chist-era

AI Integrated Framework for Intelligent Geospatial Handling and Robust Operation in MultiGIS Applications (Ai4MultiGIS)

Topic: Multidimensional Geographic Information Systems (MultiGIS)

AI4MultiGIS

D2.3 Responsible AI Framework

Work Package	2
Delivery Date	M6
Responsible Partner	ARU and UVT
Authors	Shareeful Islam and Helga Hochbauer
Distribution	*Project level*
Version	1.0

a.r.u. Anglia Ruskin University

West University of Timișoara

LUT University

UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

Deliverable Factsheet

Project Acronym:	Ai4MultiGIS
Project Title:	AI integrated framework for intelligent geospatial handling and robust operation in MultiGIS applications
Call:	Chist-Era 2023
Start date:	02/01/2025
Duration:	36 Months

Deliverable Name:	D2.3
Related WP:	WP2
Due Date:	M6

Editor:	Shareeful Islam
Contributor(s):	Shareeful Islam , Helga Hochbauer
Reviewer(s):	Akram
Approved by:	All partners

Executive Summary

The Ai4MutiGIS project aims to offer an integrated framework for MultiGIS data generation and management, aiming to enhance the overall GIS capabilities for optimized processing chain of MultiGIS applications. This deliverable, D2.3, provides a novel Responsible AI framework to support responsible development and usage of AI enabled GIS applications.

The deliverable outlines the key R-AI characteristics fairness, transparency, explainability, interpretability, ad privacy to holistically consider operationalising R-AI practice for the applications. The deliverable also investigates Requirements Specification for Responsible AI by considering these properties. Finally, the documents outline the conceptual view of R-AI and its adoption for the AI enabled GIS application.

Document History

Version	Date	Author(s)	Comments
0.1	18.03.2025	Hochbauer Helga	Reviewed the existing information, corrected some indentation errors and typos, added comments, started working on Responsible AI Framework (5)
0.2	30.03.2025	Hochbauer Helga	Updated section 4, added new requirements for 4.1
0.3	31.03.2025	Hochbauer Helga	Reviewed and updated requirements
0.4	04.04.2025	Hochbauer Helga	Added Interpretability, updated requirements, Conceptual View
0.5	11.04.2025	Hochbauer Helga	Added 5.2, started working on 5.3
0.6	14.04.2025	Hochbauer Helga	Added text and diagrams for 5.3
0.7	21.04.2025	Hochbauer Helga	Added the explanation for the diagram, the Fairness metrics and Deployment
0.8	22.04.2025	Andrei Ancuta	Reviewed subsection 5.3.4 Added section 6.
0.9	29.04.2025	Andrei Ancuta	Completed section 6 with example of 2 datasets.
0.10	29.04.2025	Helga Hochbauer	Updated the diagrams, reviewed comments and made modifications
0.11	30.04.2025	Helga Hochbauer	Added text at 5 and 4
0.12	04.05.2025	Andrei Ancuta	Redid entire section 6, dataset examples 1 and 2
0.13	22.05.2025	Hochbauer Helga	Preprocessing & Training
0.14	30.05.2025	Marc Francu	Added textual information including examples for AI and MultiGIS in Sections 5.1. and 5.2.1
0.15	06.05.2025	Hochbauer Helga	Added diagrams
0.16	10.06.2025	Hochbauer Helga	Updated diagrams after feedback
0.17	11.06.2025	Hochbauer Helga	Added Introduction and Conclusion
0.18	27.06.2025	Hochbauer Helga	Added diagram for MultiGIS integration, labelled all figures and tables, added the Metamorphic Testing section, added list of tables, figures, acronyms, wrote explanations for each diagram, formatted document
0.19	29.06.2025	Hochbauer Helga	Added diagram descriptions and contributions to other deliverables and metrics
1.0	30.06.2025	Shareeful Islam	Final version after revising all internal reviewer comments

Table of Contents

List of Figures

List of Tables

Acronyms and Abbreviations

RAI = Responsible Artificial Intelligence
RF = Requirements for Fairness
RT = Requirements for Transparency
RP = Requirements for Privacy
RXAI = Requirements for Explainability
RI = Requirements for Interpretability
ESA = European Space Agency
EEA = European Environment Agency
GDPR = General Data Protection Regulation
LPIS = Land Parcel Identification System
EU AI Act = European Union Artificial Intelligence Act
SHAP = SHapley Additive exPlanations
LIME = Local Interpretable Model-agnostic Explanations
QNN = Quantum Neural Network
GNN = Graph Neural Network
GAN = Generative Adversarial Network
Deep RL = Deep Reinforcement Learning
NeRF = Neural Radiance Fields
XAI = Explainable Artificial Intelligence
API = Application Programming Interface
REST = Representational State Transfer
GraphQL = Graph Query Language
DEM = Digital Elevation Model
IoT = Internet of Things
SLC = Single Look Complex (SAR data format)
SAR = Synthetic Aperture Radar
GRD = Ground Range Detected
NDVI = Normalized Difference Vegetation Index
VV = Vertical Transmit and Vertical Receive Polarization
VH = Vertical Transmit and Horizontal Receive Polarization
QA60 = Quality Assessment Band 60 (Sentinel-2)
MT = Metamorphic Testing
MR = Metamorphic Relation
CI = Continuous Integration
IoU = Intersection over Union

1 Introduction

1.1 Purpose and Scope

This document aims to define the **Responsible AI** concepts and the AI framework. It presents the main RAI characteristics and their requirements while exemplifying the implications of **MultiGIS** technologies. The framework is built on top of a strong ethical understanding and the process is split into multiple steps for understandability. At every stage of the process, we outline the implications of RAI concepts involved.

Each characteristic—such as **fairness, transparency, privacy, explainability,** and **interpretability**—is discussed in detail to ensure that AI systems developed under this framework are accountable and aligned with societal values. The document further contextualizes these principles within practical applications by incorporating dataset-specific examples.

1.2 Contribution to other Deliverables

Deliverable D2.3 establishes the conceptual and practical foundation for embedding Responsible AI (RAI) principles into the AI4MultiGIS project. This deliverable defines a robust RAI framework, detailing essential characteristics such as fairness, transparency, privacy, explainability, and interpretability. By contextualizing these principles through dataset-specific examples and highlighting their implications in the context of MultiGIS technologies, D2.3 guides the ethical implementation of AI methodologies across the entire project.

This framework directly informs the development of trustworthy and accountable data generation and management techniques in WP3. For example, the work on synthetic data generation (T3.2) and automated outlier detection (T3.3) leverages D2.3's insights to ensure data reliability while respecting privacy and fairness. In WP4, D2.3 contributes to the formulation of interpretable strategies for cross-model and multi-modal data integration (T4.2), ensuring that resulting tools are explainable and aligned with societal values.

Moreover, D2.3 is instrumental in shaping the policy development efforts in WP5 (T5.2) by providing a clear ethical structure that guides the implementation of responsible AI strategies. It also lays the groundwork for the framework assessment and refinement tasks (T5.4), ensuring that the final AI4MultiGIS system adheres to RAI principles.

1.3 Structure of the Document

The document is structured in a clear and methodical way, beginning with standard introductory sections such as the **Deliverable Factsheet**, **Executive Summary**, and **Document History**, followed by a comprehensive outline of the topic. It starts with **Introduction**, detailing the purpose and scope, and leads into an **Overview of Responsible AI** and its defining **Characteristics**. The core of the document focuses on the **Requirements Specification for Responsible AI**, broken down into specific dimensions like fairness,

transparency, privacy, explainability, and interpretability. Then, a detailed **Responsible AI Framework** is introduced, covering conceptual views, assumptions, and a multi-stage process including data collection, preprocessing, model training, validation, and monitoring. The final part, **MultiGIS Adoption of RAI**, includes practical dataset applications with two in-depth examples. Each application mirrors the earlier framework, reinforcing consistency. The document concludes with a dedicated **Conclusion** section.

2 Overview of Responsible AI

Responsible Artificial Intelligence (RAI) refers to the design, implementation, and operation of AI systems in a manner that prioritizes safety, ethics, accountability, and transparency. As new technologies are formed, RAI ensures they align with values and rights of humans by safeguarding fairness, privacy, security and societal welfare. Responsible AI is done holistically by integrating risk evaluation, legal compliance, and perpetual monitoring throughout the entire AI lifecycle, from data collection and model development to model deployment. Additionally, RAI emphasizes stakeholder engagement, ethical governance, and regulatory compliance with frameworks such as the EU AI Act, ensuring that AI technologies are of public good while being open and accountable (Ali et al., 2023). Through the incorporation of ongoing learning and adaptability, RAI enables AI systems to evolve responsibly, reduce emerging risks, and guarantee ethical values in a perpetually changing technology landscape.

RAI is deeply intertwined with both ethics and GIS as it deeply relies on data-driven decision-making that impacts individuals, communities and environments. RAI ensures ethical AI is applied in GIS by observing principles of fairness, transparency, privacy, interpretability and explainability to mitigate biases in spatial data analysis and decision making. With regards to ethics, RAI ensures that the application in GIS upholds fairness, transparency, privacy, interpretability and explainability as well as preventing biases on spatial data and decision-making (Saxena, Zhang, & Shahabi, 2024). The application of ethical AI practices in GIS helps to prevent some of the geospatial bias surveillance and environmental injustices by ensuring location intelligence is applied responsibly for the good of society. Moreover, RAI accounts for stakeholder participation and legal rule observance so that GIS can safeguard sensitive geospatial data. Through integrating continuous learning, RAI improves AI systems powered by GIS to respond to new issues in urban planning, climate change, and disaster response, and to make geospatial intelligence accessible and fair. Most importantly RAI shifts the attention towards humans-centred approach to serve the society with trustworthy information whilst observing ethics and building a sustainable society.

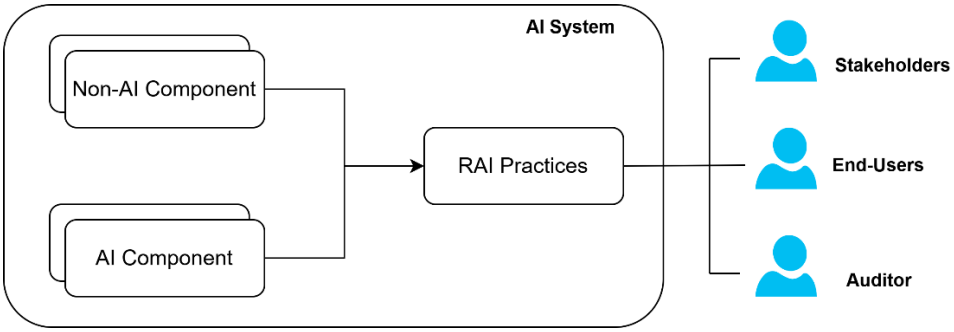

Figure 1: RAI System

Figure 1 illustrates the integration of RAI practices within an AI system composed of both AI and non-AI components. These components collectively contribute to the functioning of the system, and RAI practices are applied to ensure that ethical, legal, and societal considerations are addressed throughout its development and deployment.

