# D1.3 Interim report on legal and ethical impact assessment

| | |
|---|---|
| Related Work Package | WP1 – Project Management and Coordination |
| Related Task | Task 1.4. – Legal and Ethical Compliance |
| Lead Beneficiary | KU Leuven |
| Contributing Beneficiaries | |
| Document version | v.1.0 |
| Deliverable Type | R-Document, Report |
| Distribution level | PU - Public |
| Contractual Date of Delivery | 31/03/2025 |
| Actual Date of Delivery | 01/04/2025 |

| | |
|---|---|
| Authors | Jean De Meyere (KU Leuven)<br>Abdullah Elbi (KU Leuven)<br>Ana Maria Corrêa (KU Leuven) |
| Reviewers | Asbjørn Følstad (SINTEF)<br>Davide Moroni (CNR) |

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Directorate-General for Communications Networks, Content and Technology. Neither the European Union nor the granting authority can be held responsible for them.

## Version history

| Version | Description | Date completed |
|---|---|---|
| v0.1 | TOC preparation | 20.09.2024 |
| v0.2 | First version of the Deliverable for external review | 03.02.2025 |
| v0.3 | Integration of the EEAB comments and second version of the Deliverable for internal review | 03.03.2025 |
| v1.0 | Integration of the comments from internal reviewers and sent to the Coordinator for approval | 31.03.2025 |

## Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

## Disclaimer

## Executive summary

This deliverable introduces a comprehensive framework for assessing regulatory and ethical risks within the FAITH's Trustworthiness Assessment Framework (TAF). The assessment examines the respect of trustworthy AI characteristics by the FAITH AI_TAF. We consider the characteristics established by the EU High-Level Expert Group (HLEG): fairness and non-discrimination, technical accuracy and robustness, privacy and data governance, transparency in AI decision-making, human oversight mechanisms, societal and environmental well-being, and accountability measures. The assessment further considers those requirements in light of EU digital regulations, including the EU AI Act, GDPR, cybersecurity regulations, and anti-discrimination laws.

The deliverable consists of two complementary components. First, a **spreadsheet** template **in .xls and .ods format** provides a detailed questionnaire to gather the information necessary to complete the assessment. Second, a text document in pdf format that contains a theoretical explanation of the regulatory and ethical framework the questionnaire focuses on, guidance on how to complete the questionnaire and links to additional resources and templates **for regulatory assessments such as Data Protection Impact Assessments (DPIAs) or Fundamental Rights Impact Assessments (FRIAs).**

The assessment considers FAITH's seven large scale pilots (LSPs), focusing on seven domains: media, transportation, education, port infrastructure, waste-water treatment, healthcare and active ageing. While the assessment primarily focuses on the FAITH AI_TAF itself, it will as well take into account the impact of the FAITH AI_TAF on regulatory and ethical compliance for the LSPs. This deliverable, however, does not serve as a legal and ethical assessment of the AI system under test in each LSP. This interim report will evolve based on changes in EU legal frameworks, internal and external recommendations, and the evolution of best practices in AI governance.

## Table of Contents

1. Introduction ..... 8
1.1. Purpose of FAITH ..... 8
1.2. Purpose of this deliverable ..... 10
1.3. Input from and contribution to other deliverables and project ..... 11
1.4. Structure of the document ..... 11
2. The theoretical framework for trustworthy AI from a legal & ethical perspective ..... 13
2.1. Impact Assessment for AI Technologies in the EU ..... 13
2.2. Fundamental Rights in the EU ..... 13
2.3. Data protection Rules (GDPR) ..... 16
2.3.1. Theoretical overview ..... 16
2.3.2. Data Protection Impact Assessment – DPIA ..... 17
2.4. Cybersecurity obligations ..... 20
2.4.1. Theoretical overview ..... 20
2.4.2. In practice ..... 21
2.4.3. Relevance for FAITH ..... 23
2.5. Anti-discrimination regulation ..... 24
2.5.1. Theoretical overview ..... 24
2.5.2. In Practice ..... 24
2.5.3. Relevance for FAITH ..... 26
2.6. The EU AI Act ..... 28
2.6.1. Theoretical overview ..... 28
2.6.2. In Practice ..... 31
2.6.3. Overview of (selected) assessment framework ..... 32
2.6.4. Relevance for FAITH ..... 33
2.7. Ethical Obligations ..... 33
2.7.1. The High-level expert group on Artificial Intelligence ..... 33
2.7.2. The Fundamental Rights-Based Approach ..... 34
2.7.3. Ethical Principles and Key Requirements ..... 34
2.7.4. The Assessment List for Trustworthy AI ..... 34
2.7.5. Assessing the FAITH AI_TAF through the HLEG AI Lens ..... 35
3. The How-To: instructions for the assessment ..... 37
3.1. Who Will Carry Out or Be Involved in the Legal and Regulatory Assessment of FAITH AI_TAF ..... 37
3.2. When Will the AI system(s) impact assessment of the FAITH_AI TAF Be Conducted? ..... 38
3.3. How Will the AI system(s) impact assessment of the FAITH_AI TAF Be Conducted in Practice? ..... 38

3.4. The review by the External Ethics Advisory Board ..... 40
4. Assessment template for the FAITH AI_TAF ..... 41
4.1. Description of the FAITH AI_TAF ..... 41
4.1.1. General Information ..... 41
4.1.2. Privacy – General Information ..... 41
4.1.3. General Technological Information ..... 42
4.2. AI Act Assessment ..... 45
4.3. Legal and Ethical Risk Impact Assessment ..... 49
4.3.1. Legal Risks ..... 49
4.3.2. Ethical Risks ..... 50
4.3.3. Legal and Ethical Risk Identification ..... 50
4.3.4. Legal and ethical Risk Management ..... 60
4.3.5. Legal and ethical Risk Mitigation ..... 61
5. Conclusion ..... 62
6. References ..... 63
7. Annex ..... 66

## List of Figures

Figure 1 - Fundamental Rights in the EU

## List of Tables

Table 1 - Description of LSPs
Table 2 - FAITH AI_TAF's supporting tools
Table 3 - Fundamental Rights in the context of LSPs
Table 4 - Forms of discrimination
Table 5 - Legally Protected Grounds in the Access to and Supply of Goods and Services in EU countries based on the Country Report non-discrimination 2023 compiled by the European Equality Law Network
Table 6 - Applicability of the AI Act
Table 7 - Additional Ethical Frameworks for LSPs
Table 8 - Cluster of partners for the assessment
Table 9 - Division of the questionnaire
Table 10 - General Information Question
Table 11- General Privacy Questions
Table 12 - General Technological Questions
Table 13 - AI Act – General Information
Table 14 - AI Act Classification
Table 15 - AI Models Interactions
Table 16 - Human Oversight Assessment
Table 17 - Technical Robustness and Safety Assessment
Table 18 - Privacy and Data Governance Assessment
Table 19 - Transparency Assessment
Table 20 - Fairness Assessment
Table 21- Societal and environmental well-being assessment
Table 22 - Accountability Assessment
Table 23- Risk assessment matrix

## List of Abbreviations

| Abbreviation | Explanation |
|---|---|
| **ALTAI** | Assessment List for Trustworthy Artificial Intelligence |
| **CSA** | Cybersecurity Act |
| **CRA** | Cyber Resilience Act |
| **DPO** | Data Protection Officer |
| **DPIA** | Data Protection Impact Assessment |
| **EDPB** | European Data Protection Board |
| **EEAB** | External Ethics Advisory Board |
| **ENISA** | EU Agency for Cybersecurity |
| **FAITH** | Fostering Artificial Intelligence Trust for Humans |
| **FRIA** | Fundamental Rights Impact Assessment |
| **GDPR** | General Data Protection Regulation |
| **GPAI** | General-Purpose Artificial Intelligence |
| **HLEG** | High-Level Expert Group for AI |
| **LSP** | Large Scale Pilot |
| **NIST** | National Institute of Standards and Technology |
| **TAF** | Trustworthiness Assessment Framework |

# 1. Introduction

## 1.1. Purpose of FAITH

FAITH (Fostering AI Trust for Humans) aims to enhance awareness and foster cooperation among various stakeholders working on various aspects of trustworthiness at different stages of an AI-system lifecycle. FAITH aims to provide AI practitioners and stakeholders with a practical playbook for how to assess and build trustworthy AI systems and to measure systems' trustworthiness continuously. By adopting a human-centric, trustworthiness assessment framework (the FAITH AI_TAF), FAITH enables the testing, measuring and optimisation of risks associated with AI trustworthiness in critical domains. The developed FAITH AI_TAF will be validated throughout seven large scale pilots (LSP) activities in identified critical domains, namely robotics, education, media, transport, healthcare, active ageing and industrial processes. Table 1 describes the objective of each of these LSPs.

*Table 1 - Description of LSPs*

| Domain | Description |
|---|---|
| **LSP 1 Media** | The pilot will assess the AI trustworthiness of a solution for **supporting local journalists** in characterising if a locally developed narrative is true or false, before it takes the form of a news item and is forwarded to the PRESShub's news portal, a portal that congregates news content from 38 regional media organisations in Romania. |
| **LSP 2 Transport** | The pilot will assess the AI trustworthiness of a scalable privacy-preserving platform based on pervasive AI and video analytics that can help to **improve efficiency, safety and security on board trains and in stations** by using privacy-preserving smart cameras that capture and analyse visual data in real-time. |
| **LSP 3 Education** | The pilot will assess the trustworthiness of novel, AI-based student assessment methods of inquiry learning during laboratory work. It will evaluate **an AI-driven learning companion** that will support students in developing their learning path and provide automatic guidance. |
| **LSP 4 Port infrastructure** | The pilot will assess the AI trustworthiness of systems used for the **remote control of maritime and/or underwater UAV** (uncrewed automatic vehicle) during the missions programmed for the acquisition of data and images of the port's infrastructures; to determine automatically the state of infrastructures, the seriousness of damages and to assign a level of priority of **maintenance works**; and to determine the presence of any kind of dangerous situation that can result in a **problem for the infrastructure**. |
| **LSP 5 Wastewater treatment** | The pilot will assess the AI trustworthiness of systems used to evaluate alternative methods to **regulate nitrogen levels** in the Oslofjord in Norway and to optimise the wastewater treatment process to reduce nitrogen disposal and associated cost and risks. |
| **LSP 6 Healthcare** | The pilot will assess the AI trustworthiness of systems used for segmentation in **prostate imaging**, a particularly challenging task |

| | |
|---|---|
| | especially for the delineation of the prostate gland and its anatomic substructures. The system's aim is to provide objective, segmentation of the prostate gland eliminating intra and inter-observer variability and saving precious time for the radiologists. |
| **LSP 7**<br>**Active ageing** | The pilot will assess the AI trustworthiness of systems used for providing personalised and timely support for **improving the quality of care for the senior population**. The purpose of the pilot is to provide information about the well-being state of the senior through monitoring of daily activity parameters and his/her frailty progression, detecting changes in behaviour patterns, cognitive skills, and functional abilities through the use of AI. |

The FAITH AI_TAF methodology includes the analysis of the trustworthiness of AI systems, research on adequate risk assessment-based approach, identify suitable AI technologies and resources as well as measures to achieve trustworthiness, determines the ethical and legal requirements for outcomes, proposes psychosocial profiles to determine the trustworthiness of AI participants[1] and develops metrics and scales for measuring AI risks and trustworthiness, estimates the risks for trustworthiness. The FAITH AI_TAF focuses not only on the technical but also social and human threats and selecting appropriate technical, social, behavioural, legal and policy-related measurements and controls to mitigate risks to ensure AI trustworthiness.

The FAITH AI_TAF is a **risk-based approach** that identifies social, human and technical trustworthiness threats and vulnerabilities. It is based on four components: (i) the **NIST AI Risk Management Framework**, (ii) the **ENISA guidelines on how to achieve trustworthiness by design**, (iii) **requirements imposed by the EU legislative instruments** and **stakeholder's intelligence and (iv) users' engagement**. The FAITH AI_TAF adopt the **7 trustworthiness characteristics developed by NIST**: safety, security and resilience, explainability and interpretability, enhancement of privacy, fairness and mitigation of harmful bias, accountability and transparency and validity and reliability.

The FAITH AI_TAF is a 6 steps methodology: (1) **cartography**, (2) **threat analysis**, (3**) impact assessment**, (4) **vulnerability analysis**, (5) **risk analysis**, (6) **countermeasures.** The FAITH AI_TAF incorporates three supporting tools, and several tools have been identified as mitigation measures for specific trustworthiness characteristics.

[1] AI participants are the key players involved in the creation, design, development, and implementation of artificial intelligence systems. They encompass a range of roles including designers, developers, and data specialists, all working together to ensure that AI technologies are effective, ethical, and aligned with their intended purposes.

*Table 2 - FAITH AI_TAF's supporting tools*

| FAITH AI_TAF's supporting tools | |
|---|---|
| **FAITH AI TrustGuard** | Checklist based risk assessment for AI-based systems in isolation. |
| **FAITH AI TrustSense** | Tool used to profile the trustworthiness of AI Participants. |
| **FAITH AI Model Hub** | A metadata collection repository for AI models and datasets integrating the notion of the AI model passport and data passport. |

This deliverable aims at assessing the legal and ethical conformity of the FAITH AI_TAF and of the above-mentioned supporting tools, as well as the impact of the FAITH AI_TAF in the context of the 7 LSPs of the project. Please note that the FAITH AI_TAF is still being developed throughout the course of the project and that this section will evolve to reflect this development.

### 1.2. Purpose of this deliverable

The FAITH AI_TAF aims to be compliant with European and international initiatives such as the NIST AI Risk Management Framework and the ENISA Guidelines on achieving trustworthy AI2. Furthermore, the FAITH AI_TAF must consider requirements imposed by regulatory instruments, notably the EU legal framework surrounding artificial intelligence. This includes not only the EU AI Act3, but also privacy and data protection regulation (including the General Data Protection Regulation - GDPR4), cybersecurity regulation (including the NIS II Directive5), anti-discrimination laws and relevant sectoral regulations, depending on the actual usage of the AI system by the LSP.

This deliverable proposes a framework for assessing the regulatory and ethical risk of AI systems within the FAITH AI_TAF. The Framework will delve into the elements of trustworthiness in AI functioning: fairness, technical accuracy and robustness, privacy and data governance, transparency, human oversight, societal and environmental well-being and accountability. While parts of the FAITH AI_TAF may not fall under the legal definition of AI, they will still be assessed from a regulatory and ethical perspective. It will also provide an assessment framework and guidance for the LSPs, enabling them to further reinforce the trustworthiness assessment provided by the FAITH AI_TAF.

Throughout the course of the project, the deliverable will be updated based on relevant changes in the EU legal framework surrounding artificial intelligence. Through continuous discussion with technical partners, FAITH External Ethics Advisory Board (EEAB), technical

[2] Deliverable 2.1, p. 89

[3] Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union [2022] OJ L333/80 (NIS 2 Directive)

[4] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data [2016] OJ L119/1 (General Data Protection Regulation)

[5] Regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence [2024] (AI Act)

partners and LSPs leaders, the framework will also be further tailored to better consider practical and technical limitations that may arise throughout the project.

It is important to note that the various templates, assessments, and guidance discussed in this deliverable are offered as advice and do not constitute official legal counsel. In some cases, the activities performed by AI participants will require the intervention of legal counsel and/or ad-hoc domain experts, such as AI counsel, data protection officers or lawyers. Furthermore, the proposed guidance, assessments, and templates do not offer one-size-fits-all solutions, and tailoring may be needed for each AI user relying on the FAITH framework.

Regarding ethics, it is important to further understand that the templates, assessments and guidance discussed in this deliverable should not be understood as a simple checklist that need to be completed. Ethical considerations are complex and might require further analysis than provided by this assessment.

Following the recommendations in this deliverable will not eliminate all regulatory and ethical risks, as this is virtually impossible, but aims to contribute to legal and ethics compliance efforts by mitigating the residual risk to an acceptable level.

### 1.3. Input from and contribution to other deliverables and project

Deliverable 2.1 (FAITH Methodological Framework and Requirement Analysis v1) contains the description of the FAITH AI_TAF. Deliverable 2.1 assesses and outlines existing efforts, initiatives and results in identifying technical, legal, and policy-related efforts for AI Trustworthiness. It initiates a framework adopting a risk assessment approach in identifying, estimating and managing trustworthiness risks, capturing all its dimensions (cybersecurity, transparency, robustness, accuracy, data quality and governance, and human oversight). While similar, the objectives of Deliverable 2.1 and this deliverable are not the same. Deliverable 2.1 outlines the methodology of the FAITH AI_TAF based on its analysis of AI Trustworthiness while this deliverable focuses on the respect of legal and ethical requirements by the FAITH_AI TAF and its use within the LSPs.

The final version of this deliverable (D1.4) will rely on and extend this deliverable to provide a comprehensive overview of the regulatory and ethical issues surrounding the FAITH AI_TAF. Furthermore, it will also consider at a high-level the regulatory risks present for FAITH LSPs. This will also enable the possibility of creating links with D2.6 (Gap Analysis of the Regulatory Framework for FAITH Pilots), complementing the assessment framework developed in this deliverable with legal analysis tailored to each LSP. Finally, cross-collaboration will happen between deliverables, resulting in improved contributions for all.

