simplified structure on how **Adversarial Training** can be integrated into an AI model. Adversarial Training represents the top left component. It includes key elements of the adversarial training process, such as the AI model, clean data, adversarial data, loss function, optimizer, and the training loop. **AI Model** is the core component which is trained with both clean and adversarial data. **Data Loader** is responsible for data management, including data pre-processing, batch generation and data augmentation.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.



SUGGESTED IMPLEMENTATION PLAN



RELATIONSHIP WITH SPATIAL REQUIREMENTS

SEC.RQ.1 AI-based systems MUST be resilient against data reconstruction attacks.

USB.RQ.9 Users of AI-based systems SHOULD be able to identify, report, and correct mistakes in the decision-making of AI models.

MOD.RQ.1 The ML model MUST have a high accuracy.

MOD.RQ.8 ML models SHOULD be testable to verify they fulfil expectations on their outputs.

DAT.RQ.12 Pre-processed data MAY be enriched further before training AI models to improve robustness and performance

DAT.RQ.13 Pre-processed input data SHOULD be linked with prediction outputs of AI models to derive quantifiable explanations to users.

CLASSIFICATION

Class: Security

Tags: Evasion Attacks, Robustness, Evasion Attack Resilience, Trustworthiness

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.3.9 LABEL SANITIZATION PATTERN

INTENT

Preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted

Problems that this pattern addresses

- Errors, heterogeneity and inconsistencies in datasets (spelling and syntax errors, mistakes such as empty fields, data duplications,)
- Risk of privacy breaches, data leaks and unauthorized access to sensitive data.
- Harmful or manipulated data generated in data poisoning attacks

Aims that this pattern achieves

- Data sanitization helps protect the privacy of individuals and sensitive data subjects by removing or obfuscating personally identifiable information (PII) and sensitive details from datasets.
- Data sanitization often includes the identification and correction of errors and inconsistencies in datasets, which contributes to higher data quality and more accurate analysis.
- Many data protection regulations, such as GDPR (General Data Protection Regulation), HIPAA (Health Insurance Portability and Accountability Act), and CCPA (California Consumer Privacy Act), require organizations to sanitize data to comply with legal requirements for data privacy and security.
- Data sanitization helps detect and remove potentially harmful or manipulated data to mitigate data poisoning attacks.
- Sanitizing data by removing irrelevant or noisy features reduces the dimensionality of the data, which can lead to more efficient model training and faster inference.

APPLICABILITY

Usage scenarios

- Data sanitization should be considered in all AI-based applications, especially in healthcare and medical applications for ensuring patient privacy and complying with regulations.

Relationship with SPATIAL use case

- Data sanitization has been studied in the context of Use Case 2 - Network Traffic Analysis for Anomaly Detection to make the application more robust and resilient against poisoning attacks.

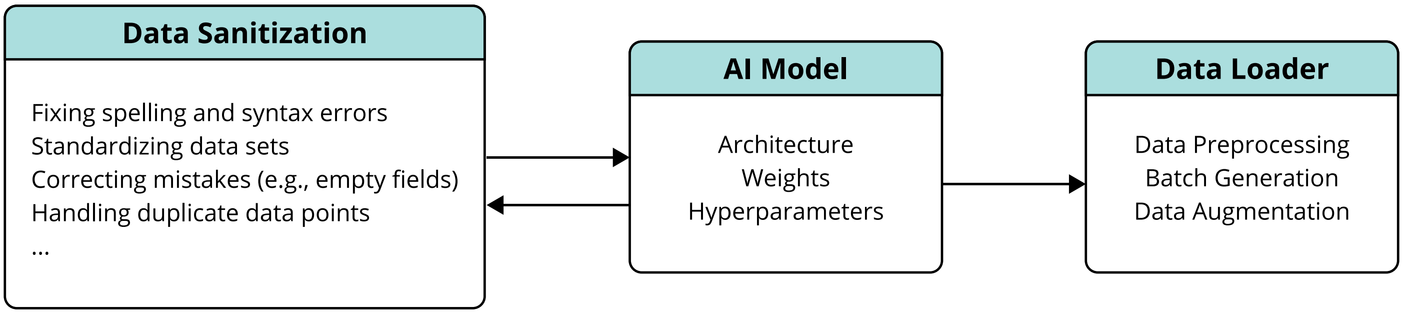
Drawbacks and limitations

- The primary purpose of data sanitization is to remove or obscure sensitive information. In the process, there is a risk of unintentional data loss. If not done carefully, important data may be irretrievably deleted.
- Sanitizing large datasets can be resource-intensive in terms of time, computational power, and storage. It may slow down data processing and analysis.

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

STRUCTURE

Presented here is an overview of the integration of Data Sanitization within an AI model framework. Positioned as the primary component in the top left, Data Sanitization encompasses critical functions such as rectifying spelling and syntactical errors, standardizing datasets, rectifying anomalies like empty fields, and pinpointing duplicate data entries. At the heart of the structure lies the AI Model, which undergoes training utilizing both sanitized and adversarial data sets. The Data Loader plays a pivotal role in data governance, handling tasks like data preprocessing, batch generation, and data augmentation to optimize the model's performance and accuracy.



