DELIVERABLE D2.1

THEMIS 5.0 Methodological Framework and Requirements Analysis

DELIVERABLE DETAILS		
SUBMISSION DATE	NAME OF THE DELIVERABLE	WORK PACKAGE
9.10.2024	THEMIS 5.0 Methodological Framework and Requirements Analysis	WP2
DISSEMINATION LEVEL	AUTHOR(S)	LEAD BENEFICIARY
PU (Public)	ATC, SINTEF, ICCS, IPT, KUL, UoS, Trustilio, IBM, also with all partners' contribution and input.	ATC

PROJECT DETAILS			
PROJECT ACRONYM	THEMIS 5.0	GRANT AGREEMENT	101121042
CALL IDENTIFIER	Human-02-01	PROJECT DURATION	01.10.2023 – 30.09.2026
PROJECT OFFICER	Anna Puig Centelles	PROJECT COORDINATOR	Gruppo Maggioli

THEMIS 5.0

QUALITY CONTROL ASSESSMENT

VERSION	DATE	DESCRIPTION	NAME	ORG
0.1	09.11.2023	Draft outline	Asbjørn Følstad Adam Doulgerakis Dimitris Apostolou	SINTEF ATC ICCS
0.2	06.02.2024	First initial draft, groundwork	All authors	All
0.3	09.04.2024	Second initial draft, groundwork	All authors	All
0.4	09.05.2024	First initial draft, requirements	All authors	All
0.5	05.07.2024	Second initial draft, requirements	All authors	All
0.6	24.09.2024	Final draft, all sections	All authors	All
0.7	03.10.2024	Review	John Beattie, Faidra Alevizou, Hara Stefanou	I4RI, IPT, MAG
1.0	08.10.2024	Final version for submission	Adam Doulgerakis, Eleni Tsalapati, Asbjørn Følstad, Filippo Remonato, Pål Vegard Johnsen	ATC, SINTEF

LIST OF AUTHORS

AUTHORS NAMES	ORG
Adam Doulgerakis, Eleni Tsalapati, Giorgos Giotis	ATC
Asbjørn Følstad, Filippo Remonato, Pål Vegard Johnsen	SINTEF
Mattheos Fikardos, Katerina Lepenioti, Dimitris Apostolou, Gregoris Mentzas	ICCS
Giannis Stamatellos, Alexandra Psaltidou	IPT
Sara Garsia, Ana Maria Corrêa Harcus	KUL
Steve Taylor, Samuel Senior	UoS
Kitty Kioskli, Theo Fotis	Trustilio
Arnaud Billion	IBM
George Karavokiros	MAG

DISCLAIMER

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This Deliverable contains original unpublished work except where clearly indicated otherwise.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.

ACKNOWLEDGEMENT

This document is a deliverable of the THEMIS 5.0 project, which has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101121042.

Table of contents

Terms and Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
AI HLEG	Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission
DE	Domain Expert
DL	Deep Learning
DSS	Decision Support System
ETA	Estimated Time of vessel Arrival
ENISA	European Union Agency for Network and Information Security
HCP	Healthcare professionals
ISO	International Organization for Standardization
GUI	Graphical User Interface
ML	Machine Learning
NIST	National Institute of Standards and Technology
RMF	Risk Management Framework
SSM	Spyderisk System Modeller
SUT	System Under Test
UCS	Use Case Scenario
UI	User Interface

EXECUTIVE SUMMARY

In WP2 of the THEMIS 5.0, the conceptual basis for the innovative approach of the project is developed, comprising the theoretical groundwork and the human-centric requirements that are summarized in this deliverable, as well as a comprehensive conceptual model to be presented in the forthcoming deliverable D2.2. 'THEMIS 5.0 Approaches and Conceptual Modelling'. The conceptual basis is established in dialogue with use-case partners as well as with users and domain experts through co-creation sessions (WP3) and are used to guide the technical architecture and development (WP4) as well as the piloting and evaluation (WP5).

This deliverable presents the theoretical groundwork for the THEMIS 5.0 approach to trustworthiness optimization of AI systems which details the concept of trustworthiness for AI systems, considers specific trustworthiness characteristics, expands on the involvement of human users for the adaptation of AI systems to human values and business targets, and discusses the use of risk management approaches for assessing AI trustworthiness.

The concept of trustworthy AI is grounded in the relevant knowledge base, including scientific literature, technical reports, standards, policy documents, and legal material. This concept is based on the notion that trustworthiness may be assessed and optimized through detectable trustworthiness criteria associated with the AI systems under test within their socio-technical environment.

Furthermore, relevant background on human-centred AI promoting the involvement of human users in optimization of AI trustworthiness is analysed, risk management approaches as means to address AI trustworthiness are discussed, and technologies supporting the analysis of AI trustworthiness characteristics and management are detailed. Finally, an overview of the ethical and legal basis for the THEMIS 5.0 approach is documented.

Furthermore, this deliverable presents the user requirements, the user journeys and the mock-ups for the THEMIS 5.0 ecosystem. Based on input from users and experts in the three case domains of THEMIS 5.0. the user requirements are identified, analysed and summarized on the basis of extensive user involvement. The requirements are instantiated in mock-ups illustrating user journeys to guide the technical development in the project.

The human-centric aspect of the conceptual work is further addressed in collaboration between WP2, and the co-creation activities performed within WP3, and includes the user requirements for the THEMIS ecosystem, the expectations of users from trustworthy AI in their line of work and us-case specific contextual information and trustworthiness considerations that will guide the development of the THEMIS 5.0 services. User journeys and mock-ups describing the different ways that users can use the functionalities of the THEMIS 5.0 ecosystem for the optimisation of AI models in the project use cases were developed to support the upcoming co-creation activities (Living Labs and Piloting).

1. INTRODUCTION

1.1. The THEMIS 5.0 project context

Artificial Intelligence (AI) is taken up for an increasingly broad range of decision-making purposes, e.g. in health (Giordano et al., 2021), logistics (Richey et al., 2023), and journalism (Moravec et al., 2024). AI is applied for work and organizational tasks previously thought to require human cognitive and social skills (Frey & Osborne, 2023) – in addition to domain-specific expertise. Such embedding of the AI, in individual, organizational and societal contexts of decision-making, holds potentially profound implications. AI decision support may impact decision-making directly, potentially affecting the outcome of the specific decision-making process. Furthermore, the introduction of AI may fundamentally alter the socio-technical context in which the decision support is introduced with implications for the decision-making process as well as the other actors involved in the process (Herrmann and Pfeiffer, 2023). This is seen in the three THEMIS 5.0 use cases where the introduction of AI systems in medical decision making, port logistics planning, and mitigation of disinformation in online media, at its best has the potential to strengthen decision-making processes to the benefit of medical patients, goods transport, and democratic society. However, uncritical introduction of AI in the same use cases may hold negative implication in terms of – for example – the robustness and fairness in decision making along with unforeseen decision impact.

Because of its expected impact, there is general agreement that implementation of AI in decision making should comply with requirements for trustworthy AI – spanning ethical, legal, and technical aspects of such implementation. Specifically, there is a need for AI to be implemented with sufficient human agency and oversight (AI HLEG, 2019), so as to enable users of AI to have sufficient control of the capabilities and outcomes of AI implementation. However, while the concept of trustworthy AI has seen substantial development over the recent years (e.g. Kaur et al., 2022), including the development of guidelines for ethical AI (Hagendorff, 2020), there is a knowledge gap on how human experts and users should be involved in the development of AI systems for decision support so as to ensure trustworthy AI where the capabilities of the AI system empower human users. Specifically, given the complexity of requirements for trustworthy and ethical AI, there is a need for support that enables developers and users to optimize the trustworthiness of AI systems.

In the THEMIS 5.0 project, the objective is to develop a platform or ecosystem that enables human-centred evaluation and optimization of AI trustworthiness. In line with established guidelines (AI HLEG, 2023) and emerging standards for trustworthy AI (NIST, 2023; ENISA, 2023; ISO, 2023), THEMIS 5.0 aims to address AI trustworthiness through a risk management approach, where key aspects of trustworthiness are considered desirable characteristics of an AI system that need to be identified, established, and maintained. Building on existing risk management approaches (NIST, 2023; ENISA, 2023), THEMIS 5.0 will become an approach for human-centred management of AI trustworthiness. At the centre of this risk management, is the measurement and analysis of AI trustworthiness in the form of detectable trustworthiness characteristics. The human-centred approach will enable the involvement of users and experts in the development and deployment of AI models and systems, allowing for AI systems to reflect human values in the context of a specific socio-technical context.

As such, THEMIS 5.0 represents an important step towards bridging the knowledge gap on involving human experts and users in the development of AI for trustworthy decision support. By enabling trustworthy AI for this purpose, *hybrid decision support* may be enabled where human and AI-based system components contribute to efficient and effective decision making aligned with human values. Furthermore, the THEMIS 5.0 approach will facilitate the development of trustworthy AI for such decision making in line with the requirements of the AI Act (EU, 2024).

1.2. THEMIS 5.0 WP2: Trustworthiness Optimisation Methodological Framework

The aim of THEMIS 5.0 to establish a human-centred framework for trustworthiness optimization in AI for decision making is innovative and requires substantial conceptual clarification and development. The trustworthiness of the AI system and underlying AI model needs to be developed and maintained from a lifecycle perspective. The assessment and optimization of trustworthiness require system trustworthiness characteristics to be detectable for inclusion in analyses and optimization considering multiple characteristics simultaneously. Users and experts need to be involved in the assessment process so as to ensure compliance with human values. Finally, for efficient and effective AI trustworthiness optimisation in complex business contexts and socio-technical environments, analysis and trustworthiness criteria and human values need to be supported by specific AI models orchestrated for input to risk assessment.

In THEMIS WP2, we develop the conceptual basis for the innovative approach of THEMIS 5.0. Specifically, the WP will (i) analyse factors affecting trustworthiness of hybrid decision support in socio-technical systems [T2.1 and T2.2], (ii) explore

AI techniques for trustworthiness evaluation and optimisation [T2.1], (iii) summarize the ethical and legal framework for AI-driven optimization [T2.1], (iv) analyse the user requirements to produce user journeys and mock-ups for a human-centric trustworthiness optimisation environment [T2.2], and (v) establish a comprehensive conceptual model defining the architectural considerations and decision-making processes needed for supporting the THEMIS 5.0 [T2.3].

The conceptual basis comprises the theoretical groundwork and human-centric requirements, summarized in this deliverable, as well as a comprehensive conceptual model to be presented in the forthcoming deliverable D2.2. 'THEMIS 5.0 Approaches and Conceptual Modelling'. The conceptual basis is established in dialogue with use-case partners as well as with users and domain experts through co-creation sessions in WP3 and are used to guide the technical architecture and development of WP4 as well as the piloting and evaluation in WP5.

1.3. Purpose and scope

This deliverable presents the theoretical groundwork for the THEMIS 5.0 approach to trustworthiness optimization of AI systems. Furthermore, it presents the users requirements for this approach, based on input from users and experts in the three case domains of THEMIS 5.0. The groundwork will detail the concept of trustworthiness for AI systems, consider specific trustworthiness characteristics, expand on the involvement of human users and AI system adaptation to human values and business targets and discuss the use of risk management approaches for assessing AI trustworthiness. The user requirements will be identified and summarized on the basis of extensive user involvement. The requirements will further be instantiated in mock-ups illustrating user journeys to guide the technical development in the project.

1.3.1. Groundwork – Section 2

The groundwork for THEMIS 5.0 is a resulting from Task 2.1 and is presented in Section 2 below. Here we detail the concept of trustworthy AI, with reference to the relevant knowledge base, including scientific literature, technical reports, standards, policy documents, and legal material. In the groundwork, we first analyse the concept of trustworthy AI as reflecting an approach to AI development where trustworthiness may be assessed and optimized through detectable trustworthiness criteria associated with the system under test in its socio-technical environment. Here, we specifically address trustworthiness criteria reflecting technical, business, and ethical/legal perspectives on trustworthy AI.

Furthermore, we detail relevant background on human-centred AI. Here, an overview is provided of human-centred explanations which will be important in THEMIS 5.0 to support the involvement of human users and experts in optimization of AI trustworthiness.

Following from this, we discuss risk management approaches as means to address AI trustworthiness and, specifically, analyse the relevance of a specific risk management tool – Spyderisk – for application as part of an AI trustworthiness optimization approach. We then detail technologies to support the analysis of AI trustworthiness characteristics and management, specifically addressing the trustworthiness characteristics of accuracy, robustness and fairness.

Finally, we provide an overview of the ethical and legal basis for the THEMIS 5.0 approach and provide a basis for lifecycle-oriented AI trustworthiness optimization.

1.3.2. User requirements – Section 3

The user requirements, the user journeys and the mock-ups for the THEMIS 5.0 ecosystem are the main outputs from task 2.2 that are presented below in Section 3. The human-centric aspect of the conceptual work is further addressed in collaboration between WP2, and the co-creation activities performed within WP3 and includes (i) the identification of the user requirements for the THEMIS 5.0 trustworthiness optimisation ecosystem, (ii) the end-users' expectations from trustworthy AI tools, (iii) use-case specific trustworthiness considerations, (iv) end-users' trustworthiness preferences, and (v) use-case specific trustworthiness related risks, followed by the development of user journeys and mock-ups describing the different ways that users can use the functionalities of the THEMIS 5.0 ecosystem for the optimisation of AI models in the project use cases.

In Section three we first present our approach first, to collect the necessary material from the use case partners and the participants of the co-creation sessions and then, to analyse the collected material for the elicitation of the user needs, expectations and preferences. Also, the use cases, the intended user types and usage scenarios of the AI systems under test in each use case are presented offering context for the in-depth understanding of the findings of task T2.2. The system requirements, stemming from the Description of Action, are then presented followed by the user requirements for the THEMIS 5.0 trustworthiness optimisation ecosystem as well as the end-users' trustworthiness preferences and expectations from trustworthy AI tools to be integrated into their respective sectors. This work also includes requirements considering legal and ethical aspects of trustworthy AI and trustworthiness related ethical, legal and business risks. Finally, the user journeys that were developed based on the previously collected and analysed material, are presented describing the different ways that users can use the functionalities of the THEMIS 5.0 ecosystem for the optimisation of AI models in the project use cases, followed by the mock-ups and mock-up dialogues that showcased and visualise the co-created user journeys.

1.3.3. Implications and recommendations – Section 4

Based on the groundwork presented in Section 2 as well as on the human-centred requirements, user journeys and mock-ups presented in Section 3, we summarize key implications and recommendations for the following research and innovation efforts of the THEMIS 5.0 project.

The recommendations presented are a summary overview of key findings in the groundwork and user requirements work. While recommendations are also made in the specific sections reporting from these two tasks, Section 0 provides a combined summary of these recommendations, followed by an overview and discussion of implications from the findings documented in this deliverable on the upcoming work in the project's implementation. Here, relevant implications concern the full conceptual framework to be established in WP2, the further co-creation activities in WP3, and the technical development in WP4.

The groundwork done in T2.1 and T2.2. has also been preliminary to T1.4. and especially for D1.3 which has been prepared in parallel to D2.1.

1.3.4. Conclusion on the key challenges guiding the work – Section 5

In the final section of the deliverable, we conclude the presented work and reflect on how it has contributed to key project challenges.

As outlined in the THEMIS 5.0 Description of Action, the work presented in this deliverable, as well as in the overall project, is guided by the following research challenges:

a) What constitutes a trustworthy AI decision support system within the dynamic reality of the socio-technical system that is operated in?

b) What are the human characteristics that contribute to the evaluation of trustworthiness of an AI decision support system?

c) How can we evaluate levels of trustworthiness and what are the effective approaches to trustworthiness optimisation in view of human preferences and values?

d) How to understand and incorporate the socio-technical system's decision-making risks in the evaluation of trustworthiness of a hybrid decision support AI system?

e) How to intelligently engage and converse with humans in the optimisation of trustworthiness of decision supporting AI systems

In the conclusion of the deliverable, we will revisit these key challenges to summarize how these have informed the presented work.

To summarize the above, the figure below presents an overview of the sections of the deliverable and how they are related.

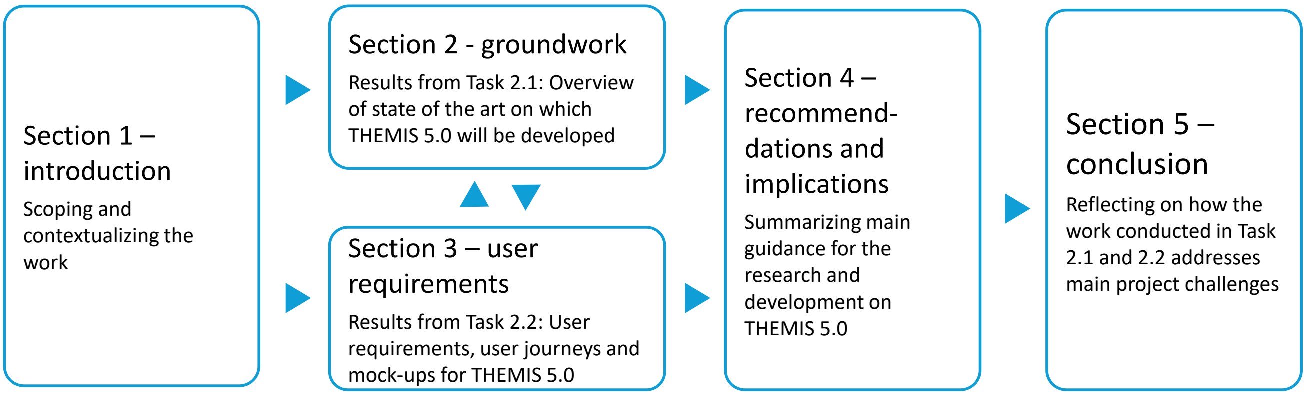

Figure 1: Overview of deliverable sections and their relation.

2. T2.1: GROUNDWORK FOR THEMIS 5.0 TRUSTWORTHINESS OPTIMISATION FRAMEWORK

To provide the needed basis for the development of the THEMIS 5.0 solutions, we have surveyed and analysed relevant literature, technologies, and approaches concerning Trustworthy AI, its assessment, and optimization. This groundwork has been conducted within the project Task 2.1. In this chapter, we present its outcomes.

2.1. Scoping Trustworthy AI for THEMIS 5.0

We start our presentation of the groundwork by providing an overview of key themes of Trustworthy AI for the purpose of THEMIS 5.0. These themes are grounded in the relevant literature. After their initial outline in this section to provide the overall scoping of Trustworthy AI for THEMIS 5.0, we present the approach for this work (Section 2.2) and detail the groundworks related to each of these themes in the subsequent sections (Sections 2.3-2.8).

2.1.1. Trustworthy AI concerns AI systems in their socio-technical environments

Trustworthy AI concerns the development and validation of AI models and systems to ensure that these provide the outcomes intended by users and stakeholders without unforeseen unwanted implications (Li et al., 2022). The current interest in Trustworthy AI is motivated by the uncertainty and risk associated with applications of AI, with potential implications at the levels of individual users, organizations, and society (Wing, 2021). Furthermore, the notion of Trustworthy AI is motivated by the acknowledgement of AI systems as embedded in socio-technical environments, accentuating the importance of considering the broader ethical and societal implications of AI systems (e.g. ENISA, 2020; 2023). Hence, while AI systems by default are developed to reliably attain defined business goals, Trustworthy AI in addition should address uncertainty and risk in a broader scope. As noted by the EC HLEG on Trustworthy AI:

> *Trustworthy AI has three components, which should be met throughout the system's entire life cycle: (1) it should be lawful, complying with all applicable laws and regulations (2) it should be ethical, ensuring adherence to ethical principles and values and (3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm* (AI HLEG, 2019)

While a colloquial understanding of trustworthiness concerns the ability of an AI system to reliably help achieve the objectives for which it has been developed, the AI HLEG definition underscores the importance of a broader concern for ethical and legal compliance, and avoidance of potential harmful implications. As noted by Zanotti et al. (2023), while Trustworthy AI needs to be reliable, mere reliability is not sufficient. The scope of Trustworthy AI also should include the vulnerability and risk associated with AI system use and the uncertainty concerning the potential ethical and societal implications of AI systems.

2.1.2. Trustworthy AI depends on verification of trustworthiness characteristics

The importance of AI trustworthiness requires a keen focus on verification (Wing, 2021). In the ISO report on Trustworthy AI, the trustworthiness of an AI system is understood as its "ability to meet stakeholders' expectations in a verifiable way" (ISO, 2020). In the NIST AI Risk Management Framework, measuring AI risk through "quantitative, qualitative, or mixed-method tools, techniques, and methodologies" is considered key to manage AI trustworthiness (NIST, 2023).

Verification of Trustworthy AI is, however, not conducted on a high-level trustworthiness construct, but rather on the different trustworthiness characteristics considered relevant for the particular AI system in question (NIST, 2023). We understand trustworthiness characteristics to mean attributes of an AI system of particular importance for assessing whether the system is technically and socially robust, and in compliance with ethical and legal requirements.