This deliverable also took inspiration from the Template & Guidance for Legal and Ethical Impact Assessment from the Themis 5.0 Project[6].

### 1.4. Structure of the document

[6] S. Garcia, A. Corrêa and G. Stamatellos, 'Template & Guidance for Legal and Ethical Impact Assessment', available at https://www.themis-trust.eu/_files/ugd/a245c2_c1fe2d866bf140e5a4e5daa8afb292ce.pdf accessed 17 March 2025

This deliverable consists of two main components: a Spreadsheet template and an accompanying Text document. These components work together to provide a comprehensive framework for assessing the FAITH AI_TAF and to provide guidance to LSPs owners.

The **Spreadsheet Template** serves as a detailed questionnaire designed to evaluate the FAITH AI_TAF and its alignment with regulatory and **ethical** standards. This template is intended to structure the assessment process by providing a systematic format for inputting information related to the evaluation of the FAITH AI_TAF. It ensures that all necessary elements of trustworthiness, compliance, and functionality are documented clearly and consistently. The template also offers prompts to guide assessors in identifying areas of potential concern and in capturing the outcomes of their assessments.

The **Text Document** complements the Spreadsheet template by offering theoretical, regulatory, and practical insights to support the assessment process. This document fulfils several key functions:

1. **Theoretical Framework**: It provides a detailed explanation of the regulations and frameworks that must be observed in the assessment process. This includes the impact of fundamental human rights as articulated in the EU Charter of Fundamental Rights and other ethical guidelines developed by the High-Level Expert Group on Artificial Intelligence. By situating these considerations within a broader theoretical context, the document helps assessors understand the importance of regulatory and ethical compliance in AI system development. **Annex I** of this document provides a summary chart containing the essential information related to the regulations analysed in this deliverable.
2. **Regulatory Overview**: The document outlines the specific regulatory obligations that the FAITH AI_TAF must meet. It also provides a high-level, step-by-step overview on how to respect those regulations and, when relevant, discusses available assessment frameworks. These regulations are:
    - The **General Data Protection Regulation (GDPR)**, which governs the processing of personal data.
    - The **EU AI Act**, which establishes requirements for trustworthiness and risk management in AI systems.
    - **Cybersecurity Regulations**, including the NIS II Directive and the Cybersecurity Act, which emphasize secure-by-design principles and the resilience of critical infrastructure.
    - **Non-Discrimination Laws**, which mandate the prevention of discriminatory outcomes.
3. **Spreadsheet Template Instructions**: The Text document explains in detail how the Spreadsheet template is to be used for assessments. It describes the type of information that must be provided in each section of the template, guiding assessors on how to accurately document their findings.

The deliverable provides a practical tool for a structured assessment as well as a theoretical guide for understanding the regulatory and ethical challenges surrounding AI. The deliverable aims at ensuring that the FAITH AI_TAF is not only compliant with existing legal requirements but also contributes to the establishment of trustworthy AI within different contexts.

## 2. The theoretical framework for trustworthy ai from a legal & ethical perspective

### 2.1. Impact Assessment for AI Technologies in the EU

In the EU, there is no overarching obligation for comprehensive "AI impact assessments," but several horizontal legal instruments impose requirements to assess and mitigate the risks AI poses to **fundamental rights and societal interests**. These obligations include:

- **Fundamental Rights Impact Assessments (FRIAs)**: Mandated by the AI Act, Article 27, for selected high-risk AI systems, focusing on their implications for fundamental rights.
- **Data Protection Impact Assessments (DPIAs)**: Required under GDPR, Article 35, when AI systems process personal data in ways that present high risks to individuals.
- **Risk Management systems for High-Risk AI systems**: Introduced by the AI Act (Chapter III), including conformity assessments to ensure compliance.
- **Cybersecurity Risk Management**: Addressed by the **NIS Directives** (Directives (EU) 2016/1148 and 2022/2555), the upcoming **Cyber Resilience Act**[7], and the **Cybersecurity Act**[8], which together establish obligations for secure design and certification of ICT products and AI-based technologies.

In the following sections, we examine these legal instruments and their relevance to the FAITH AI_TAF. While horizontal obligations are central, **domain-specific and national regulations** may introduce additional requirements, particularly in the areas of data protection and cybersecurity. It is important to note that not all parts of the FAITH AI_TAF extensively rely on AI technologies in order to perform. However, this does not mean that we should not assess those parts of the FAITH AI_TAF. Other regulatory obligations not focused on artificial intelligence such as data protection and cybersecurity obligation still apply. The FAITH AI_TAF will also integrate with AI models which might require documentation as prescribed in the AI Act. Furthermore, assessing the FAITH AI_TAF is essential, given the role it will play in trustworthy AI assessments that will result in concrete AI applications.

### 2.2. Fundamental Rights in the EU

The protection of fundamental rights is a cornerstone of the EU legal framework and plays a crucial role in the context of AI. At the EU level, fundamental rights are protected by the **Charter of Fundamental Rights of the European Union** ("the Charter")[9] , a binding instrument that guarantees a comprehensive set of rights, including **liberty and security** (Article 5 of the Charter) , **freedom of expression** (Article 10 of the Charter), **equality** (Article 16 of the Charter) , and **justice** (Article 6 and 7 of the Charter) . These rights are directly relevant to the development and deployment of AI systems, as they provide a basis for assessing whether AI technologies align with core democratic values. Fundamental rights are also protected by the European Charter of Fundamental Rights ("ECHR") and by the constitutional traditions of Member States.

[7] Proposal for a Regulation of the European Parliament and of the Council on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/1020 COM/2022/454 final (Cyber Resilience Act)

[8] Regulation (EU) 2019/881 on ENISA and on information and communications technology cybersecurity certification [2019] OJ L151/15 (Cybersecurity Act)

[9] Charter of Fundamental Rights of the European Union [2012] OJ C

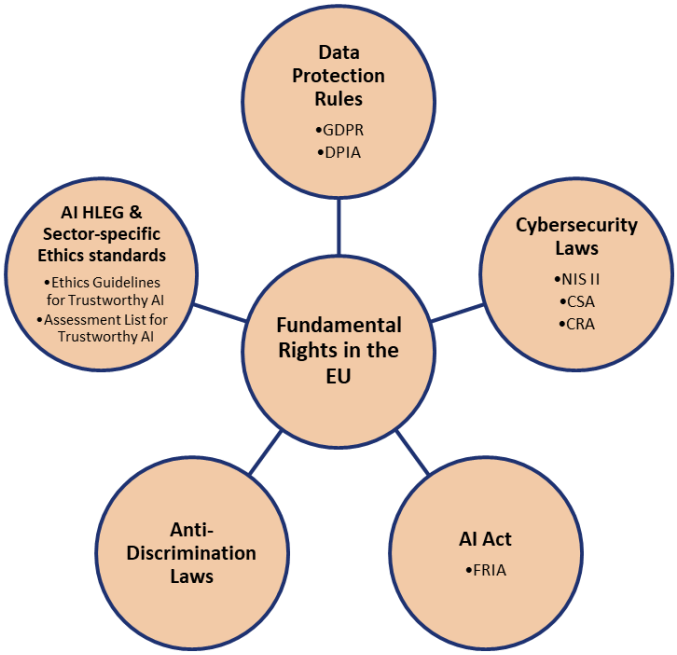


*Figure 1 - Fundamental Rights in the EU*

Both the **GDPR** and the **AI Act** share the objective of protecting those fundamental rights and freedoms: the GDPR mentions that the processing of natural personal data should "respect [data subjects'] rights and freedoms"[10]. Similarly, the AI Act states that its purpose is to "lay down a uniform legal framework for […] AI while ensuring a high level of protection of health, safety, **fundamental rights […]"**[11]**.**

Regarding **cybersecurity and robustness of AI systems**, the **NIS II Directive** states that "The availability of cyber-resilient network and information systems and the availability, confidentiality and integrity of data are vital […] for enhancing the trust of individuals and organisations in the Union's ability to promote and protect a global, open, free, stable and secure cyberspace **grounded in human rights**, fundamental freedoms, democracy and the rule of law."[12] This framework is further supported by the **Cybersecurity Act** and the forthcoming **Cyber Resilience Act**, both of which aim to **strengthen the EU's cybersecurity landscape**.

Assessing the FAITH AI_TAF through the lens of fundamental rights involves critically evaluating **whether its methodologies and outputs safeguard these rights effectively**. For example, the requirement to integrate mechanisms that ensure non-discrimination and uphold human dignity is not merely a best practice but a legal obligation under EU law. Those principles should be taken into account throughout the entire lifecycle of AI systems, influencing their design, implementation and oversight.

The significance of fundamental rights extends beyond mere legal compliance. Safeguarding fundamental rights is **integral to fostering public trust and ensuring the acceptability of AI technologies by the general public**[13] - making the safeguarding of fundamental rights a key

[10] GDPR, recital 1
[11] AI Act, recital 1
[12] NIS II, recital 70
[13] David Leslie and others, 'Human Rights, Democracy, and the Rule of Law Assurance Framework for AI systems: A Proposal' http://arxiv.org/abs/2202.02776 accessed 27 December 2024

component of developing and maintaining trustworthy AI systems. Furthermore, several domains directly concerned by FAITH LSPs have the potential to directly impact users' fundamental rights, as some of those rights are stated in Table 3[14].

*Table 3 - Fundamental Rights in the context of LSPs*

| Domain | Impact on Fundamental Rights (preliminary list) |
|---|---|
| **LSP 1 Media** | AI tools used in the context of content moderation must respect freedom of expression and information (Article 11 of the Charter). |
| **LSP 2 Transport** | AI systems used in monitoring transportation must respect the right to safety (Article 6 of the Charter) and to privacy (Article 8 of the Charter) of individuals and prevent discrimination (Article 14 of the Charter). |
| **LSP 3 Education** | AI systems used in adaptive learning systems must respect the right to education (Article 14 of the Charter) and prevent discrimination (Article 21 of the Charter). |
| **LSP 4 Port infrastructure** | AI systems used in the monitoring of maritime installations must respect the right to safety (Article 6 of the Charter) of individuals. |
| **LSP 5 Wastewater treatment** | AI systems used in the monitoring of waste-water treatment must respect the right to safety (Article 6 of the Charter) of individuals |
| **LSP 6 Healthcare** | AI systems used in diagnostics must respect the right to health (Article 35 of the Charter) and prevent discrimination (Article 21 of the Charter). |
| **LSP 7 Active ageing** | AI systems used for active ageing must respect the right to health (Article 35 of the Charter) and prevent discrimination (Article 21 of the Charter) and the right to privacy (Article 8 of the Charter). |

The FAITH AI_TAF should integrate these considerations holistically, treating fundamental rights as central to trustworthy AI. By embedding safeguards at every stage of AI lifecycle and aligning with the EU's vision of human-centric AI, the framework ensures that systems are not only technically robust but also societally legitimate, fostering trust across diverse contexts.

Other regulations might be relevant for the deployment of trustworthy AI that are not considered within this deliverable, such as copyright regulation for example. This does not mean those regulations do not play an important role, nor are they linked with exercising fundamental rights with the EU. However, they do not fit the scope of our assessment. As indicated in the initial section of this deliverable, this assessment does not replace a full compliance overview for the FAITH AI_TAF and the AI systems that relies upon the FAITH AI_TAF.

[14] Table 4 is a non-exhaustive list of potential human right impacts of FAITH's LSPs

## 2.3. Data protection Rules (GDPR)

### 2.3.1. Theoretical overview

The development and deployment of AI models resort to the processing of large amount of data. Data can be used both during the development of the model, as data is relied upon for training and during the deployment of the model when the model processes data to produce outputs. This processing of data can include personal data, e.g. "any information relating to an identified or identifiable natural person" (GDPR, Article 2(1)). In that case, the right to data protection, protected by Article 8 of the Charter, is to be respected. **It is therefore necessary to assess the conformity of the FAITH AI_TAF with data protection regulations.**

In the EU, data protection is governed by the **GDPR**, which has been in effect since May 2018. The GDPR applies to all entities that process the personal data of individuals within the EU or offer goods and services to EU residents, even if those entities are located outside the EU. Personal data, as defined by the GDPR, refers to any information that can identify a natural person, such as names, emails, and location data (GDPR, Article 4(1)). Also, anonymous data are excluded from its scope.

The GDPR establishes fundamental principles that must be followed when processing personal data:

1. **Lawfulness, Fairness, and Transparency**: Data processing must have a legal basis (Article 6) and be carried out fairly and transparently. Clear information must be provided to individuals about how their data will be used (Article 5(1)(a), Articles 12–14).
2. **Purpose Limitation**: Data must be collected for specified, explicit, and legitimate purposes and not processed further in ways incompatible with those purposes (Article 5(1)(b)).
3. **Data Minimization**: Only data that is necessary for the specified purposes should be collected and processed (Article 5(1)(c)).
4. **Accuracy**: Data must be accurate and kept up to date. Inaccurate data should be corrected or deleted promptly (Article 5(1)(d)).
5. **Storage Limitation**: Data should only be stored for as long as necessary to fulfil the purposes for which it was collected. Afterward, it must be deleted or anonymized (Article 5(1)(e)).
6. **Integrity and Confidentiality**: Personal data must be protected against unauthorized access, loss, or damage using appropriate technical and organizational measures (Article 5(1)(f)).
7. **Accountability**: Controllers must demonstrate compliance with GDPR principles and take responsibility for ensuring that all data processing activities meet its requirements (Article 5(2), Article 24).

Under the GDPR, organizations may also be required to conduct a **Data Protection Impact Assessment (DPIA)** for activities likely to pose high risks to individuals' rights and freedoms (Article 35). DPIAs involve identifying potential risks associated with data processing and implementing measures to mitigate them. They are particularly relevant when introducing new technologies, such as AI systems, that process sensitive or large-scale personal data (Article 35(3)).

The GDPR also grants individuals specific rights over their personal data, including:

- The **right to access** and obtain copies of their data (Article 15).
- The **right to rectification** of inaccurate or incomplete data (Article 16).
- The **right to erasure** or the "right to be forgotten," subject to certain conditions (Article 17).
- The **right to data portability**, allowing individuals to transfer their data between service providers (Article 20).
- The **right to restrict processing** (Article 18) and the **right to object** to specific types of data use (Article 21).

These principles and obligations have direct implications for the FAITH framework, particularly in assessing the trustworthiness of systems that process personal data. The inclusion of measures such as DPIAs ensures that privacy risks are identified and addressed proactively, reinforcing the protection of fundamental rights while fostering trust in AI technologies.

### 2.3.2. Data Protection Impact Assessment – DPIA

#### 2.3.2.1 Overview of existing DPIA Framework

Various frameworks and templates have been developed to facilitate the implementation of DPIAs across different sectors and use cases. Many **Data Protection Authorities (DPAs)** in the EU provide guidance and tools to support organizations that process personal data. These resources aim to ensure compliance with GDPR requirements while addressing specific risks associated with diverse processing activities (Article 35). Below are some examples:

- **European Data Protection Board (EDPB)**: The EDPB provides guidelines to determine whether a DPIA is necessary[15], including criteria such as large-scale data processing, profiling, or the use of new technologies.
- **French Data Protection Authority (CNIL)**: CNIL offers a comprehensive guide on Privacy Impact Assessment, including a summarization of the four essential phases of a DPIA: describing processing, assessing risks, mitigating risks, and formalizing outcomes[16]. An infographic describing these different steps has also been developed[17]. The CNIL further provides "how-to-sheets" directly aimed at AI professionals[18].
- **Irish Data Protection Commission (DPC)**: The DPC offers a DPIA template[19] that focuses on documenting decisions, engaging with data subjects, and ensuring compliance with GDPR principles such as transparency and accountability.

[15] Article 29 Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA)' (WP248 rev.01, 2017) https://ec.europa.eu/newsroom/article29/items/611236 accessed 27 December 2024

[16] CNIL, 'Privacy Impact Assessment (PIA) Methodology' (2018) https://www.cnil.fr/sites/cnil/files/typo/document/CNIL-PIA-1-Methodology.pdf

[17] CNIL, 'Guidelines on DPIA' https://www.cnil.fr/en/guidelines-dpia accessed 27 December 2024

[18] CNIL, 'AI how-to sheets' https://www.cnil.fr/fr/ai-how-to-sheets accessed 27 December 2024

[19] DPC, Sample DPIA Template, https://www.dataprotection.ie/en/organisations/know-your-obligations/data-protection-impact-assessments#sample-dpia-template accessed 25 February 2025

- **Spanish Data Protection Authority (AEPD)**: The AEPD offers a guide to managing risks to the rights and freedom of data subjects applicable to any processing activities[20]. The guide also contains guidelines for carrying out a DPIA for high-risk processing activities.
- **United Kingdom Information Commissioner's Office (ICO)[21]:**The ICO provides guidance on DPIAs[22], including specific information on when a DPIA is required[23], examples of high-risk processing activities[24] and a DPIA template[25].
- **Fraunhofer-Gesellschaft**: Fraunhofer provides extensive documentation for practically realizing DPIAs in their 2022 Conference Paper "Data Protection Impact Assessments in Practice"[26].

