This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101092889, Topic HORIZON-CL4-2023-HUMAN-01-21

LUMINOUS

Language Augmentation for Humanverse

Project Reference No	101135724
Deliverable	D5.3. H - Requirement 3 (Trustworthy AI)
Work package	WP5: Ethics Requirements
Nature	D (Deliverable)
Dissemination Level	PU - Public
Date	31/05/2024
Status	Final v1.0
Editor(s)	**Eleni Mangina (UCD)** Didier Stricker (DFKI) Muhammad Zeshan Afzal (DFKI) Daniel Perez-Marcos (Mindmaze) Florendia Fourli (Hypercliq) Oier Lopez de Lacalle (EHU/UPV) Nicoletta Cioria (Mindesk) Ander Salaberria (EHU/UPV) Kanta Shimizu (RICOH) EEAB Members
Involved Institutions	UCD; DFKI; Hypercliq; MindMaze; EHU; VICOMTECH; RICOH
Document Description	This deliverable presents the Trustworthy AI Ethics requirement for the LUMINOUS project.

CONTENTS

LIST OF TABLES

Table 1: List of Abbreviations

Term / Abbreviation	Definition
EEAB	External Ethics Advisory Board
GDPR	General Data Protection Regulation
FAIR	Findability, Accessibility Interoperability and Reusability
NFADP	New Federal Act on Data Protection
IDS	Intrusion Detection Systems
DPIA	Data Protection Impact Assessment
TFEU	Treaty of the Functioning of the European Union
TEU	Treaty of European Union
ECHR	European Convention on Human Rights
FRIA	Fundamental Rights Impact Assessment
ALTAI	The Assessment List for Trustworthy Artificial Intelligence

1 INTRODUCTION

1.1 PURPOSE OF THE DOCUMENT

WP5, Ethics Requirements, encompasses three deliverables: D5.1 for overseeing human participation, D5.2 for personal data protection**,** and **D5.3 for managing Ethical risks related to the deployment of AI algorithms.** These deliverables focus on ethical monitoring throughout the project's duration. Additionally, the project planned the appointment of an External Ethics Advisory Board (EEAB) in M1, as described in D5.1. The EEAB's role includes overseeing ethical and legal compliance aspects during the project. The EEAB provides feedback to the project's Ethics Manager, Prof. Eleni Mangina from University College Dublin, who coordinates the LUMINOUS consortium's internal ethics monitoring activities.

This deliverable addresses the ethics requirements (7) for Trustworthy AI[1] in the LUMINOUS project. D5.3 provides the high-level guidelines for the consortium members on the assessment of the human rights and Trustworthy AI ethical considerations in the development, deployment, and post-deployment phases of the project. It covers aspects such as fundamental rights impact assessment, mitigation of bias and discrimination, informed communication with research participants and end-users, and evaluation of Ethical risks. Although the deliverable has been completed at the early stages of the project, the consortium members ensure to follow the recommendations and processes through the lifetime of the project and beyond. The aim is to ensure that LUMINOUS respects human dignity considering persons' autonomy and human rights, avoids discrimination, and addresses potential risks and ethical concerns. For this purpose, the following are included:

A. **The Fundamental Rights Impact Assessment:** It analyses how fundamental rights impact assessments (FRIA) could mitigate the negative impacts that using AI can have on fundamental rights and will provide a brief overview of the current discussion on the need for fundamental rights impact assessments in this field.
B. **The Assessment List for Trustworthy Artificial Intelligence (ALTAI).** ALTAI is designed to help individuals and organizations evaluate whether AI-based systems align with the ethical requirements listed below, promoting AI technologies that benefit society and respect fundamental values. It serves as a guide for assessing AI systems to ensure that they meet the seven key requirements of Trustworthy AI, as outlined in the Ethics Guidelines for Trustworthy AI. These requirements encompass:
    1. Human Agency and Oversight.
    2. Technical Robustness and Safety.
    3. Privacy and Data Governance.
    4. Transparency.

[1] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

5. Diversity, Non-discrimination, and Fairness.
6. Societal and Environmental Well-being.
7. Accountability.

C. A detailed explanation on the measures taken to prevent, avoid, and mitigate potential bias, discrimination, and stigmatization in input data and algorithm design and outcomes.
D. A detailed explanation of how the research participants and/or end-users will be informed about:
    1. their interaction with an AI system/technology (if relevant).
    2. the abilities, limitations, risks, and benefits of the AI system/technique.
    3. the way decisions are taken and the logic behind them (if relevant).
E. An evaluation of the Ethical risks related to the AI and a description of the measures set in place to prevent/mitigate any potential negative personal/social/environmental impacts during the research, deployment, and post-deployment phase.

1.2 Theoretical Framework and Positioning

European Commission released "AI for Europe" Communication in 2018 and established the High-Level Expert Group on AI. Both initiatives emphasized the importance of protecting fundamental rights. This call for action led to the creation of the High-Level Expert Group on AI[2] by the European Commission. A group of 52 experts, including individuals from academia, civil society, and industry (including a representative from FRA), worked together in a Commission-facilitated High-Level Expert Group. In 2019, they released "Ethics Guidelines for Trustworthy AI" and "Policy and investment recommendations for Trustworthy AI"[2]. These guidelines were further refined in 2023. Their efforts sparked discussions about the importance of aligning AI-based software systems with human rights, in addition to ethical considerations. This eventually led to the creation of Ethics Guidelines that incorporate fundamental rights considerations, with a specific focus on AI. These Ethics Guidelines also include an assessment list for trustworthy AI, which has been converted into a practical checklist. This checklist is designed to assist those who are involved in developing and using AI to ensure that it meets ethical and rights-based standards.

The European Council and Commission underlined the urgency of addressing emerging trends in AI while maintaining a high level of data protection and ethical standards. They are determined to make Europe a leader in secure, trustworthy, and ethical AI. On March 13th 2024[3], the European Parliament formally adopted the EU Artificial Intelligence Act ("AI Act").

[2] https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence
[3] https://artificialintelligenceact.eu/

The AI Act is the world's first horizontal and standalone law governing AI, and a landmark piece of legislation for the EU. Based on the latest version of the EU AI Act[4] the four-point summary[5] includes:

1. **The AI Act classifies AI-based software systems according to its risk:**
   - Unacceptable risk is prohibited (e.g. social scoring systems and manipulative AI).
   - Most of the text addresses high-risk AI systems, which are regulated.
   - A smaller section handles limited risk AI systems, subject to lighter transparency obligations: developers and deployers must ensure that end-users are aware that they are interacting with AI (chatbots and deepfakes).
   - Minimal risk is unregulated (including the majority of AI applications currently available on the EU single market, such as AI enabled video games and spam filters; this is changing with generative AI).
2. **The majority of obligations fall on providers (developers) of high-risk AI systems**.
   - Those that intend to place on the market or put into service high-risk AI systems in the EU, regardless of whether they are based in the EU or a third country.
   - And, third country providers where the high-risk AI system's output is used in the EU.
3. **Users are natural or legal persons that deploy an AI system in a professional capacity, not affected end-users.**
   - Users (deployers) of high-risk AI systems have some obligations, though less than providers (developers).
   - This applies to users located in the EU, and third country users where the AI system's output is used in the EU.

4. **General purpose AI (GPAI):**
   - All GPAI model providers must provide technical documentation, instructions for use, comply with the Copyright Directive, and publish a summary about the content used for training.
   - Free and open licence GPAI model providers only need to comply with copyright and publish the training data summary unless they present a systemic risk.
   - All providers of GPAI models that present a systemic risk – open or closed – must also conduct model evaluations, adversarial testing, track, and report serious incidents and ensure cybersecurity protections.