In the literature, a range of trustworthiness characteristics are listed including reliability and robustness, fairness, security, safety, privacy, transparency, and accountability (cf. AI HLEG, 2019; EC, 2023; NIST 2023); characteristics that in part concern (a) technical aspects of the AI system, (b) aspects of the socio-technical environment, and (c) broader ethical, social, and legal considerations (ENISA, 2023).

The broad range of potentially relevant trustworthiness characteristics, and the rich nuances in their implementation, makes it unfeasible for a single project to do the research needed for trustworthiness verification and optimization on all relevant characteristics. Rather, we have selected characteristics spanning the different perspectives identified by ENISA (2023).

2.1.3. *Trustworthy AI requires a human-centred approach*

Trustworthy AI is construed as embedded in socio-technical systems, entailing uncertainties and risks requiring consideration of ethical and societal implications (AI HLEG, 2019; ENISA, 2020; 2023; NIST, 2023). In this context, human judgement is needed when mapping, measuring and managing Trustworthy AI (NIST, 2023). Specifically, there is a need for involving humans knowledgeable of the socio-technical environment and relevant stakeholder expectations.

The assessment of Trustworthy AI therefore requires humans, both in systematic verification of AI systems by stakeholders and domain experts as well as in users' regular application of the same system for their work purposes. This has two important implications. First, it is important to ensure that the means of verification for relevant trustworthiness characteristics are understandable to stakeholders, domain experts, and users – potentially drawing on the field of human-centred explainable AI (Liao & Varshney, 2021). Second, to optimize AI system trustworthiness to the needs and requirements of relevant user groups, it will be beneficial to establish profiles of users of the AI system under evaluation to align with user perceptions and values. Different user groups, such as different professional groups using the same AI system, may be represented through different user profiles. The development of user profiles may benefit from the substantial literature on personas for design of interactive systems (Salminen et al., 2022).

2.1.4. *Risk management helps ensure Trustworthy AI*

Trustworthy AI is a response to the uncertainty and risk implied in applying AI systems. As noted by the EC HLEG, Trustworthy AI entails "seeking to maximise the benefits of AI systems while at the same time preventing and minimising their risks" (AI HLEG, 2019), and risk management is widely acknowledged as a process to help ensure Trustworthy AI (ISO, 2020; NIST, 2023).

Risk management of Trustworthy AI entails mapping, monitoring, and managing vulnerabilities, risks, and mitigations pertaining to the relevant trustworthiness characteristics (NIST, 2023). Risk management as a means to strengthen AI trustworthiness is also reflected in the AI Act where risks associated with high-risk AI systems should be addressed through suitable risk management measures (Panigutti et al., 2023b).

High-level frameworks to support AI risk management have been proposed by ENISA (2023) and NIST (2023). Furthermore, tools for risk management such as Spyderisk (2024), developed by the University of Southampton, are highly relevant to support such frameworks.

2.1.5. *Technology development is needed to support Trustworthy AI*

The importance of Trustworthy AI has motivated substantial development and application of technologies to support Trustworthy AI development and verification (Kaur et al., 2022; Wing et al., 2022). In addition to technology for risk management, technology support is available for (a) assessment and verification of individual trustworthiness characteristics as well as (b) development and deployment of AI systems.

Technology support for assessment and verification of individual trustworthiness characteristics may be based on well-established approaches – such as the use of confusion matrices or measures of precision and recall for assessment of AI system accuracy – but may also benefit from emerging approaches, such as the application of machine-learning approaches to assess the technical performance of an AI system. Such emerging approaches may be important to assess AI systems deployed in real world contexts, as it is known that the robustness of AI systems may be vulnerable to changes in their context of use and that advanced approaches may be needed to capture variation in technical performance.

A range of technologies and tools are also available for the development and deployment of Trustworthy AI, for example by the large technology companies such as IBM, Google, and Microsoft. A survey of such tools is needed to provide an overview of how these can be optimally deployed or combined in AI development processes.

2.1.6. *Trustworthy AI should comply with ethical and legal requirements*

The advances of AI, and the potential societal implications, has motivated substantial industry, government, and research efforts into the ethical requirements of AI. A range of partly overlapping ethical frameworks have been developed (Hagendorff, 2020) and taken up in research and development, and even as basis for policymaking and legislation (e.g. AI HLEG, 2019). Currently, there is substantial efforts in AI legislation, not only in Europe but throughout the world. Relevant policies and legislative initiatives are found, among others, in EU, OECD, the UK, and the US. The ethical frameworks and the emerging legislation form important requirements to AI systems in general and Trustworthy AI in particular. It is critical for research and development in Trustworthy AI to overview and engage with such frameworks and legislation.

2.1.7. *Trustworthy AI should be optimized throughout the AI lifecycle.*

Trustworthy AI is highly complex, due to the broad variability in AI systems and social-technical environments in which they can be deployed (Wing et al., 2022), the multitude of relevant trustworthiness characteristics (AI HLEG, 2019; ENISA, 2023; NIST, 2023), and the range of disciplines in which Trustworthy AI has been addressed (Kaur et al., 2022). As noted by Li et al. (2022), AI trustworthiness is the result of the combined effect of relevant trustworthiness characteristics. Hence, Trustworthy AI may not be achieved by considering single trustworthiness characteristics in isolation (NIST, 2023). Satisfying different trustworthiness characteristics may entail opposing requirements, and mitigation of one characteristic may have adverse effects on others (ISO, 2020; NIST, 2023). In consequence, achieving Trustworthy AI entails a continuous optimizing process throughout the AI lifecycle. For this to be achieved, there is a need for a Trustworthiness Optimization Process, along with new knowledge on AI trustworthiness optimization.

2.1.8. *Summarizing the scoping of Trustworthy AI for THEMIS 5.0*

Above, we have detailed key themes of Trustworthy AI. These themes serve to scope the development of human-centred trustworthiness optimization of AI in THEMIS 5.0. Specifically, we have noted the need to consider AI systems in the socio-technical environment, verification through detectable trustworthiness characteristics, the human-centred aspect of Trustworthy AI, and the need for trustworthiness optimization throughout the AI lifecycle based on risk management. We have also accentuated the need to technology support and alignment with ethical and legal requirements.

Drawing on this, we propose the following vision of Trustworthy AI for the purpose of THEMIS 5.0:

Trustworthy AI concerns the verification of relevant trustworthiness characteristics for a given AI system embedded in its socio-technical environment and over its life cycle. Relevant characteristics span the technical attributes of the AI system, its impact from a business perspective, and its compliance with ethical and legal requirements. Verification may be conducted through a human-centred risk management approach involving stakeholders and users, aimed at trustworthiness optimisation in compliance with ethical and legal requirements.

The vision summarizes the scope of Trustworthy AI for THEMIS 5.0. Furthermore, it contains the elements detailed in the groundwork. The details of the groundwork will be presented immediately following the presentation of the applied methodological approach.

2.2. Methodological approach to the groundwork

The groundwork has involved all THEMIS 5.0 project partners in a comprehensive, collaborative approach, to draw on the partners existing knowledge and resources, as well as their capacity for reviewing and summarizing the state of the art.

In this section we provide an overview of (a) the collaborative approach during the groundwork, (b) the review of the relevant state of the art, and (c) an overview of the main elements of the groundwork.

2.2.1. *Collaborative approach*

The groundwork was established collaboratively, involving all project partners and spanning the first project year. The collaboration was initiated at project kick-off, through an initial detailing of key themes of Trustworthy AI and assignment of leadership for the different themes. Over the first project months, these initial themes evolved into the themes presented in Figure 2.

Collaboration and coordination were conducted through bi-weekly online status meetings, in addition to ad-hoc meetings between partners. To structure the collaboration process, the groundwork was divided into two main iterations. At the end of each iteration, the partners leading the work on each theme presented their work so far at online workshops involving the entire project.

Furthermore, the writing process was conducted in a collaborative document. Key concepts and terminology were listed in a common project glossary.

An overview of the collaborative approach is provided in Figure 2 below.

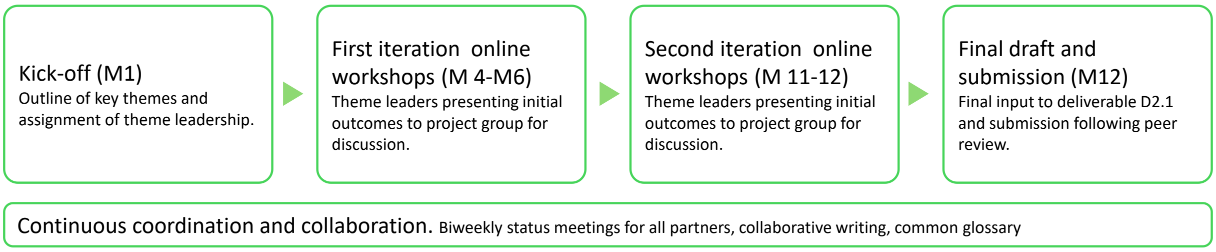

Figure 2: Overview of collaborative approach to groundwork

2.2.2. Reviewed state of the art

A large body of literature on Trustworthy AI has evolved in response to the challenges of Trustworthy AI. This literature spans different disciplines and application domains, reflecting the cross-disciplinary effort needed to understand what Trustworthy AI is and to achieve trustworthiness in AI systems. Furthermore, the general interest in Trustworthy AI has led research, industry, and government actors to develop a range of guidelines and frameworks for Trustworthy AI.

To scope and analyse Trustworthy AI as a groundwork for THEMIS 5.0, we took a starting point in current summaries of the scientific literature as well as the work reflected in white papers and standards.

Of these literature summaries, we particularly note those of Kaur et al. (2022), Wing (2021), and Li et al. (2023). The review by Kaur et al. (2022) is particularly valuable for discussing the literature from the perspective of the trustworthiness characteristics listed by the AI HLEG (2019). The review by Wing (2021) is helpful as it takes the perspective of achieving Trustworthy AI by way of formal methods for the verification of trustworthiness characteristics ('properties') of AI systems – in line with the ISO report on Trustworthy AI where trustworthiness is defined as concerning an AI system's "ability to meet stakeholders' expectations in a verifiable way" (ISO, 2020). Finally, the review by Li et al. (2023) provides an overview of literature of relevance for a lifecycle perspective on Trustworthy AI, in line with the lifecycle perspective detailed by ENISA (2020; 2023)

The work in THEMIS 5.0 is also guided by existing technical reports, white papers, and standards on Trustworthy AI. Of particular note, are the works provided by EC HLEG, NIST, ENISA, and ISO. The EC HLEG works on ethics guidelines for trustworthy AI (AI HLEG, 2019) and assessment of trustworthy AI (AI HLEG, 2020) have been highly influential on European AI research and development as well as legislation. The NIST (2023) AI Risk Management Framework provides a high-level process for managing Trustworthy AI throughout its lifecycle. The ENISA (2023) framework addresses Trustworthy AI from a cybersecurity perspective. And the ISO (2020) technical report on Trustworthy AI concerns factors that may impact the trustworthiness of AI systems.

Furthermore, the broader literature and relevant technologies were reviewed specifically for the identified key themes of this groundwork. This broader knowledge and technology base is cited inline in the text and may be found in the reference list towards the end of this report.

Finally, given the importance of regulating AI system use to prevent unwanted implications for individuals, organizations, and society, a body of ethical frameworks and regulation for AI systems development and use is emerging. Of particular interest for THEMIS 5.0 is the evolving ethical and regulatory frameworks within Europe, but we have also reviewed relevant frameworks also for relevant regions or constellations of relevance for European collaboration, such as that within OECD, the UK, and the US.

2.2.3. Main elements of the groundwork

From our collaborative groundwork process, and our review of the state of the art, we provide the needed groundwork for THEMIS 5.0. The groundwork is structured according to main themes outlined in the collaborative process. The themes correspond to the scoping of Trustworthy AI in Section 2.1.

- **The AI system in the socio-technical environment (Section 2.3).** The literature is in substantial agreement that the scope of Trustworthy AI concerns AI systems within specific technical and organizational contexts. That is, the trustworthiness of an AI system depends on, and is to be optimized with respect to a given socio-technical environment.
- **Verification through trustworthiness characteristics (Section 2.3).** There is agreement in the literature that Trustworthy AI concerns specific characteristics of an AI system (within its socio-technical environment). These trustworthiness characteristics needs to be detectable to support trustworthiness verification and

optimization. Trustworthiness characteristics representative of technical, ethical/legal, and business perspectives are detailed.

- **The human-centred character of Trustworthy AI (Section 2.4).** The ethical and socio-technical aspects of AI require Trustworthy AI to be considered as a human-centred concept. We detail the need for AI trustworthiness characteristics to be understandable to users and stakeholders, and how human-centred AI may contribute to this. Furthermore, we detail how Trustworthy AI needs to be adapted to the requirements and needs of specific users and stakeholders.
- **Risk management for Trustworthy AI (Section 2.5).** The uncertainties associated with AI systems, as well as increasing impact and implications for organizations and society, make risk management a promising approach to Trustworthy AI. We discuss relevant risk management frameworks and their feasibility for risk assessment technology support.
- **Technologies for Trustworthy AI (Section 2.6).** A range of technologies and approaches may be applicable for assessing and managing Trustworthy AI. We detail technologies that may be suitable for assessing the trustworthiness characteristics of particular interest in THEMIS 5.0. In addition, we review relevant general frameworks to support AI trustworthiness.
- **Ethical and Legal Requirements to Trustworthy AI (Section 2.7).** The ethical and regulatory basis for Trustworthy AI is evolving rapidly, both within and outside Europe. We summarize the relevant basis and provide ethical and legal requirements.
- **Trustworthiness Optimization Process (Section 2.8).** Achieving trustworthy AI requires a lifecycle perspective. We summarize the relevant basis for a Trustworthiness Optimization Process, and also move provide relevant background for detailing the notion of trustworthiness optimization.

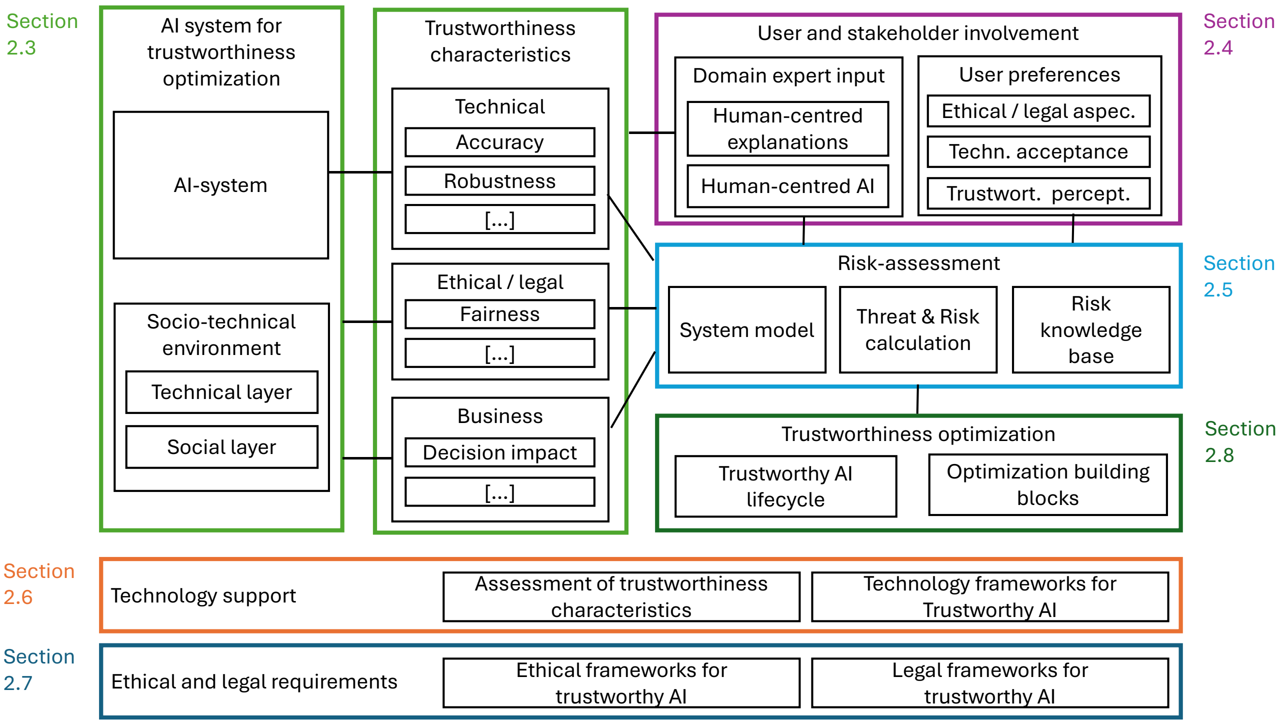

Figure 3: Overview of main themes in groundwork for Trustworthy AI in THEMIS 5.0

In addition to the detailing of key themes for Trustworthy AI, the groundwork has established a THEMIS 5.0 glossary as a living document to facilitate collaboration and development in the project (THEMIS_GLOSSARY.xlsx).

The concepts of the glossary, and their operational definitions for THEMIS 5.0 will be made subject to project internal reviews for ensure consistency and relevance for the project objectives.

2.3. The characteristics of Trustworthy AI

2.3.1. The AI system

AI systems refer to a broad range of computerized systems based on a variety of technologies. To scope the THEMIS 5.0 approach, it is important to identify what we mean by an AI system and which types of AI systems that we are particularly interested in.

Historically, AI has been used both in reference to symbolic systems based on logic and rules – such as expert systems and knowledge-based systems – and connectionist systems based on learning from datasets – such as neural networks and deep learning. Within connectionist approaches to AI, there is further broad distinctions to be made between AI for specific purposes on the one hand – such as systems for prediction or classification – and general-purpose AI on the other – such as large language models.

Current definitions of AI systems typically accentuate their ability to infer and learn, and their potential autonomy. This is reflected well in the OECD (2023) definition, which is also applied in the AI Act (EU, 2024): *"An AI system is a machine-based system 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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."*

While a number of technologies, such as different approaches to machine learning, are typically listed as AI systems, the above definition of AI is technology neutral. That is, while the definition accentuates AI systems capabilities for autonomy, adaptiveness, and inference which may impact their environment, the definition is not limited to a pre-defined list of technologies such as machine learning. This enables our understanding of AI systems to cover a variety of existing and future technologies, so as to stay relevant while the field evolves.

In THEMIS 5.0, we adhere to the OECD definition of an AI system. However, THEMIS 5.0 is particularly focused on AI systems applied for decision-making purposes, which to some extent serves to scope the range of potentially relevant systems. Furthermore, we are focused on AI systems that benefit from recent technological advancements in machine learning and deep learning, as these are seen as representing substantial societal implications in the foreseeable future and best reflecting current definitions of AI systems.

We refer to AI systems that apply the THEMIS 5.0 approach to trustworthiness optimization as the *system under test*. In doing so, we distinguish between the AI system to be optimized (the system under test) and the THEMIS 5.0 approach where trustworthiness optimization is enabled by way of AI support. When referring to the system under test we in this deliverable use the terms *AI system* and *system under test* interchangeably.

2.3.2. The socio-technical environment of Trustworthy AI

Trustworthy AI concerns AI systems embedded in a socio-technical environment, including (a) a technical layer of related technical systems and infrastructure (including computers, networks, sensors and actuators) and (b) a social layer including users and stakeholders, organizational aspects, and societal aspects. Trust with regards AI can be assessed by/from the "social" layer, through and considering the "technical" layer.

The socio-technical character of Trustworthy AI implies that (1) the AI system is developed to address specific **business goals** and assessed with regards to its **business impact** (ENISA 2020; 2023), (2) stakeholders or domain experts needs are involved in assessments of trustworthiness. Stakeholders do not pursue business goals only, but also social and environmental goals, including commons (see UN SDG). They use (legal or ethical) values and virtues to **evaluate** AI impact on the socio-technical environment.

The socio-technical environment can be **represented** as a system of **systems**, encompassing different computing layers to support AI, up to social groups such as AI users and administrators at the front line. **Models** can be used to describe the socio-technical environment, but since reality is not fixed, models will need to be constantly updated. Furthermore, some models are not fully computable, or more precisely, models are necessarily incomplete. (Korzybski, 1933)

Despite the ability to represent conceptual systems by way of graphs (see OWL, RDF), it is dubious that the frontiers of AI systems are stable and understandable as the AI act suggests (see def. Art.3) for two main reasons. (1) because an IT system represented at a certain layer of the OSI model, does not have a 1:1 counterpart at another layer. (2) because at runtime, the systems which had been recognized while at rest, may no longer be present. Therefore, it is necessary to bring some amount of agentic reflexiveness, namely a feature at the User Interface level, that enables the user to get a sense of how the computation happens, what it aims to do, and a feeling of possible misinterpretation of e.g., AI recommendations.

Bringing Trustworthiness considerations together with business considerations in the making of the AI/human decision, is a way to mitigate those inconveniences, since human evaluation from the Social Layer plays an important role to complete and enlighten blind spots of calculation made inside the Technical Layer, in order to integrate more effectively past or foreseeable Impact and make sure Business Decisions will be acceptable from a broader point of view.