#### 2.3.2.2 In Practice

A DPIA can be divided in these five different steps: (i) the initiation phase, (ii) the preparation phase, (iii) the execution phase, (iv) the implementation phase and the (v) sustainability phase[27]. The data controller is responsible for conducting the DPIA (GDPR, art. 35 (1)), with the assistance of data processors if necessary (GDPR, art. 28 (3)(f)) and with the advice of the Data Protection Officer (GPDR, art. 35 (2)). It is recommended to include representatives of data subjects if possible.

**Step 1: The Initiation Phase**

The initiation phase consists of the **threshold assessment** to determine whether a DPIA is necessary[28].The assessment focuses on any personal data processing that, considering the nature, scope, context and purposes, is likely to result in a high risk to the rights and freedoms of natural persons. While innovative technologies, including AI, have a higher chance of meeting this threshold, this is not necessarily the case. **In the case of the development and/or**

[20]AEPD, 'Risk Management and Impact Assessment in the Processing of Personal Data' https://www.aepd.es/guides/risk-management-and-impact-assessment-in-processing-personal-data.pdf accessed 25 February 2025

[21] Despite leaving the EU in 2020, the UK still follows a data protection regime similar to the GDPR.

[22] ICO, 'Data Protection Impact Assessments (DPIAs)' https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/data-protection-impact-assessments-dpias/ accessed 24 January 2024

[23] ICO, 'When do we need to do a DPIA' https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/data-protection-impact-assessments-dpias/when-do-we-need-to-do-a-dpia/ accessed 24 January 2024

[24] ICO, 'Examples of processing 'likely to result in high risk'' https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/data-protection-impact-assessments-dpias/examples-of-processing-likely-to-result-in-high-risk/ accessed 24 Janaruy 2024

[25] ICO, 'Sample DPIA template' https://ico.org.uk/media2/migrated/2553993/dpia-template.docx accessed 24 Janaruy 2024

[26]Michael Friedewald and others, 'Data Protection Impact Assessments in Practice' (2022) https://publica.fraunhofer.de/handle/publica/417238 accessed 22 January 2025

[27] ibid.

[28] ibid.

**deployment of a high-risk AI system (see below, section on the AI Act), the processing of personal data is a strong indicator of a need for a DPIA**[29].

The EDPB identified **9 criteria** to assess whether a DPIA is necessary. If two (or more) of those criteria are met, a DPIA is presumed to be necessary. Those nine criteria are:

1) The collection of sensitive personal data;
2) The **large-scale** collection of personal data;
3) The collection of data from vulnerable persons, such as minors;
4) The crossing or combination of data sets;
5) **Innovative** uses or application of **new** technological or organizational solutions;
6) Processing of personal data **for evaluation or scoring**;
7) Automated-decision making;
8) systematic monitoring;
9) When the processing prevents data subjects from exercising a right or using a service or a contract

These activities present **high risks to the rights and freedoms of individuals** either because of the physical, material or non-material damage they may cause or because of their purpose, scope and nature[30].

This evaluation should be realised **before the processing activities** take place and should be **reviewed on a regular basis and/or if any changes** in the scope and/or purpose of processing take place.

**Step 2: The Preparation Phase**

The preparation phase involves **mapping the processing activities** by describing the processing information (scope, nature, purpose, means, etc.) and **planning** the actual execution of the assessment[31].

**Step 3: The Execution Phase**

The execution phase involves **consultation with the data subject (or their representatives)** if necessary, the identification and analysis of risks to individuals' rights, including privacy breaches or unauthorised access (GDPR, art. 35 (7)), risk assessment and the evaluation of the technical and organisational measures addressing the identified risks (GDPR, art. 35 (7)(d)) in order to calculate the remaining risk. The **necessity**[32] and **proportionality**[33] should be tested

[29]CNIL, 'Carrying Out Protection Impact Assessment If Necessary' https://www.cnil.fr/en/carrying-out-protection-impact-assessment-if-necessary accessed 27 December 2024

[30] Katerina Demetzou, 'Data Protection Impact Assessment: A tool for Accountability and the Unclarified Concept of "High Risk" in the General Data Protection Regulation' (2019) 35 Computer L & Security Rev 105342

[31] Friedewald (n 13)

[32] E.g. whether measures are necessary to achieve risk reduction

[33] E.g. whether measures are proportionate to the reduction achieved

at the time. If **high-risk remains**, the processing should either be abandoned or the relevant supervisory authorities should be contacted[34].

**Step 4: The implementation of the identified measures**

The implementation phase involves the implementation and testing of mitigating measures identified in the execution phase. These measures have to be tested and the tests recorded as proof of GPDR compliance. Once this is done, **the processing can begin.**

**Step 5: The Sustainability Phase**

The sustainability phase involves the **monitoring of the processing activities** in order to identify deviation or changes[35] that could lead to a review, adjustments and/or update of the DPIA (GDPR, art. 35 (11)).

#### 2.3.2.3 Relevance for FAITH

If the FAITH AI_TAF relies on the processing of personal data, it might be necessary to perform a DPIA if the use of the FAITH AI_TAF creates high risk for data subject's rights and freedoms.

Within the FAITH LSPs that are processing personal data, leveraging existing DPIA methodologies ensures a structured approach to managing risks across pilot domains. For example:

- In **healthcare** (LSP6), DPIAs are critical for ensuring the secure handling of sensitive health data while respecting patients' rights to privacy and dignity.
- In **education** (LSP3), DPIAs can address the risks of profiling and algorithmic bias in adaptive learning systems, ensuring non-discriminatory access to educational resources.
- In **active ageing** (LSP7), DPIAs support the responsible use of personal data in caregiving technologies, balancing data-driven insights with individual autonomy and confidentiality.

## 2.4. Cybersecurity obligations

### 2.4.1. Theoretical overview

As AI systems are cyber assets within ICT infrastructure, they are vulnerable to security risks, including threats from internal and external actors. The compromission of the security of AI systems can lead to risk for fundamental rights: for example, malicious actors might target AI systems to unlawfully obtain personal data or to influence the model. **Is it therefore necessary to assess the conformity of the FAITH AI_TAF with cybersecurity regulations.**

Cybersecurity obligations in the EU aim at ensuring **the resilience of digital infrastructures, safeguarding individual rights, and protecting critical sectors**. These obligations are particularly relevant for AI systems, which are considered **cyber assets within ICT infrastructures**. The following EU frameworks establish the primary cybersecurity obligations for systems and products:

[34] Friedewald (n 13)
[35] ibid

#### 2.4.1.1 NIS Directive and NIS II Directive

The **NIS Directive**[36] was the first EU-wide legislation on cybersecurity. It established baseline requirements for operators of essential services and digital service providers to secure their networks and systems.

Its successor, the **NIS II Directive**, expands the scope to include more sectors and introduces stricter requirements. Entities categorized as "**essential**" or "**important**" are required to adopt **technical and organizational measures** proportionate to their risks, ensure supply chain security, and report significant incidents to competent authorities (Articles 20–23). The NIS II Directive entered into force on 1 October 2024.

The directive applies to sectors like **healthcare, transport, and energy**, where cybersecurity is critical to operational continuity and safety.

#### 2.4.1.2 Cybersecurity Act

The **Cybersecurity Act (CSA)** reinforces the EU Agency for Cybersecurity (ENISA) and establishes an **EU-wide certification framework for ICT products, services, and processes**. It ensures compliance with standardized rules and fosters trust in digital technologies.

Certifications under the CSA may demonstrate compliance with the requirements of the AI Act (Article 15), particularly for high-risk AI systems, by verifying adherence to cybersecurity best practices (Articles 48–56).

#### 2.4.1.3 Cyber Resilience Act

The **Cyber Resilience Act** (CRA) introduces **common requirements for products with digital elements**, focusing on secure-by-design principles. Manufacturers, importers, and distributors must conduct cybersecurity risk assessments during design, development, and operation (CRA Articles 5–10).

The CRA requires **continuous monitoring, regular updates, and vulnerability management** throughout the product lifecycle. Products meeting CRA standards also satisfy AI Act Article 15.

### 2.4.2. In practice

**Step 1: Identifying Responsible Entities and Applicable Frameworks**

**NIS II Directive**: Applies to essential and important entities operating in critical sectors such as healthcare, transport, and energy. These entities are defined by Member States based on their operational size and sectoral importance (NIS II, Articles 6 and 21).

**Cybersecurity Act (CSA):** Provides a voluntary framework for cybersecurity certifications, applicable to ICT products, services, and processes. This is particularly relevant for demonstrating compliance with cybersecurity standards (CSA Articles 48–56).

**Cyber Resilience Act (CRA)**: Applies to products with digital elements, including connected devices and AI components. Manufacturers, importers, and distributors are responsible for

[36] Directive (EU) 2016/1148 concerning measures for a high common level of security of network and information systems across the Union [2016] OJ L194/1 (NIS Directive)

conducting cybersecurity risk assessments during design, development, and operational phases (CRA Articles 5–10).

### Step 2: Mapping system Components

Once the applicable frameworks are identified, the next step involves **defining the product or system's scope and mapping its components**. This includes:

- Identifying system functionalities and interdependencies within networks.
- Understanding how the system interacts with third-party services, supply chain partners, or other connected systems.
- For example, in **healthcare (LSP6)**, the mapping should include interactions between patient databases, diagnostics, and external service providers tools such as the tools used to identify potential vulnerabilities.

### Step 3 : Conducting a Cybersecurity Risk Assessment

A cybersecurity risk assessment evaluates potential threats to system security, data integrity, and operational resilience. This process includes:

- **Identifying risks**: Assess vulnerabilities such as unauthorized access, malware, or data breaches.
- **Analysing impacts**: Consider the consequences of a breach, such as disruption of critical services, financial losses, or harm to individuals' rights and freedoms.
- **Applying risk thresholds**: Evaluate whether the identified risks meet the thresholds for further action, as outlined in NIS II, CRA, or CSA requirements.

For example, under the **CRA**, manufacturers must perform vulnerability testing throughout the product lifecycle (CRA Article 7).

### Step 4: Implementing Risk Mitigation Measures

Mitigation measures should address both technical and organizational vulnerabilities. Examples include:

- **Technical measures**: Encryption, intrusion detection systems, and secure-by-design principles.
- **Organizational controls**: Staff training, access control policies, and incident response plans.
- **Supply chain security**: Ensuring that vendors and third-party providers meet cybersecurity standards, as required under NIS II (Article 21).
- These measures should be **proportionate to the risks** identified during the assessment.

### Step 5: Documenting the Process

All cybersecurity assessments must be thoroughly documented to demonstrate compliance with EU frameworks. Documentation can include:

- **Risk assessment reports** detailing identified vulnerabilities and their impacts.

- **Mitigation measures** implemented to address risks.
- **Ongoing monitoring plans** to ensure continued compliance.

**Step 6: Incident Detection and Reporting**

Mechanisms should be in place to monitor, detect and report cybersecurity incidents:

- Incident reporting is **mandatory for essential and important entities** under NIS II (Article 23). Reports must include the nature, scope, and impact of the incident, as well as mitigation actions taken.
- Organizations should ensure real-time monitoring systems are in place to identify potential threats early.

**Step 7: Ongoing Monitoring and Updates**

Cybersecurity is an **ongoing process** that requires continuous monitoring and adaptation:

- Regularly review and update cybersecurity measures to address emerging threats.
- Deploy timely updates and patches to mitigate newly identified vulnerabilities, as indicated in the CRA (Article 7).
- Revisit assessments periodically, especially after changes in scope, functionality, or regulatory updates.

### 2.4.3. Relevance for FAITH

The FAITH AI_TAF should ensure to **respect cybersecurity requirements**. The use of the FAITH AI_TAF should not create additional cybersecurity risks. FAITH should consider these cybersecurity obligations across its LSPSs to ensure compliance and enhance system resilience. The FAITH AI TAF aims to be compliant with ENISA Guidelines[37], providing a strong foundation for meeting cybersecurity requirements.

Particular attention should be brought to **LSP2 (transport), LSP4 (maritime transport), LSP5 (waste-water treatment)** and **LSP6 (health)** as they might concern critical infrastructures designed by the NIS II directive. For instance:

- **LSP4 (maritime transport)**: Implement measures to protect operational networks within maritime infrastructures from cyber threats.
- **LSP5 (wastewater treatment)**: Apply measures such as continuous monitoring, updates and vulnerability managements to secure products and systems managing wastewater treatment facilities.
- **LSP6 (health)**: Adopt encryption and secure access controls to safeguard sensitive health data while complying with NIS II and CRA requirements. Adopt measures to comply with the Medical Device Regulation[38].

[37] See Deliverable 2.1, p. 93

[38] Regulation (EU) 2017/745 on medical devices [2017] OJ L117/1 (Medical Device Regulation)b

## 2.5. Anti-discrimination regulation

### 2.5.1. Theoretical overview

The use of AI models can lead to discriminations. Biases, for example contained within the training data, can lead to unfair treatment of individuals. Those biases should be identified and corrected in order for AI models to produce fair results. **It is therefore necessary to assess the conformity of the FAITH AI_TAF with anti-discrimination regulations.**

Anti-discrimination regulation in the European Union (EU) lacks a unified legal framework comparable to the General Data Protection Regulation (GDPR). Instead, the EU offers a partially harmonized legal landscape that establishes minimum standards for non-discrimination, leaving member states the flexibility to adopt additional protections. This means that FAITH must comply with both EU-wide frameworks and any enhanced national laws when operating across different jurisdictions.

The EU Equality Legal Framework encompasses key directives addressing discrimination in areas such as employment, education, access to goods and services, and healthcare. Examples include:

- **The Race Equality Directive**[39]**,** which prohibits discrimination based on racial or ethnic origin in employment and access to services.
- **The Gender Goods and Services Directive**[40], which ensures equal treatment between men and women in access to goods and services.
- **The Employment Equality Directive**[41], which addresses discrimination in the workplace on grounds such as religion, disability, age, or sexual orientation.

Each member state builds upon these minimum standards, allowing for additional protected characteristics. For instance, **Sweden** includes protections against discrimination based on transgender identity, a characteristic not explicitly mentioned in all member states' laws.

Furthermore, it is relevant to note that the anti-discrimination framework in the EU goes further than merely complying with formal anti-discrimination laws. The horizontal character of the Charter needs to be considered, particularly regarding the right to non-discrimination (Article 21 of the Charter). Member States constitutions and constitutional traditions also have to be considered.

### 2.5.2. In Practice

**Step 1: Identifying Applicable Regulations**

- At the **EU level**, the Equality Framework serves as the baseline, ensuring protections against discrimination on grounds such as race, sex, religion, disability, or sexual orientation.

[39] Council Directive (EU) 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin [2000] OJ L180/22 (Race Equality Directive)

[40] Council Directive (EU) 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services [2004] OJ L373/37 (Gender Goods and Services Directive)

[41] Council Directive (EU) 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation [2000] OJ L303/16 (Employment Equality Directive)

- At the **national level**, member states may introduce additional protections. Those specific protections should be identified. Table 5 provides a non-exhaustive overview of the protection offered by EU countries.

### Step 2: Assessing the Potential for Discrimination

AI systems should be evaluated for potential discriminatory outcomes. Discrimination can take different forms, as highlighted in Table 4.

*Table 4 - Forms of discrimination*

| Type of Discrimination | Description |
|---|---|
| **Direct discrimination** | Occurs when an individual is treated less favourably due to a protected characteristic. |
| **Indirect discrimination** | Occurs when neutral rules, criteria or policies cause a disproportionate adverse impact on certain protected groups with no valid justification. |
| **Multiple discrimination** | Occurs when an individual is treated less favourably due to multiple, dissociable protected characteristics. |
| **Intersectional discrimination** | Occurs when an individual is treated less favourably due to multiple, indissociable protected characteristics. |
| **Harassment** | Occurs when an undesirable behaviour linked with a protected characteristic degrades a person's dignity and/or creates an intimidating environment. |
| **Injunction to discriminate** | Occurs when a behaviour encourages discrimination based on a protected characteristic |
| **Refusal to put in place reasonable accommodations** | Occurs when reasonable accommodations based on a protected characteristic are refused. It is mostly applicable for people with health conditions or impairments. |
| **Discrimination by association** | Occurs when an individual is treated less favourably due to its link with a protected characteristic. |
| **Discrimination based on an alleged criteria** | Occurs when an individual is treated less favourably due to an alleged protected characteristics. |

### Step 3: Implementing Anti-Discrimination Safeguards

Measures to prevent discrimination in AI systems often focus on addressing risks related to data, algorithms, and oversight mechanisms:

- Training data is reviewed to ensure it is representative and unbiased with respect to protected characteristics.
- Fairness assessments are regularly conducted to identify and address potential discriminatory outcomes in decision-making processes.

- Organizations establish compliance mechanisms, such as designating teams or officers to oversee the implementation of anti-discrimination practices.

**Step 4: Monitoring and Adapting**

Effective anti-discrimination practices require ongoing monitoring and adjustments:

- Impact assessments are carried out periodically to examine the effects of AI tools on protected groups and to ensure equitable treatment.
- Policy reviews are conducted to align practices with new regulatory developments or judicial rulings related to discrimination. For example, changes in case law may necessitate updates to risk assessment methodologies.