[4] https://artificialintelligenceact.eu/ai-act-explorer/
[5] https://artificialintelligenceact.eu/high-level-summary/#:~:text=The%20AI%20Act%20classifies%20AI%20according%20to%20its%20risk%3A&text=Minimal%20risk%20is%20unregulated%20(including,is%20changing%20with%20generative%20AI).

Internationally, organizations like the Council of Europe[6] have created the first legally binding global instrument to address risks posed by AI[7]. OECD[8], and UNESCO[9] have also established their frameworks on setting standards for AI. While ethical initiatives are valuable, they are often voluntary, and a rights-based approach is crucial for effective protection.

1.3 STRUCTURE OF THE DOCUMENT

The structure of the document that follows addresses all Trustworthy AI requirements as raised in LUMINOUS ethics appraisal, namely the following:

Table 2 – Structure of D5.3

Section 1	Introduction | Theoretical Framework and Positioning
Section 2	Fundamental Rights Impact Assessment (FRIA) | the theoretical premises and the assessment for the LUMINOUS project
Section 3	The Assessment List for Trustworthy Artificial Intelligence (ALTAI)| the recommendations received for the LUMINOUS project and how they were addressed.
Section 4	Mitigating Bias and Discrimination in AI: Strategies and Safeguards
Section 5	Participant Communication and Transparency Guidelines for AI Systems
Section 6	Ethical Risk Assessment and Mitigation in AI Lifecycle
Conclusion	Summarization of the main findings and next steps
References	Listing of resources consulted for the preparation of the deliverable
Annex I	Fundamental Rights Impact Assessment (FRIA) Questionnaire
Annex II	The Assessment List for Trustworthy Artificial Intelligence (ALTAI) | LUMINOUS

[6] Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI), Factsheet: Governance for digital transformation, and Council of Europe, Recommendation CM/Rec (2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems (adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies

[7] https://www.eeas.europa.eu/delegations/council-europe/text-first-legally-binding-global-instrument-address-risks-posed-artificial-intelligence-finalised_en?s=51...

[8] https://www.oecd.org/digital/artificial-intelligence/#:~:text=Developed%20by%20the%20OECD.AI,ensure%20policy%20consistency%20across%20borders.

[9] https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

2 FUNDAMENTAL RIGHTS IMPACT ASSESSMENT (FRIA)

In the context of AI, the General Data Protection Regulation (GDPR) governs automated personal data processing within the European Economic Area, as well as data processing under EU law. The GDPR, however, doesn't apply to national security-related data processing. Together with the Law Enforcement Directive, which addresses police and judicial cooperation in criminal matters, these instruments comprise numerous provisions concerning personal data protection. They establish fundamental data processing principles, including lawfulness, fairness, and transparency. The application of EU data protection legislation depends on whether personal data is involved. Certain AI-driven applications may not involve personal data, while others employ anonymized data. In such cases, the applicability of data protection laws is either limited or unclear. The boundary between anonymised and pseudo-anonymised data is not always distinct because anonymized data could potentially be 're-identified' and requires significant effort and access to additional individual information. For the purposes of the LUMINOUS project, assessing the ethical and legal compliance of AI, the ALIGNER Fundamental Rights Impact Assessment (AFRIA) tool[10] has been utilised for all three (3) pilots, assisting then in demonstrating compliance with ethical principles and fundamental rights while deploying AI systems, **as shown in Annex I**. European anti-discrimination law plays a crucial role in upholding fundamental rights in the context of AI and related technologies. Article 2 of the TEU[11] emphasizes non-discrimination as a fundamental EU value, while Article 10 of the TFEU[12] mandates the EU to combat discrimination on various grounds. Several EU anti-discrimination directives introduce more detailed provisions.

2.1 THEORETICAL PREMISES

The FRI Assessment encompasses various fundamental rights, including human dignity, non-discrimination, data protection and privacy. To guide the FRIA, questions rooted in specific articles from the European Convention on Human Rights (ECHR)[13], its protocols, and the European Social Charter[14] were considered. This comprehensive evaluation helps ensure that any AI-based software system aligns with fundamental rights principles and ethical standards and covers the following, as identified in the respective questionnaire (available in Annex I), covering the following:

1. **Non-discrimination:** Assess whether the AI system may discriminate against individuals based on various grounds like gender, race, ethnicity, and more. Implement testing and monitoring processes to detect and rectify bias throughout the AI system's life cycle.

[10] https://aligner-h2020.eu/fundamental-rights-impact-assessment-fria/
[11] https://fra.europa.eu/en/eu-charter/article/21-non-discrimination
[12] https://eur-lex.europa.eu/EN/legal-content/glossary/non-discrimination-the-principle-of.html
[13] https://www.echr.coe.int/documents/d/echr/convention_ENG
[14] https://www.coe.int/en/web/european-social-charter

2. **Privacy and Data Protection**: Ensure that the AI system protects individuals' privacy and personal data, complying with GDPR. Implement processes to assess the need for data protection impact assessments and establish measures to safeguard personal data.
3. **Freedom of Expression and Assembly**: Consider whether the AI system might limit an individual's freedom of expression or assembly. Assess its potential impact on an individual's ability to openly express opinions, participate in peaceful demonstrations, or join unions.

2.2 THE FRIA

2.2.1 Non-discrimination

LUMINOUS project places utmost importance on ensuring the impartiality and equity of our AI-based software systems developed within the WPs. A rigorous set of activities of laborious testing and continuous monitoring for the duration of the project will be implemented to promptly identify and address any instances of bias or discrimination. We persistently commit to the regular review and updating of our processes to proactively respond to evolving challenges, ensuring a system that upholds principles of fairness and equity for all. The persistent consortium effort to uphold these principles underscores our mission to foster a technology landscape that is inclusive, just, and respectful of diverse perspectives.

2.2.2 Privacy Protection

Privacy is a fundamental concern that we vigilantly address (D5.1 & D5.2). Our AI system strictly adheres to the principles delineated in the General Data Protection Regulation (GDPR) governing the protection of personal data. We conduct thorough assessments, including detailed Data Protection Impact Assessments (DPIAs), to thoroughly evaluate the necessity and proportionality of our data processing operations. A suite of measures, including safeguards and security mechanisms, is systematically implemented to ensure robust protection of personal data.

2.2.3 Freedom of Expression and Assembly

The preservation of your freedom of expression and assembly is of top importance to us. Our AI system undergoes meticulous assessment to identify potential risks to these fundamental rights. Proactive measures are implemented to mitigate these risks, and we will welcome engagement with experts in human rights and freedom of expression during the workshops and pilots' dissemination to ensure a comprehensive evaluation. Human rights are integral to our commitment, and we are resolute in upholding them.

3 THE ASSESSMENT LIST FOR TRUSTWORTHY AI

ALTAI[15] serves as a fundamental evaluation process for Trustworthy AI self-assessment. Organizations can adapt ALTAI to their specific AI systems, incorporating sector-specific elements. It enhances understanding of Trustworthy AI, including potential AI-related risks to society, the environment, consumers, workers, and marginalized groups. ALTAI encourages the involvement of all relevant stakeholders, whether within or outside the organization. It assesses the presence of suitable solutions or processes to meet requirements, fostering responsible competitiveness. By instilling trust in AI systems' lawfulness, ethics, and robustness, ALTAI promotes responsible and sustainable AI innovation in Europe. This approach positions Europe and its organizations as global leaders in ethical and cutting-edge AI, benefiting individuals and society at large. The EEAB will review the self-assessment Trustworthy AI Assessment List to be carried out from the consortium partners per pilot that are involved with AI development as shown in Table 3.