2.3.3. Detectable trustworthiness characteristics for verification of Trustworthy AI

Verification is key to Trustworthy AI (ISO, 2020; NIST, 2023). Such verification is, however, not conducted on a high-level trustworthiness construct, but rather on the different trustworthiness characteristics considered relevant for the particular AI system in question.

In line with the NIST (2023) AI Risk Management Framework, we understand *trustworthiness characteristics* to mean attributes of an AI system of particular importance for assessing whether the system is technically and socially robust, and in compliance with ethical and legal requirements. Such trustworthiness characteristics can be related to the AI system, its basis in data, models, and algorithms, as well as its interactions with its users and broader socio-technical environment.

In much of the literature on Trustworthy AI (e.g. AI HLEG, 2019; Kaur et al., 2022), the focus is on Trustworthy AI requirements, that is, the essential aspect of an AI system for it to be considered trustworthy. In THEMIS 5.0 we apply the term trustworthiness characteristics in accordance with NIST (2023), as we find this more in line with the project's overall objective of balancing different trustworthiness characteristics for optimization purposes.

A range of trustworthiness characteristics and corresponding requirements are suggested in the literature, along with different classifications. Kaur et al. (2022) notes that existing guidelines and frameworks of Trustworthy AI by and large have converged on five main principles of Trustworthy AI, including "transparency/explainability, justice and fairness, non-maleficence/societal and environmental well-being, responsibility and accountability, and privacy". In addition, different guidelines and frameworks have identified additional characteristics. For example, the EC HLEG (AI HLEG, 2019) also accentuates 'human agency and oversight' as well as 'robustness and safety'. NIST (2023) also include 'validity and reliability', 'safety', and 'security'.

ENISA (2020; 2023), drawing on the draft NIST (2022) AI risk management framework, notes that trustworthiness characteristics may be classified in three broad groups: (a) technical characteristics, that is, characteristics in direct control of AI system designers and developers such as accuracy, robustness, and reliability, (b) socio-technical characteristics, that is, characteristics dependent on the socio-technical environment in which the system is embedded such as explainability and safety, and (c) guiding principles, that is, characteristics concerning higher level ethical norms such as fairness and accountability. The classification of trustworthiness characteristics suggested in ENISA has not been upheld in the final version of the NIST AI risk management framework (NIST, 2023), possibly due to the difficulty of consistently classifying the trustworthiness characteristics in meaningful higher-level clusters. The classification may nevertheless hold substantial value to understand different relevant perspectives on trustworthiness characteristics.

In THEMIS 5.0, we build on the classification suggested by ENISA (2020:2023), by specifically considering trustworthiness characteristics from three different perspectives:

1. **The technical characteristics of the AI system**. Here, we in THEMIS 5.0 specifically consider AI system accuracy and robustness.
2. **The impact of AI system output from the business perspective**. Here we in THEMIS 5.0 will consider trustworthiness characteristics of specific perceived relevance to stakeholders.
3. **The compliance of the AI system with ethical and legal requirements**. Here we in THEMIS 5.0 will specifically consider AI system fairness.

In the following subsections (2.3.4-2.3.6, as well as 2.6.1-2.6.3), we detail AI system trustworthiness characteristics representative of these three perspectives. An initial overview of the trustworthiness characteristics and how they are addressed in this groundwork and in the THEMIS 5.0 approach, is provided in

Table 1.

Table 1. Overview of trustworthiness characteristics and how they are addressed in this groundwork and THEMIS 5.0

Perspective	Trustworthiness characteristic	Established metrics and methods	THEMIS 5.0 method advances
Technical	Accuracy	• Correct predictions / Total predictions • Confusion matrix • ROC-curve • Precision and recall, F1 score • Loss metrics • Matthews correlation coefficient • Decomposition of accuracy in bias and variance	Method application and development for accuracy predictions in particular predictions from AI models, quantifying uncertainty • Misclassification detector • ML-based model comparisons (see Section 2.6.1)
Technical	Robustness	• Change in accuracy metric for a given perturbation	Method application and development to assess and evaluate robustness in AI models: • Adversarial examples and training • Out-of-distribution detection • Counterfactual explanations (see Section 2.6.2)
Ethical / legal	Fairness	• Group Fairness (e.g. Demographic/Statistical parity, Disparate impact, equalized odds) • Individual Fairness (e.g. Fairness through unawareness) • Causal-based metrics (e.g. counterfactual fairness)	Combine widely adopted fairness metrics and case-specific applicable metrics to support the AI trustworthiness assessment and optimization.
Business	Decision Impact on the socio-technical environment	Human-centred assessment	Human centred assessment through a risk and context management approach with tool support

2.3.4. Technical characteristics – exemplified by accuracy and robustness

The trustworthiness of AI systems depends in part on their technical characteristics. In the NIST (2023) AI risk management framework, technical characteristics – subsumed under the heading “valid & reliable” are noted as "a necessary condition of trustworthiness and a base for other trustworthiness characteristics". ENISA (2023) also notes the key importance of technical characteristics for Trustworthy AI and specifically mentions accuracy and robustness in addition to resilience and reliability.

Accuracy and robustness are given particular attention in THEMIS 5.0, due to their criticality as technical trustworthiness characteristics – as noted by ENISA (2023) and NIST (2023).

2.3.4.1 Accuracy

Accuracy may be understood as the "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true" (NIST, 2023). For AI-systems it may be important to note that this particularly concerns how well the AI system "do on new (unseen) data compared to data on which it was trained and tested" (Wing, 2021).

For balanced classification problems, accuracy may be operationalized as how close a given set of measurements are to their true value. For imbalanced classification problems, precision, recall, and F-scores are more appropriate. For regression problems, accuracy is reported through loss metrics such as mean squared error.

Wing (2021) notes that while accuracy may be considered a gold standard, it may be important to consider trade-offs between accuracy and other trustworthiness criteria, such as explainability.

The following concepts are particularly prominent for assessment and measurement of accuracy of an AI model for **a particular data set**.

2.3.4.1.1 The accuracy of classification models

Measures for evaluating classification models include a range of established metrics, including classification accuracy, the confusion matrix,

Classification accuracy: The classification accuracy is the most intuitive evaluation metric and is defined by:

$$Classification\ accuracy = \frac{number\ of\ correct\ classifications}{total\ number\ of\ classification}.$$

Confusion matrix: Within binary classification, the presentation of a confusion matrix is an easy way to understand the accuracy of a model with respect to the metrics *precision, recall, specificity and negative predictive value.* Below is an example of a confusion matrix (blue background) from a binary classification model, together with the informed metrics created from its rows and columns (green background). For classifiers that output class probabilities, the confusion matrix depends on the decision threshold. Changing the decision threshold allows trading off false and true positives. It can also help in dealing with class distributions that are weakly imbalanced.

	Predicted Positive (PP)	Predicted Negative (PN)	Informed metrics along row
Real Positive	True Positives (TP)	False Negatives (FN)	**Recall = TP/(TP+FN)**
Real Negative	False Positive (FP)	True Negatives (TN)	**Specificity = TN/(FP+TN)**
Informed metrics along column	**Precision = TP/PP = TP/(TP+FP)**	**Negative predicted value = TN/PN = TN/(FN+TN)**	

In this way, different aspects of the model accuracy can be shown in an easy way. Models can for instance be compared to each other by comparing their respective confusion matrices.

The F1-score is a function of the precision and recall given by

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} = 2 * \frac{precision * recall}{precision + recall}.$$

The F1-score ranges in [0,1] where an F1-score of 1 describes a perfect classifier.

Matthews correlation coefficient (for binary classification): Closely connected to the confusion matrix, and how to summarize the accuracy of a model based on the information given from a confusion matrix, is the Matthews correlation coefficient (MCC). The MCC is a summary quantity of the metrics of precision, recall, specificity, and negative predicted value. It is defined by:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

The MCC ranges in the interval [-1,1] and serves several nice interpretations: The value –1 means perfect misclassification, while 1 means perfect classification (perfect model). In addition, MCC = 0 is the expected value of the coin-tossing (random) classifier. According to Chicco and Jurman (2020), the MCC is more informative than the F1 score and the accuracy score for binary classification tasks.

ROC-curve: The performance of a classifier across decision thresholds can be evaluated using curves of *receiver operating characteristic* (ROC) [3]. These curves trace out true positive rate (recall) and false positive rate as the decision threshold is adjusted. If there exist performance requirements for, for example, acceptable levels of false positives, the ROC will easily detect the corresponding decision threshold as well as true positive rate. The rest of the confusion matrix can then be computed. The performance of classifiers across all decision thresholds can be summarized by the area under the ROC curve (*AUC*).

Precision and recall: For heavily imbalanced classification problems, precision and recall are more appropriate metrics because they screen out the dependence on true negatives (which drives artificially inflated accuracy). *Precision* describes how likely a positive prediction is to be correct (Saito & Rehmsmeier, 2015). *Recall* tells us how likely any given positive label is to be predicted as such. The harmonic mean of precision and recall is the F1-score. They, too, depend on the decision threshold and their trade-off can be explored using the precision-recall curve. The predictive performance across all thresholds can be summarized by the area under the precision-recall curve (*AUCPR*) or the (weighted) average precision across all thresholds (*AP*).

2.3.4.1.2 The accuracy of regression with real-valued output

There are also several established evaluation metrics of relevance to the accuracy of regression models where the outcome of the machine learning model can take any continuous value.

Mean Absolute Error (MAE): Given the model's prediction $\hat{y_i}$ and the corresponding correct value y_i for the given input x_i, the MAE is given by

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

Mean Squared Error (MSE): To prune outliers to a greater extent, we use the MSE:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Root Mean Squared Error (RMSE): The RMSE is given by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

This is by far the most used metric for regression problems with a real-valued target variable.

2.3.4.1.3 Decomposition of accuracy – Bias and variance

The accuracy of a model can in general be decomposed in the terms bias and variance. The bias reflects the expected difference between a prediction and the true value. Even though the absolute difference may have a large spread or variance, the expected difference may be close to zero, hence having a small bias. The variance of the model is basically how much the predictions spread across all the samples. This is closely connected to whether the model is underfitting (small variance) or overfitting (large variance), that is, to what degree the model has been misled to try to learn something that is just noise in the data. In relationship to this, one often discusses the bias-variance trade-off. This means that increasing the performance on one of them may go at the expense of the other.

2.3.4.2 *Robustness*

Robustness concerns the AI-systems' ability to perform as expected under varying conditions. This makes the construct of robustness broader than the construct of accuracy, and definitions vary more. The term robustness is quite recent within ML, and at the time of writing, there is no clear consensus on how to define robustness. For example, Wing (2021) notes that robustness concerns the sensitivity of a system's outcome to change in input. NIST (2023) accentuate that robustness "is a goal for appropriate system functionality in a broad set of conditions and circumstances, including uses of AI systems not initially anticipated". Kaur et al. (2022) also specifically note that robustness concerns the system's ability to deal with error at any point in the lifecycle and that the system is resilient to attacks.

In Freiesleben and Grote (2023), the authors give a thorough discussion on the nature of robustness with respect to learning systems. First, they highlight the fact that robustness is comprised of two components. The entity which is robust, called the *Robust target*, and the entity against which the target remains stable, called the *Robust modifier*. The authors explain that robustness is a causal concept meaning that the target remains stable under changes in the modifier (referred to as the domain). The more significant the changes in the domain, the more difficult it is to satisfy robustness. In a practical context it is often sufficient for the target to remain stable to a certain degree against changes within the modifier domain. Higher target tolerance implies lower demands on target robustness, and vice versa. And of course, the degree of tolerance depends on the application. For example, in a clinical setting we need high tolerance, due to the potential risk of false predictions in the model. For Email spam detection, the tolerance is lower because the impact is not as direct. With the necessary vocabulary, the formal definition of robustness as is presented in Freiesleben and Grote (2023) is what we consider as being most general and covering most cases:

2.3.4.2.1 Defining robustness

The robustness target is said to be robust to the robustness modifier if relevant interventions in the modifier, as specified by the robustness domain, do not lead to greater changes in the target than specified by the target tolerance.

All steps in a ML/AI project pipeline can be considered as robustness modifiers, such as task conceptualization, data collection and preparation, model training and evaluation, and explanation. But since the purpose of a model is prediction, deployment performance is the main modifier. According to the authors, the most established modifier is the deployment distribution. If the model is sensitive to changes in the deployment distribution – it is considered a red flag for usage. Other modifiers are often studied with respect to the deployment distribution, for example what is the impact of data augmentation on deployment performance?

2.3.4.2.2 Assessing and measuring robustness

The following approaches are relevant for assessment and measurement of robustness for a particular data set.

Robustness is, in most cases, measured based on the change in a particular accuracy metric when a particular perturbation ϕ is applied to the data set compared to the original accuracy without the perturbations. We denote A_{clean} the accuracy of the model on the clean data set, while A_{ϕ} denotes the accuracy on the perturbed data set. In Laugros et al. (2019) as well as in Braiek and Foutse (2024) the *robustness score* R_f^{ϕ} with respect to the model f and particular perturbation ϕ is given by:

$$R_f^{\phi} = \frac{A_{\phi}}{A_{clean}}$$

First, consider accuracy metrics where the accuracy is better for larger values (such as for the classification accuracy, F1-score and MCC). Then the larger the robustness score, the more robust the model is to the perturbation ϕ**.** Conversely, for accuracy metrics where a smaller value indicates better accuracy (such as MAE, MSE or RMSE), then the smaller the robustness score, the more robust the model is to the perturbation. Notice that in both cases, $R_f^{\phi} = 1$ means that the model is not affected at all by the perturbation ϕ which is the ideal case. In other words, in both cases the degree of robustness can be measured by computing the absolute difference $|1 - R_f^{\phi}|$.

Following the definition of robustness in Freiesleben and Grote (2023), the target is in this case the model accuracy, and the model is robust to the particular perturbation, created by some modifier, if $|1 - R_f^{\phi}|$ is smaller than some threshold τ, which must be specified by the user.

If one wants to evaluate the model f based on several modifiers (giving rise to different perturbations) denoted in the set S, Braiek and Foutse (2024) further defines

$$R_f^{S} = \sum_{\phi \in S} R_f^{\phi}$$

In summary, accuracy and robustness are highly connected. Here, we have talked about how a model is robust with respect to the accuracy under different modifiers giving rise to perturbations in the data. Hence, we cannot measure robustness if we have not yet defined what we mean by accuracy. However, robustness can also be with respect to other concepts in ML, for instance the robustness to explanations of a model (XAI) under different perturbations.

2.3.4.3 The challenges of the accuracy and robustness definitions for trustworthy deployment of an AI model

Note that all the definitions described above depend on historical predictions as well as how close they were to the true value. This is of course very relevant for evaluating a machine learning model in hindsight, and this can help end-users to understand to which degree one can trust the predictions by the model. The main challenge is however at deployment phase where the end-user is presented with *one single prediction* that may be used further for decision-making. To what degree can we trust this single prediction? The end-user can look at historic performances of the AI model and assume that the AI model will perform in a similar fashion. However, as a matter of fact, the performance of the AI model will vary for different input samples; some are easy to predict, and some are not. This is true no matter the performance of the AI model on a particular test data. Hence, we will need methods to quantify accuracy and robustness for one single prediction to ensure trustworthy AI. We will cover AI methodologies for this purpose in Section 2.6.

A further drawback of the robustness definition is that even though we can separate the robustness for different data characteristics (such as specific perturbations), there is not a straightforward path as to how the AI model can be trained to be robust to each type of data characteristics. We will also cover AI methodologies for this purpose in Section 2.6.

2.3.5. *Trustworthiness characteristics: business perspective – exemplified by decision impact*

The notion of Trustworthy AI presupposes a business context within a socio-technical environment where decision-making is conducted to achieve stakeholders' goals, while mitigating negative implications. When considering Trustworthy AI, it is key to assess not only the quality of the decision in itself, but also its impact from a business perspective (Business Impact) – as well as mitigation of negative implications of decisions beyond mere business considerations (Ethical Impact). Decision Intelligence recognizes that many aspects of decision-making are based on intangible elements in a complex world, elements that are impossible to capture in traditional quantitative or financial models (Pratt, 2019).

We will see to what extent impact anticipation is possible. This partly depends on the proximity of impact towards the Technical System: impacts within the Social Layer are not computable (cannot be weighed as factors in the amelioration of AI) unless they are integrated as computable information within the Technical layer.

Generally speaking, the anticipation of the Decision's possible Impact from a business perspective is taken into account in the making of the Decision. In cases where impact can be anticipated by the maker of the decision, this knowledge can be expressed in business and non-business (esp. ethical) terms. But Simon (1955) grounds the Decision Theory on the postulate of limited/bounded rationality, under which the Impact of Decision can sometimes not be known. For Von Neumann and Morgenstern (1953) "expected utility models" can reflect the rationality of the Decision-making, subject to some axiomatic conditions that are difficult to satisfy in real-world Decision systems. This approach has been further elaborated by Kahneman and Tversky, (1979) through the Prospect theory, but critics insist on the fact that it is still framed on the individualist perspective (point of reference), and subject to context blindness. Kőszegi and Rabin (2007). This means the impact of decisions from a business perspective cannot be fully anticipated.

We can then focus, considering this point of reference / contextual aspect issue, on where Impacts intervene. That can be inside (namely, in the Technical Layer), or outside the Technical System (in the Social Layer.)

- In the case where the Impact happens inside the Technical System, it may consist of any data transformation (a mere change of value, the drift of a performance indicator, the modification of a ruleset, the reconfiguration of a risk model etc.) Some software products, such as IBM BAI[1] are dedicated to the taking in account of such data changes, either in real time or in an asynchronous mode. An IT system like Spyderisk[2] permits to model impact and risk propagations within the system.
- In the case where the Impact happens in close relation to the Technical System, it can consist of the change of behaviour of a human being (e.g. the AI user). Here the human can be subject to automation bias which may then need to be addressed. (Baudel et al., 2021)
- In the case where the Impact stands outside the Technical System, it is dubious that it can be sensed, measured and directly integrated as a factor to be weighed within the automated decision process. Lemaire (1999) proposes a cognitive model of decision making in such circumstances, but the model implies feelings and emotion, which are not computable. Rather, the Impact intervening in the Social Layer has a Trustworthiness aspect that we may want to deal with. It is then necessary to bring a Human-in-the-loop to interpret the Impact and tune the Decision accordingly.

To sum up, Decision Impact can either be measured or objectivized (Baudel et al., 2023) by a machine, or at least evaluated by an individual (with or without the help of AI) through feedback loops. The taking into account of probable Impact weights in the Decision making, without certainty on how to model it may be challenging. Human evaluation happens when computation of the relevance of the decision is not possible or sufficient (because of lack or no access to data, blind side effects, system failure, suspicion of automation bias etc.). Human Evaluation fosters an Ethical evaluation and therefore can favour Trustworthiness, with the hope of a ROE (Return on Ethics) alongside the ROI in a wide sense. (Bevilacqua and al, 2023)

Based on that groundwork, in the field of Trustworthy AI, THEMIS 5.0 decides to use a specific decision protocol (called "Super Decision Engine"), a human-machine dialog in particular leveraging LLMs, to serve and ameliorate the taking into account of Impact of AI decisions or recommendations from a business perspective. More widely, IBM will explore different usages of the LLM technology within the software pipeline (chatbot, semantic enrichment, action and description of low-level code) in order to bring through the "Super Decision Engine", features useful for Trustworthiness evaluation and optimization. Several IBM assets will be leveraged and combined to explore the feasibility of bringing

[1] Business Insights - IBM Documentation
[2] https://www.it-innovation.soton.ac.uk/projects/spyderisk

some AI reflexiveness during the User Experience: ADS (Automated Decision Systems, that operate business logic and business process modelling and notation). CDM (Context Deliberation Matrix, a protocol developed within the IBM Academy of Technology for ensuring that human evaluation of the context of decision remains possible in automated decision systems); some IBM WatsonX APIs (including Langchain technologies, to ensure human-machine interface by dialog and/or graphical visualisation).

2.3.6. Trustworthiness characteristics: compliance with ethical and legal requirements – exemplified by fairness

2.3.6.1 Fairness as example of ethical and legal requirements to Trustworthy AI

The societal and ethical uncertainty and risk entailed in AI systems imply the need to assess the compliance of Trustworthy AI with ethical and legal requirements (AI HLEG, 2019; EC, 2023). Given the rapid development of legal regulation of AI, with ethical requirements serving as the basis for forthcoming law, it is necessary to consider both ethical and legal compliance.

A broad range of ethical requirements for Trustworthy AI have been proposed (Hagendorff, 2020). Furthermore, the obligation of AI service providers and users to consider such requirements is making its way into law. This is exemplified by the AI Act (EU, 2024) explicitly recalling the ethical requirements of Trustworthy AI proposed by the EC HLEG (AI HLEG, 2019):

> *[…] it is important to recall the 2019 Ethics guidelines for trustworthy AI developed by the independent AI HLEG appointed by the Commission. In those guidelines, the AI HLEG developed seven non-binding ethical principles for AI which are intended to help ensure that AI is trustworthy and ethically sound. The seven principles include human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being and accountability. Without prejudice to the legally binding requirements of this Regulation and any other applicable Union law, those guidelines contribute to the design of coherent, trustworthy and human-centric AI, in line with the Charter and with the values on which the Union is founded. […] (EU, 2024, p. 8)*

Considering compliance with ethical and legal requirements for Trustworthy AI is challenging as key concepts, such as fairness, are inherently complex and multilayered. To conduct a thorough treatment of trustworthiness criteria from an ethical and legal perspective, we in THEMIS focus on one particular ethical requirement. Fairness will therefore be given particular attention in THEMIS 5.0 as an example of trustworthiness criteria concerning the ethical and legal aspects of Trustworthy AI.