### 2.5.3. Relevance for FAITH

For FAITH, development should start with the EU Equality Framework as baseline compliance. However, to operate across the EU, the FAITH AI_TAF must address each member state's protected characteristics for equal treatment in goods and services. Providers must consider multiple discrimination types, such as direct, indirect or associative discrimination[42].

The FAITH AI_TAF should provide safeguards against discrimination and not create additional risks for discrimination.

Addressing anti-discrimination regulation is essential for FAITH's LSPs. The LSPs operate in contexts where risks of discrimination can manifest in various ways:

- **Healthcare (LSP6)**: Algorithms used for diagnostics or treatment planning are assessed to ensure that ethnicity or socio-economic status does not influence results unfairly.
- **Education (LSP3)**: Adaptive learning tools are designed to provide equitable access for students, including those with disabilities or from marginalized communities.
- **Transport (LSP2)**: Autonomous surveillance systems used within transportation should not discriminate on protected criteria.

[42] See Table 4

*Table 5 - Legally Protected Grounds in the Access to and Supply of Goods and Services in EU countries based on the Country Report non-discrimination 2023 compiled by the European Equality Law Network*

| Discrimination \ Country | AT | BE | BG | HR | CY | CZ | DK | EE | FI | FR | DE | EL | HU | IE | IT | LV | LT | LU | MT | NL | PL | PT | RO | SK | SI | ES | SE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Race & Ethnic origin | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Sex | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Age | X | ✓ | X | ✓ | X | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | ✓ | ✓ | ✓ | x | x | ✓ | x | ✓ | ✓ | ✓ | ✓ | ✓ |
| Ancestry | x | x | x | x | x | x | x | x | x | x | ✓ | ✓ | x | x | ✓ | x | x | x | x | x | x | ✓ | x | ✓ | ✓ | x | x |
| Disability | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Sexual orientation | ✓ | ✓ | X | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | x | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ |
| Religion or Belief | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ |
| Nationality | X | ✓ | X | X | X | X | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | x | ✓ | ✓ | x | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | x |
| Gender Identity | ✓ | ✓ | X | ✓ | X | ✓ | ✓ | X | X | ✓ | ✓ | ✓ | ✓ | ✓ | x | x | x | x | x | ✓ | ✓ | x | x | ✓ | ✓ | ✓ | ✓ |
| Family Status | X | ✓ | X | X | X | X | X | X | ✓ | ✓ | X | ✓ | ✓ | ✓ | x | x | x | ✓ | ✓ | ✓ | x | x | x | ✓ | X | X | X |
| Social Status | X | X | X | X | X | X | X | X | X | X | X | ✓ | ✓ | X | x | x | ✓ | x | x | x | x | x | ✓ | ✓ | ✓ | ✓ | x |
| Language | ✓ | ✓ | X | X | ✓ | X | X | ✓ | X | ✓ | ✓ | ✓ | ✓ | X | x | x | x | x | x | x | x | x | ✓ | ✓ | ✓ | ✓ | x |
| Health status | X | ✓ | X | X | X | X | X | ✓ | ✓ | ✓ | X | ✓ | ✓ | X | x | x | x | x | x | ✓ | x | x | ✓ | ✓ | x | ✓ | x |
| Political opinion | X | ✓ | X | X | X | X | X | ✓ | X | ✓ | ✓ | X | ✓ | X | x | x | x | ✓ | x | ✓ | ✓ | x | x | ✓ | ✓ | ✓ | x |
| Property | x | X | ✓ | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | x | x | x | x | ✓ | x | x | x |
| Belonging to a disadvantaged group | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | ✓ | x | x | x | x |
| Any other personal/ social characteristic | x | x | x | x | x | x | x | x | x | ✓ | x | x | ✓ | ✓ | x | x | x | ✓ | ✓ | x | x | x | x | x | ✓ | ✓ | x |
| Any other ground/criterion | x | x | x | x | x | x | x | x | ✓ | ✓ | x | x | ✓ | ✓ | x | x | x | ✓ | ✓ | ✓ | x | x | ✓ | x | x | x | x |

## 2.6. The EU AI Act

### 2.6.1. Theoretical overview

The **Artificial Intelligence Act,** published on 12 July 2024, introduces harmonized rules for the regulation of AI systems in the European Union. It aims to **protect fundamental rights** while **promoting innovation** and **fostering trust in AI technologies**. The Act entered into force on 1 August 2025. The AI Act **will be** applicable from **2 August 2026** (AI Act, art. 113), but some rules will be subject to earlier or later applicability. The Act also establishes a transitional period for certain AI systems placed on the market before the applicability of the AI Act (AI Act, art. 111). **Error! Reference source not found.** details the scheduled applicability of the Regulation.[43]

*Table 6 - Applicability of the AI Act*

| Date | Applicable Rules |
|---|---|
| **2 February 2025** | Application of rules for **prohibited AI systems** (AI Act, art. 113 (a))<br>AI Literacy (AI Act, art. 4) |
| **2 August 2025** | Application of rules for providers of **general-purpose AI models** (AI Act, art. 113 (b)) |
| **2 August 2026** | • **General applicability** of the AI Act (art. 113 AI Act)<br>• Starting from this date, **high-risk AI systems** that were placed on the market before August 2, 2026, must comply with the AI Act if **they undergo significant changes in their design**.(art. 111, 2) |
| **2 August 2027** | • Application of rules for **AI systems embedded into regulated products** (AI Act, art. 133 (c))<br>• **General-purpose AI models** placed on the market before 2 August 2025 have to comply with the AI Act (art. 111, 3) |
| **2 August 2030** | • High-risk AI systems intended to be used by public authorities have to comply with the AI Act (Art. 111, 2) |
| **31 December 2030** | • **AI systems which are components of high-risk large-scale IT systems** that were placed on the market before 2 August 2027 have to comply with the AI Act (Art. 111, 1) |

[43] The below table is not exhaustive, but contains the relevant information in the context of the FAITH Project. The complete timeline for the AI Act is available at the following address: https://artificialintelligenceact.eu/implementation-timeline/

The AI Act defines an **AI system** broadly as a "machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;" (AI Act, art. 3, (1)). This definition ensures flexibility, covering a wide array of current and emerging technologies. The Act applies to providers, and deployers of AI systems operating in the EU market, regardless of their geographic location. However, it explicitly excludes pure research and development activities unless such systems are deployed in real-world conditions, thereby ensuring that innovation and exploratory development are not hindered.

To address the diverse risks posed by AI systems, the AI Act adopts a **risk-based approach**, dividing systems into four categories based on the potential risks to health, safety, and fundamental rights:

1. **Unacceptable Risk**:

   AI practices deemed incompatible with EU values and fundamental rights are prohibited under the Act (AI Act, art. 5). These include:

   - **Subliminal, Manipulative and Deceptive Techniques (AI Act, art. 5 (1) (a))**: AI systems that exploit vulnerabilities in individuals' behaviour without their awareness to materially distort their actions.
   - **Exploitation of Vulnerabilities (AI Act, art. 5 (1) (b))**: AI systems that target individuals based on age, mental capacity, or other vulnerabilities to influence their decisions detrimentally.
   - **Social Scoring (AI Act, art. (1) (c)**: AI systems used by public authorities or private entities to systematically rank individuals based on their social behaviours when the system leads to an unfavourable or detrimental treatment of individuals or groups of individuals (i) in social contexts that are unrelated to the contexts in which the data was originally generated or collected or (ii) that is unjustified or disproportionate to their social behaviour or its gravity.
   - **Crime prediction and predictive policing (AI Act, art. (1) (d)):** AI systems that are based solely on the profiling of a natural person or on assessing their personality traits and characteristics in order to assess or predict the risk of a natural person committing a criminal offence.
   - Untargeted Scraping of Facial Images for the Creation or Expansion of Large-Scale Facial Databases (AI act, art. (1) (e)
   - **Emotion Recognition (AI act, art. (1) (f))**: AI systems used to infer emotions of a person in the areas of workplace and education institutions.
   - **Biometric Categorization (AI act., art. (1)(g)):** AI systems used to categorise persons based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation.
   - **Remote Biometric Identification systems (AI Act, art. (1)(h)):** AI systems used for the purposes of law enforcement relying on the use of 'real-time' biometric identification systems in publicly accessible spaces.

2. **High-Risk systems**:
   AI systems used in critical sectors or for safety components are subject to stringent requirements (AI Act, Annex III). These include applications in healthcare, education, public infrastructure, and employment. Providers of high-risk AI systems must comply with comprehensive obligations to ensure safety and accountability. These obligations include:
   - **Continuous Risk Management**: Providers must implement risk management systems to identify, monitor, and mitigate risks throughout the AI system lifecycle (AI Act, art. 9).
   - **High Data Quality Standards**: Training, testing, and validation datasets must be accurate, representative, and free from discriminatory biases (AI Act, art. 10).
   - **Transparency Requirements**: Providers must document and explain how the AI system works to ensure its decisions can be understood and audited (AI Act, art. 13).
   - **Human Oversight**: Measures must be in place to allow human operators to monitor and intervene in AI system operations when necessary (AI Act, art. 14).
   - **Technical Documentation**: Providers must create detailed records demonstrating compliance with the Act's provisions, including risk assessments and mitigation strategies (AI Act, art. 11).
   - **Conformity Assessment**: High-risk systems must pass a conformity assessment to verify compliance before being placed on the market (AI Act, art. 43–44).
3. **Limited Risk**:
   systems with limited risk are subject to transparency obligations to inform users about their interactions with AI (AI Act, art. 52). For example, users must be notified when interacting with AI chatbots or when content is AI-generated.
4. **Low/Minimal Risk**:
   AI systems that pose minimal risk to users' rights and safety face no specific obligations under the AI ACT, allowing for broad application without regulatory burdens. Examples include spam filters and AI-based recommendations.

The Act places particular emphasis on **high-risk systems**, recognizing their potential for significant societal and individual impact. Providers of these systems are required to maintain extensive documentation, ensure effective human oversight, and conduct pre-market conformity assessments. Furthermore, public entities, private entities providing public services deploying high-risk systems and deployers of high-risk AI systems used to evaluate a person's credit trustworthiness or to evaluate the risk and pricing in the case of life and health insurance must perform **Fundamental Rights Impact Assessments (FRIAs)** to evaluate risks to rights such as privacy, equality, and freedom of expression (AI Act, art. 27). These assessments align with the EU's broader commitment to ensuring that AI technologies respect human dignity and fundamental rights.

### 2.6.2. In Practice

**Step 1: Identifying Applicable Obligations**

Classification of AI systems under the AI Act determines their regulatory obligations:

- **Unacceptable Risk**: Includes prohibited practices such as subliminal manipulation and real-time biometric identification in public spaces (AI Act, Article 5).
- **High-Risk AI systems**: These include applications listed in Annex III, such as medical diagnostics, autonomous vehicles, and recruitment. High-risk systems require compliance with obligations such as risk management, conformity assessments, and human oversight (AI Act, art. 8–28).
- **Limited and Minimal Risk systems**: These are subject to transparency obligations, including notifying users about interactions with AI systems (AI Act, art. 52–54).

**Step 2: Establishing a Risk Management system – for high-risks AI systems not falling under an exception**

High-risk AI systems must implement a lifecycle risk management system (AI Act, art. 9). This includes:

- **Identifying Risks**: Evaluating potential threats to health, safety, and fundamental rights based on the AI system's intended use.
- **Risk Analysis**: Assessing the likelihood and severity of identified risks.
- **Mitigation Measures**: Developing technical and organizational measures to address risks, such as data security protocols and fairness audits.
- **Documentation and Monitoring**: Maintaining records of risk analyses and mitigation measures for transparency and compliance audits.

**Step 3: Conducting a Fundamental Rights Impact Assessment (FRIA) – mandatory for certain high-risk AI systems (AI act, art. 27)**

The FRIA evaluates potential impacts on fundamental rights, particularly for high-risk AI systems. FRIAs are only obligatory for public entities. However, from a trustworthy AI perspective, it is primordial to ensure the respect of fundamental rights by AI systems. This process involves:

- **Identifying Risks to Rights**: Analysing potential harms such as discrimination, privacy violations, and infringements on freedom of expression (AI Act, art. 27).
- **Stakeholder Involvement**: Consulting representatives from affected groups or civil society organizations to ensure diverse perspectives are considered.
- **Integration with Risk Management**: Aligning the FRIA with broader risk assessments to provide a holistic evaluation of risks and mitigation measures.
- **Documentation**: Recording the FRIA findings and proposed mitigation actions to ensure traceability and accountability.

**Step 4: Performing Conformity Assessments**

Conformity assessments verify that high-risk AI systems comply with the AI Act's requirements (AI Act, art. 43–44). This step includes:

- **Compliance Verification**: Demonstrating adherence to requirements such as transparency, data governance, and human oversight.
- **Certification**: Obtaining certification under EU-recognized standards, such as those outlined in the CSA framework.
- **Submission of relevant Technical Documentation**: Providing risk management records, FRIA findings, and evidence of compliance to authorities.

**Step 5: Monitoring and Post-Market Surveillance**

Post-market surveillance ensures that high-risk AI systems continue to comply with regulatory requirements after deployment (AI Act, art. 61). This involves:

- **Monitoring Performance**: Conducting regular evaluations of the AI system's functionality and impact on users.
- **Incident Reporting**: Reporting any serious malfunctions or violations of fundamental rights to regulatory authorities.
- **Updating systems**: Adapting AI systems to address newly identified risks, changes in legal requirements, or advancements in technology.

### 2.6.3. Overview of (selected) assessment framework

Several frameworks and resources provide guidance for conducting impact assessments and ensuring compliance under the AI Act. These frameworks offer practical guidance and methodologies for assessing risks and aligning with the regulatory requirements:

- **Knowledge Center Data & Society**: This resource offers a detailed template and instructions for implementing conformity assessments under the AI Act. It is particularly useful for high-risk AI systems, providing a structured approach to risk evaluation and mitigation.
- **Decision-tree based approach**: Hanif et al.[44] follows a decision-tree based approach to clearly define obligations under the AI Act, for ensuring accountability and fairness in AI systems, aligning with FRIA requirements and broader risk management processes.
- **Scenario-based approach**: Noveli et al.[45] focuses on a scenario-based approach towards AI Risk Assessment, emphasizing ethical considerations and compliance. Other methodologies, such as the Z-Inspection® framework, follow a similar method by relying on socio-technical scenarios[46].
- **European Data Protection Board (EDPB)**: The EDPB issued an opinion on certain data protection aspects related to the processing of personal data in the context of AI

[44] Hilmy Hanif and others, 'Navigating the EU AI Act Maze Using a Decision-Tree Approach' (2024) 1 ACM J Responsible Computing 21:1

[45] Claudio Novelli and others, 'AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act' (2023) https://papers.ssrn.com/abstract=4464783 accessed 27 December 2024

[46] 'Z-Inspection®: A Process to Assess Trustworthy AI' (2021) IEEE Xplore https://ieeexplore.ieee.org/document/9380498 accessed 27 December 2024

models[47] GDPR and AI models, highlighting the integration of privacy principles with AI risk assessments, particularly for systems requiring DPIAs and FRIAs.

- **Confederation of European Data Protection Organisations (CEDPO)**: The CEDPO issues a short guide on FRIA[48] and answers practical questions such as what it is, who should complete it, when should it be conducted, what the requirements are and how it interacts with DPIAs and other frameworks.
- **AI Office Methodology (forthcoming)**: A methodology under development by the European Commission's AI Office aims to provide a unified framework for risk management and conformity assessments. It is expected to align closely with the AI Act's provisions in the coming months.

#### 2.6.4. Relevance for FAITH

FAITH's objective to improve AI Trustworthiness means that the compliance of the FAITH AI_TAF with the AI Act must be observed. Not only regulatory compliance but also respect of the principles of the Act should be considered. This is to be considered not only for legal compliance reasons, but also because the AI Act share the objective of creating more trustworthy AI systems with FAITH.

For FAITH's LSPs, compliance with the AI Act is essential to ensure trust and accountability in its AI systems. Examples include:

- **Education (LSP3)**: Adaptive learning platforms powered by AI should **avoid biases** that could disadvantage students from underrepresented communities. These systems require transparency measures to explain AI-driven decisions and support human intervention where necessary.
- **Transport (LSP2)**: AI systems used to analyse passenger behaviour and manage crowd flow should undergo a Fundamental Rights Impact Assessment (FRIA) to mitigate risks such as **privacy violations or discriminatory flagging of individuals**. The system must also adhere to transparency and human oversight obligations to avoid over-reliance on AI-driven decisions.
- **Healthcare (LSP6)**: AI models recommending treatment plans must demonstrate **compliance with data quality, accuracy, and risk management requirements**. For instance, conformity assessments and robust human oversight mechanisms must ensure that the system is free from bias and supports equitable healthcare access.

