Risk Assessment Framework

For SAIGE

Draft V 1.0

List of Abbreviations:

SAIGE – Serbia Accelerating Innovation and Growth Entrepreneurship Project.
RAF – Risk Assessment Framework, this document.
AI – Artificial Intelligence.
LA – Loan Agreement.
POM – Project Operations Manual.
ESMF – Environmental and Social Management Framework.
IF – Innovation Fund.
SF – Science Fund.
PIU - Project Implementing Unit

Part 1: Introduction

This document includes definitions, explanations, and other information relevant to the assessment of risks related to AI:

1. Risk classification
2. The process;
3. Requirement for trustworthy AI;
4. Identification of risks
5. Evaluation standards and risk matrix;
6. Questionnaire for applicants;
7. Guidelines for the creation and management of the Risk Mitigation Plan;
8. Monitoring AI risks during the lifetime of the project;

For each project, risk must be governed throughout the entire project lifecycle.

This document creates a framework to enable the evaluation by the IF, SF and PIU of AI-related risks in the process of granting financing support for projects and programs under the

SAIGE project.[1] The guiding requirements for this document are outlined in the LA and further supported by the POM and grant manual documents. They require projects for research and projects for enterprise acceleration to be based, among other requirements, on Trustworthy AI (as defined in the LA). An additional requirement is that any AI Research Grant and any AI Matching Grant will implement Trustworthy AI Principles and include additional obligations for activities that implement personal data processing (data protection impact assessment) and high-risk projects (AI risk impact assessment).

The POM provides additional detail on the risk assessment process that needs to be followed, stating that "all Grant beneficiaries undergo a proportionate AI risk assessment process. Procedures for all grant programs under the SAIGE project are based on the Risk Assessment Framework, which is designed to ensure responsible and ethical use of AI throughout the grant project lifecycle." Besides provisions from the LA, the POM refers to the Serbian Ethical Guidelines and UNESCO Recommendation on the Ethics of AI, to provide a robust baseline for identifying, mitigating, and monitoring risks associated with AI systems, and fostering the development of trustworthy, transparent, and accountable AI solutions.

Each AI system funded by a Grant will be assigned a risk classification level (Low, Moderate, or High Risk) based on this process. All Grant beneficiaries will then have to comply with a set of proportionate, tiered responsibilities based on the assigned risk classification level.

Part 2: Risk Classification

All projects, on application, will be sorted into one of the following categories:

1. Excluded Activity
2. High-Risk
3. Moderate Risk
4. Low Risk

Excluded Activity (unacceptable risk)

Excluded Activities for AI include:

1. Any AI activity that includes (the following list is indicative and not exhaustive):
    a. Any activity directly or indirectly related to weapons or weapons systems;
    b. Facial recognition technologies for mass surveillance, including AI systems that create or expand facial recognition databases through the untargeted scraping

[1] The guidance set out in this document will be reviewed and updated from time to time based on stakeholder feedback. In particular, this document will be updated based on feedback and insights gained after the first round of calls.

of facial images from the internet or CCTV footage, and real-time remote biometric identification systems used in public spaces for law enforcement.

c. Social scoring systems; and
d. AI systems that are deliberately designed to deceive or manipulate people, including “deepfakes" and other technologies that can be used to create highly realistic but fabricated content, or any other AI systems that manipulate human behavior to circumvent their free will.

2. Any AI activity that directly or indirectly violates Trustworthy AI Principles or otherwise results in any diminution of due process under law, that results in bias or discrimination, that deprives an individual of their civil liberties or rights to civic participation, that infringes any freedom of expression or that misuses personal data or that causes any similar harm to any individual or group of individuals.

Identification of projects that fall within this category, for cases under point 1, will be based on an evaluation of intended and foreseeable (including unintended) uses, outputs, users, and affected stakeholders. Cases under point 2 will be evaluated based on a qualitative assessment taking into account the criteria factors detailed below in Parts 4, 5 and 6 of this document. Any project classified as unacceptable risk will be recommended for removal from the program and considered ineligible for financing. More information can be found in the Q&A document.