Table 3 – LUMINOUS partners involved in AI development per pilot.

Pilot No	Partners involved in AI
#1: Neurorehabilitation	HYPERCLIQ; DFKI; EHU; VICOMTECH; RICOH; MINDMAZE
#2: Health, Safety and Environment Training	HYPERCLIQ; DFKI; EHU; VICOMTECH; RICOH
#3 BIM and Architectural Design Review	HYPERCLIQ; DFKI; EHU; VICOMTECH; RICOH

The self-assessment is considered in parallel with all the efforts made for the development, deployment and use of AI systems within the activities of LUMINOUS project to meet the seven key requirements for Trustworthy AI[16]:

1. **Human Agency and Oversight:** AI systems should enable individuals to make informed decisions while upholding their fundamental rights. Effective oversight mechanisms should be in place, involving human input.

[15] The Assessment List for Trustworthy Artificial Intelligence, available at https://altai.insight-centre.org/

[16] Ethics guidelines for trustworthy AI, available at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

2. **Technical Robustness and Safety:** AI systems must be resilient and secure, with contingency plans for unforeseen situations. They should be accurate, reliable, and reproducible to minimize unintentional harm.
3. **Privacy and Data Governance:** Respecting privacy and data protection is essential. Proper data governance mechanisms should ensure data quality, integrity, and authorized access.
4. **Transparency:** AI system operations, data handling, and business models should be transparent. Traceability methods enhance transparency, and explanations of AI system decisions should be tailored to the stakeholders involved.
5. **Diversity, Non-discrimination, and Fairness:** AI systems should avoid unfair bias, preventing negative consequences like marginalizing vulnerable groups and increasing prejudice. Inclusivity for all, regardless of disabilities, is vital. Involving relevant stakeholders throughout the AI system's life cycle is, also, crucial.
6. **Societal and Environmental Well-being:** AI systems should benefit everyone, including future generations while being sustainable and environmentally friendly. Their broader impact on society and the environment should be carefully considered.
7. **Accountability:** Mechanisms should be established to ensure responsibility and accountability for AI systems and their outcomes. Auditability, allowing the evaluation of algorithms, data, and design processes, is especially critical for critical applications. Accessible avenues for redress should also be guaranteed.

The technical and non-technical methods need to be considered to ensure the implementation of those requirements as per European Commission guidelines. The consortium will follow the developments of the new EU AI Act rules, that will establish obligations for the providers and users depending on the level of risk attributed to the deployment of the AI-based software systems developed within the project. While the AI systems within LUMINOUS software development and employment of the applications might pose minimal risk, they still need to be assessed according to the new EU AI Act rules once they are publicly available. The use of Artificial Intelligence in the EU will be regulated by the AI Act, the world’s first comprehensive AI law. Within the LUMINOUS project, the designing and implementation of an AI software system based on machine learning (ML) involves a systematic pipeline to ensure the model's effectiveness and reliability, which is presented in the following figure.

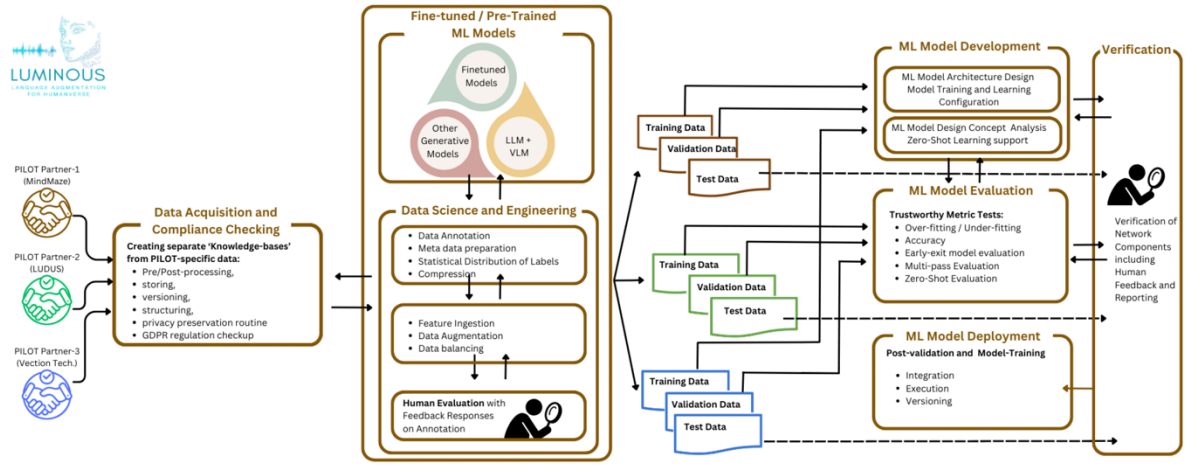


Figure 1: LUMINOUS machine learning systematic pipeline.

Regardless of the task the pipeline is, from a general perspective, equivalent for all the cases. The process starts with data acquisition and ingestion, where the data requirements are identified. Some data might be already available, while other data might require specific tasks, such as specific data generation, manual labelling, etc. Data validation and curation cleans the data and checks that it is ready to be used and as free from undesired noise as possible. Then, data pre-processing shapes the data into whichever format fits better for the AI model. The core AI model training/adaptation can happen in two different ways, that are not incompatible with each other. The in-context learning approaches work of teaching the model the objective task by manipulating the prompt (the input given to the model). This implies several techniques, such as expert-knowledge injection, Retrieval Augmented Generation (RAG), few-shots learning, Chain-of-X approaches, etc. In this case the weights of the models are not updated. In the model fine-tuning approach, the model is trained with pairs of input-output examples, to optimize it to the task. In this case the model weights are updated, so the model itself changes once the training is done. Regardless of the approach, there is a validation phase to assess the behaviour and performance of the model for each given task. This is done using quantitative metrics and the qualitative perception of the way the model completes a given task. Once the performance of the model is good enough, the model is integrated and deployed to provide its expected functionality in real-world applications.

Among the many pillars that ensure the trustworthiness of machine learning (ML) models and algorithms, "Security, Reliability, and Trust" are paramount. These three principles guarantee that artificial intelligence (AI) systems achieve the desired levels of 1) technical robustness, 2) transparency, 3) diversity with fairness, 4) societal well-being, and 5) accountability. In the LUMINOUS project, the pilot-specific (mutually exclusive and separate) diagram for data acquisition, ML model design, evaluation, and deployment illustrates distinct, sequential processes for end-to-end model deployment. This includes steps for data acquisition and

compliance checks, which assess data quality and determine necessary restrictions through proper versioning, cleansing, and structuring of datasets. Only datasets that comply with these standards (along with their annotations) are used for training and testing ML models, which learn the relationships between different high-dimensional representations. Additionally, existing ML algorithms and models can generate valuable metadata and annotations to ensure balance if annotations are disproportionately distributed. An initial human evaluation of the generated metadata ensures there are no unethical distortions before the development of any ML model begins. Samples for the test data are drawn from real-world scenarios for the situational awareness, and the development and deployment of ML models undergo several trustworthy metric tests and user surveys on the final outcomes before their practical application.
Ensuring Trustworthy AI in LUMINOUS's recommendations towards the users, it is considered crucial for its effectiveness and ethical use to follow several key considerations that the project adopts to enhance the trustworthiness of its AI recommendations:

1. **Explainability:** Making the model's decision-making process transparent by providing explanations for the recommendations.
2. **Data quality and bias mitigation:** Thoroughly assess and clean the training data to minimize biases.
3. **Ethical Considerations:** Establish and adhere to ethical guidelines in the design and use of the AI model(s). Consider the potential impact of recommendations on individuals and ensure that the model aligns with ethical principles, respecting privacy and avoiding harm.
4. **Continuous Monitoring**: Implement a robust monitoring system to continuously assess the model's performance in real-world scenarios.
5. **Security Measures**: Implement strong security measures to safeguard user data and the model itself. Ensure that data used for training and making recommendations is protected from unauthorized access, maintaining the confidentiality and integrity of the system.
6. **Regulatory Compliance:** Stay informed and comply with relevant data protection and privacy regulations. Adhering to legal standards ensures that the AI model(s) is/are deployed and used responsibly.
7. **Robustness Testing:** Subject the model to rigorous testing scenarios to evaluate its robustness. Simulate various conditions and edge cases to ensure the model behaves reliably in different situations.

LUMINOUS technical partners have filled-in the ALTAI questionnaire which is presented in the Annex II of this report. The recommendations of this exercise are presented below followed by proposed actions from the project to adopt the recommendations during the deployment and pilot testing phases.

3.1 HUMAN AGENCY AND OVERSIGHT

No recommendation for this requirement

3.2 TECHNICAL ROBUSTNESS AND SAFETY

Recommendation: Inform users as soon as possible if new threats are detected.

3.3 PRIVACY AND DATA GOVERNANCE

The LUMINOUS consortium in its work plan has a dedicated work package (WP) related to ethics dealing among others with privacy and data protection concerning the AI system, as well as an External Ethics Advisory Board (EEAB) that oversees and provides consults on these operations. There exists a rigorous mechanism that allows flagging issues related to privacy or data protection concerning the AI system. Also, a position dedicated to overseeing the Ethics managerial procedures of all WPs, and especially of WP5 (Ethics), the one of the "Ethics Manager" has been put in place. The Ethics Manager is a position of liaison among the EEAB, and the WPs operations and work being carried out. The project, in alignment with relevant standards, for its data management and governance adopts the Data Version Control protocol (DVC: https://dvc.org/). DVC is built to make ML models shareable and reproducible. It is designed to handle large files, data sets, machine learning models, and metrics as well as code.

3.4 TRANSPARENCY

Each pilot will implement an evaluation framework to design and implement a pilot plan that defines the activities and the detailed execution timeline of each pilot separately, as well as the interactions among the pilots, based on the outcomes of use cases definition. The demonstration plan will coordinate and align all pilots with each other by guaranteeing the exchange of practices and experiences. It will be developed in coherence with the key components of the LUMINOUS project and will be co-designed by the key pilot manager partners. The appropriate administrative, legal, and ethical processes will be elaborated in time under close collaboration of project partners and learning communities involved. Alongside, the task will produce the evaluation guidelines and the respective KPIs for all individual modules as well as the framework, which will walk through the technical, operational and user acceptance evaluation process. LUMINOUS will develop a set of metrics and instruments (e.g., questionnaires, interviews, etc.) for the diverse target groups and organization types participating in the demonstrators.
The technical limitations and potential risks of the AI system to end-users, such as its level of accuracy and/or error rates, will be communicated through the consent forms of pilot users. A disclaimer will be created that will communicate any technical limitations and potential risks of the solution to the end-users.

3.5 Diversity, non-discrimination and fairness

LUMINOUS is research driven project that has a Technological Readiness Level of maximum 6, meaning that the technological solution will be demonstrated in a relevant environment. The Evaluation framework and demonstration plan will specify the testing environment of the solution. These processes are to be implemented during the ML models systematic pipeline (Figure 1) through the assessment and cleansing of the training data to minimize biases. LUMINOUS ensures that stakeholders are treated fairly, as recommendations are not produced based on a one-size-fits-all approach but an inclusive one that takes into account the learning abilities of each individual learner. Diversity and representativeness of end-users in the data has been taken into consideration in all three pilot cases of the project. LUMINOUS's AI designers and AI developers are aware of possible bias and to mitigate such risks, an indicative list of educational resources are available:

- AI Ethics
- Fairness and Accountability in Machine Learning
- Responsible AI Practices
- AI and Bias Resources
- Algorithmic Justice League
- AI for Everyone
- MIT Technology Review - AI Ethics
- Partnership on AI
- AI & Ethics Podcast
- AI Now Institute
- Universal Guidelines for AI

All ALTAI recommendations will be taken into account during design and implementation processes for all pilots.

3.6 Societal and Environmental Well-being

The impact of the LUMINOUS Ai software systems will need to be assessed in terms of benefits to the society, while measuring the environmental impact. Specific dissemination, communication and training activities will take place for the duration of the project to inform the public and ensure the delivery of each pilot and impact of AI-based software systems are well understood by the stakeholders.

3.7 Accountability

LUMINOUS will design a system so that it can be audited with ease, to promote accountability and transparency. Thus, it is highly recommended to ensure traceability of the control and data flow and suitable logging mechanisms. The combination of those architectures provides high scalability and modularity since each component of the system will be developed as an

individual module that will communicate with others through events. AI systems should be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimizing unintentional and unexpected harm and preventing unacceptable harm. Consequently, developers and deployers should receive appropriate training about the legal framework that applies for the deployed systems. Indicative training seminars on AI & Legal matters are available at the same resources of Section 3.5 above.
If AI systems are increasingly used for decision support or for taking decisions themselves, it has to be made sure these systems are fair in their impact on people's lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this. Consequently, all conflicts of values, or trade-offs should be well documented and explained.

Although the ALTAI (Annex II) self-assessment has not provided any recommendations on the human oversight, the consortium will adopt two approaches on human oversight dimension:

- **"Human-in-the-loop"** which refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation and
- **"Human-in-command"** refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal, and ethical impact) and the ability to decide when and how to use the AI system in any specific situation.

The latter can include the decision not to use an AI system in a particular situation to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by an AI system.

4 Mitigating Bias and Discrimination in AI: Strategies and Safeguards

"Low data quality or poorly developed machine learning algorithms can lead to predictions that put certain groups of people at a disadvantage. Highly automated settings are prone to feedback loops, which is why high levels of automation should not be considered in areas that have an impact on people without meaningful human intervention and oversight at all stages. Algorithms are only as good as the data they are fed. The results will also be questionable if the data are out-of-date, inaccurate, incomplete, or poorly chosen. AI systems built on biased or incomplete

data may produce unreliable results that violate people's fundamental rights, such as the right to be free from discrimination"[17].

Under this perspective, LUMINOUS highlights "the need for more extensive evaluations of algorithmic bias before using them to make decisions that could affect human beings to help guard against potential violations of fundamental rights" and as such, we plan the following actions as shown in Table 4[18]:

Table *4* – AI Bias and Discrimination and how to handle them in LUMINOUS.