We, however, note that also other trustworthiness characteristics treated in THEMIS 5.0, such as accuracy and robustness, indeed are relevant as ethical and legal requirements. While we have already discussed these as technical characteristics, we have, as part of this groundwork, also considered accuracy and robustness from an ethical and legal perspective. The outcome of this consideration is presented in Annex 1.3.

Furthermore, we acknowledge that a range of other relevant ethical and legal requirements for Trustworthy AI exists, such as privacy, transparency, and accountability. We will not address these directly in THEMIS 5.0, but ensure that these can be included in the THEMIS 5.0 framework at a later point in time. Furthermore, the establishment of a THEMIS 5.0 trustworthiness optimization framework will serve to strengthen also other trustworthy characteristics such as transparency and accountability.

In the following, we will first summarize ***fairness from a philosophical ethics perspective. We will then provide an ethical-legal overview.*** A complete detailing of the ethical and legal perspective on fairness is provided in Annex 1.2.

2.3.6.2 Fairness from a philosophical ethics perspective

In an ethical and philosophical context, fairness is closely related to the principle of justice (Floridi & Cowls, 2019) as, for example, seen in Rawls' theory of justice as fairness in a liberal society (Wenar 2021).

A virtue ethics perspective is also useful when considering human-centric AI, justice, and fairness (Lin, 2021; Potter, 2002: Vallor, 2016; 2024). For example, Aristotle's concept of fairness is understood in relation to justice as tied to the principle of equity, asserting that justice involves what is lawful and fair (Pomerleau, 2023).

The principle of justice in trustworthy AI also underlines fairness and the prevention of discrimination (Thiebes et al., 2021). For Kaur et. al (2022), fairness of AI systems concerns the absence of discrimination or favouritism toward an individual or a group. In the context of the social impact of algorithmic decision-making systems Varona et al. (2022) consider fairness a culture dependent variable which may facilitate inclusion, avoidance of algorithmic constraints, and unbiased design. Gagnon et al. (2020), in the context of 'trust-by-design', consider fairness in trustworthy AI as related to the ethics of algorithms, data, and practice. Finally, Li et al. (2023) note the need to balance the aim of algorithmic fairness with the need to optimize aspects of Trustworthy AI such as accuracy and robustness as well. Towards such a balance, Li et al. argue for the need for "metrics of fairness".

Considering the above, it could be observed that the principle of fairness in AI trustworthiness involves avoiding bias and discrimination within AI systems. This involves the prevention of harm and discrimination against individuals or groups, considering factors such as race, gender, or socioeconomic status. The fairness of AI systems is susceptible to various types of biases, including data bias, model bias, and evaluation bias.

The above considerations have been also reflected in the treatment of fairness within AI ethics guidelines (Buruk 2020, Reinhardt 2023). Fairness is defined as encompassing a range of philosophical ethics ideals, including equity, impartiality, egalitarianism, non-discrimination, and justice (ALTAI, 2020). Likewise, the principle of fairness is also linked to human-centred values, which AI actors should respect throughout the AI system lifecycle, fostering a digital AI ecosystem (OECD, 2021). The concept of "Fair AI" typically encompasses non-discrimination, unbiasedness, justice, impartiality, and diversity (OECD, 2021). As noted by the OECD, unfair bias must be avoided to prevent negative implications, such as the marginalization of vulnerable groups and the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, irrespective of disabilities, involving relevant stakeholders throughout their entire lifecycle.

The question of fairness is also discussed in relation to the concepts of "Responsible AI" and "Democratic AI". This is foreseen in the context of both government and industry initiatives, such as industry specific guidelines, where the concepts of fairness, non-discrimination and justice are nearly universally addressed (Hagendorff, 2020).

2.3.6.3 Fairness: an ethical – legal overview

While the *ethical dimension* of **fairness** is particularly broad, the legal dimension is narrower. That is, the legal dimension is mainly concerned with equality and non-discrimination, while the ethical concept embraces also the societal perspective, taking into consideration equity and social justice (Hoffmann, 2019).

From a *legal perspective*, the principle of fairness is encountered in multiple EU legal frameworks. First, the fairness of AI systems is mandated by non-discrimination laws which prohibit discrimination based on specific demographic aspects. Second, the Charter of Fundamental Rights of the EU mandates equality before the law and prohibits discrimination based on a range of criteria, including 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, disability, age or sexual orientation (EU Charter of fundamental rights, art. 20 and 21). Third, discrimination based on nationality is prohibited, except when justified by specific provisions within specific fields (EU Charter of Fundamental Rights, art. 21(2)). Fourth, equality and non-discrimination are established as founding values of the EU in the Treaty on the European Union (TEU) and the Treaty on the Functioning of the European Union (TFEU) (TUE, art. 2, 3(3) and 9; TFEU art. 10).

Furthermore, specific non-discrimination directives intervene in specific domains, such as The Employment Equality Directive (Directive 2000/78), The Race Equality Directive (Directive 2000/43), The Gender Goods and Services Directive (Directive 2004/113), and The Gender Equality Directive (Directive 2006/54). Here, two types of discrimination practices are addressed: direct and indirect discrimination. Direct discrimination occurs where one person is treated less favourably than another is, has been or would be treated in a comparable situation, on any of the demographic aspects established by the relevant legislative provision. Indirect discrimination occurs where an apparently neutral provision, criterion or practice would put persons belonging to a certain group at a particular disadvantage compared with other persons, unless specific justifications provided by the law are met.

Finally, the principle of fairness is also a cornerstone principle of personal data processing, according to the General Data Protection Regulation (GDPR, Regulation 2016/679, art. 5) and is recently applied to cover also discriminatory practices (EDPB Binding Decision 2/2023).

The AI Act[3] acknowledges the discrimination risks associated with the use of AI systems, especially in more sensitive areas i.e. social scoring, education, employment, access to and enjoyment of certain essential services, migration, asylum and border control management (AI Act, Recitals 28, 31, 56, 57, 58, 60). In particular, the risk of the perpetuation of historical patterns of discrimination against certain groups, justify the classifications of the above-mentioned AI systems as high-risk.

In the AI Act, the requirement of fairness is mainly addressed by the provisions dealing with the quality of data, which set specific requirements for high-risk AI systems (AI Act, Rec. 67). It is required that datasets for training, validation and testing must be relevant, sufficiently representative, to the best extent possible free of errors and complete in view of

[3] The Regulation has been approved in its final version by the Council of the EU the 21 May 2024 (Council of the EU, 2024 Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI - Consilium (europa.eu). The text was published on the Official Journal of the EU on 12 July 2024 as Regulation (EU) 2024/1689

the intended purpose of the system and with appropriate statistical properties. (AI Act, art. 10(3)). To this end, training, validation and testing data sets used by high-risks AI systems have to undergo appropriate data governance and management practices (AI Act, art. 10(2)). Moreover, datasets have to be appropriate for the specific setting (geographical, contextual, behavioural and functional) where the system is supposed to be used (AI Act, art. 10(4)). It is worth noting that art. 10(5) establishes that to ensure bias detection and correction, special categories of personal data may exceptionally be processed subject to appropriate safeguards.

The AI Act also includes procedural fairness measures. For instance, it provides a right to lodge a complaint with a market surveillance authority for any infringement of the Regulation (art. 85), and a right to a clear and meaningful explanation of the role of the AI system in decision-making, even if limited to high-risk AI systems (art. 86).

The *multiple levels of fairness* as a legal and ethical principle presented here make its operationalization complex and exposed to risks (Pasquale, 2018; Powles and Nissenbaum, 2018). Moreover, fairness challenges are distributed along all the steps of the AI system lifecycle, involving the data, the model design and development, the metrics, the application scenario, the implementation and the broader entity where the AI system is allocated (Leslie et al., 2023). **Hence, fairness cannot be simply treated as a computational problem.** Inequalities and discrimination are first and foremost social problems, to which law and ethics respond, by providing normative and procedural safeguards. Biases cannot "be fixed" by algorithms if their roots are not investigated and understood, with the involvement of the communities and individuals concerned.

2.3.6.4 Contextual insights on fairness

Context of use is key when considering ethical and legal requirements for AI systems. This is, for example, seen in the three THEMIS 5.0 use cases: healthcare risk management, optimization of port operations and prevention of disinformation in journalism. Here, the THEMIS 5.0 system will be used to optimize fairness, accuracy and robustness of AI systems for domain-specific purposes. In the following we provide an overview of how the trustworthiness characteristic of fairness can be tailored to the specificities of the use cases.

The first use case concerns AI-empowered personalised risk management regarding radiological examination, diagnosis and treatment of pancreatic diseases[4]. The radiology field is extremely prolific in terms of AI-related research and application, due to the ability of AI to recognize patterns in images (Ueda et al., 2023). In particular, fairness is achieved when AI delivers equitable care in terms of quality and access to care, regardless of socioeconomic factors and determinants of health, such as social status, ethnic differences, education, average income, insurance status, age, gender, sexual orientation (Ueda et al., 2023; Chen et al., 2023). An example is the disproportionate underdiagnosis or misdiagnosis that convolutional neural networks (CNNs) applied to radiology images performed for Hispanic patients and Medicaid patients, compared to white patients in the US (Seyyed-Kalantari et al., 2021).

Fairness in healthcare is at stake throughout the whole AI lifecycle. Fairness failures can occur during the design phase and arise from biased data where minority groups are underrepresented, absent, or misrepresented. For instance, it is more difficult to diagnose melanoma from pictures of people with dark skin than from those with light skin and this different feature of detection needs to be considered when datasets are composed and the purposes of the system are defined (Ueda et al., 2023). When datasets are not carefully composed or selected at the design stage, during the deployment the so-called dataset shift takes place, i.e. disparate performances caused by the mismatch between the datasets and the targeted use of the AI system or by differences in diagnostic criteria across different populations or by the use of different taxonomies across healthcare systems (Chen et al., 2023).

A different type of fairness failure affects the algorithm per se, regardless of the datasets. During the development phase, the design or the learning mechanism of the algorithm can be influenced by existing healthcare disparities, leading to underperformance towards underserved groups of people (Ueda et al., 2023). For example, an algorithm in the US has been trained to predict patients' needs on the assumption that the amount of money patients spend on health care reveals their actual health needs. Whereas this can be true, it may be for any situation or every group of people because reality is complex and multifaceted: since some demographic groups spent less on healthcare the algorithm concluded that these groups were healthier than demographic groups spending more, while other factors such as healthcare accessibility and actual spending power should have been considered (Obermeyer et al., 2019). Also, human and cognitive biases can influence the fair deployment of AI systems: healthcare professionals may tend to over-rely on AI predictions if not properly trained to interpret the AI outcomes and/or due to a combination of other causes, such as numerous false

[4] See section 0

alarms, i.e. alert fatigue (Ueda et al., 2023). By contrast, patients belonging to certain groups may under-rely on AI systems, because of skepticism deriving from historical healthcare disparities (Ueda et al., 2023).

The second use case concerns the optimization of port operations by predicting the Estimated Time of Arrival (ETA) of vessels[5]. Ports archive large historical records and operate according to predictable patterns based on seasonality and types of operations. As a consequence, the ability of AI to make predictions by recognizing patterns in past data is extremely promising for many different purposes (i.e. improving efficiency, quality of port management, carbon production, and safety) and already has numerous applications worldwide (Farzadmehr et al., 2023; Skournik, 2023).

In contrast to the healthcare domain, the use of AI systems to predict vessel arrivals seems less exposed to unfair outcomes. However, a more in-depth analysis of the issue reveals that computational and even systemic biases can affect the management of port operations in terms of precedence given to types of cargo or vessels, shipping agencies and transport companies with domino effects on the access to goods by the final recipients.

The third use case regards the prevention of disinformation in journalism with the use of a collaborative platform that integrates AI-based fact-checking and hate speech-checking tools[6]. There are several steps of fact and hate speech checking that can be automatized with AI. For instance, AI can perform: i) the check-worthiness estimation, to identify claims for which there is an interest of the general public to know the truth; ii) the retrieval of evidence, to find information to verify the claim; iii) the proper assessment of the veracity/hate speech nature of a claim, with verdict prediction and supporting justification (Guo et al., 2022).

The implications in terms of fairness significantly change depending on the specific “step” performed or assisted by the AI systems and on the type of speech detection. For instance, the check-worthiness estimation of a claim, either in terms of veracity or hate speech, can be performed through different lenses, considering the view of society at large, vulnerable communities or individuals or the fact-checkers directly (Neumann et al., 2023). Fairness is impacted precisely when the use of AI tools does not benefit everyone equally. In this regard, the labelling of training data is crucial and the involvement of Knowledgeable Communities specifically consulted for their check-worthiness opinions has been studied as a measure to address differential treatments (Neumann et al., 2023).

2.3.6.5 The ethical and legal importance of metrics and standards for trustworthiness characteristics

After having identified and defined the specific elements of fairness, accuracy and robustness that Themis 5.0. technology intends to optimize - in line with the legal and ethical framework, within the relevant context and considering the interactions with other requirements - adequate metrics which reflect the definition of these requirements must be chosen. In essence, the definitions of fairness, accuracy and robustness need to be translated into detectable benchmarks.

The HLEG AI Guidelines (HLEG on AI, 2019) underline the necessity to develop multiple metrics to appropriately test and validate all the components of the AI systems. The same requirement for trustworthy AI can have multiple levels of meaning – consider for instance the complexity of the requirement of fairness - and the metrics chosen must be able to appropriately reflect such different levels of meaning that fall under a certain requirement.

The necessity of identifying adequate metrics is recognized by the AI Act which demands testing high-risk AI system against preliminarily defined metrics and probabilistic thresholds that are appropriate to the intended purpose (AI Act, art. 9 (8)) of the system, as well as an indication in the technical documentation of ”metrics used to measure accuracy, robustness and compliance with other relevant requirements...as well as potentially discriminatory impacts” (AI Act, art. 11 and Annex IV(2)(g)). Also, the instructions for use for deployers must be transparent about the metrics of accuracy, robustness and cybersecurity to be expected by the high-risk AI system (AI ACT, art. 13 (3)(b)(ii)).

The AI Act refers explicitly to metrics as the technical aspects of how to measure the appropriate levels of specific requirements, as accuracy and robustness (AI Act, art. 15(2)). These high-risk AI requirements will be subject to significant standardization. The European Commission:

- will promote the adoption of common benchmarks to measure accuracy, robustness and other relevant performance metrics (AI Act, Rec. 74 and art. 15(2)).
- will require European standardisation organizations to deliver “harmonised standards” covering the requirements for high-risk AI systems (AI Act, art. 40).

[5] See section 3.2.2 AI in Port
[6] See section 0

AI for fighting disinformation

- will be empowered to adopt "common specifications" covering the requirements for high-risk AI systems, when "harmonised standards" cannot be adopted (AI Act, art. 41).

These provisions provide for a crucial compliance mechanism. If high-risk AI systems comply with the "harmonised standards" or "the common specifications" they are presumed to be in conformity with the corresponding legal requirements (AI Act, art. 40 and 41).

The adoption of shared benchmarks and standards in the context of the AI Act implementation will be steered by the existing international initiatives of standardization bodies and global cooperation forums, when consistent with European Union values. For instance, the need to agree on shared metrics is recognized by the EU-US Trade and Technology Council (TTC) which intends to build a shared hub/repository of metrics for trustworthy AI (TTC Joint Roadmap for Trustworthy AI and Risk Management) and by extensive and ongoing efforts conducted at the OECD level (OECD.AI, Catalogue of Tools & Metrics for Trustworthy AI).

It is also realistic to recognise that not all requirements' definitions can be automated by metrics and that therefore Trustworthy AI cannot be realized entirely through technical methods (HLEG on AI, 2019).

Therefore, THEMIS 5.0. needs to make clear to its intended users how the optimization of fairness, accuracy or robustness can be automated, to what extent and according to which metrics.

Key recommendations for the optimization AI system trustworthiness characteristics, from the ethical and legal perspectives is provided in ***Table 2***.

Table 2. recommendations for optimization of fairness, accuracy and robustness of AI systems from ethical / legal perspectives

Key recommendations for the optimization of fairness, accuracy and robustness of the AI system under test – Ethical and Legal perspectives	
Premise	The ethical and legal analysis in THEMIS 5.0. has a twofold perspective, which must be clear to all the project Partners: • The design and development of THEMIS 5.0 personalised AI Trustworthiness Optimisation Ecosystem as a whole needs to abide by the legal and ethical requirements for Trustworthy AI. • The AI trustworthiness optimization of the AI system under test that THEMIS 5.0. technology aims to carry out will be focused on specific AI trustworthiness requirements (i.e. fairness, accuracy, robustness), whose understanding and implementation need to be legally and ethically compliant. These recommendations are addressed to the second point – optimization of fairness, accuracy and robustness of the AI system under test. Key legal and ethical requirements for the design and development of THEMIS 5.0. technology as a whole are included in Section 2.7 and Annex 2.
1.	The approach to trustworthiness optimization to be developed in THEMIS 5.0 initially focuses on fairness, robustness and accuracy. However, Trustworthy AI involves more dimensions and requirements than these. Therefore, it must be clear to the downstream deployers[7] of the initial THEMIS 5.0 approach which specific trustworthiness requirements THEMIS 5.0 will optimize and why. Such information is relevant also to allow the deployer to inform all the affected persons about the operation and the functionalities of THEMIS 5.0. AI systems in combination with its own AI system.
2.	The requirements for Trustworthy AI do not have a universal meaning and implementation method, instead they acquire different significance and prominence, depending on the context of use. A contextual interpretation of the requirements is therefore essential.
3.	The requirements for Trustworthy AI are not stand-alone: they can work synergistically or, on the contrary, there can be tensions among them. Their balanced interaction must be considered to achieve AI trustworthiness.
4.	From a legal and ethical point of view, trustworthiness requirements have multiple levels of meaning – consider in particular the complexity of the requirement of fairness as equality, non-discrimination, social justice, redress and procedural fairness measures etc. Hence, the THEMIS 5.0 approach should identify and define specific elements of fairness, accuracy and robustness of the AI system under test that can be optimized.
5.	After having identified and defined the specific elements of fairness, accuracy and robustness that THEMIS 5.0. technology intends to optimize - in line with the legal and ethical EU framework, within the relevant context and considering the interactions with other requirements - adequate metrics which reflect these definitions must be chosen. However, not all the elements of the requirements definitions can be automated by metrics, as Trustworthy AI is challenging to realize fully realised by only technical means. Therefore, THEMIS 5.0. needs to make clear how the optimization of fairness, accuracy or robustness can be automated, to what extent and according to which metrics.

[7] Deployer is anyone using an AI system under its authority, excluding the use in the course of personal non-professional activities

2.4. Human-centred aspects of Trustworthy AI – explainability and user preferences

2.4.1. Human-centred AI and explainability

Human-centred AI. Human-centred AI is an emerging area of research positioned within the cross-disciplinary fields of human-computer interaction and human-centred design – fields with particular concern for usability and user experience of computer technology and interactive systems. Centres of human-centred AI research have been established by universities such as Stanford University, UC Berkley, and MIT, accentuating the need for humanistic and ethical AI, enhancing rather than replacing humans (Xu, 2019). A leading theorist on human-centred AI, Shneiderman (2022), summarized this area of research and practice as concerned with applying methods of human-centred and user experience design to research and development of AI systems to ensure (a) their usefulness and benefit to users and (b) their responsible development taking into account the ethical, societal and sustainability impact of AI. In a similar vein, Riedl (2019) has accentuated social responsibility as key to human-centred AI.

In European policy-making, human-centred AI is closely related to Trustworthy AI (AI HLEG, 2019; Larsson et al., 2020). The notion of Trustworthy AI has also been embraced by researchers and theorists with an interest in human-centred AI. Trustworthy AI is seen as requiring governance structures at the levels of the team (e.g. through audit trails, verification, bias testing), the organization (e.g. through leadership, training, reviews, standards), and the industry (Shneiderman, 2020a) the latter implying trustworthiness as 'assessed by respected independent oversight' – by third parties such as accounting firms, NGO's, or research organizations. It is further argued that trustworthy AI requires both high levels of automation and high levels of human control. Successful integration of high levels of human control and automation requires careful human-centred design processes. Shneiderman (2020b; 2022) proposed high level guidelines and design patterns for supporting such processes. Such process may also build on Endsley's (2017) guidelines of automation, including 'automation transparency'.