### 2.7. Ethical Obligations

#### 2.7.1. The High-level expert group on Artificial Intelligence

Trustworthy AI extends beyond mere compliance with laws and regulations. The High-Level Expert Group on AI (HLEG AI), in its *Ethics Guidelines for Trustworthy AI*, argues that achieving

[47] European Data Protection Board, 'Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models' (2024) https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en accessed 27 December 2024

[48] Confederation of European Data Protection Organisations 'Fundamental Rights Impact Assessment : What are they? How do they work?' https://cedpo.eu/wp-content/uploads/CEDPO-micro-insight-paper-fundamental-rights-impact-assessments.pdf accessed 24 January 2024

trustworthiness requires AI systems to be lawful, ethical, and robust. These three pillars are mutually reinforcing and must coexist throughout the lifecycle of AI systems.

### 2.7.2. The Fundamental Rights-Based Approach

The HLEG AI emphasises a **fundamental rights-based approach** as the foundation for **ethical and robust AI**. As stated in the Guidelines, "respect for fundamental rights, within a framework of democracy and the rule of law, provides the most promising foundation for identifying abstract ethical principles and values, which can be operationalised in the context of AI." This framework ensures alignment with the **EU Charter of Fundamental Rights and international human rights law**, grounding the ethical principles in universally recognised standards of **dignity, freedom, and equality**[49].

### 2.7.3. Ethical Principles and Key Requirements

The HLEG on AI outlines four ethical principles critical to ensuring ethical and robust AI systems:

1. **Respect for Human Autonomy**
2. **Prevention of Harm**
3. **Fairness**
4. **Explicability**

These principles are operationalised through seven key requirements for Trustworthy AI:

1. **Human Oversight**: Ensuring AI systems support human autonomy and decision-making.
2. **Technical Robustness and Safety**: Preventing harm through robust system design.
3. **Privacy and Data Governance**: Protecting individual privacy and ensuring data integrity.
4. **Transparency**: Facilitating traceability and explainability of AI decisions.
5. **Diversity, Non-Discrimination, and Fairness**: Mitigating bias and ensuring equitable access.
6. **Societal and Environmental Well-being**: Supporting broader societal goals and minimising environmental impact.
7. **Accountability**: Establishing mechanisms for redress and minimising risks through proactive governance.

These principles should underpin the FAITH AI_TAF. This deliverable focuses on whether the FAITH AI_TAF respects those requirements and enables LSPs to operationalise these requirements effectively, ensuring adherence to the HLEG AI's standards.

### 2.7.4. The Assessment List for Trustworthy AI

The Guidelines were then further revised and translated into an **accessible and dynamic checklist** to guide developers and deployers of AI systems in implementing the principles in

[49] Nathalie A Smuha, 'The Work of the High-Level Expert Group on AI as the Precursor of the AI Act' (2024) https://papers.ssrn.com/abstract=5012626 accessed 19 December 2024

practice. The process involved discussion with over 350 stakeholders[50] and ended with the production of the **Assessment List for Trustworthy AI** (ALTAI), a 34-pages document to be used as a self-assessment checklist. The document describes the **seven requirements** elaborated by the HLEG and contains a list of questions for each requirement, as well as introductory guidance. A glossary is also available for relevant definitions and the assessment is also available as a web-based tool[51].

### 2.7.5. Assessing the FAITH AI_TAF through the HLEG AI Lens

Lawfulness is assessed by evaluating whether the FAITH AI_TAF aligns with regulatory obligations, such as those under the AI Act and GDPR. The FAITH AI_TAF should also integrate robustness by respecting cybersecurity obligations under the NIS II Directive, the Cybersecurity Act, and the Cyber Resilience Act. These standards ensure that security measures are embedded in the design of AI systems.

Ethical considerations are more expansive, demanding the FAITH AI_TAF to ensure:

- Respect for **fundamental rights**, particularly non-discrimination and individual autonomy.
- Transparency and **explainability mechanisms** allow developers and users to understand AI systems' decision-making processes.
- Adhere to the seven requirements for a trustworthy AI.

The FAITH AI_TAF's adherence to the HLEG AI's framework will be evaluated based on its capacity to foster trustworthiness through compliance, robustness, and ethical soundness, directly reflecting the principles of human-centric AI development. Relying on the ALTAI, this deliverable ensures a comprehensive understanding of ethical challenges related to AI systems.

Furthermore, additional ethical frameworks for specific sectors provide additional guidance and requirements to be followed. This is particularly relevant for FAITH LSPs for which those frameworks exist, as shown in Table 7[52].

*Table 7 - Additional Ethical Frameworks for LSPs*

| Domain | Additional Ethical Framework |
|---|---|
| **LSP 1 Media** | • The Resolution 1003 of the Parliamentary Assembly of the Council of Europe on the Ethics of Journalism adopted on 1 July 1993[53]. |

[50] European Commission, 'Assessment List for Trustworthy Artificial Intelligence (ALTAI) Self-Assessment' https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment accessed 27 December 2024

[51] European Commission, 'ALTAI - Assessment List Trustworthy Artificial Intelligence' https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence accessed 27 December 2024

[52] Table 8 contains a non-exhaustive list of ethical frameworks that could be considered by LSPs when assessing the trustworthiness of their AI systems.

[53] Council of Europe Parliamentary Assembly Resolution 1003 on Ethics of Journalism (1 July 1993) https://assembly.coe.int/nw/xml/XRef/Xref-XML2HTML-en.asp?fileid=16414

| | |
|---|---|
| | • Relevant national code of conduct or ethics[54]. |
| **LSP 3 Education** | • The Council of Europe Platform on Ethics, Transparency and Integrity in Education (ETINED) provides several guidance, such as ethical principles[55], ethical behaviour for actors in education[56] and a compendium of best practices to promote academic integrity[57]. |
| **LSP 6 Healthcare** | • The WHO guidance on Ethics & Governance of Artificial Intelligence for Health published on 28 June 2021[58].<br>• The World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Participants adopted in June 1964[59]. |
| **LSP 7 Active ageing** | • The WHO report on developing an ethical framework for health ageing[60]. |

[54] A list of 55 ethical codes from 45 countries is available here: https://www.presscouncils.eu/ethical-codes-database/codes/

[55] Council of Europe, 'Ethical Principles' https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806c90cd

[56] Council of Europe, 'Ethical Behaviour for Actors in Education' http://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806d2b6f

[57] Council of Europe, 'Best Practices in Academic Integrity' https://rm.coe.int/bpp-a-compendium-of-best-practices-eng-/1680a86621

[58] WHO, 'Ethics & Governance of Artificial Intelligence for Health' (2021) https://www.who.int/publications/i/item/9789240029200

[59] World Medical Association, 'Declaration of Helsinki' (1964) https://www.wma.net/policies-post/wma-declaration-of-helsinki/

[60] WHO, 'Developing an Ethical Framework for Health Ageing' (2017) https://www.who.int/publications/i/item/WHO-HIS-IER-REK-GHE-2017.4

## 3. The How-To: instructions for the assessment

### 3.1. Who Will Carry Out or Be Involved in the Legal and Regulatory Assessment of FAITH AI_TAF

Completing the questionnaire-based assessment described in this deliverable is a collaborative exercise that requires the **participation of multiple stakeholders** to ensure a comprehensive evaluation of the FAITH AI_TAF. A **holistic perspective** is essential to adequately assess the efficiency and relevance of the FAITH AI_TAF, integrating the **expertise and insights of all partners.**

The full questionnaire is to be completed by the partner **to the best of their expertise and capabilities**[61]. This approach guarantees that the responses are comprehensive, accurate and informed by firsthand knowledge. Some questions may call for **collaboration** among several partners, while others may necessitate input from the entire consortium.

To facilitate the filling of the questionnaire, partners are divided in **four clusters**. Each question of the questionnaire will refer to the most appropriate partner for answering the question. Table 8 describes the clusters.

*Table 8 -Cluster of partners for the assessment*

| Cluster | Partners |
|---|---|
| **Coordinating Partner** | • Idryma Technologias Kai Erevnas - FORTH |
| **Technical Partners** | • Trustilio<br>• University of Southampton - UoS |
| **LSP Partners** | • Trustilio<br>• Athens Technology Center - ATC, Freedom House Romania - FH (Media LSP)<br>• Consiglio Nazionale delle Ricerche (CNR), MERMEC (Transportation LSP)<br>• Erevnitiko Panepistimiako Institouto Systimation Epikoinonion Kai Ypologiston - ICCS, Ellinogermaniki Agogi Scholi Panagea Savva AE - EA (Education LSP)<br>• Fundacion de la Comunidad Valenciana para la investigacion, promocion y estudios comerciales de Valenciaport - VFP, APRA (Port infrastructure LSP)<br>• SINTEF, Veas Selvkost AS (Waster-water treatment LSP)<br>• FORTH, Universita degli Studi di Firenze – UNIFI, Universita de Pisa – UNIPI (healthcare LSP)<br>• Active Ageing Association - AOA, BRIDG OU (active ageing LSP) |
| **Legal Partners** | • Katholieke Universiteit Leuven – KU Leuven, author of the questionnaire |

[61] As specified in section 5.3, questions that cannot be answered can be left blank

The FAITH AI_TAF is itself an assessment framework intended to be used with a different range of AI models. As such, parts of the questionnaire will provide little information when considering the FAITH AI_TAF from a general perspective. While it is still important to collect this information, it is beneficial to also consider the impact of the FAITH AI_TAF in more concrete settings.

To do so, when LSP Partners answer the questionnaire, they should consider not only the FAITH AI_TAF itself, but as well **the (expected) impact of the FAITH AI_TAF on the trustworthiness** of the AI system(s) relied upon in the context of their specific pilot. For example, if the use of the FAITH AI_TAF results in the implementation of measures reducing certain legal and/or ethical risks identified in the questionnaire, this should be clearly indicated. This will allow to gather more information on the practical impacts of the FAITH AI_TAF. However, it is important to remember that the questionnaire does not aim at assessing the LSP system under test itself but rather the application of the FAITH AI_TAF within the LSP.

**KU Leuven will oversee the assessment**, using project documentation prepared by the relevant partners to ensure the questionnaire is completed methodically. This approach facilitates the smooth integration of contributions, fostering a coherent and structured process that reflects the shared expertise of the consortium.

### 3.2. When Will the AI system(s) impact assessment of the FAITH_AI TAF Be Conducted?

The assessment will take place in an **iterative manner** throughout the course of the FAITH project. The first phase will start following the validation of the questionnaire **after March 31st 2025**. Once all partners have answered the assessment, KU Leuven will congregate the results.

The results of the questionnaire will then be periodically reviewed throughout the project. The second iteration of the legal and regulatory assessment for the FAITH AI_TAF will involve the relevant partners through **a legal workshop** and will serve as to complete and update the information which may not be available at the stage of the first iteration.

KU Leuven **may update or adapt the current assessment** after the second iteration, depending, for instance, on the result of the first two iterations which may raise new points of attention or on developments in the field which need to be taken into consideration (e.g. new assessment methodology provided by the AI Office). The final assessment of the FAITH AI_TAF will be presented in **Deliverable 1.4** (Final report on the legal and ethical self-assessment and the activities of the EEAB) **on 31 March 2028**.

### 3.3. How Will the AI system(s) impact assessment of the FAITH_AI TAF Be Conducted in Practice?

The legal and regulatory assessment will consist of **a questionnaire conducted by KU Leuven**. The questionnaire will consist of a series of questions, divided in the following categories : **general information, AI information and identification of legal risks**. The subsequent section of the deliverable outline the questionnaire and provides additional information regarding its answering.

**All partners** will answer the questionnaire in order to **maximise the amount of information** obtained to assess the FAITH AI_TAF. Partners of the same LSP are encouraged to collaborate, delivering one questionnaire for the LSP instead of one for each partner. The questionnaire is

to be filled out to the best of the **Partner's knowledge and expertise** by taking into account the designation of the preferred cluster for each question[62]. **Partners are invited to answer as many questions as they can, but leaving answers blank is also a possibility**.

The questionnaire is composed of **168 questions, divided in 3 different sections and 13 sub-sections**. Table 9 describes the division of the questions and the estimated time[63] allocated for each section. The questionnaire does not need to be answered entirely at the same time and different persons within each organisation can answer different questions and/or sections. In the third section of the questionnaire, 'Legal and Ethical Risk Identification', each sub-section is divided in a legal and an ethical part.

*Table 9 - Division of the questionnaire*

| | Section | Number of questions | Estimated time to complete (in minutes) |
|---|---|---|---|
| **General Information** | **General Information** | 4 | 30 |
| | **Privacy Information** | 4 | 20 |
| | **Technological Information** | 27 | 120 |
| **AI Information** | **AI Act – General Information and Application** | 13 | 60 |
| | **AI Act - Classification** | 26 | 120 |
| | **AI Act – Models Interaction** | 5 | 25 |
| **Legal and Ethical Risk Identification** | **Human Agency and Oversight** | 12 | 60 |
| | **Technical Robustness and Safety** | 21 | 100 |
| | **Privacy and Data Governance** | 16 | 80 |
| | **Transparency** | 9 | 45 |

[62] See Section 5.1 of this Deliverable

[63] Current numbers are estimated. The next version of the Deliverable (Deliverable 1.4) will be able to gather feedback from Partners to improve the estimated time necessary to answer questions.

| | | | |
|---|---|---|---|
| | **Diversity, Non-Discrimination and Fairness** | 11 | 55 |
| | **Societal and environmental well-being** | 10 | 50 |
| | **Accountability** | 10 | 50 |

The questionnaire contains two separate cells to answer the question: a first cell "Answer" and a second cell "Explanation". A short and concise answer should be given in the answer cell, while the 'Explanation" cell allows for a more thought-out description. It is essential to include as much information as possible in the explanation cell. For **questions related to ethics**, a detailed description of the thought process leading to the answer is to be provided as best as possible. If possible, concrete examples should be relied upon.

### 3.4. The review by the External Ethics Advisory Board

The External Ethics Advisory Board have reviewed the first version of the deliverable, including the description of the assessment and the legal questionnaire during a workshop organised on February 17, 2025 by KU Leuven. The Deliverable has been updated to integrate the comments presented during the meeting. Additional feedback, provided by the EEAB in a written form, will be incorporated in the next version of the Deliverable (D1.4).

# 4. Assessment template for the FAITH AI_TAF

## 4.1. Description of the FAITH AI_TAF

The first section of the questionnaire aims at providing **an overview of the FAITH AI_TAF** to assess which are the applicable regulations. The purpose of this first part is to gather all pertinent information to describe the FAITH AI_TAF and its components from a technological point of view, including the intended interactions of the FAITH AI_TAF with third-party AI systems (e.g. the models LSPs are relying upon in their operations). The information contained in the **AI Model Passport**[64] proposed within the FAITH AI_TAF can serve as a blueprint for answering the questionnaire.

### 4.1.1. General Information

The first set of questions, that can be found in the Spreadsheet Questionnaire in the sheet "General Information" aims at collecting **information related to the purpose and objective of the FAITH AI_TAF**.

*Table 10 - General Information Questions*

| Question N° | Preferred cluster | Question |
|---|---|---|
| 1 | Coordinating and Technical Partners | What is the problem that the FAITH AI_TAF aims to address? |
| 2 | | Based on the problem definition, what is the intended purpose of the FAITH AI_TAF? |
| 3 | | Based on the problem definition, which are the objectives of the FAITH AI_TAF? |
| 4 | | Describe in plain language and in general terms how the FAITH AI_TAF is intended to operate. Graphic representation can be used. |

### 4.1.2. Privacy – General Information

The second part of the assessment focuses on the implication of **data protection and privacy within the FAITH AI_TAF**. This part of the assessment does not focus on a full GPDR-conformity analysis, but serve as a starting point to clearly understand what type of personal data processing is taking place while using the FAITH AI_TAF. It is found in the Spreadsheet Questionnaire under the "Privacy – General Questions" sheet.

In order to answer these questions, we refer to the section of this deliverable focusing on Data protection Rules (GDPR). Notably, in case of further questions regarding the assessment of whether the data processing activities is likely to result in high risk for data subjects, **we refer to the EDPB Guidelines.**

[64] See Deliverable 2.1, p. 76

*Table 11 - General Privacy Questions*

| Question N° | Preferred cluster | Question |
|---|---|---|
| 5 | Technical Partners | Does the use of the FAITH AI_TAF involve the processing of personal data? |
| 6 | | When using the FAITH AI_TAF, who is the data controller? |
| 7 | | When using the FAITH AI_TAF, who is the data processor? |
| 8 | Technical and Legal Partners | Is the processing of personal data in the FAITH AI_TAF likely to result in a high risk to the rights and freedoms of data subjects considering its nature, scope, context and purposes? |

### 4.1.3. General Technological Information

The final part of the general questionnaire aims to gather **technical information** regarding the FAITH AI_TAF.