Issue	Awareness and mitigation strategies	Mitigation Strategies
Recognizing AI Bias and Its Consequences AI systems are susceptible to bias, which reflects social prejudices present in the training data.	To Examine the many forms and effects of bias in AI and the significance of resolving this problem to guarantee just and equal results. Discover how to reduce prejudice and advance algorithmic justice in artificial intelligence systems.	Ensure diverse and representative training data, detecting and addressing biases in algorithms and models, testing for fairness across demographics, monitoring models post-deployment, promoting transparency and accountability, educating stakeholders, and fostering collaboration and diversity in AI development teams.
Protecting Privacy in AI AI depends on enormous volumes of data; privacy and data protection are important issues to be considered.	Examine the moral issues related to the gathering, using, and storing of data for AI applications. Learn about privacy-enhancing strategies and laws that can protect people's privacy while utilizing AI.	Ensure data anonymization, encryption, and differential privacy techniques to protect sensitive information. Adhere to privacy regulations, establish transparent data usage policies, obtain user consent, and regularly audit AI systems for compliance.
Encouraging Algorithmic Explainability and Transparency Explainability and transparency are crucial when AI	Analyse the difficulties associated with algorithmic transparency and the significance of comprehending the decision-making process used by AI. Learn about methods and	Utilise interpretable models to provide insights into model decisions. Document model architectures, disclose data sources and pre-processing steps, and implement mechanisms for users to

[17] Bias in Algorithms – Artificial Intelligence and Discrimination, p. 77
[18] The Ethics of AI: Navigating Bias, Privacy, and Algorithmic Transparency, Arcot Group, 2023

systems make decisions that affect people's lives.	structures that encourage accountability and openness in AI systems.	understand and question AI outputs.
Ensuring Proactive Design and Development of Ethical AI Proactive design and development are necessary for ethical AI.	Examine moral frameworks and rules that businesses may use to guarantee moral behaviour all the way through the AI lifecycle. Find out how varied viewpoints and interdisciplinary teams are crucial for tackling ethical problems.	Integrate ethical considerations at every stage of the AI systems' lifecycle. Identify potential ethical implications, establishing clear guidelines and principles for AI development, conducting ethical impact assessments, and engaging diverse stakeholders.
Policy and Regulation's Place in AI Ethics Policies and regulations are essential in determining the ethical framework surrounding AI.	Examine current and upcoming laws and initiatives pertaining to AI ethics. Recognize the effects of legal frameworks and the value of taking the initiative to influence ethical AI practices.	Establish guidelines that govern the development, deployment, and use of AI technologies within the project. Define ethical standards, ensuring transparency and accountability, protecting privacy and data rights, addressing bias and discrimination, and promoting safety and security.
Establishing an Ethical AI Culture Establishing an ethical AI culture is crucial for businesses.	Examine methods for encouraging moral behaviour and judgement when it comes to AI. Discover how to incorporate moral values into organizational issues and AI governance.	Promote values such as transparency, fairness, accountability, and responsible innovation throughout the lifetime of the project within the consortium. Provide regular ethics related sessions at general assembly meetings integrating ethical guidelines into AI development processes.

5 PARTICIPANT COMMUNICATION AND TRANSPARENCY GUIDELINES FOR AI SYSTEMS

Research participants and end users will be provided with a comprehensive understanding with regards to their interaction with AI systems/technologies, including their abilities, limitations, and capacities. The LUMINOUS consortium will follow the principles of transparency, human agency, and oversight, as identified in Ethical Guidelines for Trustworthy AI[19], and prioritize the relevant guidelines as stated in Ethics by Design and Ethics of Use Approaches for Artificial Intelligence for Transparency[20]:

- It MUST be made clear to end-users that they are interacting with an AI system (especially for systems that simulate human communication, such as avatars).
- The purpose, capabilities, limitations, benefits, and risks of the AI system and the decisions conveyed by it MUST be openly communicated to end-users and other stakeholders, including instructions on how to use the system properly.
- When building an AI solution, one MUST consider what measures will enable the traceability of the AI system during its entire lifecycle, from initial design to post-deployment evaluation and audit or in case its use is contested.
- Whenever relevant, the research proposal should offer details about how decisions made by the system will be explainable to users. Where possible, this should include the reason why the system made a particular decision.

Communication of interaction: It will be clearly stated how research participants/end users will interact with the AI system or technology, making sure that it provides an understandable, and transparent explanation of the user experience in a jargon-free language.
Understanding Abilities, Limitations, Risks, and Benefits: Research Participants/end users will be thoroughly informed about what the AI system can and cannot do. As such, capabilities and limitations will be clearly outlined to manage all expectations respectively. The advantages and disadvantages of using the AI system of the LUMINOUS project will be presented prior to their participation and the sign of the consent form to facilitate people make well-informed decisions.
Insight into Decision-Making Logic: The LUMINOUS consortium will provide justification of the reasoning behind any decisions made by the AI system (Explainable AI). Prior to participants' involvement and signing of the consent form, end users will be provided with all relevant information in a comprehensive and transparent manner to foster openness and confidence in the operation of the system.

[19] Ethics guidelines for trustworthy AI, European Commission, 2019
[20] Ethics By Design and Ethics of Use Approaches for Artificial Intelligence, p. 8-9, European Commission, 2021

6 ETHICAL RISK ASSESSMENT AND MITIGATION IN AI LIFECYCLE

The pilot managers have incorporated the mitigation measures and although the initial evaluation level (low, medium etc) is speculated, the purpose of this deliverable is to present the guidelines and recommendations in terms of ethical risk assessment for the consortium partners' awareness and compliance. Following the risks' assessment as identified in Data Process Impact Assessment, available in D5.2, those related to AI can be summarized in Table 5 as follows:

Table *5* – Protection of personal data and reference frameworks

Description of risk and the nature of the potential impact on individuals	**Likelihood of harm Remote, possible or probable**	**Severity of harm Minimal, significant or severe**	**Overall risk Low, medium or high**	**Mitigation Measure**
Bias and fairness: The AI algorithms may inadvertently incorporate biases, leading to unequal treatment or reinforcement of existing inequalities. Unintended bias in recommendations may result in unequal opportunities, reinforcing stereotypes, and potentially disadvantaging certain stakeholder groups.	remote	minimal	low	Ensure the training dataset for machine learning models is diverse and representative of the user population. Regularly each pilot manager will audit datasets for biases and take corrective actions to improve fairness and inclusivity.
Inaccurate recommendations/feedback: Machine learning algorithms may provide inaccurate or inappropriate recommendations, impacting the quality of the pilots. Stakeholders may receive guidance that does not align with their needs or may be directed towards irrelevant tasks.	possible	significant	medium	Incorporate human oversight into the decision-making process of the AI system. Enable intervention in cases where the AI's recommendations may need human judgement.

Limited Representation of user profiles and styles: The AI algorithms may not fully capture the diversity of user profiles. Some users may not receive recommendations that align with their unique preferences, potentially hindering their engagement and understanding of the pilot intervention.	remote	minimal	low	Implement adaptive learner profiles that continuously evolve and adjust recommendations based on user feedback and changes in learning styles. This ensures that the system remains responsive to individual needs
Lack of Explainability: The AI algorithms may lack transparency and explainability, making it challenging for users. The lack of transparency may result in a reduced ability to trust the system, potentially leading to scepticism or resistance among users.	remote	minimal	low	Enhance the explainability of AI recommendations. Provide users with clear explanations of why certain recommendations are made, highlighting the factors and criteria considered by the algorithm.