Human-centred AI and explainability. The need for explainable and comprehensible AI is accentuated in research on human-centred AI (Xu, 2019; Shneiderman, 2020a). Explainable AI has also been discussed as a means of providing needed transparency in the context of the AI Act, though the relevance of this is contested (Panigutti et al., 2023). To increase trust in AI decision making systems, decisions need to be (a) understandable to humans in a given context and (b) explicitly explained (Miller, 2019). To enable this, a range of theories and constructs on explainable AI have been proposed.

However, not all explainable AI constructs may be helpful to users. Liao et al. (2020) argued such constructs often reflect researchers' intuitions concerning explanation rather than actual user needs for explanation. This may be problematic both for lay users and expert users. For example, a model in the medical domain that predicts a patient to have a flu might explain this by listing sneezing and headache as defining features though this explanation may not satisfy a medical doctor's information needs (Liao et al., 2020). Hence, it is critical to evaluate explainable AI approaches through human-centred evaluations.

Vilone and Longo (2021), in their review of >400 studies of explainable AI and related evaluation approaches, specifically discussed 70 papers on evaluations of explainability. While some of these concerned objective metrics (e.g., explanation completeness, text quality metrics, and performance in terms of correctness or productivity), most concerned human-centred evaluations, that is, human-in-the-loop approaches. Though no high-level interpretations of the human-centred evaluations were provided, a broad range of quantitative and qualitative evaluation examples were provided. Likewise, Liao et al. (2020) listed a substantial set of studies on human-centred evaluations of explainable AI.

Explainable AI approaches that are helpful to users, may have certain common characteristics. In a social sciences-oriented review of the literature on explainable AI, Miller (2019) argued that explanations that matter to users are (a) 'contrastive', where a decision is contrasted with a counterfactual, (b) 'selected', where some particularly salient explanations are preferred, (c) referring to causes, rather than probabilities, and (d) 'social', that is, a transfer of knowledge as part of a conversation. In a similar vein, drawing on insight from interviews with AI system designers at IBM, Liao et al. (2020) proposed an Explainable AI Question Bank with questions users typically need explanations for. Specifically, questions of relevance to users in need of explanations may concern (a) System data, (b) System output, (c) System performance, (d) How the system makes predictions (global explanations), (e) Why a specific prediction was made (local explanations), and the possible impact of changes to the assumptions – such as in What if-questions. Concerning environments for human-centred explanations, Shneiderman (2020a) argued for the benefit of 'exploratory user interfaces' allowing users to actionably engage with explanation systems.

For use in THEMIS 5.0, human-centred explainable AI may be summarized as follows: Human-centred explainable AI accentuates the need for explanations that are understandable and helpful to users, and that identification of relevant explanations may require human-centred evaluations. Human-centred explainability may have certain characteristics, such as being contrastive and oriented towards causalities rather than probabilities (Miller, 2019), be proposed in response to typical user questions, such as how-, why-, and what if-questions (Liao et al., 2020), and be provided in ways that allows users to interact with explanations to adapt these to their information needs, such as through exploratory user interfaces (Shneiderman, 2020a).

An example of human-centred explainable AI for the medical domain was provided by Panigutti et al. (2023): *Doctor XAI*. The research team took as starting point that explainability is critical for decision makers to take advantage of and oversee AI systems. They noted that explainable AI concerns (a) the extraction of explanations and (b) the presentation of explanations to users, and that the latter has received limited attention. Panigutti et al. (2023) accentuated that explainable should enable 'trust calibration', that is to avoid overreliance, due to automation bias, or reluctance to trust, due to algorithmic aversion. They also noted unclear liability as a reason for distrust in AI systems.

Doctor XAI applied removal-based explanations (contrasted with perturbation-based approaches such as LIME and SHAP). For domain relevance, the explanations prioritized features with semantic relevance to the medical field, based on semantic search in medical ontologies. In the first version of *Doctor XAI*, explanations for medical diagnostics were provided with reference to features reflecting diagnostic characteristics in an easily accessible user interface. A user study with >40 doctors/nurses following the Judge-Advisor Paradigm was reported, which found explainable AI decision support to have higher weight of advice than other AI decision support, but not to impact trust or behavioural intent. The participants also provided qualitative feedback which enabled co-creation of a second version of *Doctor XAI*, exemplifying human-centred design of the explainable AI approach. The design process applied by Panigutti et al (2023) could serve as inspiration for design of human-centred explanations in THEMIS 5.0.

2.4.2. Identifying preferences of AI system user groups

In this sub-section, we present a state-of-the-art overview of AI user groups, beginning with ethical considerations and extending to technology acceptance, psychosocial aspects, and trustworthiness perceptions.

User preferences in the realm of AI is a multifaceted process involving the collection and analysis of data to construct comprehensive user groups. These groups encompass a spectrum of information, ranging from user preferences and behaviours to demographic details. The primary objective is to enhance AI applications by personalizing user experiences and optimizing overall system efficacy. One pivotal facet of AI user preferences is content personalization. This involves AI systems scrutinizing user behaviour and historical data to tailor content recommendations. Predictive analysis is integral to AI user preferences, as it empowers systems to anticipate user needs and deliver proactive services. Examples include predicting the next word in a sentence or suggesting actions based on an analysis of historical user behaviour. Within the context of security and fraud detection, identifying preferences is a crucial tool. AI algorithms analyse user interactions to detect abnormal behaviour patterns that may indicate security threats or fraudulent activities. This proactive approach is essential in maintaining the integrity of AI systems.

Ethical considerations are one major aspect of AI user preferences. In particular, privacy concerns emerge due to the collection and processing of user data. Harmonizing personalization and user privacy is imperative to prevent invasive practices. Transparency is a key ethical principle, necessitating that users are informed about the types of data being collected and the purposes for which it is used. Bias and fairness considerations are critical, as user groups algorithms may inadvertently perpetuate biases present in training data. Several legal frameworks influence AI user groups. The General Data Protection Regulation (GDPR) in the European Union (EU, 2018) regulates the processing of personal data, emphasizing user consent, transparency, and the right to be forgotten. In the United States, the California Consumer Privacy Act (CCPA) (State of California, 2018) grants specific rights to California residents regarding their personal information, shaping how companies handle user data in AI preferences.

Two important dimensions of AI user preferences are: 1) Federated learning, an emerging approach, enabling AI models to be trained across decentralized devices without transferring raw user data. This addresses privacy concerns associated with centralized data processing. 2) Explainability in AI aiming to make AI systems more transparent, helping users understand how their data is utilized in user preferences. Moreover, the integration of federated learning in AI profiling systems provides a significant improvement in user data privacy by decentralizing the training of AI models. This method reduces the risks associated with central data storage and processing, directly addressing some of the key ethical concerns previously highlighted (Yang et al., 2020). In addition, recent research emphasizes the importance of explainability in AI user groups. Making AI decision-making processes transparent and understandable is not only an ethical obligation but also a practical necessity for fostering trust among users (Doshi-Velez & Kim, 2017).

Psychological groups within AI involve a nuanced analysis of users' personality traits, preferences, and emotional states. This insight enables AI systems to tailor user experiences, content, and recommendations based on individual behaviours. However, ethical considerations, such as privacy concerns, must be addressed, emphasizing transparent and consensual practices in the analysis of psychological groups. Technology acceptance and behavioural intentions play pivotal roles in predicting user engagement with AI systems (Davis, 1989). Identifying user groups aids in understanding user acceptance and predicting behaviours, allowing AI interfaces to be tailored for enhanced user engagement. Ethically, informed consent is critical, ensuring users have control over the information used for identifying preferences and maintaining transparency in the process. As AI holds the potential to benefit numerous individuals, their willingness to embrace and effectively utilize this technology becomes crucial. A lack of acceptance can lead to diminished user uptake of AI, resulting in the underutilization of resources, an excess of AI devices, and a potential decline in technological innovation to the detriment of consumers (Kirlidog & Kaynak, 2011). Acceptance serves as a predictive measure that encapsulates a personal choice, such as making an informed purchase of AI devices. Alternatively, acceptance can be an involuntary action, such as utilizing AI chatbots that may present themselves as non-AI agents. Hence, there exist varying levels of agency involved in user acceptance. Evaluating user acceptance is essential for stakeholders to comprehend the variables necessary to maximize technology uptake in diverse circumstances. Various models have been employed to assess user acceptance of AI, including the Technology Acceptance Model (TAM; Davis, 1985, 1989).

The Technology Acceptance Model, TAM (Davis, 1985, 1989), was adapted from the Theory of Reasoned Action (Fishbein et al., 1975) and posits that external variables, such as media and social references, shape individuals' perceived usefulness (PU) and perceived ease of use (PEOU). These factors contribute to their intentions to use technology, ultimately influencing their actual system usage (Davis, 1985, 1989). PU reflects the extent to which users perceive the technology as beneficial in their everyday lives (Davis, 1989), often emerging as the most robust positive predictor of an individual's intention to use new technology (Davis, 1989; Rafique et al., 2020). On the other hand, PEOU refers to a user's perception of how effortless it would be to use a technological device (Davis, 1989). PEOU is considered to exert a weaker influence on technology acceptance than PU since it is solely relevant to the technical use of a device, a factor that has diminished in significance as users have become more accustomed to employing technology in their daily routines (Davis, 1985, 1989; Lunney et al., 2016). There are also significant differences between user groups with regards to the impact of ease of use or effort perceptions (Venkatesh et al., 2012). The TAM is frequently extended with additional variables, such as trust and knowledge, to enhance its predictive power (Kashive et al., 2021).

Trustworthiness perceptions, which encompass users' beliefs in the reliability and benevolence of AI systems, are significantly influenced by identifying groups practices. Transparent user groups and the mitigation of biases play a crucial role in building and maintaining user trust, an essential element in fostering positive user-AI interactions (Binns et al., 2018). Incorporating psychosocial aspects, such as cultural norms, social expectations, and emotional responses, into AI user groups enhances cultural sensitivity and contributes to improved user engagement (Chouldechova, 2017). Ethical considerations in this context involve addressing potential cultural biases embedded in preference algorithms and ensuring sensitivity to diverse psychosocial contexts (Barocas & Selbst, 2016). One fundamental ethical consideration is empowering users with control over their user groups. AI systems should provide clear interfaces for users to comprehend and manage their preferences, thereby ensuring a sense of autonomy in identifying user groups process (Diakopoulos, 2016). This aligns with the principle of user-centric design, emphasizing the importance of user empowerment in shaping AI interactions. Furthermore, the collaborative aspect of human-AI interaction is vital. Trust in AI systems remains a crucial aspect, and studies have shown that users’ trust can be significantly enhanced when user group practices are transparent and users are given control over their data (Binns et al., 2018). This points to the necessity of designing AI systems that prioritize user agency, allowing individuals to manage and even modify their profiles as needed. Furthermore, recent work highlights that ethical AI systems should not only avoid harm but actively promote fairness, especially in how different user groups are profiled and treated by AI systems (Jobin, et al., 2019).

Identifying preferences should inform human decision-making, but the ultimate decisions must remain collaborative, with ethical considerations emphasizing that responsibility for decisions should rest with humans. AI should be viewed as a tool for augmentation rather than replacement, underscoring the importance of maintaining human agency in decision-making processes (Bryson et al., 2017). Addressing trustworthiness perceptions and ethical considerations in AI user groups involves transparency, bias mitigation, cultural sensitivity, and user empowerment. By incorporating these principles, we can foster positive user-AI interactions and ensure that AI serves as a supportive tool within a collaborative decision-making framework.

2.5. Risk Assessment for Trustworthy AI

The THEMIS approach to risk assessment for trustworthy and responsible AI is to use the existing knowledge-based risk assessment tool Spyderisk, developed by UoS (Spyderisk, 2024, Phillips et al, 2024). This tool, currently at TRL 6, will be enhanced with knowledge on AI trustworthiness and risks.

In Section 2.5 we present the groundwork for developing such a risk assessment-based approach to trustworthiness evaluation and optimization. First, in Section 2.5.1, we summarize and compare relevant frameworks for risk management, specifically those of ENISA (2023) and NIST (2023). In Section 2.5.2 we detail how Spyderisk may be adapted to the purpose of THEMIS 5.0. In Section 2.5.3 we map the ENISA and NIST frameworks to the Spyderisk modelling concepts. Finally, in Section 2.5.4 we detail our approach in terms of knowledge extensions to Spyderisk.

2.5.1. Summary and comparison of ENISA and NIST frameworks

To assess the fit of relevant frameworks for risk management of AI systems for the THEMIS 5.0 approach, we here summarize and compare two leading frameworks – those of ENISA (2023) and NIST (2023). A more in-depth detailing of the two frameworks is provided in Annex 3.1.

The ENISA framework for AI cybersecurity practices (ENISA, 2023) highlights the need for AI-specific practices in cybersecurity. The framework applies the OECD definition of AI and accentuates the need for a lifecycle perspective on AI trustworthiness. It is structured in three layers reflecting different aspects of good AI cybersecurity practices: Cybersecurity foundations, AI-specific cybersecurity practices, and sector specific considerations. Hence, the ENISA framework covers risks directly related to the AI system as well as its socio-technical environment.

According to the ENISA framework, AI systems will have desired characteristics that contribute to the trustworthiness of an AI system. Defining AI trustworthiness as *"the confidence that AI systems will behave within specified norms, as a function of some characteristics […]"*, ENISA gives these different characteristics of AI trustworthiness as including accuracy, fairness, and robustness, in addition to accountability, explainability, privacy, reliability, resiliency, safety, security, and transparency. These AI system characteristics are classified as technical, socio-technical, and guiding principles.

ENISA recommends the ISO 2700x standards (ISO 2700x), NIST AI RMF, and ENISA's best practices for controls in risk management of general-purpose AI. With regard to practices reflecting the characteristics of the socio-technical environment, including sector specific requirements, the framework notes that fragmented recommendations, best practices, solutions, and tools may be stumbling blocks for sectoral stakeholders, and that collaboration and information sharing on sector-specific issues and mitigations between sectoral stakeholders is needed.

The NIST Artificial Intelligence Risk Management Framework, NIST AI RMF (NIST, 2023), is a framework to support the management of AI risks and foster trustworthy and responsible AI. Like ENISA, NIST applies the OECD definition of AI and accentuates a lifecycle perspective on AI trustworthiness. Risk management in NIST AI RMF is based on adaptations of risk management and assessment definitions from ISO 31000 (2018), but with an aim to identify both opportunities and threats. The NIST AI RMF acknowledges that some AI risks may be difficult to quantitatively measure, though the risk may nevertheless be relevant.

As ENISA, the NIST framework identifies a set of characteristics for Trustworthy AI, addressing technical, socio-technical and ethical / legal aspects of the AI system, though the set is not fully overlapping between the two frameworks. Key to NIST AI RMF is that a trustworthy AI system requires balancing the trustworthiness characteristics of for a specific AI system within a specific context of use – something that may entail trade-offs. However, technical characteristics, in NIST subsumed in the category Valid and Reliable, is considered a necessary condition for the other trustworthiness characteristics.

In the NIST AI RMF, it is argued that AI risk management should be incorporated into the broader risk management surrounding an AI system considering also its environment and relevant actors. To support the risk management process, an RMF core is proposed including the four functions – Govern, Map, Measure, and Manage – that are described at a process level. As such, the AI RMF gives a comprehensive non-prescriptive framework for organisations working with AI systems.

Comparing the ENISA and NIST approaches, their similarity in definitions and their consideration of the socio-technical character of trustworthy AI are noteworthy. In the ENISA framework, a layered approach addresses the ICT infrastructure, the AI system itself, and the system within a given sector or socio-technical environment. The NIST framework is a four-function framework that maps out, measures, and manages the risks of AI, whilst maintaining governance throughout. The ENISA framework categorises AI threats into types of attacks, whilst the NIST framework categorises AI harms into broad categories based on harms at the person-, organisation-, or ecosystem-level. ENISA identify 11 trustworthy characteristics of an AI system whilst NIST identify seven AI trustworthy characteristics that are broader and that ENISA's

11 fit within. Both frameworks complement each other and provide guidelines and methodology. As such, an implemented approach to AI risk assessment can be formed from them and be compatible with both.

2.5.2. Spyderisk for risk assessment and management in THEMIS 5.0

Spyderisk is an asset-based automated risk assessment and management open project following the ISO 27005 (2022) risk assessment methodology. The corresponding Spyderisk risk modelling tool is referred to as the Spyderisk System Modeller (SSM).

The core concept of SSM is shown in Figure 4 and is comprised of three main parts: a knowledge base that contains and encodes domain specific information, a GUI frontend that provides a way to easily represent and model a given system, and a validation and risk calculation engine.

When applying SSM, the user builds a model of their system under test (SUT), and the engine uses fuzzy logic and fuzzy inference with the encoded information of the knowledge base to determine and calculate the threats and risks present in the SUT, along with the threat paths that form due to threats cascading from one to another.

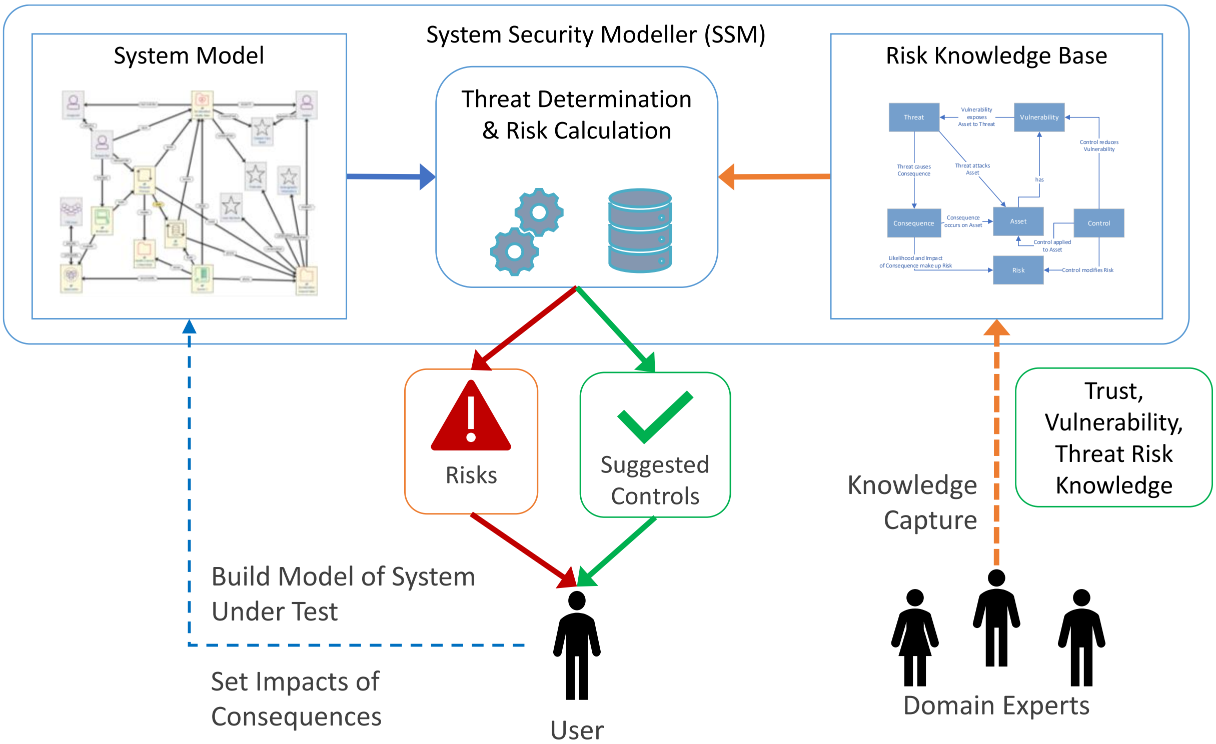

Figure 4: UoS Spyderisk System Modeller (SSM) Risk Assessment Tool

The schema of the SSM, adapted for THEMIS, is shown in Figure 5. This is derived from several risk assessment patterns, predominantly the ISO 27000 series on information security risk management, but also includes aspects such as safety derived from risk management in other domains (e.g. medicine), where safety is an important factor.