3 Responsible AI Characteristics

Building AI enabled system that operationalises RAI practices requires considering all phases of the AI life cycle and addressing the specific needs of each phase. In this context, these needs are primarily technical aspects of AI such as model decision-making process and biasness mitigation that need to be carefully investigated. Integrating RAI throughout the AI lifecycle ensures that the system is fair, accountable, and transparent, fostering trust and reliability in real-world applications. RAI is built upon five key characteristics that guarantee the ethical use of AI systems. The five characteristics are carefully selected to hold the most key principles of Responsible AI without repetition. Instead of an exhaustive checklist that may dilute focus, these characteristics effectively summarize the most salient ethical and technical concerns of AI systems. These characteristics help organize the diverse challenges presented by AI technologies so that resources are directed toward solving social problems rather than making inequities worse or cause harm. They help mitigate risks throughout the AI lifecycle, while ensuring social values are integrated as systems advance.

- Fairness: Fairness in AI refers to the principle that all individuals and groups should receive equal treatment and results from AI systems, with no discrimination or bias. It ensures that AI models do not unfairly advantage or disadvantage groups according to race, gender, age, socio-economic status, or geographical location.
  Fairness is a cornerstone of RAI, as it evaluates the risk an AI system has in trying to reinforce or worsen existing forms of discrimination. The significance of fairness in AI systems is that every person irrespective of their characteristics is treated the same and has similar experiences within the AI-driven systems (Entschew, 2024).

  In practice, fairness involves minimizing bias across the entire AI pipeline from data gathering to model deployment. Achieving fairness begins with data collection and preprocessing which should involve comprehensive bias detection. This entails scanning available data sets for any biases and inequalities that could disproportionately impact the model's outcomes. Through processes such as data balancing, oversampling, and undersampling, models are provided with more representative samples of all groups which helps in correcting these problems.

- Transparency: Transparency in AI is the ability to describe and convey how an AI system functions, the data it uses, decision-making, and the algorithms behind it. It enables AI models to function in a manner that is interpretable and accessible to all stakeholders, including users, developers, and regulators (Orobinskaya et al., 2024). Transparency is essential for trust establishment, ease of

accountability, and ensuring ethical AI use. This is vital for building trust in AI systems as users tend to trust a solution they can follow and track, while also ensuring accountability when there are problems, issues, biases, or errors.

A key element of transparency is accurate documentation of a model. This means clearly explaining how a model is developed, what datasets it is provided with, and what algorithms are available for analysis and creating predictions. The document should indicate the data source, as well as how the data is collected, processed and labelled, and what is done to control potential biases or inaccuracies within the data. Also, it is important to explain what specific training techniques such as model architecture, training strategy, and hyperparameter tuning were implemented. This allows the stakeholders to understand the foundation of the model and determine if it is ethically and socially acceptable.

- Explainability: Explainability in AI refers to the ability to make the decision-making process of an AI system understandable and interpretable by humans. Explainability enables users, developers, and stakeholders to see how inputs influence outputs, resulting in trust, accountability, and evidence-based decision-making. Explainability techniques enable breaking down complex AI models into simpler components that are easier to comprehend, allowing users to validate, debug, and optimize system performance as well as ensure ethical and transparent deployment of AI.

  Explainability techniques assists in discerning the contribution of the different data source to the predictions and analysis which aids in proper decision-making (Diwale, 2025). Some techniques include feature attribution techniques alongside the use of spatial heat maps that provide the relevance of certain data regions associated with the model output. Attention mechanisms and other form of visualizations provide insights into the focus areas of the model during analysis, while counterfactual explanations allow analysis of possible outcomes of certain scenarios.

- Interpretability: In the context of ML systems, we define interpretability as the ability to explain or to present in understandable terms to a human. [ref] It ensures reliability and robustness, confirming consistent performance despite changes. What's more, interpretability is critical when the AI is integrated into decision-making processes where understanding how the system works can directly influence actions. Interpretability enables developers to delve into the model's decision-making process, boosting their confidence in understanding where the model gets its results. (Ali et al., 2023) This allows users to understand both the predictions and their explanations.

  Interpretable explanations should provide a qualitative understanding of the relationship between the input variables and the output, while also considering the user's limitations. (Ribeiro, Singh, & Guestrin, 2016)

- Privacy: AI Privacy refers to the protection of private and sensitive data of individuals from unwanted access, misuse, or disclosure in accordance with ethical, legal, and regulatory standards. It involves having controls to protect personally identifiable information (PII) and prevent re-identification of individuals in AI-based data processing. Privacy-preserving techniques maintain user trust and avert negative impacts such as data breaches, surveillance, and identity theft.

  In AI-driven systems, privacy is of utmost importance because of the immense amount of data collected from multiple sources such as online platforms, IoT sensors, and mobile devices (Silva & Silva-Morales, 2024). The combination of various datasets escalates privacy threats since integration of several data sources might result in inadvertent disclosure of confidential data or enable

unauthorized tracking of persons' movements. In order to enhance privacy in AI systems, a number of techniques could be used. One effective approach data anonymization and aggregation, which pertain to the elimination of personal identifiers and data representation at a level broad enough to mitigate any risk of identification. Other methods like the use of differential privacy, where benign noise is inserted in the datasets to minimize revealing individual information preserving the overall data utility.

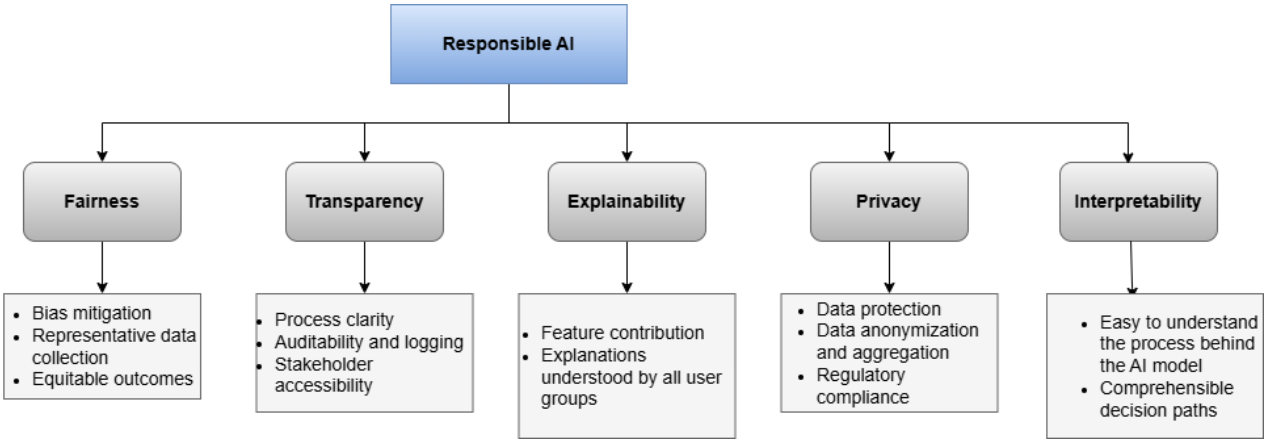

Figure 2: Responsible AI conceptual diagram

4 Requirements Specification for Responsible AI

4.1 Requirements for Fairness (RF)

To ensure fairness in AI models, it is necessary to use data that are inclusive of the population in the real world. Without proper representation, AI models can lead systems into making poor decisions that can harm selected individuals (Entschew, 2024). Representative data collection is key to fairness, as it ensures that all demographic groups are adequately accounted for, reducing the risk of bias and unfair treatment. In cases of inequity, corrective measures like dataset adjustment or refined model training can be applied. On the other hand, the failure to provide adequate data can result in biased AI systems that exacerbate discrimination and worsen social problems.

Beyond data considerations, fairness also relies on active stakeholder involvement and a human-centric design. Engaging diverse communities, domain specialists, and the affected population in the design decisions and deployment process aligns AI systems with real-world needs and averts bias. In the context of geospatial applications, collaboration is not just beneficial but essential, as it requires not only technical knowledge but also an understanding of complex spatial interactions (Marasinghe et al., 2024).

Additionally, AI models must be contextually appropriate, adapting to the specific geographical, social, and cultural characteristics of the domain they operate in.

The key requirements to ensure fairness are given below.

- **RF1:** Data shall be diverse and inclusive for all groups.
- **RF2:** Data shall collect from diverse sources to help cover different demographics, socioeconomic classes, and conditions in case data represents user groups.

- **RF3:** The system will have methods of measuring bias like stratified sampling, reweighting and oversampling to ensure adequate representation of underrepresented groups.
- **RF4:** The system shall assess fairness using established metrics, including demographic parity, equalized odds, fairness through awareness and impact disparity analysis, to measure and address potential biases.
- **RF5:** In case of inequity in data, corrective measures like dataset adjustment or refined model training can be applied.
- **RF6:** The system shall involve diverse stakeholders, including domain specialists, in the design and deployment to ensure the alignment with real-world necessities.
- **RF7:** The system shall adopt a participatory and human-centric approach to mitigate bias, enhance trust, and improve AI decision-making in urban and geospatial contexts.
- **RF8:** The system must ensure contextual appropriateness by adapting AI models to the specific geographical, social, and cultural characteristics of the target domain.

4.2 Requirements for Transparency (RT)

Transparency in AI systems is essential to sustain accountability and trust. To achieve transparency, it is crucial to consider both auditability and logging, ensuring that all AI processes are traceable and verifiable. Auditability along with logging guarantees that every process or activity within an AI system's lifecycle, from data pre-processing to model deployment, is recorded and can be audited when needed (Fernsel, Kalff, & Simbeck, 2024). This allows tracing how decisions are made and enables stakeholders to examine, reproduce, or challenge model predictions when required. Not having an audit trail creates an inability to justify AI decisions which poses risks of bias and errors. Lack of proper documentation may cause problems in the organization's ability to comply with regulations like GDPR and other AI governance policies, which can even damage the reputation.

The key requirements to ensure transparency are given below.

- **RT1:** The system shall maintain a comprehensive audit trail that records all processes, from data pre-processing to model deployment, to ensure traceability and accountability.
- **RT2:** There should also be systematic logging of data changes, model modifications, and inference conclusions to examine and reproduce AI decisions when required.
- **RT3:** There should be version control mechanisms to track changes made to models and datasets.
- **RT4:** Incorporation of explainability tools to clarify how models generate their predictions, supporting interpretability and stakeholder understanding.
- **RT5:** The system shall extract and review system logs periodically to verify that AI models are functioning as expected and to detect potential issues.
- **RT6:** The system shall implement a robust validation process, including continuous impact assessments and real-world testing, to ensure performance reliability.
- **RT8:** Organisational strategies shall be established to address data scarcity and sustainability challenges in AI implementation.

4.3 Requirements for Privacy (RP):

AI technologies frequently analyse vast amounts of sensitive user data, which makes privacy a primary concern. To safeguard privacy, it is essential to consider both data anonymization and aggregation, ensuring that sensitive information remains protected while enabling AI models to learn effectively. Data anonymization and aggregation techniques are important to ensure personally identifiable information (PII) can be protected while still allowing AI models to learn from data patterns (Asthana et al., 2025). Anonymization enables removing or masking identifiable details whereas aggregation combines data points in such a way that makes individual re-identification impossible. AI models without proper anonymization expose sensitive user data that leads to privacy violation and legal lawsuits due to regulations [ref]. In addition to this, data that is not protected poses a risk of being misused, resulting in destruction of trust from users towards AI systems.

The key requirements to ensure privacy are given below.

- **RP1**: Anonymization techniques, such as pseudonymization and generalization, must be implemented to remove or mask personally identifiable information (PII) while preserving data utility.
- **RP2**: Applying data aggregation methods ensures that data points are combined in a way that prevents individual re-identification while maintaining dataset functionality.
- **RP3**: Techniques like differential privacy should be incorporated to minimize the risk of data leakage and unauthorized access.
- **RP4**: Regular security audits must be conducted to assess vulnerabilities, including potential de-anonymization attacks, ensuring data privacy and regulatory compliance.