7 CONCLUSION

D5.3 provides a general overview of the ethical assessment for AI within the project. It outlines the processes that have followed to safeguard that AI will be implemented through all the project phases in a manner safeguarding fundamental rights and addressing efficiently all seven criteria for Trustworthy AI. Furthermore, the document encompasses a precise consideration of perspectives related to bias and discrimination, as well as an in-depth analysis of specific AI risks, accompanied by corresponding mitigation measures. The Ethics Manager will collaborate with the LUMINOUS External Ethics Advisory Board and the consortium to ensure strict adherence to these processes. The Ethics Manager will take proactive measures for additional ethical monitoring when deemed necessary.

8 REFERENCES

[1] Charter Of Fundamental Rights Of The European Union, available: https://www.europarl.europa.eu/charter/pdf/text_en.pdf
[2] Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI), Factsheet: Governance for digital transformation, and Council of Europe, Recommendation CM/Rec (2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems (adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies), available at https://rm.coe.int/09000016809e1154
[3] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast), OJ L 204, 26.7.2006, pp. 23-36, available at https://www.eumonitor.eu/9353000/1/j4nvk6yhcbpeywk_j9vvik7m1c3gyxp/vitgbgik8rzi
[4] Ethics guidelines for trustworthy AI, available at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[5] European Commission, White Paper on Artificial Intelligence – A European approach to excellence and trust, COM (2020) 65 final, Brussels, 19 February 2020, p. 2. 25 Ibid., p. 12, available at https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-int elligence-feb2020_en.pdf
[6] European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, 4 November 1950, ETS 5, available at https://www.echr.coe.int/documents/d/echr/convention_ENG
[7] Overview of the application of the Charter, see FRA (2018a), Applying the Charter of Fundamental Rights of the European Union in law and policy making at national level, Luxembourg, Publications Office. Getting The Future Right Artificial Intelligence And Fundamental Rights, available at https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights
[8] High-level expert group, available at https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai
[9] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, pp. 1-88, available at https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32016R0679
[10] Recommendation on the Ethics of Artificial Intelligence available at https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence
[11] The Assessment List for Trustworthy Artificial Intelligence, available at https://altai.insight-centre.org/
[12] Bias in Algorithms –Artificial Intelligence and Discrimination, available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf

[13] The Ethics of AI: Navigating Bias, Privacy, and Algorithmic Transparency, Arcot Group, 2023, available at https://www.linkedin.com/pulse/ethics-ai-navigating-bias-privacy-algorithmic-transparency
[14] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, OJ L 180, 19.7.2000, pp. 22-26, available at https://www.eumonitor.eu/9353000/1/j4nvk6yhcbpeywk_j9vvik7m1c3gyxp/vitgbgi6x6z8
[15] Council Directive 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, OJ L 373, 21.12.2004, pp. 37-43, availability at https://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:373:0037:0043:en:PDF
[16] European Parliament resolution of 14 March 2017 on fundamental rights implications of big data: privacy, data protection, nondiscrimination, security and law-enforcement (2016/2225(INI)), para. 1 available at https://www.europarl.europa.eu/doceo/document/TA-8-2017-0076_EN.html
[17] Ethics By Design and Ethics of Use Approaches for Artificial Intelligence, p. 8-9, European Commission, 2021 available at: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf
[18] EU-AI-Act: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

ANNEX I: FUNDAMENTAL RIGHTS IMPACT ASSESSMENT (FRIA)

FUNDAMENTAL RIGHTS IMPACT ASSESSMENT (FRIA) OF THE LUMINOUS PROJECT

FRIAs template for all pilots is provided below with a plan to provide any updates, if necessary, for each pilot at the Annex of D6.1, D6.2 and D6.3 respectively for each pilot.

Fundamental Rights Impact Assessment

Created by:
https://aligner-h2020.eu/fundamental-rights-impact-assessment-fria/
ALIGNER

Template #1: Fundamental Rights Impact Assessment

Fundamental Rights Impact Assessment Template	
Name	Florendia Fourli
Organisation/Position	CEO Hypercliq IKE
Date	22/04/2024
Contributors	Eleni Mangina (UCD), Muhammad Zeshan Afzal (DFKI), Daniel Perez-Marcos (Mindmaze), Florendia Fourli (Hypercliq), Oier Lopez de Lacalle (EHU/UPV), Nicoletta Cioria (Mindesk), Ander Salaberria (EHU/UPV), Marco Bianchi (LUDUS), Didier Stricker (DFKI)
AI system assessed	LUMINOUS
Detailed description of the technology and input data	Creation of Language Augmented extended reality (XR) systems and applications, where natural language-based communication and Large Language Models (LLM) redefine the future interaction with novel XR technology and enhances understanding of the users' situation and environment even in situations that are encountered for the first time. Input data consists of user interactions with the XR system including voice commands/dialogue, guestures, eye movements and usage data (e.g. scores in games).
Detailed description of the purposes and context of use	The LUMINOUS system will be piloted in three different application areas: Neurorehabilitation, Health, Safety and Environment training and BIM & architectural design review

1. Presumption of innocence and right to an effective remedy and to a fair trial Everyone charged with a criminal offence must be presumed innocent until proved guilty according to law. Everyone whose rights and freedoms are violated has the right to an effective remedy before a tribunal. Everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal previously established by law, including rights: ❖ to be informed promptly of the nature and cause of the accusation; ❖ to bring their arguments and evidence as well as scrutinise and counteract the evidence presented against them; and to obtain an adequately reasoned and accessible decision.		
Challenge	Evaluation	Estimated impact level
1.1 The AI system does not communicate that a decision/advice or outcome is the result of an algorithmic decision	N/A	-
1.2 The AI system does not provide percentages or other indication on the degree of likelihood that the outcome is correct/incorrect, prejudicing the user that there is no possibility of error and therefore that the outcome is undoubtedly incriminating	N/A	-
1.3 The AI system produces an outcome that forces a reversal of burden of proof upon the suspect, by presenting itself as an absolute truth, practically depriving the defence of any chance to counter it	N/A	-
1.4 There is no explanation of reasons and criteria behind a certain output of the AI system that the user can understand	N/A	-
1.5 There is no indication of the extent to which the AI system influences the overall decision-making process	N/A	-
1.6 There is no set of measures that allow for redress in case of the occurrence of any harm or adverse impact	N/A	-
2. Right to equality and non-discrimination Everyone is equal before the law. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, Everyone should be protected against discriminatory decisions or policies, including automated decision-making based on sensitive data.		
Challenge	Evaluation	Estimated impact level
2.1 The AI system targets members of a specific social group	Neurorehabilitation Pilot: Stroke affects men and women equally. Inclusion criteria of planned studies do not discriminate by sex/gender or any other social condition. The AI system does consider gender diversity via the selection of representative training data.	Low
2.2 There are no mechanisms to flag and correct issues related to bias, discrimination, or poor performance	The AI system will not provide such mechanisms but relevant tests will be carried out on the prototypes to be created.	Medium
2.3 The AI system does not consider the diversity and representativeness for specific population or problematic use cases	The AI system does consider the diversity and representativeness for all users via the selection of represenative training data.	Low
3. Freedom of expression and information Everyone has the right to freedom of expression, including freedom to hold opinions, communicate and acquire information ❖ State negative obligation not to interfere and positive obligation to facilitate the exercise of the right		
Challenge	Evaluation	Estimated impact level
3.1 There is no mechanism to limit the deployment of the AI system to suspected individuals	N/A	-
3.2 The data stored, recorded, and produced are not easily accessible to concerned individuals	The data stored, recorded and produced by the prototypes will be easily accessible to concerned individuals upon request.	Low