*Table 12 - General Technological Questions*

| Question N° | Question Category | Preferred cluster | Question |
|---|---|---|---|
| 9 | Model information | Technical Partners | Which AI models/types of AI models are used within the framework of the FAITH AI_TAF? |
| 10 | | | Will the AI model(s) to be used within the framework of the FAITH AI_TAF be developed throughout the project and/or will existing AI models be used? |
| 11 | | | If the FAITH AI_TAF does not rely on an AI model, what technology does the FAITH AI_TAF use? |
| 12 | Data Information and Quality | | Will the FAITH AI_TAF rely on personal data processing? How will the FAITH AI_TAF detect and react if the user shares personal data? |
| 13 | | | What are the sources of the input data that will be used for the FAITH AI_TAF? |
| 14 | | | What, if any, training, validation and testing data will the FAITH AI_TAF use? You can rely on examples to describe each category of data with examples |

| | | | |
|---|---|---|---|
| | | | |
| 15 | | | What are the sources of the training, validation and testing data that will be used in FAITH AI_TAF? |
| 16 | Output : (un)intended uses and human supervision | | What is the intended output produced within the FAITH AI_TAF? What will the users receive as final output? |
| 17 | | | What is the intended use of this output within the FAITH AI_TAF? Is the output intended to be used directly by the participant directly receiving it? |
| 18 | | | Which expertise is expected by the humans receiving the output of the FAITH AI_TAF in order to use it in accordance with its intended purpose? |
| 19 | | | Will the reasons behind the output of the FAITH AI_TAF be explained to the human(s) directly receiving it? |
| 20 | | Technical and Legal Partners | Are there any foreseeable risks of misuse of the FAITH AI_TAF ? If so, which one? |
| 21 | Deployment Setting | LSP and Technical Partners | Is there a specific setting where the FAITH AI_TAF is intended to be deployed, and/or is it intended as a technology that is adaptable to several different settings? |
| 22 | | | If there is one or a limited set of specific settings where the FAITH AI_TAF is intended to be deployed, what are the specific features of those settings? |
| 23 | | | What are the geographical area and language context in which the FAITH AI_TAF will be deployed? |
| 24 | | | If the FAITH AI_TAF is intended to be used in different settings, are there any plans to adapt the technology depending on the deployment of the FAITH AI_TAF? If yes, describe this plan. |

| | | | |
|---|---|---|---|
| **25** | | | Describe the interested parties involved in the design / development of FAITH AI_TAF and the specifics of their involvement. Refer to documents developed in the context of the project. |
| **26** | | | Describe the impact the involvement of each group of interest parties has or may have on the design and/or development of the FAITH AI_TAF. |
| **27** | | | For any specific setting the FAITH AI_TAF is intended to be deployed : |
| **27a** | | | Describe the interested parties directly involved in the deployment of the FAITH AI_TAF? |
| **27b** | | | Describe the parties not directly involved, but potentially affected by the deployment of the FAITH AI_TAF? |
| **27c** | | | Is there any foreseeable impact the FAITH AI_TAF can have on a wider community? |
| **28** | | | If the FAITH AI_TAF is intended to be used in different settings : |
| **28a** | | | Are there any plans to adapt the FAITH AI_TAF depending on the expected parties (individuals and/or groups of individuals) directly involved in the deployment of the system? |
| **28b** | | | Are there any plans to adapt the FAITH AI_TAF depending on the expected parties (individuals and/or groups of individuals) not directly involved, but affected by the deployment of the system? |
| **28c** | | | Are there any plans to adapt the technology depending on the foreseeable wider community on which the FAITH AI_TAF can have an impact? |
| **28d** | | | If yes, describe these adjustments. If no, justify why such adjustments are not needed. |

## 4.2. AI Act Assessment

This next section assesses whether or not the FAITH AI_TAF or its use by LSPs owners result in the applicability of the AI Act. The subsequent set of questions therefore aims at establishing (1) whether the FAITH AI_TAF correspond to **the definition of artificial intelligence within the AI Act**, e.g. "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments[65]".

*Table 13 - AI Act - General Information*

| Question N° | Preferred cluster | Question |
|---|---|---|
| 29 | Technical Partners | Does the FAITH AI_TAF includes components that qualify as an AI system? |
| 30 | | Is the FAITH AI_TAF composed of a chains of different AI systems? If yes, why? |
| 31 | | Is the FAITH AI_TAF based on (one or more) general-purpose AI models (GPAI models)? |
| 32 | | Will the FAITH AI_TAF be 'placed on the market' or 'put into service' during the project lifecycle? |
| 33 | | When will the FAITH AI_TAF be 'placed on the market' or 'put into service'? |
| 34 | | Is the FAITH AI_TAF used throughout the project research, testing or development activities that take place before the FAITH AI_TAF is placed on the market or put into service? |
| 35 | | Who are the providers of the FAITH AI_TAF |
| 36 | | Who are the deployers of the FAITH AI_TAF? |
| 37 | | Are the providers of the FAITH AI_TAF established in the EU? |
| 38 | | Are the deployers of the FAITH AI_TAF established in the EU? |
| 40 | | Are the expected interested parties of the FAITH AI_TAF (see question 27) located in the EU? |
| 41 | Legal Partner | Based on the information above, is the FAITH AI_TAF in the scope of application of the AI Act? |
| 42 | Consortium | Given the overall objective of the FAITH AI_TAF regarding AI Trustworthiness, which aspects of AI Act compliance and/or AI HLEG Requirements should be observed by the FAITH AI_TAF on a voluntary basis? |

[65] AI Act Act, article 3, (1)

It is important not only to assess whether the FAITH AI_TAF or its use within a LSP enters the scope of the AI Act, but also to **correctly classify the FAITH AI_TAF within the AI Act itself**. The following questions aim at classifying the FAITH AI_TAF within the categories of the AI Act.

*Table 14 - AI Act Classification*

| Question N° | Question Category | Preferred Cluster | Question |
|---|---|---|---|
| 43 | Prohibited AI system | Technical Partners | Does the FAITH AI_TAF deploy subliminal, manipulative and/or deceptive techniques? |
| 44 | | | Does the FAITH AI_TAF exploit vulnerabilities of natural persons' based on their age, disability or socio-economic situation? |
| 45 | | | Does the FAITH AI_TAF create or expand facial recognition database through the untargeted scrapping of facial images? |
| 46 | | | Does the FAITH AI_TAF result in social scoring of individuals by public authorities? |
| 47 | | | Does the FAITH AI_TAF rely on emotion recognition and manipulation in the workplace and/or education institutions? |
| 48 | | | Are there any safeguards against using the FAITH AI_TAF as a prohibited AI beyond its intended purpose? |
| 49 | | Legal Partner | Does the FAITH AI_TAF fit any of the prohibited uses of AI described in the AI Act?<br><br>If yes, then the use of the FAITH AI_TAF is prohibited within the EU if the FAITH AI_TAF classifies as a prohibited AI system |
| 50 | High-Risk AI | Technical and LSP Partners | Is the FAITH AI_TAF intended to be embedded within a product, meaning the system is physically integrated into a product? |
| 51 | | | Is the FAITH AI_TAF intended for remote biometric identification systems and not used for the sole purpose of verifying a person's identity? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| 52 | | | Is the FAITH AI_TAF intended for a safety component of a critical infrastructure? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |

| | | | |
|---|---|---|---|
| **53** | | | Is the FAITH AI_TAF intended for use in education and vocational training? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| **54** | | | Is the FAITH intended for use in employment, workers management and access to self-employment? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| **55** | | | Is the FAITH AI_TAF intended for use in the context of the access to and enjoyment of essential private services and essential public services and benefits? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| **56** | | | Is the FAITH AI_TAF intended for use in the context of law enforcement? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| **57** | | | Is the FAITH AI_TAF intended for use in the context of migration, asylum and border control management? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| **58** | | | Is the FAITH AI_TAF intended for use in the context of the administration of justice and democratic processes? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| **59** | | | Is the FAITH AI_TAF developed and put into service for the sole purpose of scientific research and development? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| **60** | | | Is the FAITH TAIF performing a narrow procedural task? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| **61** | | | Is the FAITH AI_TAF intended to improve the result of a previously completed human activity? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| **62** | | | Is the FAITH AI_TAF intended to detect decision-making patterns and is not meant to replace or influence previously completed human assessments without proper human review? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| **63** | | | Is the FAITH AI_TAF intended to perform a preparatory task for an activity described in questions 50-58bis ? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |

| | | | |
|---|---|---|---|
| 64 | | | Is the FAITH AI_TAF intended for use in an activity described in questions 50-58bis **and** performs profiling of natural persons? In the context of an LSP, will the FAITH AI_TAF be connected to such usage? |
| 65 | | Legal Partner | Does the FAITH AI_TAF fit any of the high-risk categories of AI described in the AI Act? |
| 66 | | | Does the FAITH AI_TAF benefit from an exception for classifying as high-risk under the AI Act? |
| 67 | AI with transparency obligations | Technical Partners | Is the FAITH AI_TAF directly interacting with persons? |
| 68 | | | Does the FAITH AI_TAF generate AI synthetic content? |

Finally, AI systems are usually part of a larger network of components and their interaction and relationships should be taken into account within this assessment. This should be considered for the FAITH AI_TAF.

*Table 15 - AI Models Interactions*

| Question N° | Preferred cluster | Question |
|---|---|---|
| 69 | Technical Partners | Describe how the FAITH AI_TAF is intended to interact with third-party AI systems. |
| 70 | | Describe the information that will be made available to users of the FAITH AI_TAF within their own AI systems. |
| 71 | | In the course of the FAITH project, will the providers/deployers of the third-party AI systems be able to modify to FAITH AI_TAF intended purpose? |
| 72 | | Will it be possible to prevent a third party AI system from using the FAITH AI_TAF in a manner that is not compliant with legal and ethical considerations? |

| | | |
|---|---|---|
| **73** | LSP and Legal Partners | What are the advantages brought by the FAITH AI_TAF when it comes to legal and ethical considerations? |

## 4.3. Legal and Ethical Risk Impact Assessment

The second part of the assessment focuses on **legal and ethical** risk. The goal of this section is to identify regulatory risks arising within the FAITH AI_TAF. The final goal is to assess the overall impact of the FAITH AI_TAFs for LSPs, while ensuring that the FAITH AI_TAF itself do not create any additional legal, regulatory or ethical risks.

Doing so requires analysing those risks towards specific requirements for trustworthy AI. As the FAITH AI_TAF aims at improving AI Trustworthiness within the seven key requirements of AI, we will input the questionnaire with a high-level overview of risks within **the seven characteristics of trustworthy AI highlighted by the HLEG: (i) human oversight, (ii) technical robustness and safety, (iii) privacy and data governance, (iv) transparency, (v) fairness (including non-discrimination and diversity), (vi) social and environmental well-being, (vii) accountability.**

Here, it is important to consider **that legal and ethical obligations are not always entirely similar**. While compliance with regulatory obligations ensures ethical AI, it is not always sufficient to merely comply with regulations such as the AI Act. While legal compliance ensures AI meets minimum regulatory standards, ethical responsibility considers the larger societal impacts of AI.

### 4.3.1. Legal Risks

In the context of legal analysis, risks pertain to **protecting rights and interests enshrined in the law**, such as health, safety, and fundamental rights. Within FAITH, the legal risks associated with the operation of the FAITH AI_TAF are **those that may affect rights and principles outlined in the seven key requirements for trustworthy AI**. Before considering those requirements in more detail, we need to establish what we mean by 'risk' from a legal perspective: **when does legal risk arise**?

This is a nuanced question, and a comprehensive exploration exceeds the scope of this guide. However, to conduct an effective impact assessment of FAITH AI_TAF, we must clarify the concept of legal risk as it applies here. In the EU Legal Framework, legal risks have to be comprehended in the more general settings of recent EU regulation for digital services, which follows a risk-based approach.

The **"rights-based approach"** posits that a right is either violated or not, adhering to a binary logic[66]. As such, risk is framed solely as **the potential violation of a right,** with no consideration

[66] Karen Yeung and Lee A Bygrave, 'Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship' (2022) 16 Regulation & Governance 137

for gradations or probabilities[67]. Assessing legal risks therefore consists of relying on a risk-based approach where the risk relies in the breaching of a given right.

Conversely, the **"risk-based approach"** endorsed by the GDPR and more recently by the AI Act, equates risk with "harm," where the likelihood and severity of the harm form the basis for measuring the risk. Within this framework, rights are deemed at risk when foreseeable harm or damage is both probable and measurable.

Given the alignment of the EU's regulatory framework with the risk-based approach, this is the perspective adopted in FAITH's impact assessment. However, we also expand on this framework to incorporate scenarios that might not be fully addressed within it. The next questions of the questionnaire aims at collecting information and assessing potential inferences to rights, even though they might not consist of a violation *per se*. Recognizing and evaluating these interferences provides valuable insights into how technology impacts rights and supports informed adjustments to system design[68].Therefore, legal risk is conceived in FAITH's legal impact assessment as **the risk of interference with the rights and principles encompassed by the seven requirements for trustworthy AI**.

### 4.3.2. Ethical Risks

We complement this identification and assessment of legal risks with a **high-level ethics overview** that further develop the fundamental rights-based approach we have focused on in this deliverable. We do so by relying on the ALTAI prepared by the HLEG on AI[69].

The ALTAI aims to help organisations better understand the requirements for trustworthy AI and to identify what risks an AI system might generate and how to minimise those risks. It aims to allow AI providers and deployers to critically approach the potential impact of their AI systems on society, the environment, consumers, workers and citizens. During our assessment, we rely on the questionnaire developed by the HLEG on AI within the ALTAI to **identify potential risks** caused by the FAITH AI_TAF or by its use in the context of LSPs. The questions have been slightly reworked as to fit within the context of the FAITH Project and the rest of the questionnaire, consisting mainly in minor grammatical and syntactical adaptations.

This questionnaire is designed to assist the Consortium in i) **Identifying potential interferences**, ii) **Evaluating the identified interferences**, and iii) **Managing these interferences** to ensure they do not escalate into violations that could ultimately compromise fundamental rights and/or ethical requirements for trustworthy AI.

### 4.3.3. Legal and Ethical Risk Identification

#### 4.3.3.1 Human Oversight

Human oversight is a key requirement under the AIA for high-risk AI systems, as outlined in Article 14. While it is voluntary for non-high-risk systems, its implementation is encouraged through the development of specialized Codes of Practice (AI Act, article 56).

[67] Gianclaudio Malgieri and Cristiana Santos, 'Assessing the (Severity of) Impacts on Fundamental Rights' (2024) https://papers.ssrn.com/abstract=4875937 accessed 13 January 2025

[68] ibid.

[69] See Section 5.7.4 of this Deliverable

Human oversight safeguards autonomy and facilitates accountability throughout the AI lifecycle. It ensures human judgment is preserved by regulating the allocation of tasks between humans and AI, and it emphasizes that oversight must go beyond superficial approval of AI outputs. To be meaningful, oversight must address potential cognitive biases, allocate tasks to qualified personnel, and provide sufficient time and resources for effective decision-making.

The ALTAI states that AI systems should support human agency and human decision-making. AI systems should support the user's agency and uphold fundamental rights, both aspects which should be underpinned by human oversight. The ALTAI identifies two requirements for human oversight:

- **Human agency and autonomy**, dealing with how AI systems affect human decision-making, perception, affection, trust and independence.
- **Human oversight,** focusing on the necessary oversight measures, including human-in-the-loop (HITL), human-on-the-loop (HOTL) and human-in-command (HIC) approaches.

*Table 16 - Human Oversight Assessment*

| Question N° | Preferred Cluster | Question |
|---|---|---|
| **74** | Technical Partners | How will human oversight be implemented for the FAITH AI_TAF? |
| ALTAI | | |
| **75** | Technical, LSP and Legal Partners | Is the FAITH AI_TAF designed to interact with, guide, or take decisions affecting humans or society? |
| **76** | | Could the FAITH AI_TAF generate confusion for some or all end-users or subjects on whether they are interacting with a human or FAITH AI_TAF? |
| **77** | | Could the FAITH AI_TAF affect human autonomy by generating over-reliance for end-users? |
| **78** | | Could the FAITH AI_TAF affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way? |
| **79** | | Does the FAITH AI_TAF simulate social interaction with or between end-users or subjects? |
| **80** | | Does the FAITH AI_TAF create a risk of human attachment, stimulating addictive behaviour, or manipulating user behaviour? |
| **81** | | Is the FAITH AI_TAF self-learning or autonomous, overseen by a human-in-the loop, overseen by an human-on-the loop or overseen by a human-in-command? |

| | | |
|---|---|---|
| **82** | | Have the humans received specific training on how to exercise oversight of FAITH AI_TAF? |
| **83** | | Are detection and response mechanisms in place for undesirable adverse effects of FAITH AI_TAF? |
| **84** | | Is there a 'stop button' or procedure in FAITH AI_TAF to safely abort operations when needed? |
| **85** | | Does the FAITH AI_TAF implement specific oversight and control measures to reflect its self-learning or autonomous nature (if applicable)? |

#### 4.3.3.2 Technical Robustness and Safety

Technical robustness and safety are critical elements of trustworthy AI systems, ensuring that they operate reliably and can handle errors or unexpected conditions without causing harm[70]. Ensuring safety is part of the requirements of the AI Act, and in the case of personal processing, of the GDPR.

The above section focuses on cybersecurity regulation in the EU highlighting the importance of proper security measures. Both the CSA and the CRA proposes cybersecurity frameworks that are compliant with article 15(1) of the AI Act, while the NIS II directive further specifies the requirement for AI systems and tools.

The ALTAI highlights the importance of **dependability**, e.g. the ability to deliver services that can be justifiably trusted and **resilience**, e.g. robustness when facing changes. It highlights four main issues regarding technical robustness and safety : (i) **security**, (ii) **safety**, (iii) **accuracy** and (iv) **reliability, fall-back plans and reproducibility**.