High Risk

“High Risk AI Systems” means AI systems that tend to directly or indirectly violate Trustworthy AI principles throughout their lifecycle. This includes, but is not limited to, AI systems in the following areas designated as high-risk under Serbia’s Ethical Guidelines for Development, Implementation and Use of Robust and Accountable Artificial Intelligence:

1. The AI system is intended to be used as a safety component of a product, or the AI system is itself a product that has the function of a safety system, and this product is required to undergo a third-party conformity assessment under Serbian law
2. Biometric identification and classification of persons.
3. Management of critical infrastructures and their operation.
4. Education, vocational training, and qualifications.
5. Employment, human resources, management, and access to self-employment.
6. Healthcare.
7. Access to and use of public and social services and basic private services.
8. Law enforcement.
9. Migration, Asylum, and Border Control Management.

If a project falls within the list of high-risk areas noted above, there is a presumption that it will be classified as High Risk. However, the final risk classification will depend on a holistic

qualitative risk assessment in accordance with the Trustworthy AI principles, as per part 5: “Identification of risks”, and part 6: “Evaluation standards and risk matrix”.

Moderate Risk

Projects will be classified as Moderate Risk under the following conditions:

1. The project is not classified as Excluded or High Risk, in accordance with the guidelines listed above.
2. The project meets the relevant conditions to be classified as Moderate Risk under Part 5: “Identification of risks”, and Part 6: “Evaluation standards and risk matrix” of this document.

Low Risk

Projects will be classified as Low Risk under the following conditions:

1. The project is not classified as Excluded or High Risk, in accordance with the guidelines listed above.
2. The project meets the relevant conditions to be classified as Low Risk under Part 5: “Identification of risks”, and Part 6: “Evaluation standards and risk matrix” of this document.

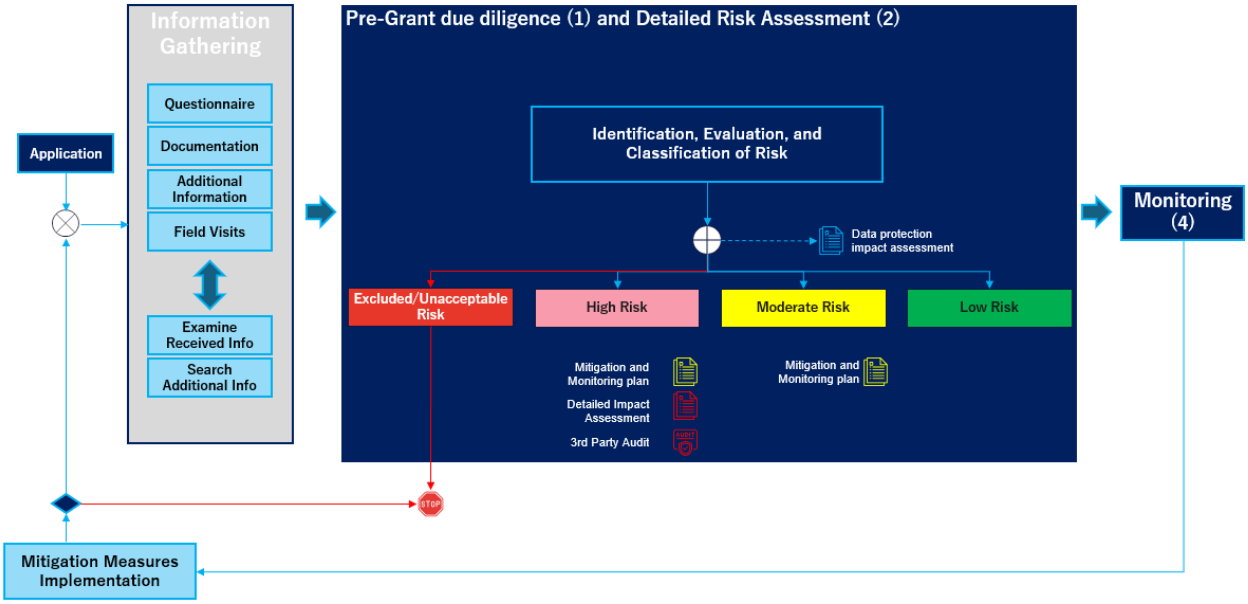
Requirements

Applicants will be required to adhere to the following requirements based on their assigned risk classification level:

1. **Low risk** – Applicants should adhere to the minimum requirements:
    - No AI Excluded Activities should be conducted;
    - [If Personal Data is processed]: Applicant shall conduct a data protection impact assessment and comply with applicable data protection laws;
    - Comply with Trustworthy AI principles; and
    - Re-assess risk rating at regular intervals.
2. **Moderate risk** – In addition to the above minimum requirements, applicants will need to prepare a risk mitigation and monitoring plan (which will be signed off by the IF or SF, as applicable), including implementation review.
3. **High risk** – In addition to the requirements applicable to Low and Moderate Risk projects, High Risk AI System activities will require a detailed AI risk impact

assessment to be conducted by the applicant before implementation, along with mandatory 3rd party audits.

Diagram 1. Process overview



Part 3: The Process

The risk assessment process includes several stages:

1. Pre-grant Due Diligence
2. Detailed AI Risk Assessment
3. Development of Mitigation Measures
4. Progress Monitoring

Stage 1: Pre-grant Due Diligence

At the outset, applicants are required to complete a self-assessment questionnaire to identify potential *risks posed by their AI systems*. This questionnaire should consider risks that arise throughout the lifecycle of the AI system, from *data collection to model training, system design, and deployment*.

The questionnaire will include questions relevant to understanding the project's context, goals, use case, risks, design approach, trustworthiness, and other important information about the project and AI system design. Answers to all questions are desirable, but it is possible that, given the early stage of project design, an applicant may not have answers to

all questions. In case an applicant doesn't provide an answer to a question, this lack of answer will be evaluated on a case-by-case basis. The questionnaire will be provided during the application phase. These submissions are reviewed by the AI specialist, and the project is classified in one of the possible categories. This initial stage provides an essential baseline for understanding and addressing project-specific risks. The AI specialist can request an interview, additional information or documents as part of the process of evaluation process.

Excluded activities are not eligible for financing and will be rejected. Low-risk applications do not undergo any further AI risk assessment, while Moderate and High-risk applications pass to the next stage of AI risk assessment – detailed AI risk assessment.