SUGGESTED IMPLEMENTATION PLAN

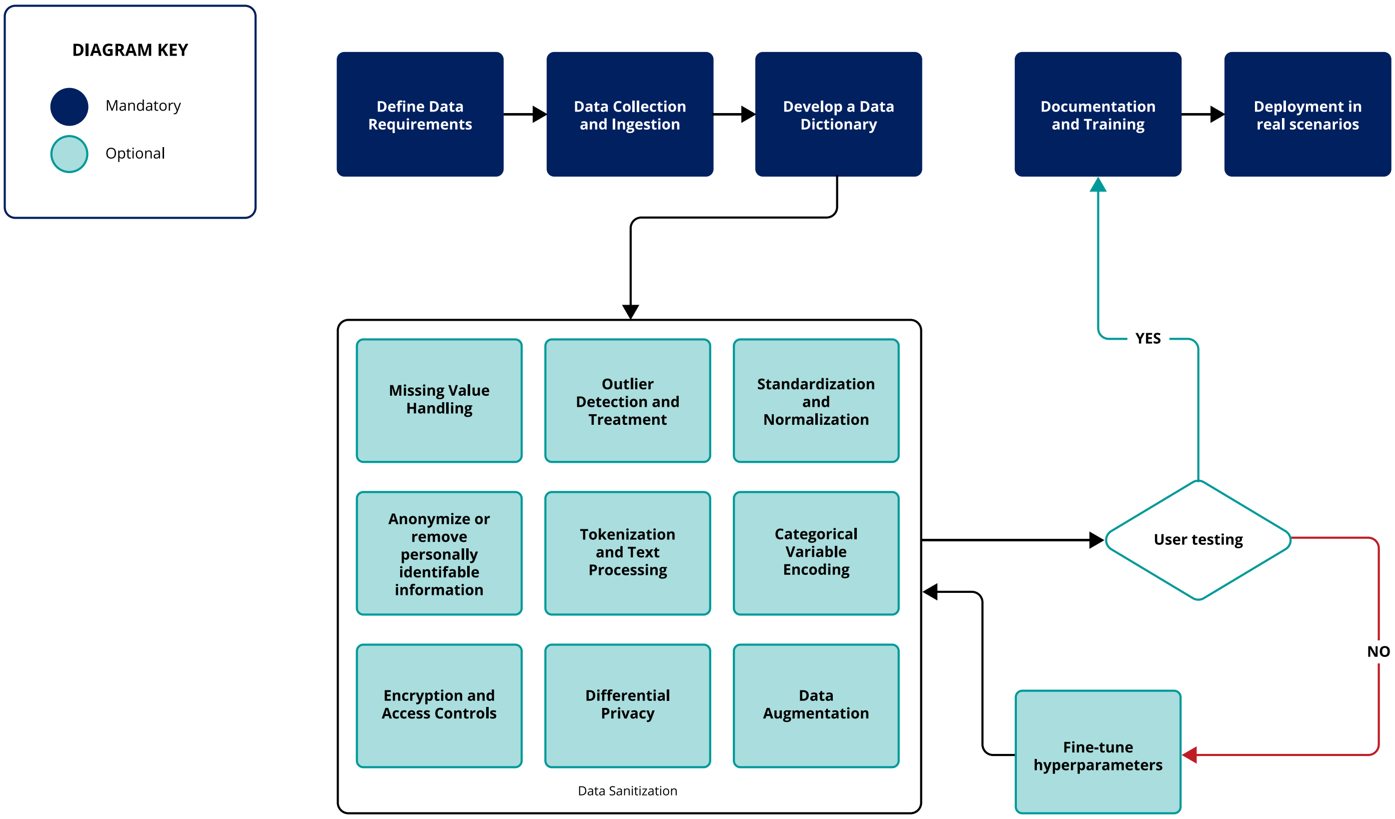


DIAGRAM KEY
Mandatory
Optional
Define Data Requirements
Data Collection and Ingestion
Develop a Data Dictionary
Documentation and Training
Deployment in real scenarios
Missing Value Handling
Outlier Detection and Treatment
Standardization and Normalization
Anonymize or remove personally identifiable information
Tokenization and Text Processing
Categorical Variable Encoding
Encryption and Access Controls
Differential Privacy
Data Augmentation
Data Sanitization
YES
User testing
NO
Fine-tune hyperparameters

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

RELATIONSHIP WITH SPATIAL REQUIREMENTS

SEC.RQ.1	AI-based systems MUST be resilient against data reconstruction attacks.
USB.RQ.9	Users of AI-based systems SHOULD be able to identify, report, and correct mistakes in the decision-making of AI models.
MOD.RQ.1	The ML model MUST have a high accuracy.
MOD.RQ.8	ML models SHOULD be testable to verify they fulfil expectations on their outputs.
DAT.RQ.12	Pre-processed data MAY be enriched further before training AI models to improve robustness and performance
DAT.RQ.13	Pre-processed input data SHOULD be linked with prediction outputs of AI models to derive quantifiable explanations to users.

CLASSIFICATION

Class: Data Integrity

Tags: Data Cleansing, Privacy Compliance, Poisoning Attack Mitigation

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.3.10 ENHANCED INTERPRETABILITY PATTERN

INTENT

Increase understandability of XAI outputs by adding exploratory and customizable elements to the interaction.

Problems that this pattern addresses

- Different XAI outputs provide information at different level of understanding
- Interpretability is subjective
- No one XAI solution suits all problems

Aims that this pattern achieves

- This pattern focuses on providing users with the flexibility to select their preferred type of explanation, thereby catering to diverse needs and enhancing user autonomy in the interpretative process.
- Dedicated to refining the clarity and comprehensibility of explanations, this measure aims to make interpretive outputs more accessible and understandable to a broader user base.
- Concentrates on developing more dynamic and engaging interfaces for XAI solutions, fostering a more interactive and user-friendly experience in engaging with explainable AI systems.

APPLICABILITY

Usage scenarios

- This pattern may be used when the explanation is intended for heteregenous users.
- This may also be used when the explanations needs to be flexible.
- This may also be used when intended for a generic audience.

Relationship with SPATIAL use case

- We use this for the Emergency e-calling system use-case to understand the needs of the healthcare experts from an explainability perspective.
- We may also extend this further to suit the explanations for the patients, who may vary in needs and expertise.