*Table 17 - Technical Robustness and Safety Assessment*

| Question N° | Preferred Cluster | Question |
|---|---|---|
| **86** | Technical Partners | What are the cybersecurity measures in place for the FAITH AI_TAF? |
| **87** | | Is the FAITH AI_TAF certified for cybersecurity (e.g. the certification scheme created by the Cybersecurity Act in Europe) or is it compliant with specific security standards? |
| ALTAI | | |
| **88** | Technical Partners | Could the FAITH AI_TAF have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use? |

[70] See Data protection Rules (GDPR)

| | | |
|---|---|---|
| 89 | | How exposed is the FAITH AI_TAF to cyber-attacks? |
| 90 | | Does the FAITH AI_TAF maintain its integrity, robustness and overall security against potential attacks over its lifecycle? |
| 91 | | Does the FAITH AI_TAF undergo red-team/penetration test activities? |
| 92 | | Does the FAITH AI_TAF provide information to end-users about the duration of security coverage and updates? |
| 93 | LSP Partners | Are risks, risk metrics and risk levels related to the FAITH AI_TAF defined in each specific use case? |
| 94 | Technical Partners | Are the possible threats to the FAITH AI_TAF identified (design faults, technical faults, environmental threats) as well as the possible consequences? |
| 95 | | Does the FAITH AI_TAF include fault tolerance via, e.g. a duplicated system or another parallel system (AI-based or 'conventional')? |
| 96 | | Does the FAITH AI_TAF incorporate a mechanism to evaluate when changes merit a new review of its technical robustness and safety? |
| 97 | | Could the FAITH AI_TAF's low level of accuracy result in critical, adversarial or damaging consequences? |
| 98 | | Do measures ensure that the data (including training data) used to develop the FAITH AI_TAF is up-to-date, of high quality, complete and representative of the environment the system will be deployed in? |
| 99 | | Is the FAITH AI_TAF's accuracy monitored and documented? |
| 100 | | Could the FAITH AI_TAF's operation invalidate the data or assumptions it was trained on, and could this lead to adversarial effects? |
| 101 | | Does the FAITH AI_TAF ensure that end-users and/or subjects properly understand the level of accuracy to be expected? |
| 102 | | Could the FAITH AI_TAF cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility? |
| 103 | | Does the FAITH AI_TAF implement verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of its reliability and reproducibility? |
| 104 | | Does the FAITH AI_TAF include tested failsafe fallback plans to address errors of whatever origin and governance procedures to trigger them? |

| | | |
|---|---|---|
| **105** | | Does the FAITH AI_TAF include a proper procedure for handling cases where it yields results with a low confidence score? |
| **106** | | Does the FAITH AI_TAF use (online) continual learning? |

#### 4.3.3.3 Privacy and data governance

Privacy and data governance are essential to ensure that all data processed within AI systems is handled securely, lawfully, and ethically. These principles are vital for maintaining user trust and compliance with regulatory standards such as the **GDPR**[71].

**Data governance** goes beyond ensuring privacy; it establishes the **overarching framework for managing data assets responsibly across their entire lifecycle**. This includes defining roles and responsibilities, implementing policies for data quality, and creating mechanisms to monitor compliance and manage risks. Effective data governance ensures that data remains accurate, secure, and accessible, empowering organizations to make informed decisions while minimizing misuse or inefficiencies. Additionally, a robust governance structure provides transparency and accountability, critical for building stakeholder trust in AI systems.

The ALTAI emphasises the **importance of privacy as a fundamental right** that is particularly affected by AI systems as well as the necessity of proper data governance capabilities in order to effectively preserve privacy.

*Table 18 - Privacy and Data Governance Assessment*

| Question N° | Preferred Clusters | Question |
|---|---|---|
| | *The following questions should only be answered if the answer to question 8 highlights a high risk for data subjects' rights and freedom.* | |
| **107** | Technical Partners | Describe the **type** of the personal data processing |
| **108** | | Describe the **scope** of the personal data processing |
| **109** | | Describe the **context** of the personal data processing |
| **110** | | Describe the **purposes** of the personal data processing |
| **111** | | How will you/have you engaged with relevant stakeholders in relation the processing activities? |
| **112** | | Describe compliance and proportionality measures |
| **113** | | Describe the **source of risk and nature of a potential impact on individuals**. Include associated compliance and corporate risks as necessary |
| **114** | | Identify **additional mitigating measures** to reduce or eliminate risks as identified as medium or high risk |

[71] See Section 5.3 of this Deliverable

| | | |
|---|---|---|
| **115** | | Has a DPIA been conducted, signed off, approved and is it correctly recorded? |
| ALTAI | | |
| **116** | Technical Partners | Is the impact of the FAITH AI_TAF on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection taken into account? |
| **117** | | Are mechanisms established that allow flagging issues related to privacy concerning the FAITH AI_TAF, depending on the use case? |
| **118** | | Which following measures, some of which are mandatory under the GDPR, are in place for the FAITH AI_TAF: DPIA, designating a DPO that is included in the AI development, oversight mechanisms for data processing, data minimisation and privacy-by-design and by default? |
| **119** | | Are the right to withdraw consent, the right to object and the right to be forgotten implemented into the development of the FAITH AI_TAF? |
| **120** | | Are privacy and data protection implications considered for all data collected, generated, or processed throughout the FAITH AI_TAF's lifecycle? |
| **121** | | Are privacy implications considered for non-personal training data and other processed non-personal data? |
| **122** | | Is the FAITH AI_TAF aligned with relevant standards (e.g., ISO, IEEE) and widely adopted protocols for daily data management and governance? |

#### 4.3.3.4 Transparency

Transparency ensures that users and stakeholders can understand the functioning of the tools and interact with them appropriately. This understanding goes beyond providing information; it requires knowledge to be relevant, clear, concise, and tailored to the audience. Transparency requirements can be supported by measures such as (i) comprehensive documentation, (ii) user-friendly design, (iii) disclosures about AI usage, and (iv) implementing safeguards for emotion recognition and profiling.

The ALTAI encompasses four elements for transparency within the goal of achieving trustworthy AI: (i) **traceability** (e.g. the proper documentation of the data and systems yielding the FAITH AI_TAF's decision, (ii) **explainability** (e.g. the ability to explain the technical process and reasoning behind the decision made) and (iii) **communication** (e.g. whether the FAITH AI_TAFs' capabilities and limitations have been clearly communicated).

*Table 19 - Transparency Assessment*

| Question N° | Preferred cluster | Question |
|---|---|---|
| **123** | Technical Partners | Are the instructions to use the FAITH AI_TAFs provided to the user? |
| **124** | | Are there any disclaimers regarding the use of the FAITH AI_TAF? |
| **125** | | Is the FAITH AI_TAF relying on emotion recognition? If yes, how does the FAITH AI_TAF inform the exposed parties regarding the emotion recognition performed by the FAITH AI_TAF? |
| **126** | | Is the FAITH AI_TAF relying on profiling? If yes, what is the logic behind the profiling and how it this explained to interested parties? |
| ALTAI | | |
| **127** | Technical Partners | Are measures in place that address the traceability of the FAITH AI_TAF during its entire lifecycle? |
| **128** | | Are the decision(s) of the FAITH AI_TAF explained to the users? |
| **129** | | Are the users continuously surveyed to verify if they understand the decision(s) of the FAITH AI_TAF? |
| **130** | | In cases of interactive use of the FAITH AI_TAF, is it communicated to users that they are interacting with an automated tool instead of a human? |
| **131** | | Are there mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the FAITH AI_TAF? |

#### 4.3.3.5 Fairness

Fairness is grounded in the principle that all individuals are of equal moral worth. Within the European Union, laws protect the right of every person to equal treatment and respect. Article 14 of the ECHR states that: "The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status." Discrimination occurs when decisions, actions, or institutional structures fail to uphold this equality[72].

[72] See Section 5.5 of this Deliverable

The EU's non-discrimination laws are fragmented but provide a harmonized baseline through the EU Equality Legal Framework. This framework mandates a minimum level of protection across all Member States, while allowing individual countries to expand these protections in their national laws.

The ALTAI states that inclusion and diversity need to be enabled throughout the entire AI's Lifecyle in order to achieve trustworthy AI. Several obstacles are identified regarding the fairness of an AI system from an ethical point of view:

- The inclusion of **unfair bias**, caused by historic bias, incompleteness and/or bad governance models that can lead to unintended (in)direct prejudice and discrimination against certain groups of people.
- The lack of **accessibility** which prevents certain categories to access and effectively rely on the tools offered. In order to produce AI systems that are user-centric and inclusive, inclusion of Universal Design Principles[73] and relevant accessibility standards are to be considered.
- The lack of stakeholder participation throughout the life cycle of the AI system.

*Table 20 - Fairness Assessment*

| Question N° | Preferred Cluster | Question |
|---|---|---|
| **132** | Technical, LSP and Legal Partners | Are there aspects of the FAITH AI_TAF that can result in direct and indirect discrimination? Which ones? |
| ALTAI | | |
| **133** | Technical, LSP and Legal Partners | Is there a strategy or set of procedures in place to avoid FAITH AI_TAF creating or reinforcing unfair bias, both regarding the use of input data and the algorithm design? |
| **134** | | Is diversity and representativeness of end-users and/or subjects in the data taken into account in FAITH AI_TAF? |
| **135** | | Are educational and awareness initiatives in place to ensure AI literacy efforts to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the FAITH AI_TAF? |
| **136** | | Is there a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance in FAITH AI_TAF? |
| **137** | | Is there a commonly used definition of fairness that is implemented in any phase of setting up the FAITH AI_TAF? |

[73] European Committee for Standardization, 'Accessibility: Design for All' https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/accessibility/design-for-all/ accessed 27 December 2024/

| | | |
|---|---|---|
| **138** | | Does the FAITH AI_TAF accommodate the variety of preferences and abilities in society? |
| **139** | | Is the FAITH AI_TAF usable by those with special needs, disabilities, or those at risk of exclusion? |
| **140** | | Are Universal Design principles incorporated during every step of planning and development of FAITH AI_TAF? |
| **141** | | Is the impact of the FAITH AI_TAF on potential end-users and subjects evaluated? |
| **142** | | Does the FAITH AI_TAF's design and development process include mechanisms for participation from a wide range of stakeholders? |

#### 4.3.3.6 Societal and environmental well-being

Societal and environmental well-being focuses on creating AI systems that actively promote societal progress while minimizing harm to the environment. These systems should prioritize addressing pressing societal challenges, such as enhancing healthcare accessibility, improving disaster response, and mitigating climate change. They must also consider the broader impact on communities, ensuring that technological advancements do not exacerbate existing inequalities or create new ones. For example, AI deployment should aim to support underserved areas and contribute to reducing disparities rather than perpetuating them.

Moreover, **the environmental footprint of AI systems requires close scrutiny**. Developers should adopt energy-efficient practices and sustainable materials, aiming to balance innovation with ecological responsibility. Such considerations are essential to align AI progress with global sustainability goals.

The ALTAI highlights the importance of considering society as a whole, as well as the environment and sentient beings' interests when developing AI systems. It further stresses out the importance for AI systems to take into account their impact on the long term, on society in general, the environment as well as on democratic processes. The ALTAI distinguishes between (i) **the impact on the environment**, (ii) **the impact on work and skills** and (iii) **the impact on society at large and democracy**.

*Table 21- Societal and environmental well-being assessment*

| Question N° | Preferred Clusters | Question |
|---|---|---|
| **143** | Consortium | Is there a strategy in place regarding potential societal and/or environmental impact of the FAITH AI_TAF? |
| **144** | Technical and LSP Partners | Will the FAITH AI_TAF have impact on workers within a specified organisation? |
| ALTAI | | |

| | | |
|---|---|---|
| **145** | All Consortium | Are there potential negative impacts of the FAITH AI_TAF on the environment? |
| **146** | Technical Partners | Where possible, did you establish mechanisms to evaluate the environmental impact of the FAITH AI_TAF's development, deployment and/or use (for example, the amount of energy used and carbon emissions)? |
| **147** | Technical and LSP Partners | Does the FAITH AI_TAF impact human work and work arrangements? |
| **148** | | Is the way paved for introducing the FAITH AI_TAF in your organisation by informing and consulting with impacted workers and their representatives (trade unions, (European) work councils) in advance? |
| **149** | | Are measures adopted to ensure that the impacts of the FAITH AI_TAF on human work are well understood? |
| **150** | All Consortium | Could the FAITH AI_TAF create the risk of de-skilling of the workforce? |
| **151** | Technical and LSP Partners | Does the FAITH AI_TAF promote or require new (digital) skills? |
| **152** | Consortium | Could the FAITH AI_TAF have a negative impact on society at large or democracy? |

#### 4.3.3.7 Accountability

Accountability is a fundamental pillar of trustworthy AI, ensuring that **all actors** involved in the lifecycle of the system—from design to deployment—**are responsible for their decisions and actions**. It requires establishing mechanisms to trace and address potential issues, fostering trust among users and stakeholders.

A **robust accountability framework** maintains transparency and ensures that the AI models operate within ethical and legal boundaries. This includes defining clear roles, embedding accountability into design and implementation, and proactively addressing risks through regular assessments.

The ALTAI states that accountability requires to put in place mechanisms that ensure responsibility for the development, deployment and use of AI systems. This requires a strong risk management framework. The ALTAI further specifies that accountability focuses on (i) **auditability** and (ii) **risk management**.

*Table 22 - Accountability Assessment*

| Question N° | Preferred Clusters | Question |
|---|---|---|
| **153** | Technical Partners | Are the FAITH AI_TAF's development and deployment activities documented? How? |

| | | |
|---|---|---|
| **154** | | What are the risk management mechanisms in place regarding the FAITH AI_TAF? |
| ALTAI | | |
| **155** | Technical Partners | Are mechanisms established that facilitate the auditability of the FAITH AI_TAF (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)? |
| **156** | | Is the FAITH AI_TAF auditable by third parties? |
| **157** | | Is any kind of external guidance or third-party auditing processes foreseen to oversee ethical concerns and accountability measures? |
| **158** | | Is risk training organised, and if so, does this also inform about the potential legal framework applicable to the FAITH AI_TAF? |
| **159** | Legal Partner | Is there an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas concerning the FAITH AI_TAF? |
| **160** | Technical Partners | Is there a process to discuss and continuously monitor and assess the FAITH AI_TAF's adherence to this Assessment List for Trustworthy AI (ALTAI)? |
| **161** | | Is there a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the FAITH AI_TAF? |
| **162** | | For applications of the FAITH AI_TAF that can adversely affect individuals, are there redress by design mechanisms put in place? |

### 4.3.4. Legal and ethical Risk Management

Based on the information provided in the questionnaire, KU Leuven will be able to assess and identify potential interferences created by the FAIT TAF. The next step is to **assess those legal and ethical risks**. This requires measuring the identified interferences using consistent criteria. In line with EU law, risk assessment combines **severity**, which reflects the magnitude of an event, and **likelihood**, which indicates the probability of its occurrence. While legal and ethical risks cannot be quantified mathematically, they can be categorized qualitatively as **low, medium**, or **high**.

**Assessing severity** involves understanding the event causing the interference and **predicting its potential impact**. This requires consideration of **objective factors**, such as the infringement of legal norms, its reversibility, duration, and scope. Additionally, **subjective elements**, including societal, group, and individual perceptions, as well as **adverse effects** such as economic loss, time implications, and harm to well-being, are crucial for determining gravity. **Evaluating likelihood**, on the other hand, is relatively more straightforward and focuses on

estimating the probability of the identified event occurring, ranging from remote to more likely than not.

By combining these dimensions of severity and likelihood, risks can be assessed in a structured manner. This approach ensures enables comprehensive understanding of the potential impacts and probabilities associated with the FAITH AI_TAF.

Table 23 helps to perform the assessment in a structured manner:

*Table 23- Risk assessment matrix*[74]

| SEVERITY OF THE INTERFERENCE | High interference | Low risk | High risk | High risk |
|---|---|---|---|---|
| | Middle interference | Low risk | Medium risk | High risk |
| | Minimal or no interference | Low risk | Low risk | Low risk |
| | | Low | Medium | High |
| | | Likelihood of the interference | | |

### 4.3.5. Legal and ethical Risk Mitigation

After identifying and assessing the interferences with the seven requirements for trustworthy AI that the TAF may pose, the next step is to **manage the identified legal and ethical risks**. This involves prioritising interferences classified as medium or high risk to ensure they do not escalate into violations of fundamental rights.

Based on the legal and ethical risk assessment, the FAITH consortium and/or LSPs should adopt targeted risk management measures. These can include **design modifications, organizational policies, and improved transparency or redress mechanisms**. For example, if the FAITH AI_TAF or one of the LSPs disproportionately impact non-English speakers due to language bias in input data, the data must be recalibrated. Similarly, if human oversight mechanisms are insufficient for operators to act on system outputs, the framework must ensure adequate response times.