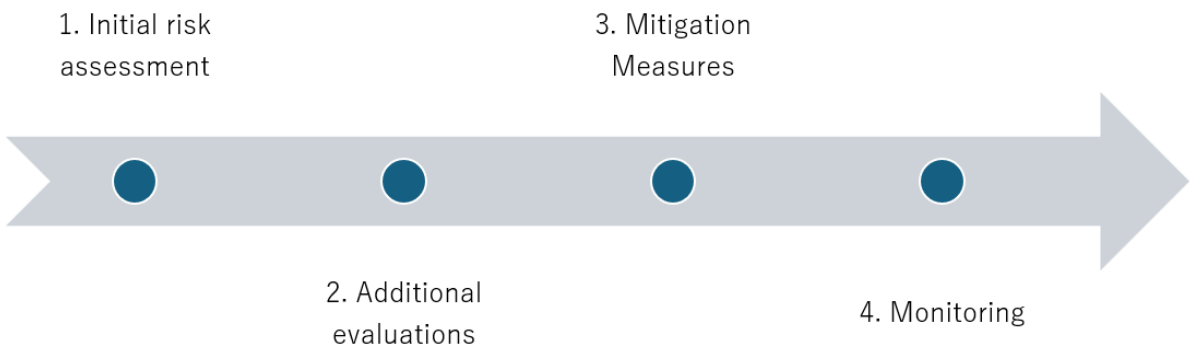
Stage 2: Detailed AI Risk Assessment

This stage involves the AI specialist conducting an in-depth analysis of the identified risks for Moderate or High Risk classified projects and their alignment with Trustworthy AI principles. Stakeholder consultations, including interviews with applicants, are conducted by the AI specialist to validate risk classification. Besides interviews with grantees, the AI specialist can conduct field visits, request interviews with potential users and stakeholders, or conduct other activities to investigate risks. For all those activities, the AI specialist will keep records in the form of memos. Particular attention is given to AI applications involving sensitive sectors such as healthcare, education, critical infrastructure, and law enforcement, which may require additional risk mitigation mechanisms that account for sectoral or use-case-based risks.

Stage 3: Development of Mitigation Measures

The IF and SF, respectively, for the grants that fall under their purview, and on advice from the AI specialist, will require Grantees to comply with the adoption of tailored mitigation measures based on the project's risk classification and proposed measures. Several Key performance indicators (KPIs) will be established by the AI specialist before implementation as part of the mitigation and monitoring plan. A detailed impact assessment is carried out before implementation starts for High-risk projects.

Diagram 2. Process flow



Stage 4: Progress Monitoring

This final stage ensures ongoing compliance with the mitigation plan and further risk management. Key performance indicators (KPIs) established in the previous phase are used by the AI specialist to monitor the AI system's performance during implementation. In cases of non-compliance, corrective actions may be imposed, ranging from enhanced oversight to finding adjustments or project termination.

The goal of this Risk Assessment Framework is to map, evaluate, classify, and govern risks related to AI to ensure safe, secure, and trustworthy AI.

Grantees may be required to conduct additional activities to enable progress monitoring, including providing additional documentation or information, or enabling field visits and exchange of questions and answers with the AI specialist.

Part 4: Trustworthy AI

Trustworthy AI means, collectively, the principles and requirements for robust and accountable AI systems, processes, and applications throughout the lifecycle of the AI based on Serbia's Ethical Guidelines for Development, Implementation and Use of Robust and Accountable Artificial Intelligence, which include, at a minimum, the following:

1. Explainability and verifiability;
2. Dignity;
3. Prohibition to cause damage;
4. Fairness and non-discrimination;
5. Human agency, oversight, determination, and control;
6. Technical reliability, safety, and security;
7. Privacy, personal data protection, and data management;

8. Transparency;
9. Diversity, non-discrimination, and equality;
10. Social and environmental well-being;
11. Accountability and responsibility;

Additionally, “Trustworthy AI” principles incorporate the following principles, as included in other good practice frameworks recognized by Serbia, including the “UN General Assembly Resolution on Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development” A/78/L.49 (UNGA Resolution) and the “UNESCO Recommendation on the Ethics of Artificial Intelligence” SHS/BIO/PI/2021/1 (UNESCO Recommendation), including, inter alia:

12. Proportionality and do no harm;
13. Due process;
14. Freedom of expression;
15. Sustainability;
16. Awareness and literacy; and
17. Multi-stakeholder and adaptive governance, and collaboration.

Any project, for its entire lifecycle, must satisfy the listed requirements. A description of the relevant principles will be provided as a separate Q&A document to help applicants better understand their obligations.

Part 5: Identification of Risks

Risks are evaluated against the Trustworthy AI principles. Those are:

1. **Transparency** (including transparency, explainability, and verifiability) – risk that the AI system will not satisfy transparency requirements.
2. **Discrimination, Bias, and Fairness** (including fairness, dignity, diversity, non-discrimination, equality, multi-stakeholder and adaptive governance, and collaboration) – risk that an AI system will fail to satisfy fairness requirements and will create biased outcomes for relevant stakeholders or groups or allow discrimination against marginalized or other groups.
3. **Human Centricity, Oversight, and Agency** (including human agency, oversight, determination, and control, due process, freedom of expression, awareness, and literacy) – risks that an AI system may cause harm or create unintended outcomes due to a lack of human oversight.
4. **Harmful AI System** (including prohibition to cause damage, proportionality, and “do no harm”) risk that the AI System might create harm to life, health, or property, human rights, democratic values, or the rule of law.