4.4 Requirements for Explainability (RXAI):

AI models, especially complex like deep learning, tend to operate as “black boxes”, limiting the user’s ability to understand the rationale of a model’s decision. For true explainability, AI systems must provide user-friendly explanations that are clear, actionable, and accessible to stakeholders. The outcomes of any AI model are helpful to stakeholders only if the explanations made available are actionable and usable (Kulaklıoğlu, 2024). This gap should be overcome with adequate interpretations. To produce such explanations, AI decisions need to be explainable to the level that the audience can understand. When users cannot decipher AI suggestions, they are likely to disregard or abuse the technology, which, in turn, decreases the technology’s efficacy. For Explainability we can use a variety of techniques such as: SHAP, LIME, Partial dependence plot, anchors.

The key requirements to ensure explainability are given below.

- **RXAI1**: AI models must provide explanations that allow stakeholders to understand the rationale behind decisions.
- **RXAI2**: Explanations should be meaningful, clear, and applicable, ensuring that stakeholders can use them effectively for decision-making.
- **RXAI3**: The system must provide both local and global explanations to ensure users can understand individual decisions while also gaining insights into the overall model behaviour.

- **RXAI4**: Global explanations should be presented in an interpretable format so that non-technical users can understand and provide meaningful feedback.
- **RXAI5**: AI solutions must be able to provide counterfactual explanations that allow users to explore alternative scenarios. (Oluoch, 2024)

4.5 Requirements for Interpretability

Interpretability is crucial in the context of AI applications because it assures that both the inner working process of the chosen AI model and the results can be understood by the user. Additionally, this assures reliability by helping to achieve the consistency of the results. Therefore, Interpretability represents the ability to understand the model's mechanics, or how it makes predictions based on its inputs. Technique like Methods such as SHAP, LIME, and attention heatmaps should be applied to highlight input features that significantly impact AI decisions. In comparing with explainability, interpretability mainly focuses on the degree to which a human can understand the internal mechanics of a model. It is about the transparency and simplicity of the model itself, allowing users to comprehend how input features directly influence the output. For example, a linear regression or decision tree model is considered interpretable because its structure and parameters can be directly examined and understood.

The key requirements to ensure interpretability are given below.

- **RI1:** Provide a clear and detailed explanation of the model's internal workings, including how it processes inputs and generates outputs.
- **RI2:** Clearly specify all model parameters and architectural configurations, such as learning rates, number of layers, activation functions, and any regularization techniques used.
- **RI3:** Document any fine-tuning or optimization strategies applied, including the datasets used, evaluation metrics, and the rationale behind chosen modifications.
- **RI4**: The system shall log key decision-making steps and intermediate results in a traceable manner, enabling post-hoc auditing and accountability.

5 Responsible AI Framework

5.1 Conceptual View

The aim of Responsible AI is to lessen biases, advance equality, correctness, and streamline the interpretation and justification of findings (Kuanr, Rout, & Mohapatra, 2025). As previously mentioned, the characteristics of Responsible AI include fairness, transparency, privacy, explainability, interpretability, ensuring AI systems are ethical, trustworthy, and considerate of their societal impact. The need for responsible AI has arisen due to a significant lack of understanding regarding the major challenges and risks associated with the deployment and use of these advanced technologies. As AI systems continue to evolve and become more integrated into various aspects of society, it has become increasingly clear that without proper oversight and consideration of ethical, social, and technical implications, the potential for unintended consequences and harmful effects grows substantially. RAI is aligned with the idea of AI developing ethical principles and human values to reduce biases, simplify and explain results, and ensure security and resilience.

The interaction between human and AI plays a crucial role in determining the effects the growth of AI has on society. This relationship is complex and constantly evolving, therefore when building an AI tool, we must consider the ethical and moral issues that may arise. To prevent unintended consequences, we must assure a human-centered orientation while developing AI systems. This involves prioritizing the characteristics of RAI to ensure that the developed technologies are aligned with societal values.

“Grey box” AI

When we talk about a "black box" we typically mean a process that cannot be modified and is based on unknown knowledge. Black box AI fits this broad definition in some cases by being inaccessible for inspections and analysis. What's more, opaque training data and unclear algorithms can lead to such a "black box". This kind of AI model does not provide any insight on how they arrived at a certain conclusion making it impossible for us to understand the logic behind the model. Therefore, the black box AI lacks explainability and transparency and should be avoided in critical applications.

In contrast to the aforementioned "black box" AI, there is also "white box" AI, which is designed in such a way that its internal logic can be openly examined and understood. While this model offers greater explainability, it often suffers from lower accuracy. From a trustworthiness standpoint, the "white box" model is more desirable; however, it typically does not deliver the desired outcomes due to its reduced accuracy.

The final solution would be a “grey box” AI. Standing between the two extremes, it can provide valuable results while still being partially analysable. This approach requires a balance between interpretability and predictive power, making it suitable for applications where both accuracy and some level of explainability are required. Moreover, it allows for controlled adjustments and improvements, ensuring that models remain adaptable to new data without losing their transparency. Within this broader category, XAI techniques can be used to enhance interpretability by either designing models to be inherently more explainable or by applying post-hoc methods to improve transparency. (Ali et al., 2023)

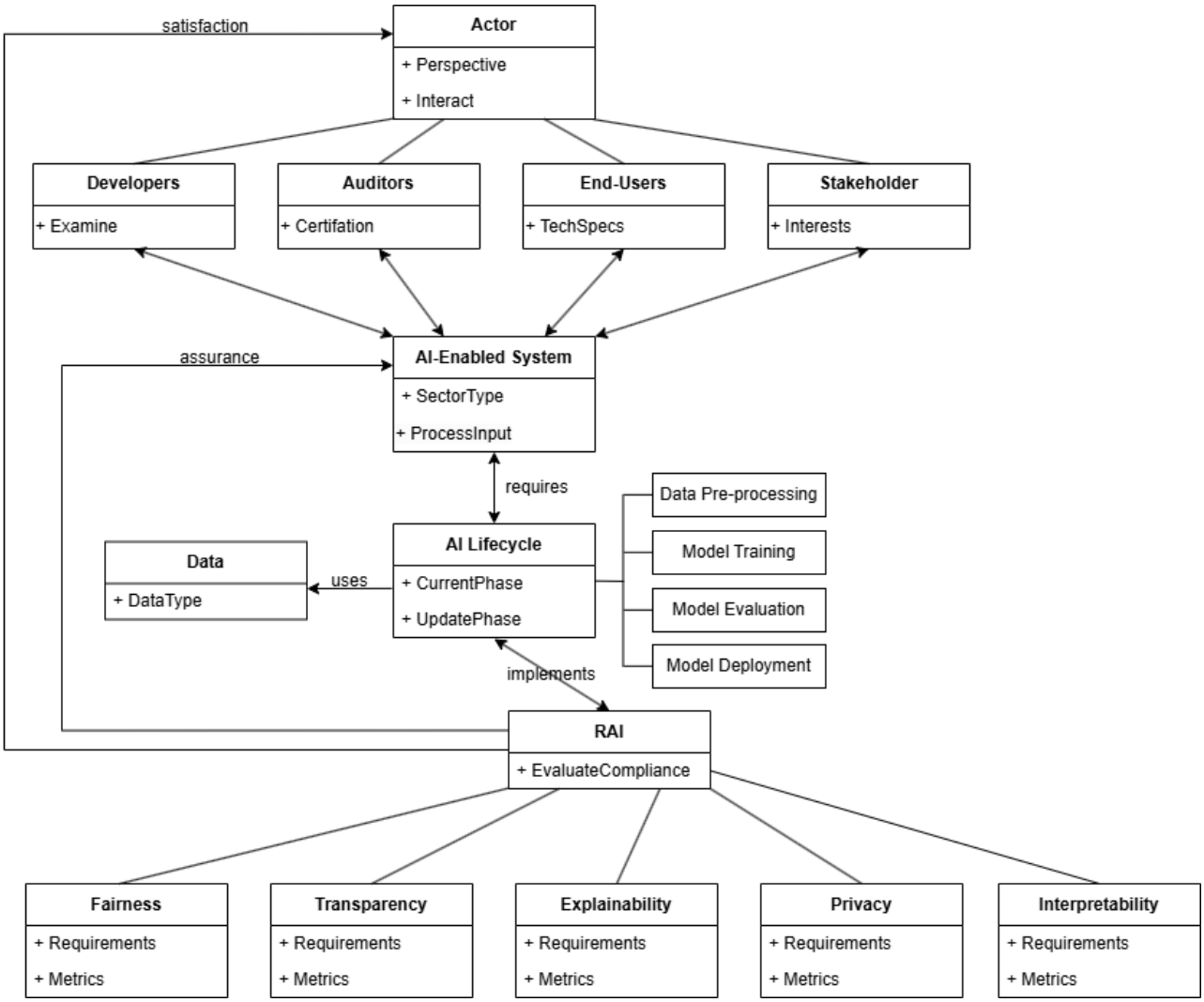

Figure 3: Responsible AI Governance Framework: Interactions Between Actors, Lifecycle, and Ethical Principles

The **actor** represents the individuals or groups that interact with the system. **I**n the above diagram we have four roles that these actors can take when using the AI system. Each of them has different interests and perspectives that must be considered when modelling the system.

- Developers – Examine the system. In the context of GIS the developers typically include a multidisciplinary team with professionals from all covered areas such as: AI/ML engineers (for model design and explainability), GIS specialists (for spatial data accuracy), data engineers (for real-time data pipelines), and specialists from the specific domain of the system.
- Auditors – Provide certification and compliance checks. Auditors are typically independent or semi-independent actors tasked with reviewing the system's integrity, fairness, and compliance.
- End-Users – Use the system and rely on its technical performance. The users may vary, from NGOs and researchers to residents and conservation agencies. The tools may be limited for the public while allowing greater access to advanced analytics, data exports, and species modelling features for scientists, environmental authorities, and policy makers.

- Stakeholders – Have business, ethical, or social interests in the system. In a GIS context, the key stakeholders typically include government agencies, NGOs, academic and research institutions, AI/ML developers, policy makers, local communities, and private sector partners.

AI-enabled System is the core functional system that processes the input received and works with a specific sector type. Within GIS, AI can be used to perform clustering, classification, segmentation and regression tasks on multispectral imagery. Combined with other data (i.e., MultiGIS) such as sensor data (e.g., numeric, image/video), textual information (e.g., reports, articles) or imaging sources (e.g., digital elevation models, LiDAR, and aerial, architectural designs and plans) AI models can be enriched and tailored for specific case studies (e.g., architectural designs, digital elevation models, multispectral imagery for urban planning; research articles, sensor data – images and numeric, digital elevation models for invasive crayfish spread modelling; LiDAR and multispectral images for heritage sites).

Data is a foundational component of the AI system, and it is used by the AI Lifecycle to train, evaluate and deploy the models. The data can be classified by a specific type, and it is crucial in the assurance process, ensuring the model behaves correctly.

The **AI Lifecycle** is split into four stages: Data Preprocessing, Model Training, Evaluation and Deployment.

RAI is based on four key pillars: Fairness, Transparency, Explainability and Privacy. Each of these pillars contains requirements and metrics to ensure responsible behaviour from the AI system. RAI ensures that the AI Lifecycle follows ethical and responsible guidelines, evaluating compliance in areas like fairness and privacy.