4. Right to respect for private and family life and right to protection of personal data Everyone has the right to respect for their private and family life, home and communications. ❖ Self-development without state interference. ❖ Everyone has the right to the protection of personal data concerning them. ❖ Personal data must be processed fairly for specified purposes and on a legitimate basis. ❖ Rights of access and rectification. ❖ Independent oversight.		
Challenge	Evaluation	Estimated impact level
4.1 There are no mechanisms for the user to exercise control over the processing of personal data	The users of the AI system will be adequately informed of the processing of their personal data before use of the system.	Medium
4.2 There are no measures to ensure the lawfulness of the processing of personal data	Measures to ensure the lawfulness of the processing of the personal data by the AI system have been established (e.g. GDPR compliance by design).	Low
4.3 There are no procedures to limit the access to personal data and to the extent and amount necessary for those purposes	Measures to limit the access to personal data by the AI system have been established (e.g. data anonymization, local/secure data processing).	Low
4.4 There is no mechanism allowing to comply with the exercise of data subject's rights (access, rectification and erasure of data relating to a specific individual)	Mechanisms allowing to comply with the exercise of data subject's rights have been established.	Low
4.5 There are no specific measures in place to enhance the security of the processing of personal data (via encryption, anonymisation and aggregation)	Specific measures are in place to enhance the security of the processing of personal data (e.g. anonymisation).	Low
4.6 There is no procedure to conduct a data protection impact assessment	The AI system has to follow procedures to conduct data protection impact assesments when and as applicable by each involved organisations data protection policies.	Medium

LUMINOUS
LANGUAGE AUGMENTATION FOR HUMANVERSE

Template #2: AI System Governance

Created by: ALIGNER

AI System Governance	
Name	Florendia Fourli
Organisation/Position	CEO Hypercliq IKE
Date	22/04/2024
Contributors	Eleni Mangina (UCD), Muhammad Zeshan Afzal (DFKI), Daniel Perez-Marcos (Mindmaze), Florendia Fourli (Hypercliq), Oier Lopez de Lacalle (EHU/UPV), Nicoletta Cioria (Mindesk), Ander Salaberria (EHU/UPV), Marco Bianchi (LUDUS), Didier Stricker (DFKI)
AI system assessed	LUMINOUS
Detailed description of the technology and input data	Creation of Language Augmented extended reality (XR) systems and applications, where natural language-based communication and Large Language Models (LLM) redefine the future interaction with novel XR technology and enhances understanding of the users' situation and environment even in situations that are encountered for the first time. Input data consists of user interactions with the XR system including voice commands/dialogue, guestures, eye movements and usage data (e.g. scores in games).
Detailed description of the purposes and context of use	The LUMINOUS system will be piloted in three different application areas: Neurorehabilitation, Health, Safety and Environment training and BIM & architectural design review

1. Human autonomy								
Component	Minimum standards to be achieved	Initial impact estimate		Additional mitigation measures implemented	Final assessment		Responsible department	Timeline
		Challenge no.	Impact level		Final estimated impact level	Further actions		
Human agency	The task allocation between the AI system and the user allows meaningful interactions	[1.2]	High	The AI system interacts with the users via natural language.	Low		Luminous technical team	Dec-26
		[1.5]	High	The user can play an active role in the actions of the AI system by influencing the systems outputs via the language interface.	Low		Luminous technical team	Dec-26
	There are procedures to describe the level of human involvement and the moments for human interventions	[1.5]	-					
		[2.2]	Medium	Procedures are described in the use cases	Low		Luminous technical team	Aug-24
		[4.1]	Medium					
Human oversight	The AI system does not affect human autonomy by interfering with the user decision-making process	[1.2]	-					
		[1.3]	-					
		[1.5]	-					
		[4.1]	Medium	The AI system does not interfere with the user decision-making process as it only provides suggestions for actions	Low		Luminous techninal team	Dec-26
	There are mechanisms to prevent overconfidence or over-reliance in the results offered by the AI system	[1.1]	-					
		[1.2]	-					
	There are mechanisms to detect and correct wrong outputs	[1.6]	-					
		[2.2]	Medium	Mechanisms to detect and correct wrong outputs have been considered at design phase and will be impemented	Low		Luminous technical team	Dec-26
		[2.3]	Low					
	There are mechanisms to safely abort an entire operation when needed			Yes				

2. Transparency								
Component	Minimum standards to be achieved	Initial impact estimate		Additional mitigation measures implemented	Final assessment		Responsible department	Timeline
		Challenge no.	Impact level		Final estimated impact level	Further actions		
Traceability	There are mechanisms to ensure the traceability of the input data used by the AI system and its outcomes			Tracebility of input data is possible				
Explainability	It is possible for the user to understand and explain the reasons and criteria behind a certain output of the AI system	[1.4]	-	The AI system provides adequate expaination of it's outputs via the language interface	Low		Luminous technical team	Dec-26
Communication	There are procedures enabling the user to communicate to the public that decisions are taken on the basis of an algorithmic process	[1.3]	-	n/a				
	There are procedures enabling the user to explain to the public the purposes, characteristics, limitations, and shortcomings of the AI system			n/a				
	There are procedures enabling the user to make the data stored, recorded, and produced available to concerned individuals	[3.2]	Low	n/a				
3. Diversity, non-discrimination and fairness								
Component	Minimum standards to be achieved	Initial impact estimate		Additional mitigation measures implemented	Final assessment		Responsible department	Timeline
		Challenge no.	Impact level		Final estimated impact level	Further actions		
Unfair bias avoidance	There are procedures to test and evaluate the diversity and representativeness of the used datasets, also for specific social group or use cases	[2.3]	Low	Data collection protocols to be submitted to and approved by the local Ethics Committee of the organisations carrying out the studies	Low		CHUV, UCL	Dec-26
	There are procedures to test and evaluate the diversity and representativeness of the algorithm used, also for specific social groups or use cases	[2.3]	Low	Yes			Luminous technical team	Dec-26
	There are procedures to evaluate whether specific social groups are disproportionally affected by the AI system	[2.1]	Low	Yes			Luminous technical team	Dec-26
	There are mechanisms to flag and correct bias, discrimination or poor performance	[2.2]	Medium	Yes by design (ref. ML Pipeline)			Luminous technical team	Dec-26

4. Democracy and societal wellbeing								
Component	Minimum standards to be achieved	Initial impact estimate		Additional mitigation measures implemented	Final assessment		Responsible department	Timeline
		Challenge no.	Impact level		Final estimated impact level	Further actions		
Social impact	There are procedures to ensure that the social impacts of the AI systems are well understood by the public			Project disemmination procedures			Luminous technical team	Dec-26
Society and democracy	There are procedures to assess the broad social impact of the AI system (e.g., chilling effect, power asymmetry, trust, …)			n/a				
	There are mechanisms to limit the deployment of the AI system to groups of individuals on the basis of suspicion/objective criteria	[3.1]	-	n/a				