5. **Data Management and Privacy** (including privacy, personal data protection, and data management) – risks that personal and/or proprietary data will not be handled in compliance with the law, or that data used for AI development or deployment will not be properly handled.
6. **Accountability, Responsibility, and Legal Considerations** (including accountability and responsibility, and legal considerations) – risks of low accountability or responsibility awareness or procedures. Missing considerations of mandatory legal requirements for the use of the system in a specific sector.
7. **Technical robustness, metrics, and safety** (technical reliability, metrics for success, monitoring and maintenance of the system, documentation, and respect for the project management process best practices, safety, and security) - failing to satisfy technical robustness and security measures for the model, relevant data, and outputs. Failing to set metrics for measuring development, implementation, and use, and to ensure that the planned scope is achieved, and that auditability, traceability, and monitoring are preserved.

They should be dealt with during the design and development phase. The idea is that diligent, trustworthy, and proper project design will mitigate any risks that may arise during the project lifecycle.

For each of the above Trustworthy AI requirements, there will be an evaluation of some of the specific risks that might arise as a result of proposed project design. To properly evaluate the listed risks, specific activities and design approaches will be evaluated through the intake questionnaire that applicants will be required to complete. For example:

1. Is an AI system among the areas that fall under the excluded or high-risk activities?
2. What is the context where the AI system will be used?
3. What is the quality of governance at the organization and at the system level?
4. What is the management of data used for model development and for system operation, and for general data governance?
5. What are the plans and tools for the development and tracking of performance at the component level and at the system level?
6. Is there continuous monitoring of performance?
7. What is the level of human agency, autonomy, and human oversight?
8. How is resilience to attacks and security implemented?
9. Outlook on general safety?
10. What accuracy is planned, and how is it achieved?
11. How privacy is protected
12. How is auditability and traceability secured?
13. How is transparency achieved and secured?
14. Activities on the avoidance of unfair bias
15. Level of accessibility?

16. Is stakeholders' participation secured and enabled?
17. Is there a plan for risk management?
18. How is accountability secured?
19. What methodology and tools are selected and used for metrics and validation?
20. What is the impact on individuals and groups?

This list is not an exhaustive one and will include edits as required during the program development. For this purpose, there will be a checklist to ensure that the most important risk factors are taken into consideration.

Part 6: Evaluation standards and risk matrix

The risk evaluation process is grounded within the overall evaluation of the context of project implementation: is the AI system among the list of excluded activities or high-risk activities, what are the areas of implementation, what business function, basic users, and stakeholders, and other context-relevant factors. Context is of great importance and influences further risk evaluation. Application in medical or other sensitive sector or business activity, or higher scale and scope or other elements, push for higher risk awareness and sensitivity.

This Part outlines the risk classification matrix to be used by the AI specialist in classifying risk in relation to each of the Trustworthy AI Principles (which have been aggregated for ease). The matrix will have three dimensions: severity, likelihood, and materiality. Below is the explanation and tabular representation.

Table 1. Specific risks matrix element with values

Mark	Likelihood	Severity	Materiality (on selected trustworthy principle)
0	Not expected	No meaningful impact	Not materially important
1	Expected rarely	Small impact	Slightly materially important
2	Expected occasionally	Moderate impact	Materially important
3	Expected often	Heavy deviations	Crucial importance

Likelihood of risk – returns the numerical representation of the possibility and frequency of the possibility that the listed risk can occur or the possibility that the evaluated activity can contribute to the risk after any existing mitigation measures are applied.

- "0" – The risk is "not expected" to happen. The risk is not anticipated to occur under current circumstances or system design.
- "1" – The risk is "expected rarely" to happen. The risk could occur in exceptional or unusual situations, but is unlikely in normal operation.