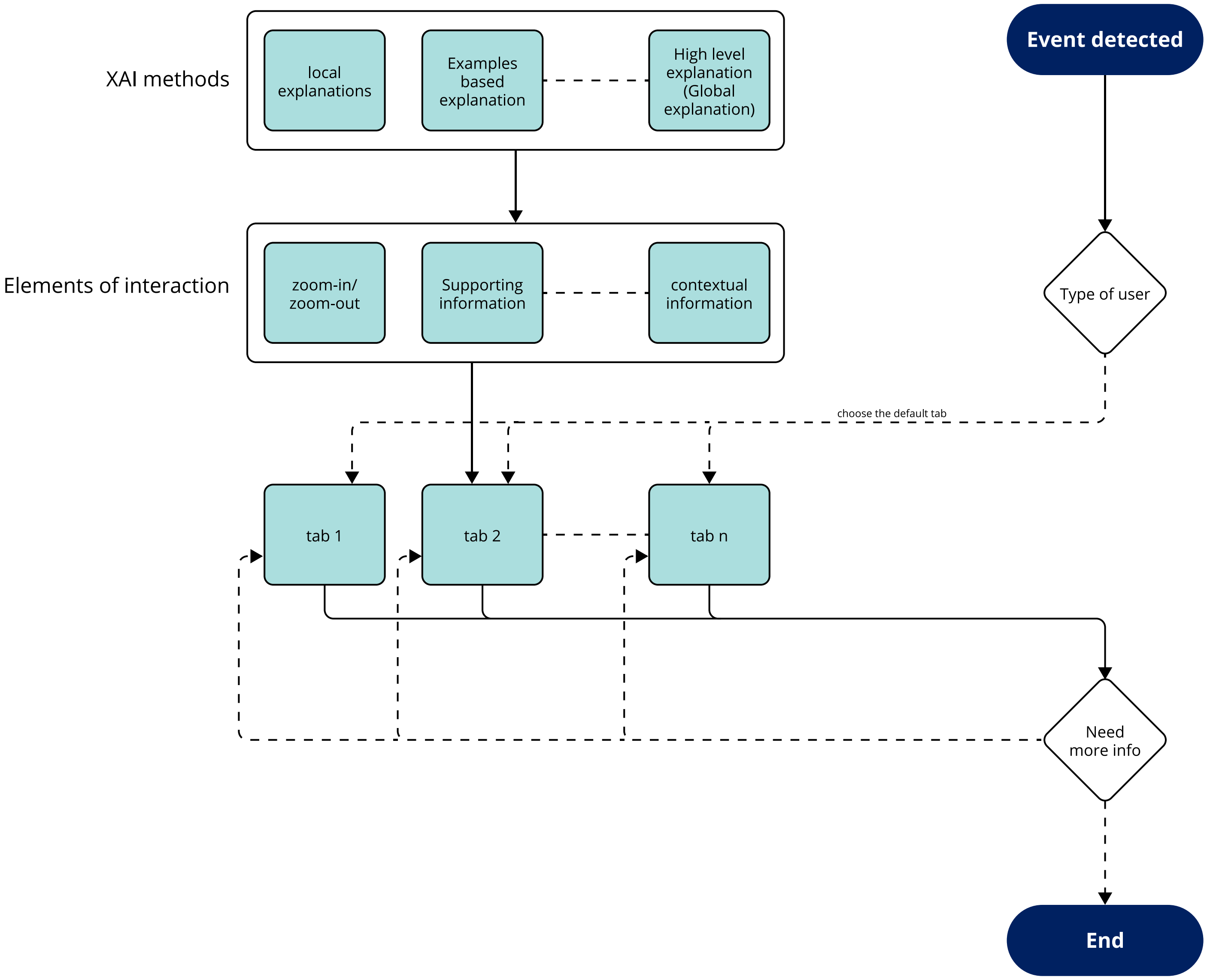
Drawbacks and limitations

- User-driven tailoring of explanations carries a potential trade-off between flexibility and the ability to give the target audience the best explanation possible.
- The selection of XAI methods to be used in the solution might need more information about the intended audience.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

STRUCTURE

Explanations are generated through various methods, ranging from detailed local explanations to more overarching global ones. In terms of presentation, these explanations are visualized and augmented with interactive features, enabling users to alter the display for more effective and intuitive access to pertinent information. This interactivity might include options like zooming for enhanced visibility or incorporating text-based context and guidance to assist in navigating and comprehending the explanation's components. Different methods of explanation are organized into distinct tabs within the user interface of the explanation tool. This arrangement allows users to easily switch between and evaluate the clarity of explanations according to their personal preferences.



SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

SUGGESTED IMPLEMENTATION PLAN

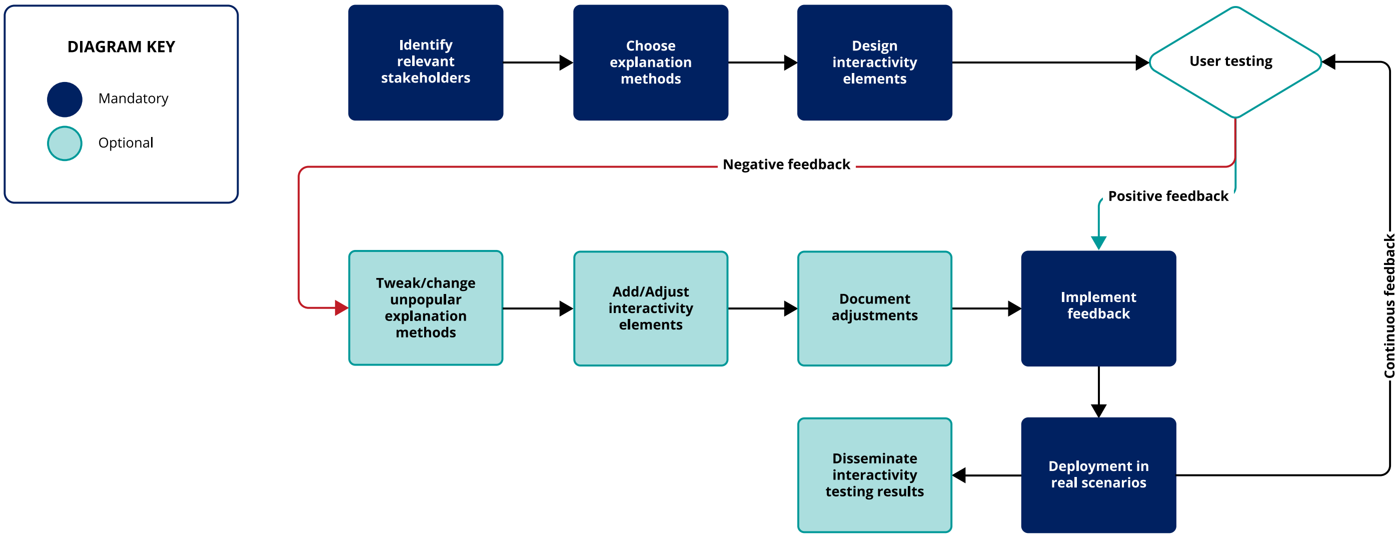


DIAGRAM KEY
Mandatory
Optional
Identify relevant stakeholders
Choose explanation methods
Design interactivity elements
User testing
Negative feedback
Positive feedback
Tweak/change unpopular explanation methods
Add/Adjust interactivity elements
Document adjustments
Implement feedback
Continuous feedback
Disseminate interactivity testing results
Deployment in real scenarios

RELATIONSHIP WITH SPATIAL REQUIREMENTS

MOD.RQ.6	ML models' predictions SHOULD provide high-level of explainability and should be understandable by humans.
MOD.RQ.12	ML models' predictions SHOULD be interpretable by humans and provide meaning in the context of their designed functional purpose.
USB.RQ.1	AI-based systems MUST provide comprehensible, uniform, and easy-to-use interfaces.
USB.RQ.3	An AI-based system SHOULD have functionalities that guide users in the usage of the system and provide help in case of problems.
USB.RQ.4	All decisions and outputs of AI-based systems SHOULD be as consistent as possible and follow pre-specified and interpretable formats.
USB.RQ.6	AI-based systems MUST provide explanations for individual decisions of the deployed AI models that have to be adapted to the respective technical expertise and domain knowledge of the users.

CLASSIFICATION

Class: Explainability

Tags: Flexibility, Interactive, User-Oriented

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

4.3.11 ITERATIVE RESILIENCE IMPROVEMENT AGAINST EVASION ATTACK PATTERN

INTENT

Improve and validate the resilience of ML-based systems against evasion attacks through a process of iterative empirical assessment and continuous improvement. This approach involves systematically evaluating the system's ability to withstand various evasion tactics, identifying vulnerabilities, and applying targeted enhancements.

Problems that this pattern addresses

- There is no fool-proof defense against evasion attacks. Different defense approaches need to be tested and combined to achieve best robustness depending on context.
- ML-based systems must be resilient against well-known adversarial attacks, such as evasion attacks which is the most popular.
- The resilience of ML-based systems needs to be quantified and validated, i.e., we must demonstrate it reaches expected security requirements.

Aims that this pattern achieves

- Enabling to make design and implementation choices for the ML system driven by security criteria, i.e., resilience against evasion attacks.
- Improve overall resilience of ML systems against evasion attacks.
- Provide evidence of resilience Improve overall resilience of ML systems against evasion attacks.

APPLICABILITY

Usage scenarios

- When an ML system is potentially threatened by evasion attacks and the risk of attackers launching such attacks is high. This should be inferred through threat modelling.
- When the risk and consequences for performance degradation of the ML system are important.
- The design pattern can be applied during implementation of the ML system, prior to deployment and during operations, respectively to improve the secure design and to validate a sufficient level of resilience

Relationship with SPATIAL use case

- UC 2: Resilience against adversarial ML attacks is a key requirement in cybersecurity applications such as traffic analysis for anomaly detection
- UC 4: Resilience against adversarial ML attacks is a key requirement in cybersecurity applications such as MalDoc and APK detection system

Drawbacks and limitations

- Induce delay and implementation and deployment of ML systems because of extra process to improve and validate security
- Resilience improvements come with a trade-off with other ML requirements such as performance, accuracy and privacy.
- Additional cost for security monitoring and resilience validation during operations