5.2 Assumptions

5.2.1 Limitations

Although the available sets of RAI characteristics provide us with a strong ethical base for our desired tool, we must recognize the limitations and issues that may appear.

A first example of such a limitation would be the relation between fairness and privacy. The implementation of privacy measures has proved to affect the model's accuracy, especially for minority groups. This could lead to the reduction of fairness because the model's predictions for underrepresented groups may become even worse. (Sanderson, Douglas, & Lu, 2024)

On the same note we must consider the relation between the Transparency and Privacy concepts because the increase of transparency can negatively affect privacy. Revealing the data about the system such as neural network architecture and training procedure and datasets can facilitate the potential deanonymization of individuals. In addition, there is a great threat of targeted adversarial attacks to cause malfunctions. (Sanderson, Douglas, & Lu, 2024)

Another significant issue is bias, which can affect the final results, and the trustability of the tools implemented. To ensure the effectiveness and positive impact of AI technologies, it's crucial to minimize bias as much as possible. Additionally, to avoid bias we must be able to recognize the different types of bias.

Bias can be present both in the algorithm of the AI system and in the data used to train and test it. Although bias can never be fully eliminated, we should make sure to minimize algorithmic bias through ongoing research and responsible data collection representative of a diverse population. (IBM, 2022)

The main types of bias and their characteristics are presented below.

- Availability bias – refers to the overestimation of events with greater recency in memory. Base Rate Fallacy represents the tendency to focus on specific information (specific cases) rather than on general information. Within GIS this can refer to image resolution availability (some areas on the globe have lower satellite resolution than others) or training models for specific areas with data gathered from other regions (due to lack of available data for the targeted area). For MultiGIS it can extend to sensor and text data as well, e.g., train a model for urban planning in Romania with data from the UK; predict invasive crayfish spread between disconnected rivers by using a spread model along a river.
- Algorithmic bias appears when systematic errors in machine learning algorithms produce unfair outcomes. This can be caused by biases in training data, algorithm design, proxy data and evaluation. When algorithmic bias goes unaddressed, it can perpetuate discrimination and inequality, create legal and reputational damage and erode trust.(Jonker & Rogers, 2023) For instance, this can mean training models with few training examples due to lack of frequency in real life (e.g., natural imbalances such as the types of geographical or anthropomorphically features in an image; or artificial imbalances caused by data gathering techniques such as number of crayfish observations across European countries). Within MultiGIS algorithmic bias extends beyond images and applies to numeric and text information as well.
- Representation bias occurs when the data used to train a model does not adequately reflect the diversity of a certain real-word situation.

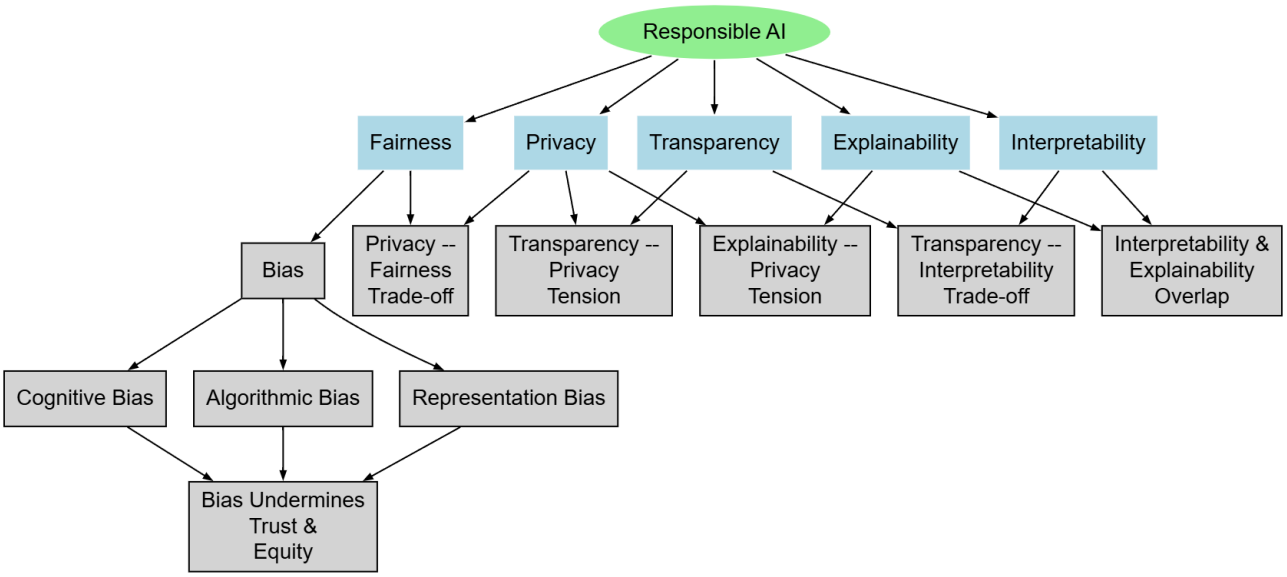

Figure 4: Responsible AI Limitations

Figure 4 presents a conceptual framework for Responsible AI, emphasizing five core principles: Fairness, Privacy, Transparency, Explainability, and Interpretability. It illustrates how these principles are interdependent and often in tension—for example, enhancing transparency may compromise privacy, and efforts to ensure interpretability may reduce model complexity or performance.

At the center of these tensions lies bias, which emerges in cognitive, algorithmic, and representational forms. These biases, if unaddressed, undermine trust and equity in AI systems. The diagram underscores the need for balanced trade-offs and integrated design to ensure that AI systems align with ethical and societal values.

5.2.2 Potential

While balancing the concepts of responsible AI can present notable challenges, it also opens new horizons for innovations in AI design. With the growth of societal emphasis on data protection and trustworthy resources, there is a significant potential in developing models that are both privacy-preserving and fairness-aware. Advances in adaptive noise mechanisms, fairness-constrained learning algorithms, and representation learning can help mitigate bias amplification even under differential privacy constraints.

5.3 Process

This section defines the proposed approach while emphasizing the concepts of RAI implemented. The process is separated on five stages: Data collecting, Preprocessing, Model training, Testing and Validation, and Post-deployment Monitoring and Auditing. Each of these stages will be thoroughly explained and analysed in detail, highlighting how Responsible AI principles are integrated throughout the entire pipeline.

Figure 5: Responsible AI Process Steps

5.3.1 Data collecting

Firstly, when developing the data collecting process, we must recognize the types of data collected: Spatial, temporal, textual, image, sensor data; real and synthetic GIS data. The data used in training AI models can directly influence the fairness, transparency, privacy, explainability, interpretability of the model, therefore the data should be collected from trustable sources. In our case the information will be collected from IoT sensors, edge-cloud devices, remote sensing, scraping, satellite imagery. Additionally, we will collaborate with trustable partners such as British Geological Survey, Met Office, Environment Agency, UK Soil Observatory, to assure the quality and correctness of the data. In the process of collecting data one must acknowledge the ethical implications regarding privacy in geolocation data, transparency, fairness and compliance with regulations such as the General Data Protection Regulation (GDPR).

5.3.2 Preprocessing

At this stage, the collected data must be “prepared” for the training and further processing. This involves cleaning the data with methods based on automated outlier detection, uncertainty quantification, human-in-the-loop feedback for refinement. These methods help ensure the data is accurate and consistent, minimizing anomalies and potential sources of confusion during model development.

Next, we shall focus on the labelling phase. This often includes the use of human-in-the-loop for uncertainty annotation. Human annotators shall assist in labelling data particularly those flagged by uncertainty estimation methods. In addition, human feedback should be incorporated to iteratively enhance labelling accuracy and model performance over time.

During preprocessing, it is crucial to address representation bias, which occurs when the data utilized for training a model fails to properly represent the diversity found in a specific real-world scenario. It can be caused by certain groups, scenarios, or features are overrepresented or underrepresented in the dataset. To mitigate the representation bias we may use techniques such as data auditing, balanced sampling and human-in-the-loop validation.

Data auditing refers to analysing the dataset distribution to identify imbalances. We shall focus on the key variables such as sensor types, demographics or classes, depending on our use case.

As previously mentioned, it is crucial for RAI to include human-in-the-loop techniques, to ensure the RAI characteristics are fulfilled. In this case we shall involve human experts in reviewing flagged data, especially from underrepresented categories, to improve label quality and reduce hidden biases.

Additionally, we shall apply stratified sampling or data augmentation techniques to ensure more equal representation of underrepresented classes. In this context we can integrate AI models to generate synthetic data for the less represented individuals.

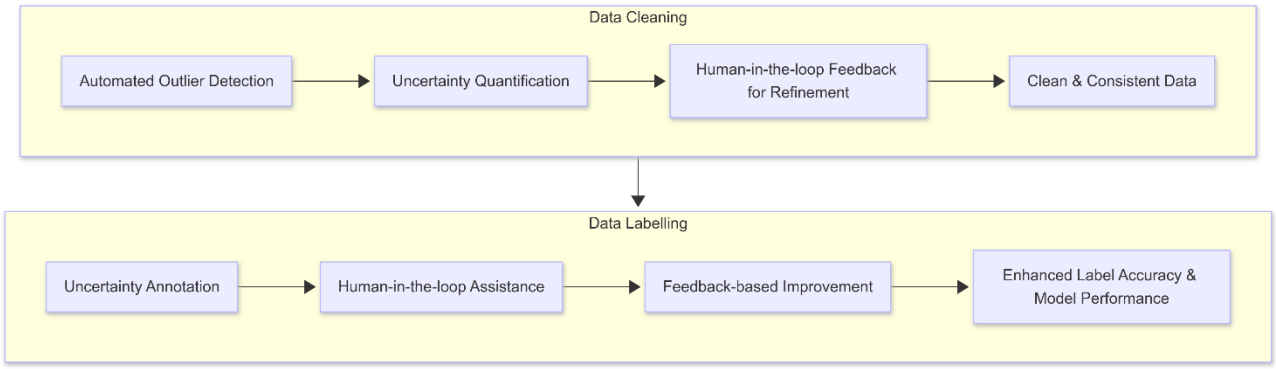

Figure 6: Responsible Preprocessing: Data Cleaning and Labelling overview

5.3.3 Fair and Transparent Model Training

In this stage of our process we may introduce a variety of advanced algorithms, including Quantum Neural Networks (QNNs), Graph Neural Networks (GNNs), Generative Adversarial Networks (GANs), Deep Reinforcement Learning (Deep RL), and Neural Radiance Fields (NeRFs). A Responsible AI framework is employed to address fairness in these algorithms, which actively incorporates principles of equity and accountability. Furthermore, spatial issues such as autocorrelation and the Modifiable Areal Unit Problem (MAUP) are also recognized and tackled successfully so that these do not invalidate the performance of the model.

To ensure geographic generalization and mitigate the risk of overfitting to specific regions, several strategies are implemented. Spatial cross-validation techniques are used alongside ensemble methods to enhance the robustness of the models. Federated learning further contributes to this goal by allowing models to learn from decentralized data, reducing overfitting while maintaining generalization capabilities.

A key concern during model training is the risk of algorithmic bias—systematic errors that unfairly disadvantage certain groups or regions. To mitigate this, fairness-aware learning techniques such as reweighting, adversarial debiasing, and fairness constraints are applied. Additionally, bias detection tools are integrated into the pipeline to monitor model outputs across different demographic and geographic subgroups, ensuring more equitable performance throughout the system.