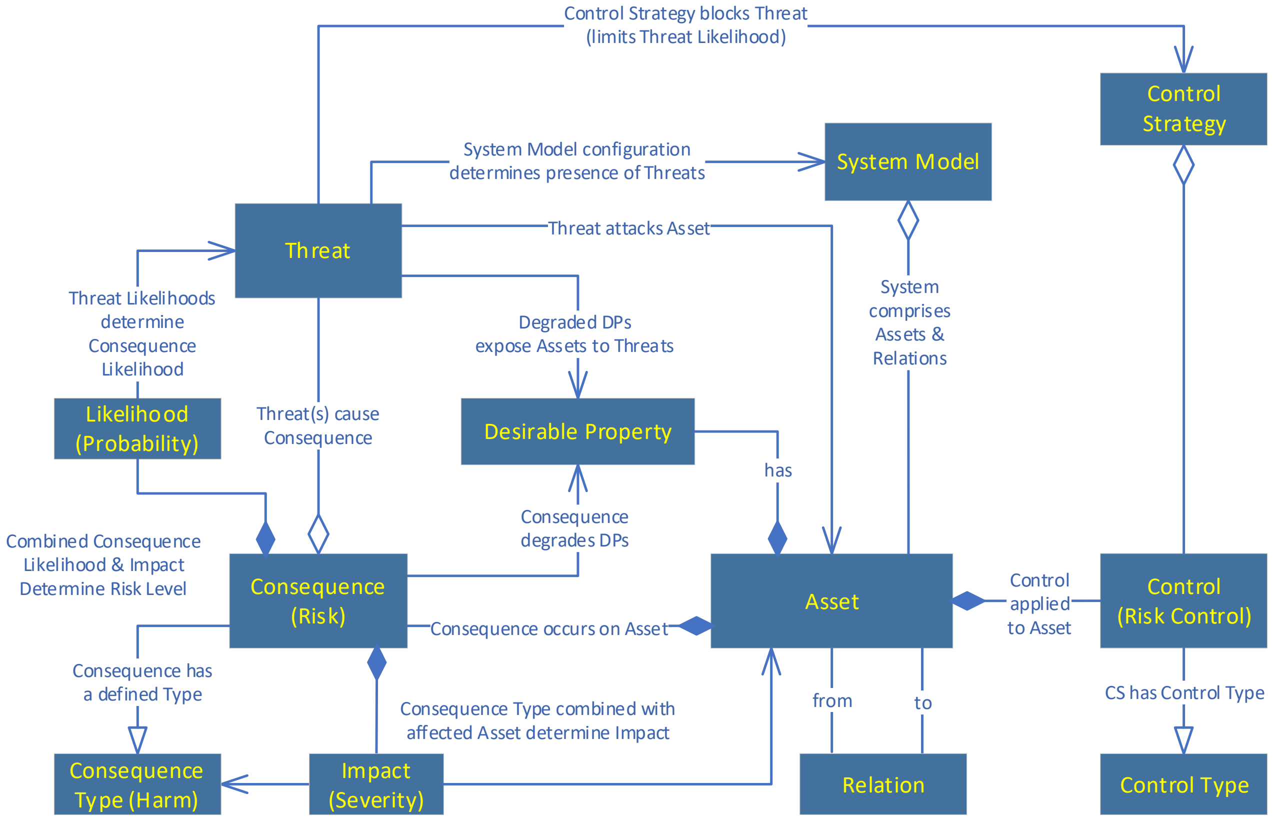

Figure 5: SSM Knowledge Schema (adapted for THEMIS)

In line with ISO 27005 (2022), Spyderisk concerns system Assets and the Harms that may occur if those Assets are exposed to Threats. Definitions of key Spyderisk concepts are provided in Annex 3.2.

For a given domain, the Spyderisk knowledge base includes the encoded threats, assets, consequences, controls, and default likelihood, and impact values. The knowledge base is a generalised abstraction, not so specific to a system model that it cannot be used for others within the given domain, but specific enough to contain the assets, threats, etc., from the domain and with a generality that it can be used for many different system models, each of the given domain. Of the three main components of Spyderisk, the knowledge base is domain-specific, with the other two parts being domain-agnostic. The knowledge base currently packaged with Spyderisk is a cyber-physical systems knowledge base pertaining to cybersecurity. It already contains numerous Assets that are relevant to AI systems, e.g. software processes, ICT hardware, data, computer networks, people, places & jurisdictions.

A key contribution of THEMIS is to extend the SSM knowledge base towards assets, threats, harms, risks, consequences and desirable properties that are inherent in an AI system or may be affected by AI systems. The extension of SSM knowledge will be a key ongoing task throughout THEMIS. The knowledge extension process named “domain modelling”, as it concerns modelling the knowledge from a specific domain of expertise and is shown in Figure 6.

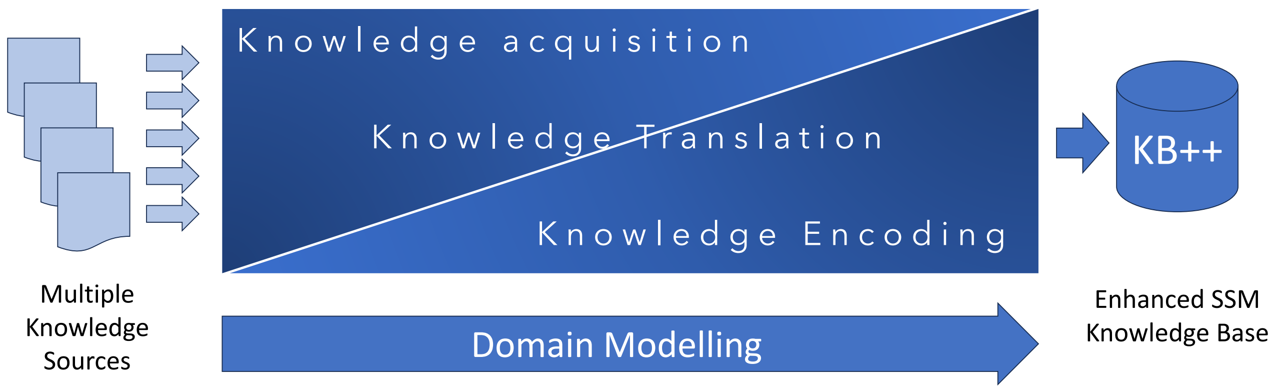

Figure 6: SSM Knowledge Extension

Knowledge acquisition is the first step and can take multiple forms, e.g. literature survey, consultation, discussion within the project, but is essentially the process of locating sources of knowledge, assessing their relevance and collating the knowledge. Collation often involves clustering similar themes from different sources and cross-referencing to determine correlations or to explore differences. Key questions include: what to include, what are important harms and assets to model?

Knowledge Translation is the second step and involves translating the acquired knowledge into the format of the SSM knowledge schema, i.e. determining which elements of knowledge can be modelled as Assets, Threats, Consequences, Desirable Properties, Controls etc. It also involves understanding interrelationships between the new knowledge and existing knowledge already in the knowledge base, and in some cases resolving inconsistencies / contradictions between the existing and new knowledge.

Knowledge Encoding is the final step and involves practical steps to encode the new knowledge into the physical representation of the knowledge base (RDF triples), unit testing for correct behaviour, bug fixing, regression testing for overall consistency with the existing knowledge, with associated consolidation, bug fixing, conflict resolution as needed.

A core principle of domain modelling is that abstraction is needed as a domain model specific and specialised to one given system means that it can only ever be used for that system, but abstracting away the specifics of assets and relationships in example systems to get their generic system-independent forms means the domain model is usable for many systems across all the domain in question.

Related to the principle of abstraction and domain knowledge, a key consideration alluded to above in the Knowledge Translation section is that there are interrelationships between knowledge domains in the knowledge translation and encoding. An example of such domains is shown in Figure 7.

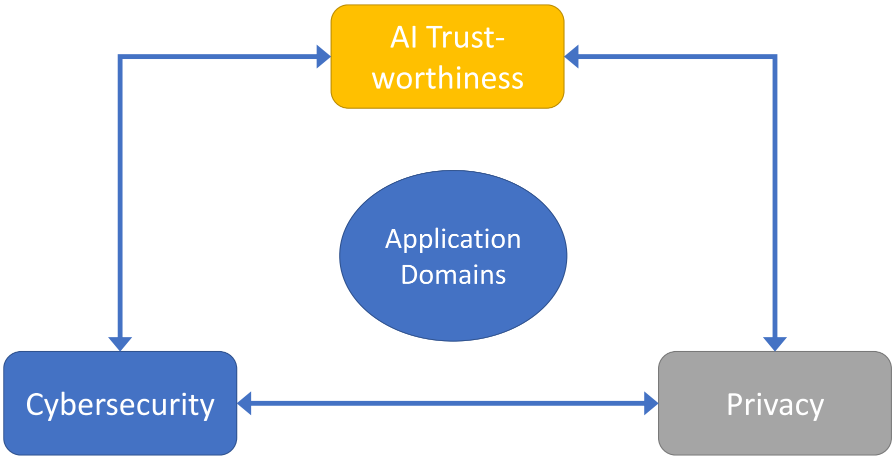

Figure 7: Current / New Risk Knowledge Domains

Here, SSM already has extensive encoded knowledge regarding computer processing, data storage and usage, ICT hardware, networking, plus a significant amount in human and social aspects such as human roles, rights (focusing on privacy), institutions, physical spaces and regulation. This contains threats and controls associated with cybersecurity threats and attacks, the consequences of the threat on the different asset types and how those threats can propagate through systems made from these asset types. The domain model representing this is named the "Network Domain Model".

THEMIS 5.0 is extending the Network Domain Model towards AI trustworthiness, risks and harms etc. Thus, there are cross-domain interrelationships that need to be represented in the knowledge base. This is a natural reflection of the enhancements associated with the knowledge extensions and represents the complexities in real world situations where AI systems are deployed and used. Examples of such cross-domain interrelationships are as follows.

- AI models are instances of computer processes, reading data as input and writing data as output, running on ICT hardware, communicating over computer networks and interacting with people.
- Cybersecurity threats affect AI quality / accuracy / robustness / fairness.
- AI accuracy can affect the Integrity of data resulting from AI model results.
- AI models can unknowingly expose personal information, harming the privacy of citizens.
- Cybersecurity Controls on e.g. data security may reduce risks of unauthentic training data for AI models.

These are initial examples but illustrate different cause and effect relationships between the domains, e.g. cybersecurity threats affect AI models, which may, in turn, affect AI system trustworthiness characteristics, but Cybersecurity controls may be able to limit harm from AI threats. Patterns such as these will be actively monitored in the domain modelling exercises. A key observation made so far is that data is a key link across domains. This is primarily due to its universality of data and the multiple forms data can take, plus the severity of indirect impacts of data compromises, especially on people. For example, data is used for AI training so compromises in training data affect the model results (also data).

2.5.3. Mapping ENISA and NIST to Spyderisk modelling concepts

Through the examined frameworks, an initial set of AI desirable properties, AI harms, AI threat types, and AI controls have been identified. In THEMIS, the focus is on robustness and accuracy, decision impact, and fairness; and these are the key classes of desirable property of the AI.

2.5.3.1 Mapping Spyderisk to ENISA

It is possible to model most of the AI asset types identified by ENISA in Spyderisk, and there is already good cyber-physical coverage for data, actors/stakeholders, and processes. Extensions needed cover the specifics of AI models and data, and initial work towards this is discussed later.

The framework by ENISA approaches risk assessment in a three-layer manner. Spyderisk is compatible with this as each layer is included within the Spyderisk process. Regarding the first layer, for a user modelling a system in Spyderisk to produce a correctly functioning system model, they must specify the cyber-physical components present within the system. This then means that the user will inherently model the ICT infrastructure of the first layer, which is already covered in Spyderisk. The second layer of the ENISA framework is sector-agnostic AI risk management and provides requirements for threats, risks and controls specific to AI. Relevant to the sector-specific third layer, some AI threats may be encoded that are sector-specific. Additionally, covering the third layer, the modeller of a system shall be able to adjust consequence impact levels related to key parts of the sector-specific AI system. These impacts will then affect the risk levels so that sector-critical risks are highlighted and prioritised.

2.5.3.2 Mapping Spyderisk to NIST

The methodology of Spyderisk is also compatible with the NIST AI RMF and the four functions that it uses. Spyderisk helps to make risk assessment more accessible to non-expert audiences, and so the use of it works towards the accountability, transparency, and reporting structures of the Governance function, as well as the processes and procedures related to the mapping, measuring, and managing of AI risks. For the Map function, designing an AI-related system model within Spyderisk and running risk assessments for it will give the context of risks concerning the AI system. For the Measure function, provided appropriate impact input levels are given, the automated risk assessment of Spyderisk will quickly and efficiently calculate the likelihood and risk levels for the threats and consequences that it identifies, including AI-related risks and consequences. Additionally, Spyderisk evaluates the trustworthy characteristics of the AI system, and through repeating the risk assessment over the lifecycle of the AI system, identified AI risks can be tracked over time, both furthering the Measure function. Finally, for the Manage function, the risk assessment provides and recommends controls that will address identified threats and risks. This will also be the case for AI-related threats and risks, and so the user performing the risk assessment will know the appropriate controls to apply to manage the risks of their AI system.

2.5.3.3 How Spyderisk can be used in ENISA- and NIST-Based Assessments

Both the ENISA and NIST frameworks have a large focus on the compromise of AI trustworthiness characteristics, actions to reduce the likelihoods of these compromises occurring, transparent reporting on the AI risks present, and accountability for the actions and decisions made regarding them. Both frameworks also focus on the whole lifecycle of the AI system and surrounding ICT system, from design to deployment, and its evolution over time.

Within Spyderisk, risk is represented by the occurrence of compromised desirable properties. This compromise is, in turn, represented by consequences that have a likelihood and result from threat mechanisms. As both the frameworks and Spyderisk focus on these compromises, Spyderisk can be used in risk assessments following both frameworks. This is unsurprising as they all follow accepted conventions regarding the fundamentals of risk assessment. Additionally, as Spyderisk provides and recommends controls to address threats and risks and provides an approachable way for non-domain experts to assess the system and related risks, in this context, it can again be used for risk assessments within these frameworks and adds to the accountability and transparency.

Spyderisk requires the user to explicitly model the cyber-physical system and will require the explicit modelling of the AI-system components, and therefore the full cyber-physical socio-technical context of the AI system and associated ICT system can be included within the Spyderisk system model. As such, the importance of considering the full ICT context the AI system is in and the importance of considering the AI system as a socio-technical system, as highlighted by both frameworks, are inherently met within Spyderisk.

Spyderisk can be used throughout the lifecycle of an AI system. At the design stage of a system, it can be used to assess the risks and pre-emptively mitigate them and can also assess predicted risks at a point in the system operation. Further, it can assess the risks live, during the lifetime operation of the system, by using external sensors to adjust metrics related to vulnerability as the system operation evolves, and as a result different consequences become more likely and thus the resulting risks may change.

Finally, this leads to the optimisation of desirable properties, which ENISA highlights results in trade-offs between different ones as the optimisation of one can negatively affect another. Through Spyderisk, different optimisation scenarios can be tried, and risk assessments run, so that these optimisations and trade-offs are known and planned for, before any changes are made to the real-life system. This means that an informed decision can be made and the most appropriate set of optimisations for a system and context can be chosen.

2.5.4. Spyderisk knowledge extensions for Trustworthy AI assessment and optimization

This section describes the initial work towards domain modelling of AI-related assets, threats, harms, consequences etc for Spyderisk. This work will be updated as the project progresses. This first pass here aims to contain the essentials needed to model an important subset, which is AI-based harms due to issues with the integrity of training data or a malicious user of an AI model. The process used here follows that described above and will be adapted as necessary in further additions.

Knowledge was elicited from sources that included: *A Taxonomy of Trustworthiness for Artificial Intelligence* (Newman, 2023), *Ethics guidelines for trustworthy AI* (HLEG, 2019), The ENISA *Multilayer Framework for Good Cybersecurity Practices for AI* (ENISA, 2023), NIST *Artificial Intelligence Risk Management Framework* (NIST, 2023), *Securing Machine Learning Algorithms* (ENISA, 2021) and *AI Cybersecurity Challenges - Threat Landscape for Artificial Intelligence* (ENISA, 2020). Using these, an initial set of key concepts were determined that are needed to model an AI system over its lifecycle.

Examining the material mentioned, the core assets needed to model a simple AI system are as shown in ***Table 3***.

Table 3. AI Assets Required

Assets Required	Superclass from SSM Domain
AI Model	Software Process
AI Model Parameters/Weights	Data
AI Model Input Data	Data
AI Model Predictions	Data
AI Training Algorithm	Software Process
Hyperparameters	Data
Training Data	Data
Testing Data	Data
Validation Data	Data

In this initial work, we assume the “AI Model” and “AI Training Algorithm” assets are subclasses of the “Process / General Process” asset that already exists in the Network Domain Model concerned with cybersecurity. The Process asset type is a software process that runs on hardware, and optionally consumes data as input and produces data as output. That this asset type already exists means there are already some threats applicable to it in the Network Domain Model and provides a point of integration between the domains of cybersecurity and AI.

Other assets required are all data, and in this initial work, they are modelled as subclasses of the existing “Data” asset in the Network Domain Model. This represents a dataset that can be related to a software process, stored within databases / filesystems on ICT hardware and transmitted over computer networks. Data also has relationships with humans, e.g. humans interact with data via software processes or for example personal data relates to a person as a data subject. As with the software process, the existing properties of the Data asset type in the Network Domain Model can be inherited from the specific data types required for machine learning.

New relationships linking the AI process and data assets and their functionality together are proposed, as shown in Figure 8: Domain representation of AI system assets and relationships ***Figure 8***. These new relationships will enable a user of the SSM tool to construct models of an AI SUT, that encompasses different phases of the lifecycle, e.g. training, testing and use within the operating environment.

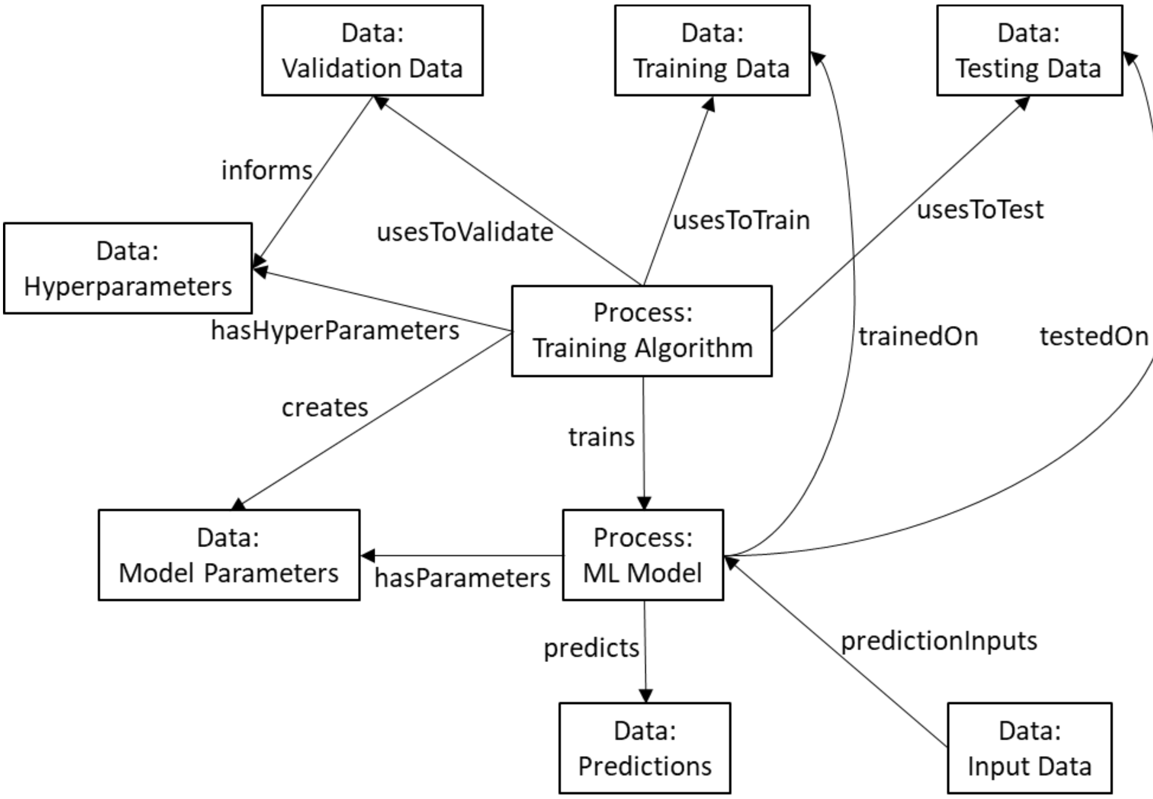

Figure 8: Domain representation of AI system assets and relationships

This diagram represents a high-level schematic of an AI system that encompasses both the training and deployment phases of the system. The top half represents the training iterations, where training data is used to determine the model parameters, testing data is used to test the model accuracy, hyperparameters are used to define the training process, and validation data are used to guide the hyperparameter values. The bottom half represents the usage phase, where input data is received by the AI model, which, in conjunction with the model parameters, results in predictions being made by the AI Model based on this input. These relationships enable different system models to be constructed representing model creation via training, model operation, or the two phases combined, modelling a continuous training/re-training + usage process and enabling the modelling of issues in one phase affecting the other.

SSM domain models contain threats, and the presence of a threat in a system model is determined by a *matching pattern*, which is part of the threat's specification inside the domain model. The matching pattern consists of assets and relations that need to be present in the user's system model to enable the threat to be considered present in that system. An example of a matching pattern is shown in Figure 9. Here, the pattern is denoted by the Asset types (a process and data) and the relation between them ("usesToTrain"). The new asset types and relations in Figure 8 determine the prototype elements from which the matching patterns are constructed.