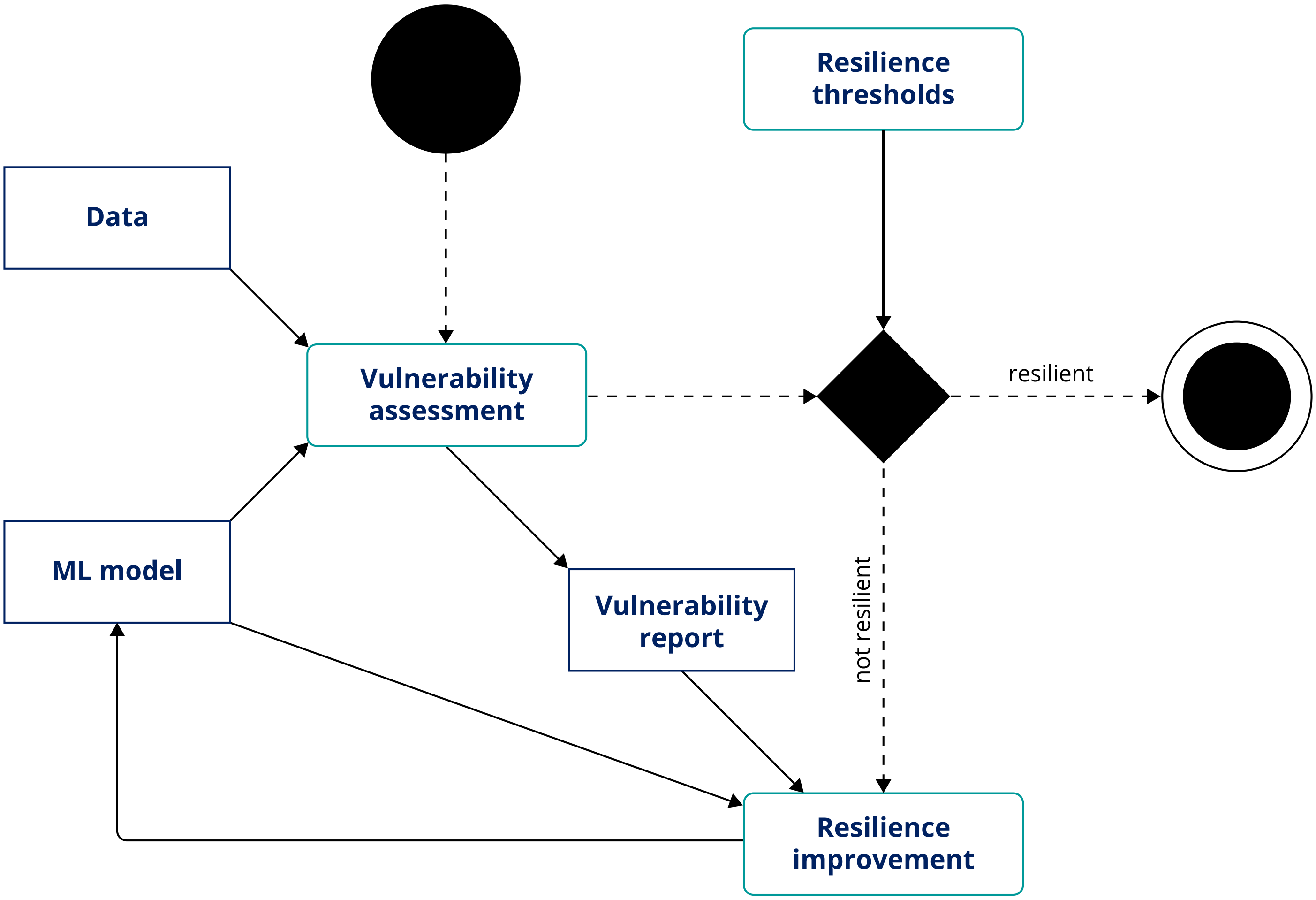
SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

STRUCTURE

The iterative resilience improvement process starts with a vulnerability assessment step, which takes as input a small test Data set and the ML model to improve. It computes several resilience metrics, i.e., impact, complexity, detectability and produces a vulnerability report. The computed resilience metric values are compared with thresholds, which have been previously defined, to decide whether the ML model is resilient enough or not. These resilience metric tests can be augmented with accuracy metric tests or explainability metric tests, if one wants to validate more than just resilience, and aim to meet some trade-offs.

If the ML model is resilient enough (based on resilience metrics thresholds), the iterative resilience improvement process is over. If it is not, the resilience improvement step is applied, using as input the vulnerability report from vulnerability assessment. Resilience improvement is applied onto the ML model, and it modifies it. It can consist in a) feature selection, to remove features manipulated by evasion attacks, b) adversarial training or c) query rate limiting. Resilience improvement produces a new ML model, expected to be more robust, which can restart the iterative resilience improvement process with the vulnerability assessment step.

The process is repeated until the ML model meets the resilience requirements based on the resilience metrics thresholds.



SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

SUGGESTED IMPLEMENTATION PLAN

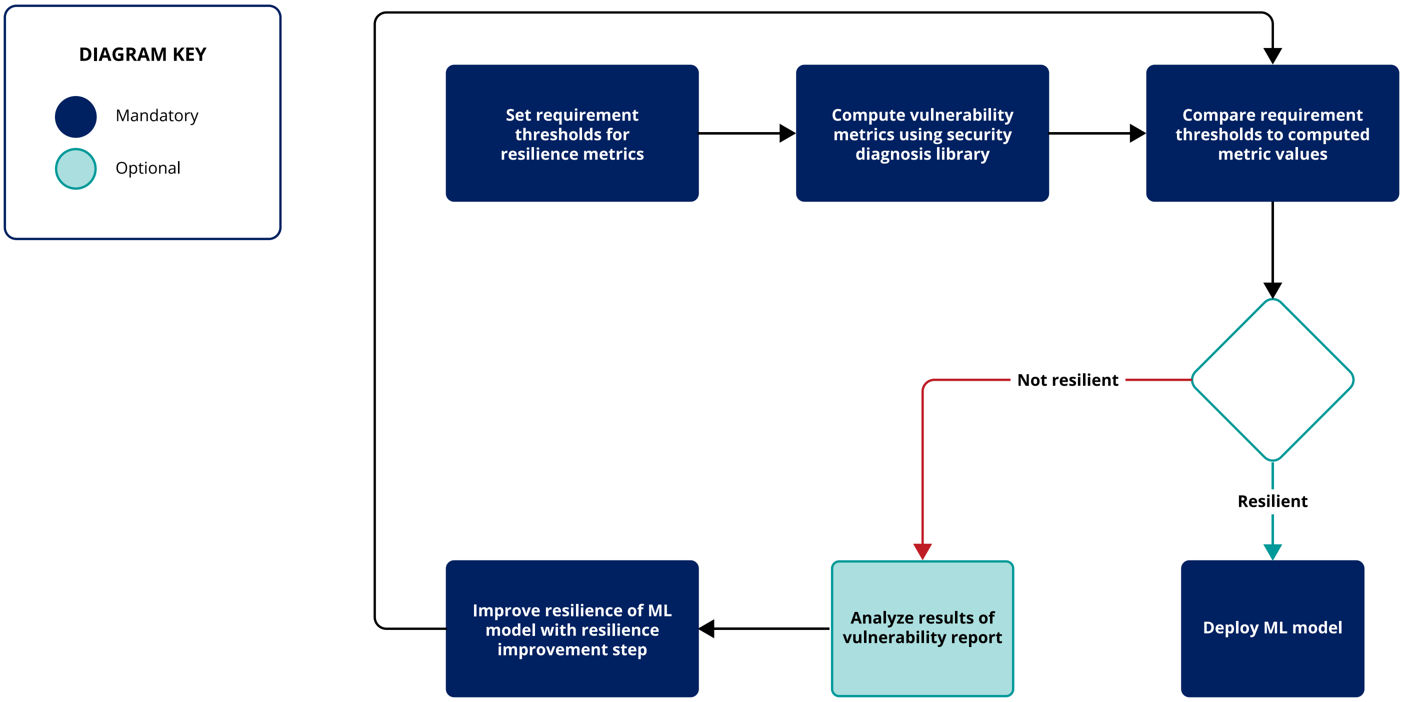


DIAGRAM KEY
● Mandatory
○ Optional
Set requirement thresholds for resilience metrics
Compute vulnerability metrics using security diagnosis library
Compare requirement thresholds to computed metric values
Not resilient
Resilient
Improve resilience of ML model with resilience improvement step
Analyze results of vulnerability report
Deploy ML model

RELATIONSHIP WITH SPATIAL REQUIREMENTS

MOD.RQ.8	ML models SHOULD be testable to verify they fulfil expectations on their outputs.
MOD.RQ.10	ML models MUST provide objective evidence that requirements and a specific intended use have been fulfilled.
MOD.RQ.13	ML models SHOULD withstand unexpected adverse events, unexpected changes and malicious attacks in their environment or use.
LEG.RQ.6	According to the AI Act, there MAY need to be a testing process to identify risks and determine appropriate mitigation measures, and to validate that the system runs consistently for the intended purpose, with tests made against prior metrics and validated against probabilistic thresholds.
LEG.RQ.14	According to the AI Act, AI systems MAY need to be designed with the option that allows deployers to monitor the robustness and cybersecurity measures of the system.
SEC.RQ.1	AI-based systems MUST be resilient against the evasion attacks.

CLASSIFICATION

Class: Security

Tags: Continuous Assessment, Evasion Attack Resilience, Validation

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

DESIGN PRINCIPLES

5.1 INTRODUCTION TO SPATIAL DESIGN PRINCIPLES

The SPATIAL project's seven design principles are integral to the development of reliable, ethical, and effective AI systems. These principles were defined through a collaborative effort involving experience from the project and a series of workshops with every consortium member of the SPATIAL project. Each principle serves a specific purpose and is led by a different consortium member, ensuring a comprehensive and multi-faceted approach to AI system design and implementation. Each principle was meticulously crafted, reflecting a blend of practical experiences from the SPATIAL project and collective wisdom from consortium members, ensuring a robust and holistic approach to AI system development.

The main objective of the SPATIAL design principles is to guide the development and implementation of AI systems in a manner that is ethical, reliable, secure and user-centric. These principles aim to ensure that AI technologies are designed and operated with a high standard of data quality, privacy, and security, while also being fair, transparent, and compliant with legal and regulatory frameworks. The overarching goal is to create AI systems that are not only technically proficient and high-performing but also trustworthy and beneficial to society, respecting individual rights and promoting inclusivity and fairness. By adhering to these principles, the SPATIAL project seeks to foster the development of AI technologies that are aligned with human values and capable of positively impacting a wide range of stakeholders.

5.2 BACKGROUND RESEARCH

Over recent years, most organizations linked to technology policy have formulated or supported various AI principles. These guidelines, aimed at ensuring ethical, rights-focused, and socially beneficial AI, are evolving as quickly as the technology itself, highlighting the critical need for comprehensive understanding. In preparation for the design stage of the SPATIAL design principles, we have reviewed the most comprehensive survey to date "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI"[14] which is a research publication from the Berkman Klein Center for Internet & Society, authored by Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar, and released in 2020. This paper represents a significant scholarly effort to understand and compare various ethical and rights-based frameworks that have emerged in response to the rapid development and spread of AI systems. The research focuses on thirty-six prominent AI principles documents as presented in Figure 5, analyzing them to identify common themes and trends.