Finally, to address bias migration, techniques such as stratified sampling across different geographic zones are employed. Data augmentation techniques are particularly aimed at areas with low sample sizes, enhancing the representativeness of the training data. Human-in-the-loop feedback offers continuous enhancement, while techniques such as uncertainty quantification and synthetic data generation are utilized to increase enhance the data for underrepresented areas. This multi-faceted approach is aimed at creating more evened and functional AI systems.

5.3.4 Testing and validation

5.3.4.1 Metrics

In this section, we focus on the application of the previously discussed ethical principles during the testing and validation phase. A critical step is to establish performance metrics for this stage, which include uncertainty reduction, prediction accuracy, and the handling of spatial autocorrelation. Furthermore, the tools developed will undergo testing across diverse datasets to ensure geographic robustness, specifically through two distinct pilot case studies in the UK and Romania.

For verifying spatial consistency, we use the World of Crayfish visualization platform integrated with GIS tools like QGIS and GRASS because it allows for effective visualization of predictions over maps.

Regarding the Fairness, there are multiple metrics that must be taken into consideration during the Testing and Validation stage:

- **Spatial Fairness-** refers to the concept of ensuring that algorithmic decisions or data-driven outcomes do not unfairly favor or disadvantage individuals or groups based on their geographic location. We shall consider the variety of concepts based on spatial fairness such as: Need-Based Spatial Fairness, Distance-based fairness, Zone-based fairness.

- **Need-Based Spatial Fairness:** Ensures resource allocation is proportional to actual need rather than being skewed toward areas with higher data density.
  The lack of data about the less accessible areas can crate bias. For example, if we consider the use case about the population of crayfish, we can observe how data tends to be concentrated in easily accessible areas while remote habitats may have fewer observations. Instead of allocating resources based on data volume, we shall apply a need-based fairness model by considering **habitat**

vulnerability, such as pollution levels or invasive species risk and giving higher priority to **under-sampled but ecologically important areas**.

- **Distance-based fairness** is relevant in any location-aware system where **proximity** plays a key role, such as in nearest-neighbour queries, spatial search, or geospatial resource distribution.
  In this setting, location bias occurs when individuals are impacted by their distance to a reference point. (Shaham, Ghinita, & Shahabi, 2022)

- **Zone-based fairness** is applicable in scenarios where spatial range queries are the norm. Basically, we look at how we shall ensure spatial fairness with respect to coordinate values, instead of distances. This setting is broader, as it can provide fairness with respect to any reference point. (Shaham, Ghinita, & Shahabi, 2022) Some general examples of scenarios where this metric is applicable are loan analysis or insurance pricing.

- **Spatial Disparity Score** - Identifies unequal treatment or prediction accuracy across different geographic zones.

Spatial disparities occur when socio-economic outcomes differ across places. The extent of spatial disparities, and the ranking of different areas, can differ a lot across different outcomes, even those that seem closely related. This illustrates that spatial inequality is inherently multidimensional, as regions exhibiting favourable outcomes in one domain (e.g., employment) may simultaneously experience disadvantages in others (e.g., health or education), highlighting the complexity and heterogeneity of spatial disparities across socio-economic indicators. (What Works Centre, n.d.)

  - **Representation Parity** - Representation parity ensures that geographically remote, sparsely populated, or isolated regions are not left out of data and model results. Without representation parity, models risk overlooking critical information about places that are already neglected due to lack of access, infrastructure, or visibility (Sanderson et al., 2024). This may result in inadequate service provision or biased conclusions about these areas. By using all-encompassing datasets, structured paradigms, and ecological output tracking systems, representation parity seeks to correct data neglect by extending beyond urban, populous, or well-connected areas, ensuring ample samples from all habitats. Thus, designing for representation parity proactively defends against systematic omission, exclusion, and discrimination, thereby creating models that are more relevant and applicable everywhere.
  - **Demographic Parity** - Demographic parity guarantees that all socioeconomic categories are treated fairly and equally and are protected from discrimination in the model or algorithm (Mougan et al., 2023). For instance, the model's outcome—be it approval rating, level of service, or intervention intended to be provided—should not depend on sensitive aspects such as income, educational level, or job type. This fairness measure is critical for models that determine resource allocation, policy issuance, or the provision of health services, education, or employment opportunities. Without such conditions, the model is likely to perpetuate bias, enabling economically advantaged groups further while deepening the gap. Enforcing demographic parity enables the construction of systems that serve people from all socioeconomic strata fairly by averting discrimination and preventing the exacerbation of social and economic inequalities.

- **Equalized Odds** - Equalized odds requires balance in a model's errors by ensuring that the chances of false positives and false negatives remain constant across various demographic groups, such as regions, races, genders, and other subgroups. This criterion is crucial when prediction mistakes could result in misallocating aid, misdiagnosing a patient's condition, or incorrectly distributing resources. For example, a model might tend to predict successful outcomes more often for urban regions while failing more often in rural ones, even when underlying conditions are similar. Such disparities can erode trust, perpetuate disadvantages, and cause measurable harm. Equalized odds ensures that a model's outcomes and risks of error are fairly distributed across the population, including vulnerable groups. Fulfilling the principle of equalized odds enables fairness and responsibility to all segments of society, ensuring accountability for the risk of error.
- **Outcome Disparity** - Monitoring outcome disparity involves tracking whether the effects achieved by a model differ significantly from the predicted effects across different regions or demographic groups of interest. While a model may undergo staged testing and appear fair—such as achieving representation or demographic parity—its interventions may not result in equitable outcomes in actual practice. Such shortfalls may arise from previously disregarded contextual factors, region-specific differences, gaps between expectations and reality, or blind spots within the model itself. Monitoring outcome disparity is essential for exposing systematic biases that may remain hidden during the design phase but emerge during deployment [ref]. With ongoing evaluation of outcome disparity, organizations can identify where models are failing to provide equitable value, adjust their strategies, and modify models to correct disparities between predicted and actual impacts, thereby upholding fairness in practice.

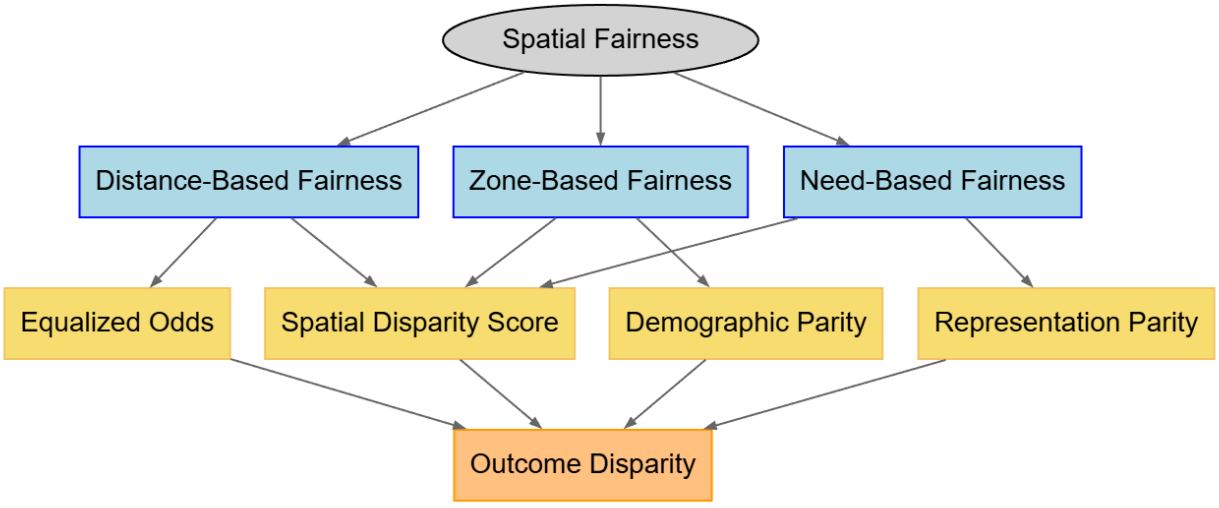

Figure 7: Fairness Metrics for RAI

Additionally, we must consider the other RAI characteristics and their metrics. Transparency is critical for trust and governance. Without measurable transparency, it's hard to prove that the system is auditable or traceable. Using the method of Documentation Completeness Score we can ensure and prove that our AI system is transparent.

Documentation Completeness Score is a qualitative or semi-quantitative assessment of the completeness and accessibility of model documentation, using standardized frameworks such as Model Cards, Datasheets for Datasets, or Factsheets.

Regarding **privacy**, the framework mandates the implementation of **regular and systematic privacy audits** to safeguard sensitive and personally identifiable information throughout the AI lifecycle. By embedding audit checkpoints at key stages (e.g., after data ingestion, model training, and deployment), the system ensures continuous monitoring and rapid response to emerging threats or compliance gaps.

To assess the degree to which an AI system adheres to the principles of explainability and interpretability, specific evaluation metrics must be established that capture the system's ability to provide human-understandable justifications for its outputs. Explainability metrics focus on evaluating how effectively the system conveys the rationale behind its predictions. These include the fidelity of local explanations, which measures how accurately an interpretability method (e.g., SHAP, LIME) reflects the model's actual behaviour in a local decision context, and the coverage of explanation techniques, indicating the proportion of model components equipped with such tools. In contrast, interpretability metrics are primarily concerned with the structural and cognitive transparency of the model itself. Key metrics include the model complexity index—such as the number of parameters or tree depth in decision trees—which serves as a proxy for interpretability, and the feature importance stability score, which measures consistency of feature rankings across different data splits or perturbations.

5.3.4.2 Metamorphic Testing

In this section we will discuss Metamorphic Testing, a software testing technique that addresses the issue of test oracle problems. The test oracle problem arises when determining the expected outcome of a test case is difficult or not possible. MT circumvents this by focusing on the relations between the outputs of multiple executions of the program under test, using altered inputs. These relations are known as Metamorphic Relations (MRs). (Devens, 2023)

To effectively address the software testing problem, MT adopts a distinctive approach that sets it apart from conventional testing strategies. Instead of concentrating solely on individual outputs, MT investigates multiple executions of the program, providing a broader and more comprehensive understanding of its behaviour. It checks whether the inputs and outputs of these various executions satisfy certain metamorphic relations, which are necessary properties of the intended program's functionality. A metamorphic relation transforms existing test cases into new ones. If the program's behaviour across these sources and follow-up test cases violates the metamorphic relation, the program must be faulty. (Segura et al., 2020)

To implement MT in MultiGIS systems, the following methodological integration can be adopted:

1. **Define Metamorphic Relations (MRs)**:

   Based on domain knowledge and physical/geospatial constraints, define MRs such as *"temporal consistency," "spectral invariance,"* or *"spatial coherence."*

2. **Automate Input Transformations**:

    a. Extend the preprocessing pipeline to generate follow-up test cases using MR-defined transformations (e.g., cloud-simulation overlays, synthetic vegetation growth, or elevation shifts).
3. **Compare Output Responses**:
    a. Use spatial metrics (e.g., pixel-level IoU, boundary accuracy) to compare original and transformed outputs.
    b. Evaluate violations of metamorphic relations as indicators of potential faults or ethical inconsistencies (e.g., geographic bias).
4. **Incorporate MT into Continuous Integration (CI)**:
    a. Integrate MT workflows into the model training and deployment pipelines.
    b. Ensure MT checks are embedded in **post-deployment monitoring** to detect model drift in evolving geospatial environments.
5. **Log and Audit Failures for RAI Compliance**:
    a. Document MT failures as part of the **blockchain-based spatial audit system** discussed in the framework.
    b. Use audit data to trigger **human-in-the-loop interventions** or automatic retraining.