[74] Kovačević N and others, 'Application of the Matrix Approach in Risk Assessment' (2019) 2(3) Operational Research in Engineering Sciences: Theory and Applications 55

## 5. CONCLUSION

The assessment methodology and templates presented in this deliverable provide a structured framework for evaluating regulatory and ethical risks in relation to the FAITH AI_TAF. The incorporation of feedback from technical partners, LSP leaders, and the External Ethics Advisory Board will ensure a practical approach while enhancing legal and ethical compliance.

The comprehensive questionnaire addresses fundamental rights, data protection, cybersecurity, and fairness considerations through a unified approach. This holistic perspective enables LSP owners to identify potential risks early and implement appropriate mitigation measures and to evaluate the impact of the FAITH AI_TAF on their own activities. The information collected through the questionnaire within this assessment will allow for a better understanding of the FAITH AI_TAF and its impact on AI models relied upon by Partners within LSPs. It will result in the final version of the Deliverable (D1.4) that will identify any potential legal and ethical risk for the FAITH AI_TAF and propose measures to diminish and/or avoid those risks.

Moving forward, the framework will be updated based on the evolving EU legal landscape for AI regulation, practical insights gained from LSP implementations, technical advances in AI trustworthiness assessment, and emerging best practices. The final version of this deliverable (D1.4) will incorporate these learnings to provide an enhanced assessment framework and questionnaire. This will contribute to FAITH's broader objective of fostering trustworthy AI development while ensuring that legal and ethical considerations remain at the forefront of technological innovation.

# 6. REFERENCES

## 6.1. LEGISLATION

Charter of Fundamental Rights of the European Union [2012] OJ C326/391

Council Directive (EU) 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin [2000] OJ L180/22

Council Directive (EU) 2000/78/EC establishing a general framework for equal treatment in employment and occupation [2000] OJ L303/16

Council Directive (EU) 2004/113/EC implementing the principle of equal treatment between men and women in the access to and supply of goods and services [2004] OJ L373/37

Directive (EU) 2016/1148 concerning measures for a high common level of security of network and information systems across the Union [2016] OJ L194/1

Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data [2016] OJ L119/1

Regulation (EU) 2017/745 on medical devices [2017] OJ L117/1

Regulation (EU) 2019/881 on ENISA and on information and communications technology cybersecurity certification [2019] OJ L151/15

Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union [2022] OJ L333/80

Proposal for a Regulation on horizontal cybersecurity requirements for products with digital elements COM/2022/454 final

Regulation on harmonised rules on Artificial Intelligence [2024]

## 6.2. SECONDARY SOURCES

Article 29 Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA)' (WP248 rev.01, 2017)

CNIL, 'AI how-to sheets' https://www.cnil.fr/fr/ai-how-to-sheets accessed 27 December 2024

CNIL, 'Carrying Out Protection Impact Assessment If Necessary' https://www.cnil.fr/en/carrying-out-protection-impact-assessment-if-necessary accessed 27 December 2024

CNIL, 'Guidelines on DPIA' https://www.cnil.fr/en/guidelines-dpia accessed 27 December 2024

CNIL, 'Privacy Impact Assessment (PIA) Methodology' (2018)

Council of Europe, 'Best Practices in Academic Integrity' https://rm.coe.int/bpp-a-compendium-of-best-practices-eng-/1680a86621

Council of Europe, 'Ethical Behaviour for Actors in Education' http://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806d2b6f

Council of Europe, 'Ethical Principles' https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806c90cd

Council of Europe Parliamentary Assembly Resolution 1003 on Ethics of Journalism (1 July 1993)

Demetzou K, 'Data Protection Impact Assessment: A tool for Accountability and the Unclarified Concept of "High Risk" in the General Data Protection Regulation' (2019) 35 Computer L & Security Rev 105342

European Committee for Standardization, 'Accessibility: Design for All' https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/accessibility/design-for-all/ accessed 27 December 2024

European Data Protection Board, 'Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models' (2024)

European Equality Law Network https://www.equalitylaw.eu/ accessed 27 December 2024

Friedewald M and others, 'Data Protection Impact Assessments in Practice' (2022)

Garcia S.,Corrêa A and Stamatellos G., 'Template & Guidance for Legal and Ethical Impact Assessment', available at https://www.themis trust.eu/_files/ugd/a245c2_c1fe2d866bf140e5a4e5daa8afb292ce.pdf accessed 17 March 2025

Hanif H and others, 'Navigating the EU AI Act Maze Using a Decision-Tree Approach' (2024) 1 ACM J Responsible Computing 21:1

Kovačević N and others, 'Application of the Matrix Approach in Risk Assessment' (2019) 2(3) Operational Research in Engineering Sciences: Theory and Applications 55

Leslie D and others, 'Human Rights, Democracy, and the Rule of Law Assurance Framework for AI systems: A Proposal' http://arxiv.org/abs/2202.02776 accessed 27 December 2024

Malgieri G and Santos C, 'Assessing the (Severity of) Impacts on Fundamental Rights' (2024) https://papers.ssrn.com/abstract=4875937 accessed 13 January 2025

Novelli C and others, 'AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act' (2023)

Smuha N, 'The Work of the High-Level Expert Group on AI as the Precursor of the AI Act' (2024)

WHO, 'Developing an Ethical Framework for Health Ageing' (2017)

WHO, 'Ethics & Governance of Artificial Intelligence for Health' (2021)

World Medical Association Declaration of Helsinki (1964)

Yeung K and Bygrave L, 'Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship' (2022) 16 Regulation & Governance 137

'Z-Inspection®: A Process to Assess Trustworthy AI' (2021) IEEE Xplore https://ieeexplore.ieee.org/document/9380498 accessed 27 December 2024

## 7. Annex 1: SUMMARY CHART – LEGAL IMPACT ASSESSMENTS

| Impact Assessments / Questions | GDPR<br>Data Protection Impact Assessment (DPIA) | AI Act<br>Risk Management Assessment | AI Act<br>systemic risks arising from GPAI model | AI Act<br>Fundamental Rights Impact Assessment (FRIA) | Cybersecurity risk management<br>NIS Directives – Focus on NIS 2 | Cybersecurity risk management<br>Cybersecurity Act | Cybersecurity risk management<br>Cyber Resilience Act |
|---|---|---|---|---|---|---|---|
| **Summary** | The DPIA is a specific undertaking that must be carried out when the **processing** of **personal data** is likely to **result in a high risk to the rights and freedoms of natural persons** (GDPR, art. 35).<br><br>The high risk is particularly associated with the **use of new technologies** | The risk management system is a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system aimed at the identification, evaluation and | GPAI models classified as **'systemic risk' GPAI models** must undergo an **identification, assessment and mitigation of the systemic risks** arising from their development, placing on the market or use at the Union | The FRIA aims to identify the specific risks to the rights of individuals or groups of individuals likely to be affected by selected high-risk AI system, and to identify measures to be taken if these risks | **Essential and important entities operating in critical sectors** must put in place cybersecurity **risk-management**, based on a list of minimum measures to be implemented. | The EU's Cybersecurity Act (Regulation (EU) 2019/881, CSA) introduced a voluntary **cybersecurity certification framework**, covering ICT products, services and processes, which | The upcoming Cyber Resilience Act (CRA) sets **common cybersecurity requirements for products** with digital elements. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | management of the reasonably foreseeable risks that the high-risk AI SYSTEMS poses to health, safety and fundamental rights. | level, with identification of the possible sources of these risks. (AI ACT, Art. 55 (1)(a)(b)). | materialise (AI ACT, art. 27). | | includes the ones using AI technologies. | |
| **Material Scope** | The scope of the DPIA is any operation or set of operations on personal data (i.e. "processing", GDPR art. 4(2)) that, taking into account the nature, scope, context and purposes, **is likely to result in a high risk to the rights and freedoms of natural persons.**<br><br>In the context of FAITH, a high risk may occur when personal aspects as preferences or interests | **All high-risk AI systems** must undergo the risk management introduced by Art. 9 of the AI ACT.<br><br>The only **exceptions** are the **high-risk AI systems covered by Union harmonizati** | The risk management required by art. 55 AI ACT concerns **only GPAI models with systemic risks**.<br><br>An AI model is a GPAI model when<br><br>i) is trained with a **large amount of data** using **self-** | The scope of the FRIA is:<br><br>- **all high-risk AI systems** listed in AI ACT, Annex III, if the **deployer** is a **public law body** or if the high-risk AI SYSTEMS falls into the notion of **public service provided by a private entity**. | The risks to be managed are those posed to the **security of networks and information systems** that the entities use for their operations or the provision of their services (NIS 2, art. 21).<br><br>Network and information | A cybersecurity certification scheme is "a comprehensive set of rules, technical requirements, standards and procedures that are established at Union level" and which serves to assess the cybersecurity | **'Product with digital elements'** is both **software** (like mobile apps or operating systems) or **hardware** products (like laptops and smartphones) **and their software or hardware components,** placed on the |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | are evaluated in order to infer "personal profiles" or, depending on pilots 'sectors, when special categories of data are processed, such as data concerning health (in the healthcare LSP), or when minors are involved (in the education LSP) | **on legislation**, such as **vehicles or marine equipment**, for which the requirements for high-risk AI SYSTEMSs will be integrated and considered directly by the sectorial legislation. | **supervision** at scale;<br>ii) displays significant generality;<br>iii) is capable of competently **performing a wide range of distinct tasks**, with the potential to be integrated into a variety of downstream systems or applications (AI ACT, art. 3 (63)).<br>**In the context of FAITH.**, it may be the case of the large language models, LLMs, | In the context of FAITH, cases potentially relevant are AI SYSTEMSs used in the area of biometrics, including emotion recognition and access to healthcare services.<br>- high-risk AI SYSTEMSs intended to **evaluate the creditworthiness/ credit scores of natural persons or the risks and prices of life and health insurance** of natural | systems include:<br>i) **any device or group** of interconnected or related devices, one or more of which, pursuant to a programme, **carry out automatic processing of digital data**, and<br>ii) **digital data** stored, processed, retrieved or transmitted by the device or group of devices, defined under i), for the purposes of their operation, | of specific ICT products, ICT services or ICT processes (CSA, Art. 2(9)).<br>- 'ICT product' is an element or a group of elements of a network or information system;<br>'ICT service' is a service operating on information through a network and information system and<br>- 'ICT process' is a set of activities performed to design, develop, deliver or | market separately (CRA, art. 3(1)).<br>AI SYSTEMSs are comprised in this broad definition.<br>Products with digital elements which are made **available on the market,** and which are intended to be used or reasonably used in connection to a device or **network** must undergo a cybersecurity risk assessment. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [redacted] | | | incorporated in the AI chatbot.<br><br>systemic risk **pertains to the specific capabilities of the GPAI model**, matching or exceeding the capabilities of the most advanced GPAI model, and that can have a significant impact on the Union market (AI ACT, art. 3(65)). | persons (AI ACT, Annex III, points 5b and c), regardless of the private or public nature of the deployer or the service offered. | use, protection and maintenance (NIS2, art. 6(1)).<br>AI SYSTEMSs and their assets are therefore "network and information systems". | maintain an ICT product or ICT service (CSA, Art. 2(12), (13), (14)). | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Personal Scope** | Conducted by data controllers and might involve Data Protection Officers (DPOs), data subjects and supervisory authorities for consultations.<br><br>In the context of FAITH. it is possible that two or more Partners determine jointly the purposes and means of the data processing, for instance the LSP Partner and the technical Partner working closely together: in this scenario the Partners would be "joint controllers" and would be equally responsible for the DPIA. | The obligation lies on **the provider of the high-risk AI SYSTEMS** (AI ACT, art. 16, 43).<br><br>Provider is anyone – person, public authority, company etc. - who develops, places it on the market or puts into service an AI SYSTEMS or a general-purpose AI model whether for payment or free of charge (AI | The obligation lies on **the provider of the GPAI model** with systemic risk, who bears a crucial responsibility along the AI value chain, given that the model can be used by a wide range of downstream AI SYSTEMSs. | Only deployers that are (AI ACT, art. 27):<br><br>1) **bodies governed by public law,** typically public administration;<br><br>2) **private entities providing public services**, such as performing tasks in the public interest (AI ACT, Rec. 96);<br><br>3) private or public entities offering private or public services | The entities must be **identified by Member States among** those operating in **critical sectors** listed by Art. 2 and 3 and by the Annexes of the NIS 2, such as **energy, transport** and **healthcare**. | **Manufacturers and providers** of ICT products, services and processes may rely on a European cybersecurity certification scheme to ensure compliance with the cybersecurity requirements covered by the scheme. | The common cybersecurity rules are mainly addressed at the **manufacturers of products with digital elements**. The other economic operators who participate in making the product available on the market, such as importers and distributors, are responsible for verifying that the manufacturers have respected their |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | ACT, art. 3(3)). | | - typically **banking or insurance entities (**AI ACT, Rec. 96) - intended to perform credit, risk or price evaluations regarding natural persons (AI ACT, Annex III, points 5b and c). | | | obligations (CRA, Chapter II). |
| **Practicalities** | The controller has to conduct an assessment that contains:<br>1) A **detailed description** of the processing activity;<br>2) A necessity and proportionality test for the processing activity; | The risk management system under Art. 9 of the AI ACT takes a **product safety approach to the regulation of AI SYSTEMSs.** | Providers have to evaluate the model to identify the risks, also through adversarial testing by standardised protocols and AI SYSTEMSs reflecting the | The assessment consists of:<br>1) **a description** of all the **processes in which the AI SYSTEMS will be used,** specifying the time, the frequency and | The risk management concerns:<br>i) the adoption of appropriate and proportionate **technical, operational and organisational risk** | The CSA outlines the framework, for instance in terms of objectives, assurance levels and minimum elements, within which European cybersecurity | The cybersecurity level of the product must be **appropriate to the risks identified** by the manufacturer thanks to the cybersecurity |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [redacted] | 3) An **evaluation of the risks** envisaged to the rights and freedoms of data subjects ;<br>4) Risk mitigation measures to be adopted. | This means that **AI SYSTEMSs are treated as products** that must undergo risk assessment and management procedures and implement specific requirements to protect **health and safety, but also fundamental rights.**<br>The risk management system is aimed at identifying and managing **"the known** | state of the Art (AI ACT, Art. 55(1) (a)).<br>In practice, providers have to continuously assess and mitigate the risks, also through risk management policies, post-market monitoring throughout the whole lifecycle and in cooperation with the downstream actors (AI ACT, Rec. 114). | the persons or group likely to be affected in a specific context;<br>2) a **risk analysis,** identifying specific risks of harm impacting the fundamental rights of the affected persons and groups;<br>3) **a risk management phase,** putting in place human oversight safeguards and procedural measures to be adopted in case a certain risk occurs, | **management** measures; and<br>ii) the prevention and minimisation of the impact of incidents on recipients of services.<br>Risk management measures must consider physical, environmental, human and interference risks (so called **all-hazards approach**). They can be technical and organizational . | certification schemes can be established.<br>The EU Agency for cybersecurity (Enisa) develops draft certification schemes, upon request of the European Commission or the EU Member States, with the support of Ad-Hoc Working Groups and in collaboration with the EU Commission, EU countries, and relevant stakeholders. | risk assessment.<br>The cybersecurity risk assessment consists of (CRA, art. 13):<br>- **an analysis of cybersecurity risks** considering the intended purpose and reasonably foreseeable use, the conditions of use and the length of time the product is expected to be in use;<br>- an indication of how the cybersecurity |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [redacted] | | **and the reasonably foreseeable risks" that** the high-risk **AI SYSTEMS** can pose to health, safety or fundamental rights when the system **is used following its intended purpose**, taking into consideration the experience of the expected deployers and the presumable context of use. | | such as redress mechanisms.<br>(!) If a DPIA is required, the two assessments must be coordinated. | | | requirements are implemented in practice. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Temporal Scope** | Initiated **before the start** of the processing, **reviewed periodically** and when **there is any change** that may impact its results. | It is a **continuous process**, which requires **systematic review and updating**. Compared to the DPIA and also to the FRIA, the iterations of the risk management system seem even stricter, because it should accompany every step of the AI SYSTEMS lifecycle. | Art. 55 of the AI ACT does not specify when risk management has to start. An indication is contained in Rec. 114 which states that the model evaluation in view to the identification of systemic risks takes place **"prior to its first placing on the market"** and that the assessment and mitigation of risks is **"continuous".** | **Before putting the AI SYSTEMS to its first use.** Like the DPIA, the FRIA needs to be reperformed when any of the elements relevant to the assessment **change** or become **outdated** (AI ACT, art. 27 and rec. 96). Then, if a specific risk occurs, the risk management measures, including internal governance arrangements and complaint mechanism, | As soon as the directive is transposed into national legislation by Member States (by 18 October 2024) the concerned entities will be obliged to implement the risk management measures. | Under the CSA, the **cybersecurity certification is voluntary**, implying that it can be performed also after the commercialisation of the relevant ICT product, service or process (CSA, Art. 56). In practice, to meet the certification scheme requirements, the ICT product, service or process must be designed and developed in accordance | The cybersecurity risk assessment starts during the **design phase** of the product and accompanies its **development and production**. Documentation and updating must be ensured during a specific "support period" determined by the manufacturer, based on the period of use expected for the product |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [redacted] | | | | need to be activated. | | with such requirements. | |