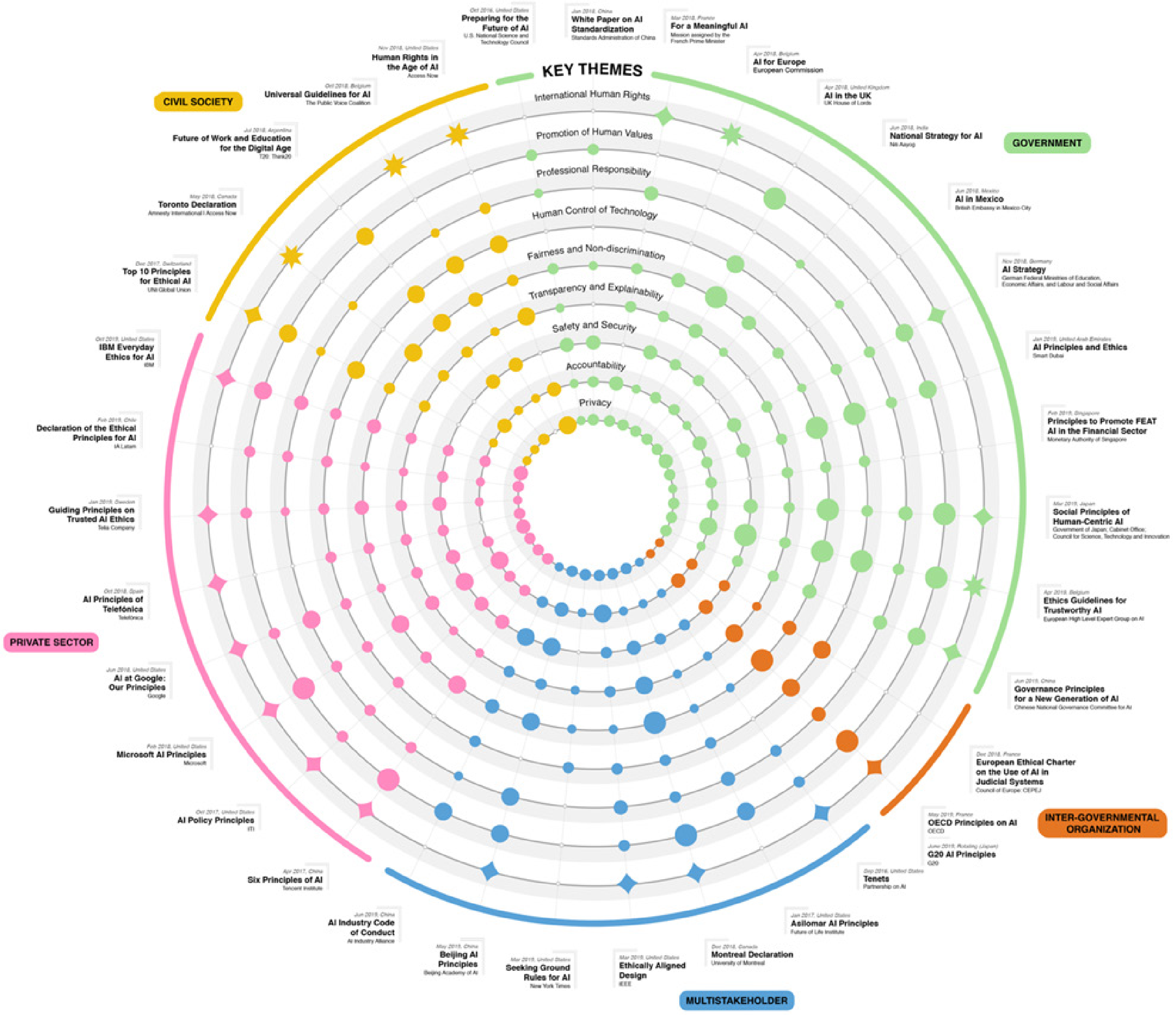


Figure 5 thirty-six prominent ai principles documents analyzed by berkman klein center for internet & society

This analysis revealed a growing consensus around eight key thematic trends: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and the promotion of human values. This detailed analysis helped us to provide a clearer understanding of the consensus and divergences among different ethical and rights-based approaches to AI and as a result ,supported us in defining principles presented in Section 5.3. The SPATIAL project's principles are not only vital in the development of ethical, reliable, and effective AI systems but also unique for their inclusion of the experiential element, which is often missing in other AI ethical frameworks, such as the ones discussed in the survey cited above. This experiential element is crucial because it reflects practical, real-world considerations that are essential for the successful deployment and acceptance of AI technologies

5.3 PROPOSED SPATIAL PRINCIPLES

Over recent years, most organizations linked to technology policy have formulated or supported various AI principles. These guidelines, aimed at ensuring ethical, rights-focused, and socially beneficial AI, are evolving as quickly as the technology itself, highlighting the critical need for comprehensive understanding. In preparation for the design stage of the SPATIAL design principles, we have reviewed the most comprehensive survey to date “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI”[14] which is a research publication from the Berkman Klein Center for Internet & Society, authored by Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar, and released in 2020. This paper represents a significant scholarly effort to understand and compare various ethical and rights-based frameworks that have emerged in response to the rapid development and spread of AI systems. The research focuses on thirty-six prominent AI principles documents as presented in Figure 5, analyzing them to identify common themes and trends.

Principle 1 — Ensure high-quality and accurate data

Theme: Data quality, representativeness, and management

The data used for training and evaluating AI models must be representative of the real-world scenario the AI system aims to address. This requires collecting diverse data that encompasses various demographics, geographies, and relevant contextual factors. Additionally, it is essential to ensure the data is unbiased which could introduce unfairness or inaccuracies into the AI models.

Furthermore, implementing robust data management practices is crucial for maintaining data integrity. This involves data cleansing where inconsistencies and inaccuracies in the data are corrected and may involve data integration from different sources to capture the full context. All transformations and aggregations of raw data must be documented to ensure transparency and reproducibility. Data must be managed consistently and in compliance with regulations and organizational requirements. Regular monitoring and assessment of the quality of data is essential to identify and correct issues such as inconsistencies and biases.

Principle 2

Guarantee data privacy and security

Theme: Data privacy and security

This principle mandates privacy and security groundworks for data at rest, in transit, and in use. All the sensitive data must be removed in preprocessing procedures. Access to the data in multiparty computations must respect the least privileged approach. Data at rest and in transit must be encrypted and all the data processing must happen in controlled and isolated environments to reduce potential attack frames.

The AI system must be resilient to attacks targeted at the model, data, or the system itself. This principle also demands all the development tooling (such as libraries) and deployment platforms to pass security audits. Users must be in control of their sensitive data and the system must be able to permanently remove personal if required by data owners. Trained models should be as isolated as possible and all the interactions with the model and the platform monitored.

Principle 3

Achieve robust ai model performance

Theme: Model performance and robustness

This principle mandates the provision of high accuracy in the AI model's predictions, emphasizing that the model must generalize well to unseen data for robustness and optimal performance. The principle underscores the importance of the model's ability to adapt, ensuring that it maintains its level of performance under various circumstances. It advocates for scalability, asserting that AI models should handle varying data volumes, devices, and services effectively.