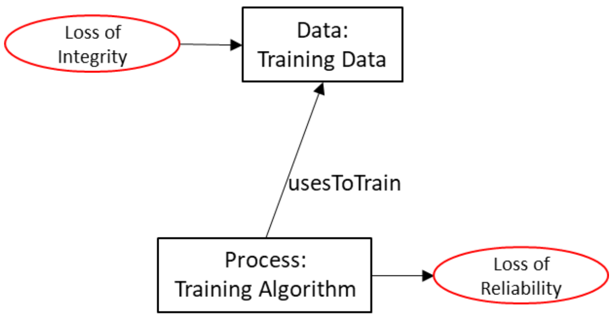

Figure 9: Example Threat Matching Pattern

In our initial work on Spyderisk knowledge extensions for THEMIS 5.0, we have focused on the accuracy and fairness of an AI system. We associate the model accuracy and fairness with the prediction data as that is the output of the model. This output may directly or indirectly affect other assets of all kinds in the system (e.g. people, institutions, downstream processes). Hence, if these predictions are inaccurate or unfair, there is likely to be a propagated consequence on connected downstream entities.

We have used threats to model how the training data becoming incorrect (either unintentionally or maliciously) affects the accuracy and fairness of the model predictions, and this has required consideration of the path all the way from the model training to the model use. We have further modelled how a user of the AI model maliciously altering the input affects the model predictions. For this, we need to only consider the model usage phase.

This modelling has required addition of desirable Properties to assets in the domain model. The relevant of relevant DPs is shown in ***Table 4***.

Table 4: Relevant Existing and New Desirable Properties for an AI system

Desirable Property	*Asset*	*Description*	*New?*
Integrity	Data	The data is correct and fit for purpose	N
Authenticity	Data	The data is what it claims to be, i.e. it is neither forged nor altered in a way designed to induce false behaviour in other assets consuming the data.	N
Fairness	**Data**	**Equality or equity is represented within the data, with issues such as harmful bias or discrimination not present**	**Y**
Reliability	Process	The asset will perform tasks correctly, with no functional errors, assuming the asset is not supplied with corrupt or inaccurate information as input	N
Fair processing	**Process**	**Inputs are processed such that the outputs resulting from them are not unduly detrimental, do not lack equality or equity, and do not contain issues such as harmful bias or discrimination.**	**Y**
Accuracy	**Data**	**Correct when compared with reality**	**Y**

Equivalent Consequences have also been modelled. These are reflected as losses (or degradations) of DPs and are shown in ***Table 5***.

Table 5: Relevant Existing and New Consequences for an AI system

Consequence	*Asset*	*Description*	*New?*
Loss of integrity	Data	Alteration or corruption of data such that its use will produce incorrect outputs or outcomes. It is not specified whether this is malicious or not (compare with Authenticity).	N
Loss of authenticity	Data	Deliberate Integrity Loss: forging or alteration of data in a way designed to induce false behaviour in other assets consuming the data.	N
Loss of fairness	**Data**	**Equality or equity is not represented in the data, with issues such as harmful bias or discrimination present.**	**Y**
Loss of reliability	Process	The device, process or human is liable to make errors with an unacceptable frequency or extent. Caused by internal failings including lack of expertise, software bugs, etc., by using forged, corrupt or inaccurate information as input, or by dependency on some other asset that is not reliable	N
Loss of fair processing	**Process**	**Inputs are processed such that the outputs are unduly detrimental, lack equality or equity, with issues such as harmful bias or discrimination.**	**Y**
Loss of accuracy	**Data**	**Data is incorrect when compared with reality**	**Y**

We use threats to model how one consequence can cause another consequence, thus enabling the propagation of consequences throughout the SUT. The modelling process uses a state mapping diagram as illustrated in Figure 10 and Figure 11. Here, the red ovals are consequences, the black rectangles are threats, and the green ovals are controls. Figure 10 covers the training phase of an AI system and Figure 11 covers the deployment/usage phase. These state mapping diagrams show how one consequence leads to another via a threat and so each threat can then be defined in terms of its causes, consequences, and controls. The chain of cause and effect from a threat to a Consequence to a downstream threat and the Consequence it causes, is named a Threat Path.

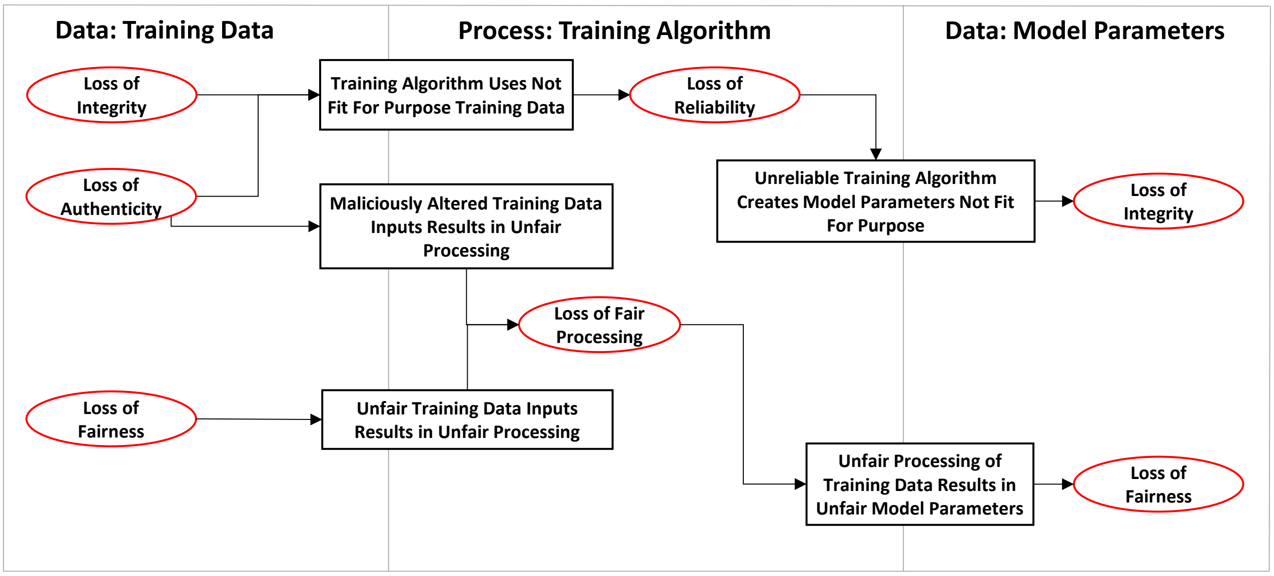

Figure 10: AI system training phase state mapping diagram

Figure 10 illustrates multiple Threat Paths that relate losses of integrity, authenticity and fairness on training data to the losses of reliability and fair processing in the training algorithm, and then propagated to integrity and fairness losses in the resulting model parameters.

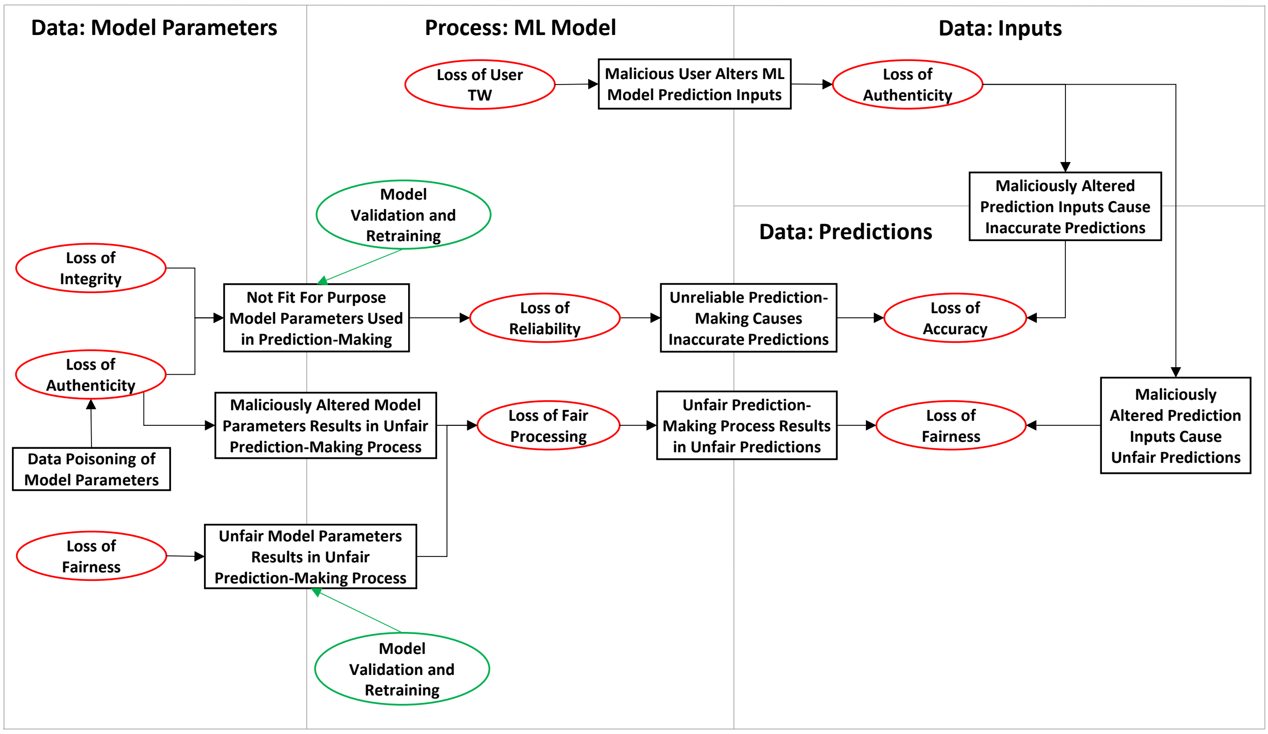

Figure 11: AI system usage phase state mapping diagram

Figure 11 illustrates Threat Paths that cover relationships between the data model parameters generated for the AI training and their effects on the usage of the AI model, as well as during usage where untrustworthy users (expressed by "Loss of UserTW") deliberately attempt to corrupt model inputs.

For each Threat, a specification has been created, which consists of a matching pattern, optional causing Consequences and resulting Consequences. For an example specification showing how loss of integrity and reliability may impact accuracy, see Annex 3.3.

To further demonstrate the applicability of Spyderisk to the purpose of THEMIS 5.0, a simple Spyderisk system model has been created that contains an AI training algorithm, AI model, training and testing data, and a server that hosts everything. The server has an application process on it that creates the training data.

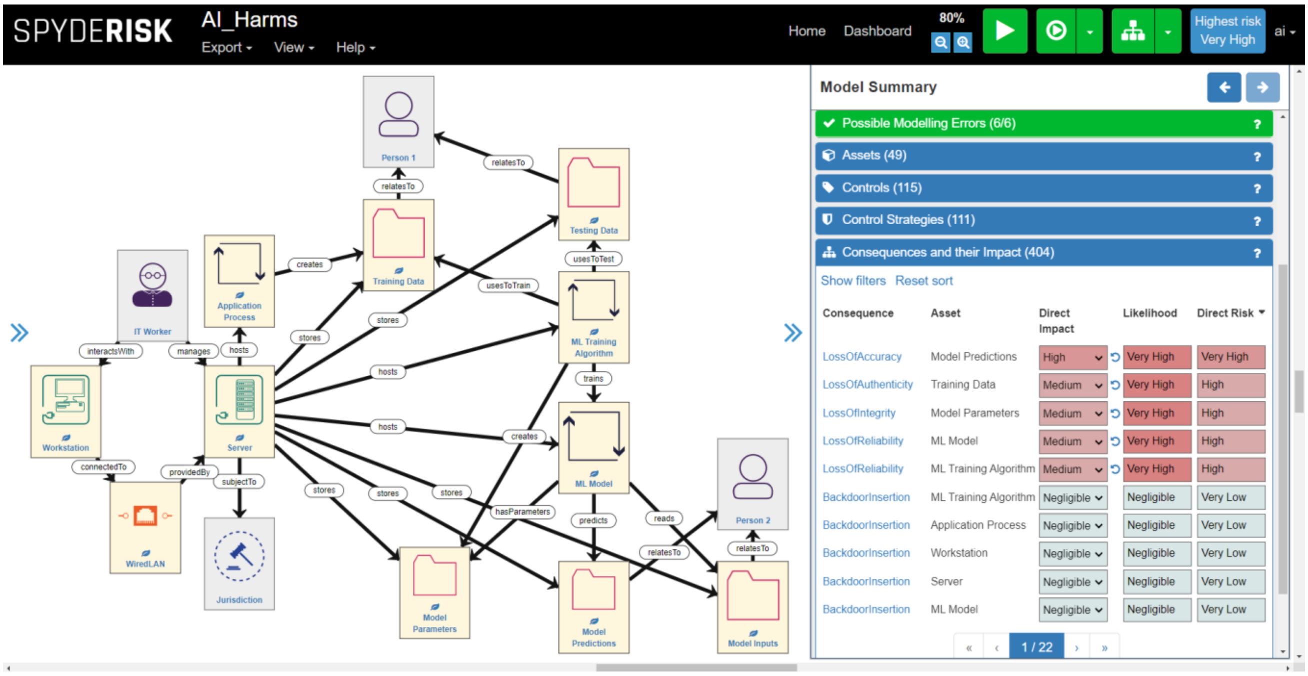

Figure 12: Spyderisk system model showing encoded AI harms.

Here we model a malicious user gaining access to the application process that creates the training data and altering the training data. This is modelled by reducing the UserTW desirable property on the application process, meaning that the trustworthiness of the users of the process creating the training data is suspect. The effect of this is that the likelihood of the consequence Loss of Authenticity of the training data becomes very high, as seen in the figure. This then propagates through to the AI Training Algorithm, with it losing reliability, and this then causes the Model Parameters to lose integrity. Due to this, the AI Model loses reliability, and this finally causes a loss of accuracy in the Model Predictions.

2.5.5. Conclusions

The THEMIS 5.0 approach to risk assessment is to base its work on the existing Spyderisk simulator from UoS and extend its knowledge base towards AI trustworthiness, threats, risks, asset types and controls to manage risks.

Key source material has been examined and compared to Spyderisk with the conclusion that they are compatible. The source material has also suggested aspects for knowledge extension to Spyderisk, and initial steps have been taken to model AI threats & risks and to integrate this new knowledge with existing knowledge in the realm of ICT hardware, software, data, networks, human interaction, physical spaces, cybersecurity and human rights such as privacy. To illustrate the extensions, we have described a simple system model in the Spyderisk UI.

The knowledge extension is an ongoing task and additional sources will be examined such as ISO/IEC 23894:2023, as well as relevant regulation such as the EU AI Act.

2.6. AI Technologies for Trustworthy AI

In this section, we survey suitable AI/ML technology and resources for achieving trustworthiness during deployment. Relevant methodologies will be those which help detect and verify trustworthiness characteristics, not by evaluating the model performance in hindsight, but to assess the trustworthiness of a prediction as it comes. We will also outline certain training procedures to make an AI model robust to certain data perturbations or data distributions in general.

2.6.1. Accuracy

When addressing technologies for trustworthy AI, we will first consider approaches for assessing the accuracy of particular AI-based predictions. Then, as a potentially innovative approach, we will thereafter discuss how an AI model could be applied for the assessment of accuracy through model comparisons.

2.6.1.1 The accuracy of particular AI-based predictions

While we in Section 3.3.4, addressed approaches for assessing accuracy for a given dataset, a risk assessment may benefit from considering the accuracy of a **specific prediction** coming from an AI model.

Below are methods to assess and evaluate the accuracy of a prediction coming from an AI model to make the process more trustworthy.

2.6.1.1.1 Methods for uncertainty quantification

Uncertainty in a prediction is often described using *prediction intervals*, where the true value is presented as residing within some interval with a given probability, often 95 %. This means that the prediction itself is more uncertain, the wider the interval is.

There are several ways to quantify uncertainty. Bayesian modelling, bagging, bootstrapping or Monte Carlo drop-out are all methods with this purpose, by estimating prediction intervals. See Abdar et al. (2021) for a review of such uncertainty estimation methods.

Within classification problems, the prediction interval will be in terms of the range of the class prediction probability. Even though prediction intervals are useful for quantifying uncertainty, it is not straightforward to use it in practice, and perhaps even so for classification tasks.

2.6.1.1.2 Misclassification detection

Hendrycks and Gimpel (2016) consider ML models for multiclassification problems with softmax output layers. The authors describe a pattern in which the maximum softmax probability tends to be smaller for misclassified samples than for correctly classified samples. In other words, the uncertainty in the prediction can be based on the magnitude of the maximum softmax probability. Moreover, a simple misclassification detection procedure is to flag predictions as incorrect whenever the maximum softmax probability is smaller than some threshold T.

In Granese et al. (2021), they find that the probability of misclassification for any sample is x, $Pe(x)$, is lower bounded by $1 - \hat{g}(x)$ given by:

$$Pe(x) \leq 1 - \hat{g}(x) = 1 - \sum_{y \in Y} P^2_{\hat{Y}|X}(y|x).$$

A misclassification detector, $D_\alpha(x, \gamma)$ is further introduced given by

$$D_\alpha(x, \gamma) = 1[1 - \hat{g}(x) > \gamma \hat{g}(x)]$$

for some parameter γ in which $D_\alpha(x, \gamma) = 1$ means the sample is flagged as incorrect. The authors show that the detector is optimal, given γ, in terms of minimizing the Type I + Type II error. The parameter γ can be set based on what is the optimal value from historical predictions.

Notice that the two methods mentioned above only need the softmax output values of the model.

In Johnsen and Remonato (2024), another misclassification detection procedure is proposed by looking not only at the output layer, but also at the hidden layers. In particular, they assume that the random vector of pre-activations for correctly classified samples at each layer follows a Gaussian distribution. The Mahalanobis distance is used to construct a hypothesis test, and reject a sample as correct for a p-value less than some significant level. The authors further suggest performing hypothesis tests at each layer, yielding a p-value per layer, and finally combine the p-values into one single p-value such as acquiring information from each layer with the intention to give a more robust rejection procedure than when only looking at one layer.

2.6.1.1.3 Accuracy estimates for individual predictions as input to risk assessment

By quantifying the uncertainty of a prediction, one also opens the door to taking risk into account. That is, if an action is based on a particular AI prediction, the action to be taken may be based on the uncertainty of the prediction and moreover, the uncertainty may be used as input to a risk assessment procedure.

2.6.1.2 AI-based assessment of accuracy through model comparisons - reflections on a possible innovative approach

While the most common ways to measure and assess accuracy are the mathematical approaches outlined in Section 2.3.4, potentially with those discussed above in Section 2.6.1.1, there is a possibility of using an ML-based approach to that end as well. As part of the groundwork, we have discussed a possible innovative approach towards this end and include it here to conclude our treatment of accuracy.

Let us assume that we have a ML model, hereinafter Model 1, that is trained to perform a particular task, either classification or regression. As is standard practice in ML, we can assume that Model 1 has been trained on the train dataset, and there are validation and test dataset available. Let us call x the inputs to the model (the features), y the correct labels. The labels can be either discrete classes (classification) or target values (regression). Further, let us call $\hat{y}_1$ the labels predicted by Model 1, i.e. the output of Model 1, on a suitable set like the test set.

We can then train a second model, call it hereinafter Model 2, which will act as an assessor for Model 1. The way we envision this working is the following: Model 2 will be trained to predict the true labels y on the same training set as Model 1 but using both x and $\hat{y}_1$ as its *input* features. That is, we extend the original input x with the output of Model 1 and use this augmented input to train Model 2. After training we can compare the output of Model 2 on the test set, call it $\hat{y}_2$, with the output of Model 1 on the same set.

We then have three possible cases:

1) The performance (accuracy) of Model 2 is significantly worse than that of Model 1.
2) The performance of Model 2 is comparable to that of Model 1.
3) The performance of Model 2 is significantly better than that of Model 1.

In the first case, if the performance of Model 2 is worse than that of Model 1, it means adding the predicted labels from Model 1 to the input has not helped increase accuracy. This may mean that the output of Model 1 may not contain useful information, or may be adding more noise than information, therefore damaging the performance of Model 2. Consequently, this may indicate that the output of Model 1 is not accurate.

In the second case, since Model 2 has not been able to improve on the performance of Model 1 despite the added information contained in the labels, we could infer that Model 1 exploited all the available information in the original input x. Therefore, this may indicate that the output of Model 1 may be accurate, or at least as accurate as the input allows it to be. Note, however, that this reasoning is valid only in the case that both models exhibit acceptable performance.

In the third case, a substantial improvement in the performance of Model 2 in respect to Model 1 means that the output of Model 1 contained information that is useful for the task at hand, which is a positive factor in the assessment of Model 1. At the same time, since the performance of Model 2 is significantly better, it shows that there is room for improvement in respect to Model 1. This could mean for instance that Model 1 may benefit from additional training, or that the input x contains some noise which the addition of the predicted labels helped mitigate by improving the signal-to-noise ratio. Either way, since Model 2 has better performance, it is now possible to include it in the pipeline, chaining it after Model 1, effectively constructing a new model composed of Model 1 and Model 2 together. This new model will have in this case significantly better performance than the original Model 1 alone. In this respect, the procedure here outlined can be used to not only assess, but also improve the performance (accuracy) of a ML model.