Ethical and Practical Benefits:

- **Fairness**: Detects whether the model treats different regions or demographics consistently under slight modifications (e.g., urban vs. rural treatment consistency).
- **Explainability**: Uncovers latent model behaviour patterns, aiding stakeholders in understanding model logic under varying conditions.
- **Reliability**: Enhances model robustness by identifying edge cases and promoting adaptive retraining.

5.3.5 Post-deployment Monitoring & Auditing

During this step of the process monitoring for model drift due to environmental/geopolitical changes is critical, therefore we shall use monitoring techniques based on Digital Twin synchronization, feedback loops, and continual updates from edge devices. The monitoring should be continuous, and the issues found should be solved immediately.

Additionally, this step shall include a human-in-the-loop approach for feedback and corrections. It reinforces model reliability by engaging expert oversight and ensures the adherence of the ethical principles discussed. Stakeholders such as scientists, policymakers, and local communities must be involved in iterative feedback loops to ensure that the model aligns with ethical and social norms.

The AI model is deployed via interface-based mechanisms, primarily REST or GraphQL APIs, to support modular integration, fine-grained access control, and traceability.

While REST APIs are common, they may lead to issues such as over-fetching or under-fetching. AI systems consist of various components that are traditionally, accessed through multiple REST APIs or tightly coupled systems. This can lead to over-fetching or under-fetching data, and scalability issues since REST APIs often force clients to retrieve unnecessary information or make multiple calls for different datasets. Therefore, another variant for this would be GraphQL because of its dynamic querying, real-time updates, and the centralized API management. A single GraphQL schema abstracts multiple backend services, making the platform modular and future-proof. (Singh, 2024)

To offer maximum transparency and traceability in data processing, audit logs are created with a spatial context using a blockchain-based decentralized data system. This approach not only enhances visibility but also guarantees the integrity of the logs, making them auditable and secure. In addition, stakeholder engagement, ethical oversight, and data governance mechanisms should be embedded throughout the lifecycle to align system behaviour with social and environmental values.

Layer	**Recommended Technique**	**Justification**
Deployment Interface	API-based (REST or GraphQL)	Provides structured and secure access to model outputs. Facilitates modular integration with GIS systems and supports logging, version control, and traceability.
Fairness and Feedback	Human-in-the-loop validation	Enhances reliability by incorporating domain expertise. Enables correction of edge cases, ensures contextual relevance, and supports iterative improvement.
Model Monitoring	Digital Twin synchronization or drift detection	Enables continuous assessment of model performance in dynamic environments. Helps detect changes due to evolving spatial, environmental, or social factors.
Audit and Integrity Layer	Blockchain-based spatial audit logging (Optional)	Adds transparency and traceability to AI decision-making. Especially valuable in high-stakes contexts where accountability and data integrity are critical.

Table 1: Post-deployment Monitoring and Auditing

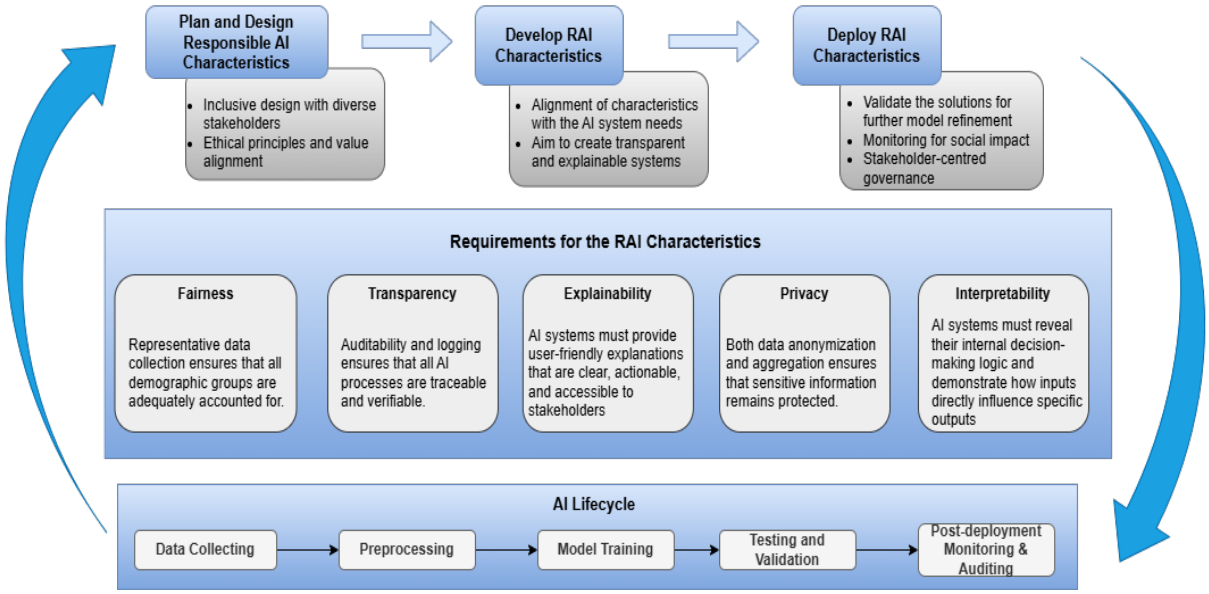

Figure 8: Responsible AI Lifecycle with respect to each specific characteristic

Figure 8 conceptualizes how Responsible AI (RAI) principles must be systematically embedded across all stages of the AI lifecycle, from data collection to post-deployment auditing. It emphasizes that fairness, transparency, explainability, privacy, and interpretability are not isolated values but functional requirements that must guide both technical design and socio-ethical alignment. Each RAI characteristic addresses specific risks—such as bias, opacity, and data misuse—by introducing mechanisms like representative data sampling, audit trails, and interpretable model architectures. The process is iterative and stakeholder-driven, ensuring continuous validation, governance, and impact assessment, especially after deployment. Ultimately, the model advocates for operationalizing ethics in AI development through structured integration rather than post hoc adjustments.

6 MultiGIS Adoption of RAI

6.1 Dataset Types in the Context of MultiGIS

MultiGIS integrates a diverse array of datasets to enhance geospatial analysis, risk assessment, and informed decision-making. These datasets serve as the backbone of various analytical models, enabling organizations to monitor, evaluate, and respond to complex spatial and cyber threats across infrastructure systems. The primary dataset categories include:

- **Spatial Data**: This includes satellite imagery, vector layers (e.g., roads, rivers, land use), digital elevation models, and topographical maps. Such data is crucial for environmental monitoring, urban planning, and identifying physical vulnerabilities in critical infrastructure. For example, satellite data

can be used to assess flood zones or wildfire-prone regions in real-time. (Ugliotti, Daud, & Iacono, 2025)

- **Temporal Data**: Time-series data collected from Internet of Things (IoT) devices, such as environmental sensors or vehicular telemetry, provides insight into how infrastructure or environmental conditions evolve over time. This helps in detecting anomalies or forecasting risk trajectories.

Integrating these heterogeneous data sources allows MultiGIS to build a multidimensional view of risk that spans both physical and digital realms. The synergy between spatial and cybersecurity datasets enables more holistic threat modeling, situational awareness, and resilience planning.

6.2 Datasets for RAI

Aligning with Responsible AI (RAI) principles, MultiGIS emphasizes the ethical and responsible use of data throughout the AI lifecycle. This ensures that AI applications embedded in geospatial and cybersecurity contexts are not only functional but also fair, transparent, and aligned with societal values. Core commitments include:

- **Bias Minimization**: MultiGIS implements algorithmic audits and training data assessments to reduce representational and measurement biases. For example, spatial datasets are checked for urban-rural balance, while cybersecurity logs are reviewed for labeling imbalances.
- **Transparency**: All AI model development stages—from data collection to model tuning—are logged, and metadata is documented. Stakeholders can trace decisions back to specific data sources or model versions, which is essential for explainability and accountability.
- **Explainability**: Techniques like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) are used to make AI decisions interpretable to human operators. These tools help identify which features (e.g., "attack vector = network") contributed most to a model's prediction.

These principles are not merely ethical add-ons but legal necessities in many contexts. The EU Artificial Intelligence Act classifies cybersecurity and critical infrastructure as high-risk areas, requiring strict adherence to transparency, traceability, and human oversight protocols.

6.4 Dataset Example: Sentinel-2 Satellite Imagery for Crop Type Classification

6.4.1 Data Collecting and Definition

What is Sentinel-2?

Sentinel-2 is a satellite mission from the European Space Agency (ESA) that captures high-resolution images of the Earth's surface. The satellite collects data in 13 different spectral bands (including visible, near infrared, and shortwave infrared), which makes it very useful for observing vegetation, water bodies, and land use changes.

What is this dataset used for?
In this example, we'll use Sentinel-2 imagery to classify different types of crops (e.g., wheat, maize, soybean) based on how they reflect light in different bands. This helps in **agriculture monitoring**, **yield estimation**, **land use mapping**, and even **climate change tracking**.

Where can you view or process the data?
You can work with this dataset in several ways:

- **Google Earth Engine (GEE)** — A free online platform that lets you process satellite imagery in the cloud using JavaScript or Python. You don't need to download anything.
- **Copernicus Open Access Hub** — If you want to process the data locally, you can download it from https://www.copernicus.eu/en/access-data/conventional-data-access-hubs

How is the data collected?

- Sentinel-2 satellites capture images using onboard multispectral cameras. The images are taken every 5 days over the same location (depending on cloud cover), with a resolution of up to 10 meters per pixel.
- The raw data goes through calibration and correction to remove errors caused by the atmosphere, sensor limitations, or satellite angle.
- The processed images are then made freely available via the Copernicus programme — a joint initiative of the European Commission and the ESA.

6.4.1 Data Collecting

The CORINE Land Cover dataset is collected through a combination of **satellite imagery analysis** (primarily Sentinel-2, Landsat imagery) and manual validation by national agencies. In the MultiGIS context:

- **Data Source Trustworthiness**: CLC data is sourced from the European Environment Agency (EEA), a reputable and compliant body ensuring adherence to GDPR, privacy regulations, and open-access policies (Copernicus Land Monitoring Service, 2024).
- **Spatial and Temporal Dimensions**: The dataset includes spatial (geographic polygons) and temporal aspects (year of data acquisition) — matching the requirement for diverse data types (spatial and temporal).
- **Ethical Considerations**: The data is anonymized at a macro level (land cover, not individuals), hence inherently respects personal privacy. However, usage must still comply with ethical geospatial data standards, especially when integrating with finer-grained IoT datasets.

6.4.2 Preprocessing

Preprocessing represents a critical step in ensuring that satellite imagery is suitable for use in crop type classification tasks. For Sentinel-2 data, this phase involves several key procedures that aim to improve data quality, minimize noise, and enhance the relevance of inputs for subsequent analysis.

- ***Defining the Area of Interest (AOI)***

The initial stage involves the delineation of the geographical region under investigation. This Area of Interest (AOI) may correspond to agricultural parcels, administrative boundaries, or any spatial extent relevant to the classification task. In platforms such as Google Earth Engine, this can be accomplished either manually or by importing geospatial files (e.g., shapefiles or GeoJSON datasets).