5. Privacy and data governance								
Component	Minimum standards to be achieved	Initial impact estimate		Additional mitigation measures implemented	Final assessment		Responsible department	Timeline
		Challenge no.	Impact Level		Final estimated impact level	Further actions		
Respect for privacy and data protection	There are mechanisms for the user to exercise control over the processing of personal data	[4.1]	Medium	Yes and users are informed before use				
	There are measures to ensure the lawfulness of the processing of personal data	[4.2]	Low	Yes			DPO of each organisation	
	There are measures to minimise the amount of personal data processed	[4.3]	Low	Yes			DPO of each organisation	
	There is a mechanism allowing to comply with data subjects' rights	[4.4]	Low	Yes			DPO of each organisation	
Quality and integrity of data	There are specific measures to enhance the security of the processing of personal data (via encryption, anonymization and aggregation)	[4.5]	Low	Yes			Luminous Technical team	
	There are processes to ensure the quality and integrity of data			Yes			Luminous Technical team	
	The AI system is aligned with relevant standards (ISO, IEEE) for data security, management and governance			Yes			Luminous Technical team	
Access to data	There are procedures to limit the access to personal data	[4.3]	Low	Yes				
Governance	There is a procedure to conduct a data protection impact assessment	[4.6]	Medium	Yes			DPO	
	A data protection officer has been appointed			Yes				
	There are mechanisms to allow reporting of processing activities to the supervisory body			Yes				
International data transfers	There are mechanisms to control the transfer of personal data to third countries			No international data transfers				

6. Technical robustness and safety								
Component	Minimum standards to be achieved	Initial impact estimate		Additional mitigation measures implemented	Final assessment		Responsible department	Timeline
		Challenge no.	Impact level		Final estimated impact level	Further actions		
Security	The potential vulnerability of the AI system has been assessed			Yes			Luminous Technical team	
	There are mechanisms to ensure the integrity and resilience of the AI system against potential cyberattacks			Yes			Luminous Technical team	
Fallback and general safety	There is a fallback plan for adversarial attacks or unexpected situations			Yes			Luminous Technical team	
Accuracy	There is an assessment of the level of accuracy required in relation to the envisaged use			Yes via evaluation procedures described in the workplan				
	There are mechanisms to evaluate and ensure that the used datasets are comprehensive and up to date			Yes via data collection protocols				
Reliability and reproducibility	There are procedures to evaluate the reliability and reproducibility of the AI system's aspects (inputs and outputs), also in specific contexts			Yes by design				

7. Accountability								
Component	Minimum standards to be achieved	Initial impact estimate		Additional mitigation measures implemented	Final assessment		Responsible department	Timeline
		Challenge no.	Impact level		Final estimated impact level	Further actions		
Competence	There are clear programs to provide information on the role of the operator, the competencies required to operate the AI system and the implications of operator error			Users will be informed during the evaluation phase				
	There are safeguards against incompetent operation of the AI system			Yes				
Misuse awareness	There is an assessment of the likelihood of misuse of the AI system and of its possible outcomes			No				
	There are ethics education and security awareness programs to sensitise the users to the potential risk of misuse			No				
Auditability	There are legged and traceable procedures to enable independent audit, also in order to remedy to identified issues in the AI system			No				
Ability to redress	There are measures that allow redress in case of the occurrence of any harm or adverse impact	[1.6]	-	Yes				
	There are procedures to provide information to affected parties about opportunity for redress			Yes				

ANNEX II: THE ASSESSMENT LIST FOR TRUSTWORTHY AI (ALTAI)

THE ASSESSMENT LIST FOR TRUSTWORTHY AI (ALTAI) OF THE LUMINOUS PROJECT

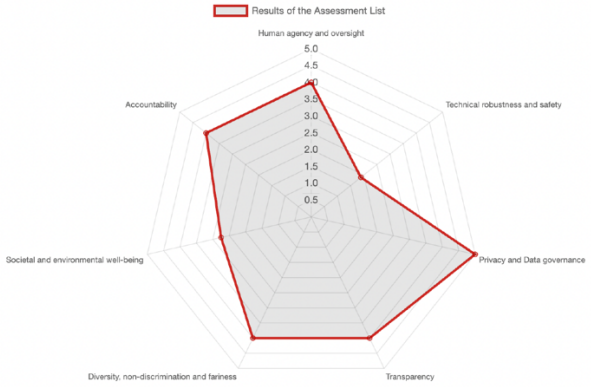
The ALTAI self-assessment for all pilots is provided below with a plan to provide any updates, if necessary, for each pilot at the Annex of D6.1, D6.2 and D6.3 respectively for each pilot.

The Assessment List for Trustworthy AI (ALTAI)[21]

[21] See recommendation from EEAB

Self assessment results

The requirements not completed score 0.



Recommendations

Human agency and oversight

No recommendation for this requirement.

Technical robustness and safety

Red-team/pentest the system

Inform users as soon as possible if some new threats are detected.

Privacy and Data Governance

Consider establishing mechanisms that allow flagging issues related to privacy or data protection concerning the AI system.

Transparency

Consider explaining the decision adopted or suggested by the AI system to its end users.

Diversity, non-discrimination and fairness

Test for specific target groups or problematic use cases.

Assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness).

Put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system.

Depending on the use case, ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system.

You should establish clear steps and ways of communicating on how and to whom such issues can be raised.

Consult with the impacted communities about the correct definition of fairness, such as representatives of elderly persons or persons with disabilities.

Ensure a quantitative analysis or metrics to measure and test the applied definition of fairness.

You should ensure that information about, and the user interface of, the AI system is accessible and usable also to users of assistive technologies (such as screenreaders).

You should involve or consult with end-users or subjects in need for assistive technology during the planning and development phase of the AI system.

You should assess whether there could be groups who might be disproportionately affected by the outcomes of the system.

You should assess the risk of the possible unfairness of the system onto the end-user's or subject's communities.

Societal and environmental well-being

Consider the potential positive and negative impacts of your AI system on the environment and establish mechanisms to evaluate this impact.

Define measures to reduce the environmental impact of your AI system's lifecycle and participate in competitions for the development of AI solutions that tackle this problem.

Inform and consult with the impacted workers and their representatives but also involve other stakeholders. Implement communication, education, and training at operational and management level.

Take measures to ensure that the work impacts of the AI system are well understood on the basis of an analysis of the work processes and the whole socio-technical system.

Provide training opportunities and materials for re- and up-skilling measures.

Accountability

To foresee 3rd party auditing or guidance can help with both, qualitative and quantitative risk analysis. In addition, it can contribute to generate trust in the technology and the product itself.

AI systems should be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. Consequently, developers and deployers should receive appropriate training about the legal framework that applies for the deployed systems.

If AI systems are increasingly used for decision support or for taking decisions themselves, it has to be made sure these systems are fair in their impact on people's lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this. Consequently, all conflicts of values, or trade-offs should be well documented and explained

Involving third parties to report on vulnerabilities and risks does help to identify and mitigate potential pitfalls

A risk management process should always include new findings since initial assumptions about the likelihood of occurrence for a specific risk might be faulty and thus, the quantitative risk analysis was not correct and should be revised with the new findings.

Acknowledging that redress is needed when incorrect predictions can cause adverse impacts to individuals is key to ensure trust. Particular attention should be paid to vulnerable persons or groups.

ANNEX III: EEAB RECOMMENDATION

15-05-2024

Dear Coordinator,

Dear Ethics Manager of the LUMINOUS project,

We hereby confirm that we have read Deliverable D5.3 AI-Requirement and provided feedback to the consortium partners. As an Ethics Advisory Board, we have noticed that the ALTAI checklist has some weaknesses, focused on the fact that the answers to the questions are not verified and supported by evidence. So, our recommendation is placed on addressing these weaknesses and explaining how they are planning to handle them throughout the project, providing a thoughtful and precise analysis of the technology and its deployment context, especially the domain-specific monitoring aspect. It is recommended to be taken into account during the implementation of the technologies within each Pilot over the future months of the development of the project.

Chair of the EEAB
Georgia Livieri
Ethics Expert