Additionally, the principle demands that AI models must be resilient against adversarial ML attacks, including evasion and poisoning, through rigorous testing, threat modelling, and security monitoring processes. It promotes extensive testing using well-defined performance metrics to ensure consistency in diverse environments. Lastly, the principle highlights the necessity for AI models to provide objective evidence of fulfilling specific intended use, ensuring reliability and accountability within a given time interval and under defined conditions.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

Principle 4 Eliminate bias, uphold fairness

Theme: Ethical and fair AI practices

The Fairness and Bias Mitigation principle emphasizes the importance of ensuring that AI systems are designed and trained to be fair and unbiased. Bias can emerge in AI systems due to biased training data or flawed algorithms, leading to discriminatory outcomes. Addressing this principle involves identifying and mitigating biases related to race, gender, ethnicity, religion, sexual orientation, and other protected characteristics. To implement ethical and fair AI practices, developers adopt a multifaceted approach.

Firstly, they prioritize diverse and inclusive datasets, actively seeking perspectives from different demographic groups to diminish biases and promote fair outcomes. Rigorous evaluation methods, including statistical analysis, stakeholder feedback, and third-party audits, are employed to identify and rectify biases within both training data and algorithms. During algorithm development, techniques such as re-sampling, re-weighting, and adversarial training are applied to mitigate biases, with a focus on designing fairness-aware machine learning algorithms that minimize disparate impact across diverse groups. Additionally, the calibration of model predictions, considering uncertainty and avoiding reliance on predefined thresholds, is emphasized to enhance effective communication of results.

By embedding the values proposed by this principle, AI systems yield more equitable outcomes, as their decisions are not influenced by unfair biases, what increases stakeholders' trust on AI systems and contributes to social harmony reducing existing inequalities and promoting inclusivity.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

Principle 5 Mandate transparency and explinability

Theme: Explainability and transparency

This principle mandates the provision of clear and comprehensible explanations for AI decisions and operations, a critical factor in building trust and understanding among all users. The developers must maintain thorough, accessible documentation covering all aspects of AI training, deployment, and operations, ensuring transparency and accountability.The principle demands that AI systems are designed with user-centric interpretability, ensuring that their outputs and decisions are meaningful and relevant in real-world contexts.

This is vital for enhancing user engagement and trust. Furthermore, the inherent testability and verifiability of AI models are obligatory, underpinning their reliability and performance. Essential to this principle is the incorporation of robust mechanisms for human oversight, including intervention and override capabilities.

Principle 6 Optimize for user-centric design

Theme: Harmonizing User Experience through Comprehensive Design

This design principle encapsulates the essence of user accessibility and usability, fostering a harmonious interaction between users and AI-based systems. This involves the incorporation of features such as a dark mode for user interface customization, accommodating diverse cultural and linguistic preferences, and ensuring user-friendly interfaces. The adoption of a microservice architecture contributes to infrastructure resilience, while API specifications facilitate smooth integration with external services. Continuous Integration/ Continuous Deployment (CI/CD) practices play a pivotal role in the deployment of microservices, ensuring that updates are seamlessly rolled out with minimal disruption to users.

SPATIAL project is funded by the European Union’s Horizon 2020
research and innovation programme under grant agreement N° 101021808.

Transparency regarding the health and psychological impact of AI system usage is crucial, and the design should incorporate mechanisms to communicate these aspects to users openly. Privacy measures, when correctly implemented, not only safeguard user data but also contribute to user comfort and trust in the AI system. The requirement for fast response times aligns with the overall goal of good performance, enhancing the system's usability.

Principle 7 Comply with legal and ethical standards

Theme: Legal and regulatory compliance

This principle mandates adherence to prevailing legal frameworks for AI systems, specifically the General Data Protection Regulation and the AI act. This principle encompasses processes, documentation, and safeguards to protect individual rights and mitigate potential risks. The GDPR demands adherence to data protection principles and individuals' rights, such as providing accessible information and responding to data access and removal requests.

The AI Act addresses other aspects of AI systems, encompassing risk management, testing processes, data governance, transparency, and human oversight. This principle mandates comprehensive documentation, including system architecture and accountability details, fostering transparency and traceability. Furthermore, the principle underscores the importance of accountability in AI systems, necessitating clear documentation of the ML solution's purpose, inherent risks, and non-functional requirements. By embedding these legal and regulatory considerations into the design process, this principle establishes a responsible foundation, fostering trust, transparency, and the protection of fundamental human rights in the development and deployment of AI systems.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.

CONCLUSIONS

In this document and the SPATIAL project overall, a significant effort was dedicated to developing 7 design principles and 11 design patterns, all of which were informed by the experience of building the SPATIAL platform and exploring four distinct use cases. Despite what might seem like a limited number of cases and a single platform, the scope and diversity of these use cases provide a substantial foundation for deriving relevant and effective principles and patterns.

The four use cases covered a range of scenarios - from privacy-preserving AI on edge networks to resilient cybersecurity analytics. Each use case presented unique challenges, offering rich insights into various aspects of AI system development. This diversity ensured that the design principles and patterns developed were not theoretical but rather grounded in practical, real-world applications.

The SPATIAL design patterns were tailored to address the unique challenges of AI-centric architectures in security domains. Unlike general software design patterns, these were specifically crafted to guide developers in creating secure, transparent, and accountable AI-driven systems. This focus on AI-specific challenges underscores the relevance and applicability of these patterns to current and future AI developments.

The development of these principles and patterns within the context of the SPATIAL platform and the four use cases demonstrates the project's commitment to creating a robust framework for AI system development. The deep exploration of specific AI challenges ensured that the principles and patterns developed were not only comprehensive but also deeply informed by real-world applications and challenges.

In conclusion, the SPATIAL project's development of design principles and patterns, though based on a specific platform and a limited number of use cases, offers a versatile and practical toolkit. These principles and patterns provide a solid foundation for guiding the development of secure, transparent, and accountable AI systems in various domains. While future expansions to more use cases and platforms could enrich these principles and patterns further, the current framework represents a significant contribution to the field of AI system development.

SPATIAL project is funded by the European Union's Horizon 2020
research and innovation programme under grant agreement N° 101021808.