- ***Cloud Filtering and Scene Selection***

Given the susceptibility of optical satellite imagery to atmospheric interference, particularly cloud cover, it is essential to identify and exclude scenes with significant obstructions. This may involve:
Filtering imagery by predefined cloud coverage thresholds (e.g., less than 20%),
Or applying automated cloud masking algorithms, such as those based on the Sentinel-2 **QA60 bitmask** or the **S2_CLOUD_PROBABILITY** product, which facilitate the removal of cloud-contaminated pixels.
In cases where imagery contains persistent cloud cover, alternate acquisitions from temporally adjacent dates may be used to replace compromised scenes.

- ***Temporal Compositing***

To enhance spatial completeness and reduce noise, composite images can be generated by aggregating multiple scenes over a defined temporal window. Techniques such as median or best-pixel compositing are commonly employed to synthesize representative imagery of a particular phenological phase, such as peak greenness during the growing season.

- ***Band Selection***

Sentinel-2 provides data across 13 spectral bands; however, not all are pertinent to vegetation or crop analysis. For classification purposes, a subset of relevant bands is typically selected:
B4 (Red) and **B8 (Near-Infrared)** for vegetation analysis,
B2 (Blue) and **B3 (Green)** for vegetation-soil differentiation,
B11 and B12 (Short-Wave Infrared) for moisture and biomass assessment.
Focusing on these bands reduces computational complexity and enhances model performance.

- ***Radiometric Normalization***

To harmonize reflectance values across multiple acquisitions—accounting for differences in solar angle, atmospheric conditions, or acquisition times—pixel values may be normalized to a standard scale. This ensures consistency across scenes and improves the comparability of input features.

- ***Derivation of Vegetation Indices (Optional)***

Vegetation indices, such as the **Normalized Difference Vegetation Index (NDVI)**, provide robust indicators of vegetative vigor and biomass. These indices can be calculated from spectral bands (e.g., Red and NIR) and appended as additional input layers for classification models.

- ***Preparation of Labeled Training Data***

In supervised classification workflows, it is necessary to assemble labelled datasets corresponding to known crop types. These reference data may be derived from:
-Ground-truth surveys,
-Official agricultural records (e.g., the Land Parcel Identification System – LPIS),
-Manual annotation using very high-resolution imagery.
Labelled data should be representative of the diversity of crops, geographies, and conditions present within the AOI to ensure robust model generalization.

6.4.3 Fair and Transparent Model Training

Following preprocessing, the prepared Sentinel-2 imagery is used to train a supervised classification model capable of identifying crop types based on their spectral and temporal characteristics. In accordance with Responsible AI principles, the training process is designed to ensure fairness, transparency, and applicability across varied geographic contexts.

- ***Model Selection***

For this application, a supervised classification algorithm such as **Random Forest** is well-suited due to its high performance in image classification tasks and its native support within Google Earth Engine. The model utilizes reflectance values from selected Sentinel-2 bands (e.g., red, near-infrared, short-wave infrared) and vegetation indices (e.g., NDVI) as input features to distinguish between different crop types.

- ***Labelled Training Data***

Training requires a collection of labeled examples that correspond to known crop types. These labels may be obtained from official agricultural databases, historical surveys, or manually digitized using high-resolution imagery. It is important that the training data cover a range of crop classes and spatial conditions within the Area of Interest (AOI), ensuring the model is exposed to the full variability present in the landscape.
To mitigate bias, the labeled samples should be well-distributed across the AOI and balanced across classes. For example, if the dataset includes maize, wheat, and soybean fields, each class should be adequately represented to prevent the model from disproportionately favoring the most common class.

- ***Addressing Geographic Diversity***

Spatial diversity in training data contributes to better generalization of the model across different regions and growing conditions. Sampling should include variability in terrain, field size, irrigation practices, and land management to support a robust classification outcome. This geographic spread also helps mitigate the risk of spatial bias in model predictions.

- ***Training Execution and Output***

The labeled training data are used within Google Earth Engine to train the classifier directly on the processed Sentinel-2 imagery. Once trained, the model is applied across the AOI to produce a crop classification map, where each pixel is assigned a crop type based on spectral characteristics.
Model outputs can be visualized within the platform and exported for further analysis or integration into external systems. Additionally, feature importance rankings provided by the Random Forest algorithm offer insights into which spectral bands and indices most influenced the classification results, supporting model interpretability.

6.4.4 Testing and Validation

After the classification model has been trained and applied, the results must be reviewed to assess performance and identify potential areas for improvement. In this context, testing and validation focus on visual inspection and qualitative assessment, as well as an evaluation of fairness and spatial consistency.

- ***Visual Assessment of Classification Output***

The classified map is first reviewed visually within Google Earth Engine by comparing the output against recent Sentinel-2 true-color imagery and the training sample locations. This allows for a basic verification of whether crop boundaries appear reasonable and whether classifications align with known land use patterns.

- ***Comparison with Training Data***

Where available, the classification results can be overlaid with the training data to identify potential misclassifications or inconsistencies. This process supports the early identification of systematic errors (e.g., a particular crop type being misclassified in a specific region).

- ***Geographic Consistency***

To evaluate whether the model performs equally well across the full Area of Interest, attention is paid to variation in output across different geographic zones. This includes examining how the model handles:
Varying field sizes,

Differences in image quality due to cloud cover or terrain,
Regions with lower training data density.
Such checks support a basic audit of spatial fairness, ensuring that no specific areas are consistently misrepresented in the final output.

- ***Interpretation and Feature Relevance***

Google Earth Engine allows users to retrieve the **feature importance rankings** from the Random Forest classifier, providing insight into which spectral bands and vegetation indices were most influential. This contributes to the interpretability of the model and supports transparent decision-making in further refinement.

6.4.5 Post-deployment Monitoring and Auditing

Following deployment, it is important to ensure that the crop classification model remains relevant, accurate, and responsive to new data. In this context, post-deployment monitoring focuses on maintaining alignment between model outputs and real-world agricultural dynamics.

- ***Monitoring with Updated Imagery***

The model can be periodically reapplied to new Sentinel-2 imagery (e.g., monthly or seasonal intervals) to detect changes in crop patterns or confirm ongoing consistency in classifications. This allows for continuous alignment with actual field conditions and supports ongoing use in agricultural monitoring.

- ***Expert Review and Feedback***

Subject matter experts (e.g., agronomists or GIS analysts) may visually review classification results at regular intervals to identify areas of misclassification, anomalies, or unexpected shifts in output. This qualitative feedback can inform model refinement or highlight where retraining may be necessary.

- ***Result Transparency***

Model outputs, including classification maps and feature importance summaries, should be clearly documented and accessible. Where appropriate, summary statistics and known limitations should be communicated alongside outputs to support informed interpretation by end users.

- ***Ethical Considerations***

Even with non-sensitive data, ethical oversight is important. Transparency in how data is used, openness about classification limitations, and responsiveness to stakeholder concerns all contribute to a more responsible and accountable use of AI in agricultural applications

6.5 Dataset Application Example 2: Sentinel-1 SLC - Radar Data for Flood Detection

This section demonstrates the application of the Responsible AI pipeline using Sentinel-1 radar data for flood detection and mapping. The use of Synthetic Aperture Radar (SAR) imagery is particularly suitable in this context due to its ability to penetrate clouds and operate regardless of lighting conditions—making it ideal for monitoring sudden and large-scale environmental events such as river overflows or storm-induced flooding.

6.5.1 Dataset Overview and Data Collection

Sentinel-1
Sentinel-1 is a radar imaging mission operated by the European Space Agency (ESA). It captures Synthetic Aperture Radar (SAR) data in dual polarizations (typically VV and VH), allowing users to detect changes in surface conditions, including water accumulation, with high accuracy.

Use Case
In this experiment, Sentinel-1 Ground Range Detected (GRD) data is used to classify flooded versus non-flooded areas following a flood event. This use case supports disaster response, risk assessment, and resource planning.

Data Access

- The dataset can be accessed and processed in **Google Earth Engine**, where Sentinel-1 imagery is pre-processed (orthorectified and calibrated) and readily available.
- Alternatively, raw data can be downloaded for local processing via the **Copernicus Open Access Hub**: https://www.copernicus.eu/en/access-data/conventional-data-access-hubs

Data Characteristics
Each Sentinel-1 GRD scene provides amplitude (intensity) values in decibels (dB), reflecting surface roughness and moisture. Radar imagery is available approximately every 6–12 days, with a spatial resolution of ~10 meters, making it well-suited for regional flood assessments.

6.5.2 Preprocessing

Preprocessing of radar data is essential for ensuring that the classification model receives consistent and relevant input. The following steps are typically performed:

- ***Defining the Area of Interest (AOI)***

Users define a spatial region of interest, such as a floodplain or administrative boundary affected by a recent storm. This can be drawn manually or loaded using vector files.

- ***Selection of Pre- and Post-Event Images***

Two images are selected:

- A **pre-flood** image, ideally from a few weeks before the event.
- A **post-flood** image captured immediately after the suspected flood event.

These allow for comparative analysis to detect areas where water presence has changed.

- ***Speckle Noise Reduction***

SAR images often contain speckle noise, which can obscure patterns. A **temporal multi-image mean** or a **boxcar filter** can be used to smooth the radar backscatter and highlight surface changes.

- ***Change Detection***

A difference image is computed by subtracting the pre-flood from the post-flood image. Pixels with significantly reduced backscatter values—typical of open water—are flagged as potential flooded areas.

- ***Thresholding and Masking***

A classification threshold is applied to the difference image or VV/VH ratio to isolate flooded zones. Terrain masking (using a Digital Elevation Model) may also be applied to exclude steep areas where radar shadows are common.

6.5.3 Fair and Transparent Model Training

In this case, the model aims to distinguish **flooded** from **non-flooded** areas based on radar backscatter characteristics. The training process follows these principles:

- ***Model Type***

A binary classification model such as **Random Forest** is trained using labeled points or polygons derived from known flooded and dry areas. In Google Earth Engine, this can be implemented directly using pre-processed Sentinel-1 inputs.

- ***Labelled Training Data***

Training data may be obtained from government-released flood maps, field surveys, or visual interpretation of radar images and historical flood patterns. It is important that the samples represent both urban and rural areas to ensure geographic balance.

- ***Bias Mitigation***

To prevent bias, the model should be trained with samples drawn from different terrain types (lowlands, riverbanks, urban settings), ensuring equal performance across diverse geographic zones.

- ***Transparency***

Feature importance analysis helps identify which radar characteristics (e.g., VV backscatter, change detection values) are most influential in classification. This improves transparency and facilitates human oversight.

6.5.4 Testing and Validation

After classification, the results are evaluated through qualitative and comparative means:

- ***Visual Comparison***

The classified flood map is compared to the post-flood radar image to confirm that major water bodies and inundated zones have been correctly identified.

- ***Overlay with Known Maps***

Where available, outputs are compared with official flood extent maps or open datasets (e.g., Copernicus Emergency Management Service) to assess reliability.

- ***Geographic Consistency***

Outputs are examined across different subregions to ensure consistent classification quality. Particular attention is given to rural and low-income areas, where lower data density could affect accuracy.

6.5.5 Post-deployment Monitoring and Auditing

Once the flood classification model has been deployed, it can be operationalized in an ongoing manner by applying it to newly acquired Sentinel-1 radar imagery during subsequent flood events. This enables dynamic monitoring of flood-prone areas by leveraging the satellite's consistent temporal coverage and all-weather imaging capabilities.

- ***Ongoing Monitoring***

The classification script can be reused or scheduled to run after each significant rainfall event, automatically detecting changes in flood extent over time.