- "2" – The risk is "expected occasionally" to happen. The risk may arise under certain conditions or due to known vulnerabilities, but it is not frequent.
- "3" – The risk is "expected often" to happen. The risk is likely to materialize frequently under typical use or environmental conditions.

Severity of risks – Severity refers to the numerical representation of the magnitude or seriousness, scope, and scale of the consequences if a risk happens. It is also calculated after the implementation of any existing mitigation measures. It reflects the level of negative impact a risk event can have on individuals, organizations, or society. In AI risk management, severity can often be assessed in terms of harm to rights, safety, finances, reputation, or broader systemic effects. It is also important to consider the specific context where the AI system will be implemented, including industry, task, users, stakeholders, and the overall use environment.

- "0" – No meaningful impact. No significant harm or disruption, or negligible effect on objectives or stakeholders.
- "1" – Small impact. Minor, localized, or easily reversible harm; limited to minor inconvenience or operational slowdowns.
- "2" – Moderate impact. Noticeable disruption or harm may require intervention, result in moderate harm to Trustworthy AI principles or moderate financial, legal, or reputational damage, or cause distress to affected individuals or groups, or the impact is happening within a sensitive use environment or has larger impact on vulnerable social groups.
- "3" – Heavy deviations. Severe or systemic harm could result in significant harm to Trustworthy AI principles, significant legal violations, major financial loss, critical safety incidents, or irreversible damage to individuals, organizations, or society at large.

Materiality or risks – refers to the numerical representation of the significance or importance of a risk in relation to a specific trustworthy AI principle after the implementation of any existing mitigation measures. A risk is considered material if it has the potential to affect the achievement of key objectives, compliance with legal or ethical standards, or stakeholder trust. Materiality is context-dependent and should be aligned with organizational values, regulatory requirements, and the expectations of stakeholders.

- "0" – Not materially important. The risk does not meaningfully affect the Trustworthy AI principle; negligible or irrelevant in a given context.
- "1" – Slightly materially important. The risk has minor relevance for the Trustworthy AI principle. May warrant monitoring, but does not require immediate action.
- "2" – Materially important. The risk has high relevance for a principle. It can affect compliance, stakeholder trust, or the achievement of objectives.

– "3" – Crucial importance. The risk is fundamental for a principal. Threatens core compliance, standards, or the foundation of trustworthiness.

Considering that risk is evaluated so early in the process of development of an AI system, applicants may only have answers to some of the risk evaluation questions if the grant is approved. At the same time, to enable fair evaluation of AI risks while also promoting the execution of activities that will decrease risk, at this stage each risk is evaluated as "inherent risk" – i.e. the risk level based on key controls being applied in the current project design and environment. During the mitigation and monitoring stage (Part 7 below), any "residual risks" left after mitigation measures are applied may be subject to further actions as needed.

Total risk is calculated by multiplying the parameters' severity*likelihood*materiality. The lowest mark is 0 and the largest is 3. This leads to the possible combination for the "weight" of one risk to be 0, 1, 2, 3, 4, 6, 8, 9, 12, 18, 27. For each calculated risk, a mitigation measure will be proposed, and the final risk will be calculated, including the implementation of the mitigation measure.

Classification of risks can change during the project time in cases where the applicant doesn't respect and implement assigned mitigation measures or in cases where circumstances change during the project period. In the second case, changes can go in the direction of an increase or decrease of risk, depending on which circumstances have changed.

For evaluation, an AI specialist will use the identification of possible risks that arise from the project and can be recognized in such an early stage, as well as risks that are recommended for evaluation in part 5 of this Document: "Identification of Risks".