- ***Expert Feedback***

  Environmental analysts play a critical role in the post-deployment phase by systematically reviewing model outputs to ensure ecological validity and alignment with ground realities. Through expert interpretation, they can identify areas where the model may have misclassified environmental features or failed to capture key patterns. Their domain-specific insights enable the detection of nuanced inaccuracies that automated systems might overlook. By providing targeted feedback, analysts contribute to iterative model refinement, enhancing both predictive accuracy and contextual reliability in real-world environmental applications.

- ***Ethical Oversight***

Outputs should be shared with relevant stakeholders, including disaster response teams and local communities, with clear documentation on model limitations and confidence levels. Transparency in reporting helps maintain accountability in high-stakes applications such as flood detection.

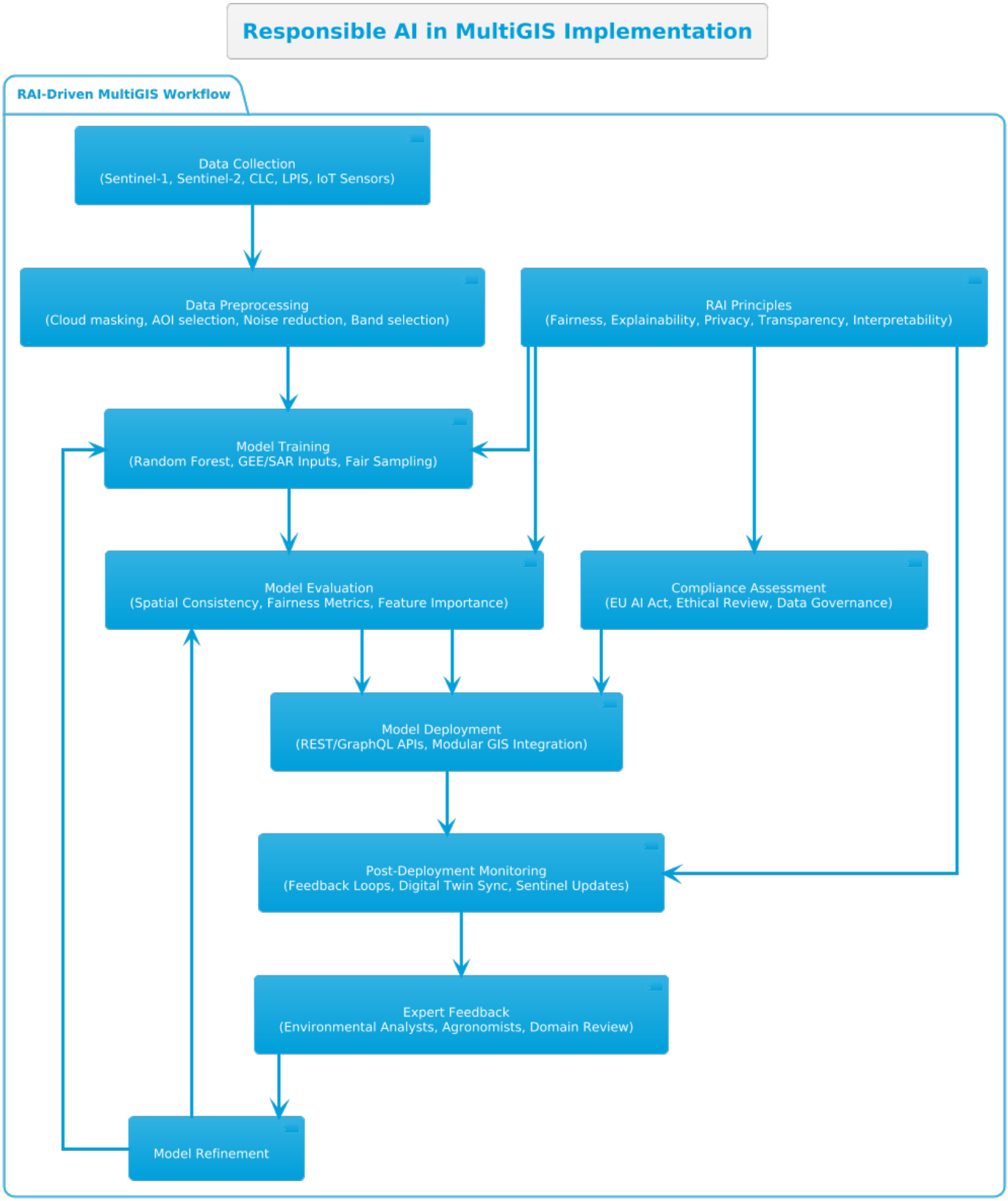

Figure 9: Responsible AI in MultiGIS

7 Conclusion

In conclusion, it is imperative to embrace Responsible AI (RAI) principles as we construct our framework. The call for responsible AI has emerged from a notable gap in our understanding of the significant challenges and risks associated with deploying and utilizing these advanced technologies. As artificial intelligence programs continue to advance and more firmly intertwine themselves in the social fabric, it is more obvious than ever before that, if not watched with careful attention and careful consideration of the technical, social, and ethical ramifications, the potential for unwanted side effects and adverse results grows exponentially. RAI embodies the vision of creating AI that not only reflects ethical principles and human values but also actively works to minimize biases, clarify outcomes, and bolster security and resilience.

To prevent unintended outcomes, we must maintain a steadfast human-centered focus in the development of AI systems. This involves prioritizing the characteristics of RAI to ensure that the developed technologies are aligned with societal values.

The integration of ethical AI practices in Geographic Information Systems (GIS) plays a crucial role in addressing issues of geospatial bias, surveillance, and environmental injustices, ensuring that location intelligence is applied responsibly for the greater good. Furthermore, RAI emphasizes the active participation of stakeholders and adherence to legal standards, enabling GIS to protect sensitive geospatial data with integrity. By fostering a culture of continuous learning, RAI empowers AI systems driven by GIS to adeptly respond to emerging challenges in urban planning, climate change, and disaster response, making geospatial intelligence both accessible and equitable. Most importantly, RAI shifts our focus toward a human-centered approach that strives to provide society with reliable and trustworthy information, all while championing ethical practices and nurturing a sustainable future.

In this document we have described the notions of Responsible AI as well as the AI framework. It states the major RAI features and specifications and gives examples of the implications of MultiGIS technologies. Additionally, the framework is constructed upon a solid ethical knowledge and the process is segmented into a few steps for understandability. In each step of the process, we outline the implications of RAI concepts engaged.

References

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What

we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Asthana, S., Mahindru, R., Zhang, B., & Sanz, J. (2025). Adaptive PII mitigation framework for large language models. arXiv. https://doi.org/10.48550/arXiv.2501.12465

Bhouri, H. (2025). Navigating Data Governance: A Critical Analysis of European Regulatory Framework for Artificial Intelligence. In IntechOpen eBooks. https://doi.org/10.5772/intechopen.114342

Diwale, V. (2025, January 15). ML and AI Model Explainability and Interpretability. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2025/01/explainability-and-interpretability/

Devens. (2023, May 15). Metamorphic Testing: A new horizon in software testing. Medium. https://medium.com/@mailtodevens/metamorphic-testing-a-new-horizon-in-software-testing-6fdec595dba8

Entschew, E. M. (2024). Discriminatory data yields discriminatory systems: When AI biases harm human beings. In Edward Elgar Publishing eBooks (pp. 37–65). https://doi.org/10.4337/9781803928241.00009

Fairness in prediction and allocation. (2023). In Cambridge University Press eBooks (pp. 676–693). https://doi.org/10.1017/9781108937535.032

Fernsel, L., Kalff, Y., & Simbeck, K. (2024). Assessing the Auditability of AI-Integrating Systems: A Framework and Learning Analytics Case Study. arXiv. https://doi.org/10.48550/arXiv.2411.08906

IBM. (2022). Everyday ethics for artificial intelligence. IBM. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

Jonker, A., & Rogers, J. (2023). What is algorithmic bias? IBM Think. https://www.ibm.com/think/topics/algorithmic-bias

Kulaklıoğlu, D. (2024). Explainable AI: Enhancing interpretability of machine learning models. Human Computer Interaction, 8(1), 91. https://doi.org/10.62802/z3pde490

Kuanr, M., Rout, C., & Mohapatra, P. (2025). Navigating Responsible AI: A Case Study with Recommender System, 18, 867–876. https://doi.org/10.5281/ZENODO.13988088

Marasinghe, R., Yigitcanlar, T., Mayere, S., et al. (2024). Towards Responsible Urban Geospatial AI: Insights From the White and Grey Literatures. Journal of Geovisualization and Spatial Analysis, 8, 24. https://doi.org/10.1007/s41651-024-00184-2

Mougan, C., State, L., Ferrara, A., Ruggieri, S., & Staab, S. (2023). Beyond demographic parity: Redefining equal treatment. arXiv. https://doi.org/10.48550/arXiv.2303.08040

Orobinskaya, V. N., Mishina, T. N., Mazurenko, A. P., & Mishin, V. V. (2024). Problems of Interpretability and Transparency of Decisions Made by AI. In 2024 6th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA) (pp. 667–671). https://doi.org/10.1109/SUMMA64428.2024.10803745

Oluoch, I. (2024). Crossing Boundaries: The Ethics of AI and Geographic Information Technologies. ISPRS International Journal of Geo-Information, 13(3), 87. https://doi.org/10.3390/ijgi13030087

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier (Version 3). arXiv. https://doi.org/10.48550/arXiv.1602.04938

Sanderson, C., Douglas, D., & Lu, Q. (2024). Implementing responsible AI: Tensions and trade-offs between ethics aspects (Version 4). arXiv. https://doi.org/10.48550/arXiv.2304.08275

Sanderson, R., Franklin, R., MacKinnon, D., & Matthews, J. (2024). Left out and invisible?: Exploring social media representation of 'left behind places'. GeoJournal, 89(1). https://doi.org/10.1007/s10708-024-11010-z

Saxena, N. A., Zhang, W., & Shahabi, C. (2024). Spatial Fairness: The Case for Its Importance, Limitations of Existing Work, and Guidelines for Future Research. arXiv. https://doi.org/10.48550/arXiv.2403.14040

Segura, S., Towey, D., Zhou, Z. Q., & Chen, T. Y. (2020). Metamorphic Testing: Testing the Untestable. IEEE Software, 37(3), 46–53. https://doi.org/10.1109/MS.2018.2875968

Shaham, S., Ghinita, G., & Shahabi, C. (2022). Models and mechanisms for spatial data fairness (arXiv Version 2). arXiv. https://doi.org/10.48550/arXiv.2204.01880

Silva, M. J., & Silva-Morales, M. (2024, November 19). Geospatial Data & AI: Ethics, Governance & Decision-Making. Ecolonical LAB. https://ecolonical.org/geospatial-data-ai-ethics-governance/

Singh, A. (2024, December 4). Transforming generative AI platforms with GraphQL: The latest in scalable, modular, and robust architecture. Medium. https://medium.com/@anand94523/transforming-generative-ai-platforms-with-graphql-the-latest-in-scalable-modular-and-robust-a69ab8278c16

What Works Centre for Local Economic Growth. (n.d.). Understanding spatial disparities. https://whatworksgrowth.org/insights/understanding-spatial-disparities/

Ugliotti, F. M., Daud, M., & Iacono, E. (2025). Spatial insights for building resilience: The Territorial Risk Management & Analysis Across Scale framework for bridging scales in multi-hazard assessment. Smart Cities, 8(1), 27. https://doi.org/10.3390/smartcities8010027