Table 2. Specific risk matrix

Trustworthy AI Principle	No	Risk Discussion	Likelihood	Severity	Materiality	Classification
Transparency	1	Risk consideration 1	3	1	2	Low
	2	Risk consideration 2	3	2	3	High risk
		RESULT (avg)				
Discrimination, Bias, and Fairness	3		1	2	1	Low
…	4		3	2	1	Moderate

The threshold for classification of risks is as follows:

1. Low risk – 0 to 4,
2. Moderate risk – 6 to 12.
3. High risk – 18.
4. Unacceptable risk – 27.

Diagram 3. Risk classification scale

The overall risk evaluation for any given principle (per row) is based on the combination of the following elements:

1. If any risk is classified as excluded, the application will be dismissed.
2. If any risk is classified as high-risk, the entire project will be classified as high risk.
3. In case all risks are low, the project will be classified as low.
4. In case all risks are moderate, the project will be classified as moderate.
5. In case of a combination of low and moderate, the classification will be based on the share and specific of moderate risks, and specifics of the case, especially taking into consideration the time and the possibility of mitigating those risks.

It is very important to stress that the risk assessment is the assessment of the project in its totality, in the current moment, and future. The proposed matrix can be waived in case it can't comprehensively evaluate all risks. **The AI specialist and the evaluation team reserve the right to provide a different classification, based on the project in its totality.**

Part 7: Guidelines for the creation and management of risk mitigation and monitoring plans

For any risk or element of the project that will be monitored, a plan should include a list of measures that would include:

1. Project ID number
2. What Risk – Identify What Risk Is being Mitigated/Monitored. This provides information on the specific risk being addressed, along with a brief explanation. This should ensure understanding of the context and rationale for the mitigation action.
3. What Element Should Be Monitored? This specifies the process, system component, data, or activity that requires ongoing observation. It focuses on monitoring efforts on the relevant aspects, ensuring early detection of issues and designating what is the target of the measure.

4. What Is the Expected Mitigation Activity from the Applicant, and What Is the Expected Result? It should define the concrete actions required from the applicant to mitigate the risk, along with the anticipated outcome. The purpose is to provide clear instructions and set measurable expectations for risk treatment.
5. Who Is Responsible for Reporting? It should assign a specific individual or role responsible for tracking the mitigation activities with applicants so that accountability is established and ensure that reporting is consistent and timely.
6. Time for Check / Period for Check. Define the frequency of monitoring and reporting activities. This is important to ensure risks are managed proactively and that mitigation effectiveness is regularly evaluated and adjusted as needed.

If required, some additional fields could be added during the process.

Table 3. Mitigation plan elements

ID	What risk	What element	What activity from the Applicant	Person responsible for reporting	Time for a check
1	Risk this	Data quality check	Improve data quality checks	Person 1	One month
2					
…					

Fulfilment of the risk mitigation plan will be monitored as part of standard monitoring activities, as well as in periods that are set within the mitigation plan.

Part 8: Monitoring AI Risks during the lifetime of the project

Monitoring is required to provide information about progress on the respective mitigation requirements for the project. Monitoring is a continuous activity, and an AI specialist can request further information during the whole period of the project. It should include both standard and specialized checks. After monitoring, a monitoring memo should be included in the risk assessment file.

The standard monitoring plan involves regular checks on the applicant's activities. It should be done at least once during the project time. Projects differ in duration, so the regular check will be done sometime around the end of the first half of the project duration. The monitoring would evaluate respect for the provided measures and investigate the existence and maintenance of the initially confirmed documentation. It should include checking all of the evaluated risks and any additional risks that are discovered during the project. In the event

that any residual risks are identified after mitigation measures are fully implemented, the AI specialist may require the applicant to implement further follow-on actions as needed.

Specialized plans for cases of moderate or high risk will be monitored in accordance with established mitigation plans.

It is also possible to create special monitoring controls in cases where it is required for any reason (some new scientific breakthrough that would completely change the perspective on some topic, reveal of some hidden facts during the process of evaluation, or other event that would require the monitoring process of any element.

For high-risk cases, a third-party audit is required. Auditor will be decided on a case-by-case basis.