2.6.2. Robustness

Assessing the robustness of a single prediction from an AI model may be important for risk assessment, just as it may be so for single-prediction accuracy assessments discussed in 2.6.1. Previously, in Section 3.3.5, we addressed approaches for assessing robustness with regard to a given dataset. Below are methods to assess and evaluate the robustness of a **specific prediction** from a machine learning model, or how to be trained to be robust to particular data distributions.

2.6.2.1 Adversarial examples and training

2.6.2.1.1 Adversarial examples

One procedure to assess the robustness of a machine learning method within a classification is how the AI model performs on so-called **adversarial examples**. These are samples where small, but efficient perturbations are applied to the original sample such that the AI model misclassifies, even though the original sample is correctly classified. A good example of such an adversarial sample is the one described in Goodfellow et al. (2015) for a classification model (called GoogLeNet) trained on the ImageNet dataset. An image of a panda with softmax confidence of 58 % is perturbed with noise, consequently leading the model to misclassify the image as a gibbon.

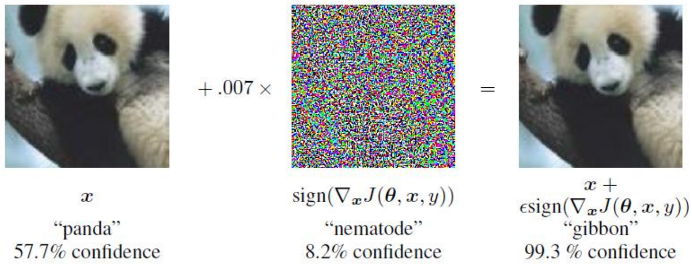

Figure 13: From Goodfellow et al, page 3 showing an adversarial example in which perturbations lead to misclassification.

From Figure 2, the added noise does not fundamentally change the important aspect of the sample picture, namely that we can still see an image of a panda. Hence, for a good classification model, it should be able to distinguish the signal from the noise. As this was not the case here, one can question whether the model is considering the right features in the image sufficient to observe a panda.

There are many ways to perturb the sample. One efficient method, the Fast gradient sign method (FGSM), is given in Goodfellow et al by adding the term $\boldsymbol{\eta}$ to the original input $\boldsymbol{x}$ which creates the perturbed sample $\boldsymbol{x}'$:

$$\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\eta} = \boldsymbol{x} + \epsilon sign\left(\nabla_{\mathrm{x}} J(\theta, \boldsymbol{x}, y)\right),$$

where ϵ is a step size, $\nabla_{\mathrm{x}} J(\theta, \boldsymbol{x}, y)$ is the gradient of the model with respect to the input sample $\boldsymbol{x}$, θ is the parameter vector in the model and sign() indicates the sign-function. The authors motivate the reason for this being efficient by looking at the corresponding change in the output of the activation function for a particular node in the layer after the input layer. As there is a linear operation between the weights and the sample, and in the case where the samples are high-dimensional such as for images, one can show that even though only a tiny perturbation is applied to each variable (e.g. pixel) of the sample, the perturbations add up to make a significant change in the activation function.

The reason for applying the sign() to the gradient, and not only the gradient, is according to the authors, to make adversarial examples that are difficult for the model to learn (as the sign() function is not differentiable).

2.6.2.1.2 Mitigating adversarial examples through adversarial training

Adversarial examples can potentially be addressed through **adversarial training**. Goodfellow et al. (2015) proposed a procedure in which the training of a machine learning model includes **adversarial training**, with the purpose of being robust to adversarial examples. They define the loss function

$$\tilde{J}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \alpha J(\theta, \boldsymbol{x}, y) + (1-\alpha) J\left(\theta, x + \epsilon\, sign\left(\nabla_x J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right)\right).$$

The parameter α weights the importance of regular training (first term) compared to the adversarial training (second term). The authors used $\alpha = 0.5$ in all experiments. The authors observed a reduced test error from 0.94 % for the original model to 0.84 % for the adversarially trained model. The authors observed that the performance for the original samples in the validation data stagnated during training, while the corresponding loss on the adversarial validation data did not. For that reason, the authors trained the model with the early stopping criterion on the adversarial validation data. Additionally, by making the hidden layers a bit wider in the neural network, the authors were able to reduce the test error even further to an average of 0.782 % (based on five runs). Adversarial training was explored even further, as they aimed to identify the reason why the test classification accuracy is worse than the *robust accuracy* which includes adversarial perturbations. They show that there exists an algorithm that can solve in most settings the unknown parameters θ^*, achieving robust accuracy including ℓ_∞-perturbations (such as the one proposed by Goodfellow)

The FGSM method is in fact applied in (Maham et al., 2024) for fake news detection. The authors show that the classification accuracy in fake news detection improves (with around 2 %) when including adversarial training for the experiments and data sets in their work.

2.6.2.2 Out-of-distribution examples and training

2.6.2.2.1 Out of distribution examples

The robustness measures discussed above, make only sense when the perturbations applied to the data samples are not changing the true nature (class) of the sample, such as perturbations that change the class of an image in a classification problem. Hence, equally important it is to ask the question: In which scenarios should we not expect the AI model to be robust? In real life, the data samples can be so different from the training data used for constructing the AI model that they cannot be considered relevant or correctly understood by the model. For example, an AI model to detect different fish species in Norway should only be expected to predict with acceptable accuracy images of fish species in Norway. The model is irrelevant and should perform unexpectedly when predicting fish species that are not present in Norway. Ideally, one wants to ensure that the deployment of the model is within the same environment as when it was trained. This is, however, not always possible. A non-native species may enter the Norwegian sea. The non-native species is an example of an *out-of-distribution* (OOD) sample. Procedures to automatically detect the presence of OOD samples in an ML deployment setting are called *out-of-distribution detection*.

In this case, we cannot expect the model to be robust in terms of making accurate predictions. Instead, we can make procedures that can accurately detect out-of-distribution samples before they are deployed by the model itself. From this perspective, robustness concerns accurately detecting out-of-distribution samples together with the handling of those samples under deployment, such as flagging a prediction as untrustworthy if the corresponding sample is deemed to be out of distribution.

2.6.2.2.2 Out-of-distribution detection

Importantly, we cannot make a ML model itself robust to out-of-distribution examples. Rather we need to have a separate procedure that can robustly detect out-of-distribution samples once they occur, and act accordingly.

There are several proposed methods to flag out-of-distribution samples.

In (Liang et al., 2020) a procedure called ODIN is proposed for neural network models. The procedure takes advantage of two observations already discussed, namely the difference in the probability distribution of maximum softmax probabilities for correct vs incorrect predictions, as discussed in Hendrycks and Gimpel (2018), as well as the effect of *adversarial samples,* as discussed in Goodfellow et al. (2015). In particular, the two-step procedure is as following for a multiclassification problem with C classes:

1. Given the unnormalized feature vector $f(\boldsymbol{x}) \in R^C$ prior to the softmax computation at the output layer. Compute the softmax probability $S_i(\boldsymbol{x};T)$ for each class i using *temperature scaling parameter* $T \in (0,\infty]$*:*

$$S_i(\boldsymbol{x};T) = \frac{\exp\left(\frac{f_i(\boldsymbol{x})}{T}\right)}{\sum_{j=1}^{C}\exp\left(\frac{f_j(\boldsymbol{x})}{T}\right)}$$

2. Perform *Fast* gradient sign method (FGSM) perturbation of sample **x** into perturbed sample $\widetilde{\boldsymbol{x}}$, however in **the opposite direction** ($\eta = -\epsilon sign\left(-\nabla_{\mathrm{x}} \log S_{\hat{y}}(\boldsymbol{x};T)\right)$ which will force the maximum softmax probability, $S_{\hat{y}}(\boldsymbol{x};T)$, to become larger (the model gets more convincing of its prediction). Compute the resulting $S\,(\widetilde{\boldsymbol{x}};T)$.

The OOD-detector is then flagging a sample **x** as OOD whenever $\max_i S\,(\widetilde{\boldsymbol{x}};T) \leq \delta$.

The authors show, via Taylor expansion, that the intuition behind 1. is that the value of $S_i(\boldsymbol{x};T)$ for sufficiently large T is largely dependent on the expression $U_1 = \frac{1}{N-1}\sum\left(f_{\hat{y}}(x) - f_i(x)\right)$ in which larger U_1 will give a larger $S_i(\boldsymbol{x};T)$. As we have from Hendrycks and Gimpel, in-distribution samples tend to provide larger maximum softmax probabilities $f_{\hat{y}}(x)$ which again yields larger U_1. Hence temperature scaling helps separate in-distribution and out-distribution samples with respect to the probability distribution $\max_i S\,(\boldsymbol{x};T)$.

The intuition behind 2. has the same purpose. In short, the FGSM perturbation was observed to give larger positive jumps in the maximum softmax probability for in-distribution samples than OOD samples. Hence, it further separates the two probability distributions, but with respect to the perturbed samples $\max_i S\,(\tilde{\boldsymbol{x}};T)$.

Lee et al. (2018) presented a procedure with particular focus on neural network architectures. Regardless of the use case (vision, NLP, tabular), most often, the *penultimate layer* placed before the softmax output layer of the neural network is a flattened layer (a vector of a predefined dimension). The authors show that the probability density distribution of the penultimate vector for the *in-distribution* samples of each known class c can be estimated as a multivariate Gaussian distribution. Assuming equal covariance, Σ, for all clusters. The sample mean, $\hat{\mu}_c$, is computed for each class c together with the sample covariance, $\hat{\Sigma}$. The authors moreover introduce the Mahalanobis score to indicate the confidence to which a sample is in-distribution given by:

$$M(x) = \max_c -(f(x) - \hat{\mu}_c)^T \widehat{\Sigma^{-1}}(f(x) - \hat{\mu}_c)$$

$M(x)$ quantifies the distance from the input sample x to the closest multivariate Gaussian cluster corresponding to a class c with observed penultimate vector $f(x)$. The intention is that the further absolute value distance from this cluster, the more it indicates the sample x is out-of-distribution. The authors generalize for several layers, assuming a Gaussian distribution in each layer.

The authors show that the method is efficient based on several experiments, and when comparing it with other out-of-distribution detection methods.

2.6.2.3 Counterfactual explanations

As another method of evaluating the robustness, we can use a concept first introduced in the context of *explainability*, namely *counterfactual explanations.* The purpose of counterfactual explanations is to inspect the minimal change in input that is required to change the classification of the model. The concept is introduced by (Wachter et al., 2018) where a method for finding counterfactuals for each prediction is defined as a mathematical optimization problem. Namely, given input x and output $\hat{f}(x)$ from model, we seek for the perturbation x' with desired output $y' = \hat{f}(x')$ that minimizes the loss function:

$$L(x, x', y', \lambda) = \lambda(f(x') - y')^2 + d(x, x'),$$

Where $d(x, x')$ quantifies the input difference between x and x', which we want to be as small as possible.

Using counterfactual explanations both for explainability purposes, but also for robustness evaluation is a user-friendly concept for persons without expertise in AI. This is because they can use their domain knowledge to decide whether the change in prediction from $\hat{f}(x)$ to y' can be explained reasonably as a consequence of perturbing x to x'. A drawback of counterfactual explanations, for explainability purposes, is that there can be numerous examples of perturbations giving the same prediction y'. However, for robustness evaluation, this is, as a matter-of-fact, valuable information for the degree of the model's robustness. One may define the stability of a prediction as how many different perturbations of the input that yield the same change in output to y'.

2.6.3. Fairness

Building on our treatment of fairness in Section 2.3.6, we here provide an overview of fairness definitions and associated metrics. In THEMIS 5.0 such general metrics, along with case-specific applicable metrics, will be combined to support AI trustworthiness assessment and optimization.

2.6.3.1 Definitions

According to Kaur et al. (2022), "fairness of a system ensures that there is an absence of any discrimination or favouritism toward an individual or a group based on any inherent or acquired characteristics that are irrelevant in the context of decision making." This definition distinguishes between group and individual fairness.

According to EC HLEG on AI (AI HLEG, 2019), fairness is both substantive and procedural. Substantive fairness refers to equality before the law, non-discrimination and avoidance of biases, equal and just distribution of opportunities and costs, and proportionality between means and ends. Procedural fairness refers to redress against AI-enabled or assisted decisions, accountability of the system and its human operators, explicability of the decision-making process followed by the AI system, and participatory design of the AI system.

Both substantive and procedural fairness entails an individual dimension, e.g., equality before the law and non-discrimination, redress against decisions, and a societal/group dimension, e.g., equal and just distribution of opportunities and costs, participatory design of the AI system. According to the Assessment List for Trustworthy Artificial Intelligence (AI HLEG, 2020), fairness refers to a variety of ideas known as equity, impartiality, egalitarianism, non-discrimination and justice. Fairness embodies an ideal of equal treatment between individuals or between groups of individuals. This is what is generally referred to as 'substantive' fairness.

But fairness also encompasses a procedural perspective, that is the ability to seek and obtain relief when individual rights and freedoms are violated. Fair AI denotes AI that shows no discrimination toward people from any group. Its output should have little correlation with the traits of individuals, such as gender and ethnicity. The fairness of an algorithm is defined as "the absence of any prejudice or favouritism toward an individual or a group based on their intrinsic or acquired traits in the context of decision making" (Liu et al., 2022).

Fairness is closely related to the rights to Non-discrimination, Solidarity and Justice (Díaz-Rodríguez et al., 2023). Although there are many different interpretations of fairness, the European Commission advocates for having both: (a) a substantive dimension of fairness that ''commits to ensure equal and just distribution of benefits and costs, commits to free from unfair bias, discrimination and stigmatization, implies respecting the principle of proportionality between means and ends and a careful balancing of competing interests and objectives''. (b) a procedural dimension allowing to ''contest and seek redress against decisions taken by AI systems or who operates them''. To achieve this, the entity responsible for the decision must be identifiable, while decision making processes should be explainable.

As stated from the above definitions, the two main types of fairness are the procedural and the distributive (outcome).

Procedural fairness is a concept inherited from administrative law concerned with equality of treatment within the process that carries out a decision. It relates to the information that must be considered in decision-making. In short, it is concerned with the fairness of the process itself, while also emphasizing the transparency and consistency of the procedure.

Substantive (Distributive Outcome) Fairness is the term used to define equality ('fair result') of the outcomes themselves, regardless of the decision-making process. The principle here is to measure the outcomes of models and aim for equitable distribution of benefits and burdens among individuals or groups, ensuring that outcomes are just and not discriminatory.

2.6.3.2 Fairness metrics

Regarding fairness metrics, their categorization can be divided based on the methodological approach and the criteria levels. For the former, the metrics can be split between observational (statistical) and causality-based metrics, while the latter in group (statistical) or individual (similarity-based) fairness (Castelnovo et al., 2022).

Observational metrics assess fairness by examining patterns and statistical relationships in the observed data. This kind of metric is more widely used in practice due to its generalizability and focuses on the associations between variables such as predictions and protected attributes. In contrast, causality-based metrics incorporate reasoning to assess fairness, taking into account the causal relationships between variables, including protected attributes. They require causal models (e.g. causal graphs, structural equations), address issues of indirect discrimination, and can distinguish between fair and unfair sources of disparity.

Group fairness metrics make sure all the groups of society are treated equally without any discrimination by the AI system. In other words, it focuses on ensuring fairness across relevant groups. It aims to prevent disparate impact on protected groups by ensuing statistical parity or equality of error rates. Moreover, group fairness observational metrics can be further categorized through the criteria of independence, separation, and sufficiency (Barocas et al., 2023). On the other hand, Individual Fairness metrics ensure that similar types of individuals get similar predictions. It focuses on ensuring fairness at the level of the individual. It aims to ensure that individuals with similar relevant qualifications receive similar system outcomes.

Even though the literature lacks a clear association between the fairness types and their metric categories, their definitions suggest that observational metrics apply only to distributive fairness, while causality metrics could be relevant in both types. The remainder of this section presents some of the most common metrics categorized into group fairness (statistical measurements), individual fairness (similarity measurements), and causal-based measurements.

2.6.3.2.1 Group fairness metrics

Demographic Parity/Statistical Parity Difference: Demographic parity aims to ensure equal representation of different demographic groups in the outcomes of a model. Specifically, it requires that the proportion of positive outcomes e.g., being approved for a loan, or receiving a job offer, is the same across different demographic groups, regardless of their membership in those groups. In simpler terms, it means that the distribution of favourable outcomes should be consistent across different demographic categories, such as race, gender, or age (Zemel et al., 2013). The statistical parity of a hypothesis h is:

$$SP(h) \coloneqq p(\hat{Y} = 1 | S = 1) - p(\hat{Y} = 1 | S = 0)$$

Where Y corresponds to a decision variable, S to some common sensitive variable, and the ideal value of the metric is 0 (McNamara et al., 2019).

Equalized odds/Equality of odds: An outcome is fair if false positive and true positive rates are equal across groups. In other words, both the probability of incorrect positive predictions and the probability of correct positive predictions should be the same across protected and privileged groups. Given an output variable Y, an input variable X, and a protected variable Z, a predictor $\hat{Y}$ satisfies equality of odds if $\hat{Y}$and Z are conditionally independent given Y . This means that, for all possible values of the true label Y , $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable:

$$P(\hat{Y} = \hat{y} | Y = y) = P(\hat{Y} = \hat{y} | Z = Z, Y = y)$$ (Zhang et al., 2018).

Equal Opportunity Difference: The difference of true positive rates (or sensitivity/recall) between the unprivileged and privileged groups (AI fairness 360). A smaller equal opportunity difference indicates less disparity in predictive performance between groups, suggesting a fairer model. The ideal value of the metric is 0.

Average Odds Difference: The average difference of false positive rate (fallout) and true positive rate (sensitivity/recall) between unprivileged and privileged groups (AI fairness 360). Essentially, it evaluates whether there is a disparity in both the model's ability to correctly identify positive outcomes (true positive rate) and its tendency to incorrectly classify negative outcomes as positive (false positive rate) across different demographic categories. A smaller average odds difference indicates less disparity in predictive performance between groups, suggesting a fairer model. The ideal value of the metric is 0.

Disparate Impact: The ratio of favourable outcome for the unprivileged group to that of the privileged group (AI fairness 360). It examines whether AI systems disproportionately harm certain groups, irrespective of intent. It aims to uncover biases in outcomes, intentional or not (Chen et al., 2023). Formally, it is often calculated using a statistical measure such as the disparate impact ratio, which compares the proportion of positive outcomes for the privileged group to that of the unprivileged group. A value significantly different from 1 indicates disparate impact, with higher values indicating greater disparity. The disparate impact of a hypothesis h is:

$$DI(h) = \frac{p(\hat{Y} = 1 | S = 0)}{p(\hat{Y} = 1 | S = 1)}$$

Where Y corresponds to a decision variable, S, to some common sensitive variable, and ideal value of the metric is 1 (McNamara et al., 2019).

2.6.3.2.2 Individual fairness metrics

Fairness through unawareness: The corresponding principle here is that model Y achieves fairness through unawareness if the protected attribute A (i.e. sex) is not utilized to make predictions (Feldman et al., 2021). In other words, it is achieved by ignoring sensitive attributes, such as race, gender, or age, in the training and decision-making of a machine learning model. Despite its intentions, it might indirectly perpetuate bias present in data. (Chen et al., 2023)

Causal-based metrics

Counterfactual Fairness: It evaluates whether a decision or outcome would still be considered fair if certain aspects of the situation were different. According to counterfactual fairness, an outcome is fair if an automated decision made about an individual belonging to a sensitive group would have been the same were that individual was a member of a different group in a closest possible alternative (or counterfactual) world. For example, a predictor Y is counterfactually fair if for any attributes X = x and sex A = a:

$$P(\hat{Y}_{A\leftarrow a} = y | X = x, A = a) = (\hat{Y}_{A\leftarrow a'} = y | X = x, A = a)$$ (Feldman et al., 2021).

All of the above definitions and metrics can play an important role in assessing fairness in THEMIS 5.0, making it a complicated endeavour. While many of the existing metrics exist for fairness, their applicability is usually highly affected by the context in which they are implemented. With this in mind, we anticipate that we will utilize the aforementioned metrics for assessing fairness and also explore the possibility of applying or developing case-specific metrics to support the AI System owners with meaningful measurements.

2.6.4. Technologies and optimization for Trustworthy AI

The abundance of conceptual principles, guidelines, and methods has been recently accompanied by many concrete software tools that attempt to address the need to move from 'what' to 'how', i.e., to move beyond ethical AI guidelines to concrete operational mandates and tools that enable better oversight mechanisms in the way AI systems are developed and deployed. Various survey papers review the related technologies and tools. For example, Morley et al. (2020) review tools and methods in order to help translate principles into practice, while Li et at. (2023) introduce a framework that consolidates the existing fragmented approaches to trustworthy AI into a unified, systematic approach. This approach encompasses the entire lifecycle of AI systems, spanning from data acquisition to model development, system development and deployment, and ultimately to continuous monitoring and governance. Toreini et al. (2022) focus on four categories of system properties that are considered instrumental in achieving the policy objectives of AI trustworthiness, namely fairness, explainability, auditability and safety & security (FEAS). They further review the main technologies and tools with respect to these four properties, for data-centric as well as model-centric stages of the machine learning system life cycle. Liu et al. (2022) concentrate on six dimensions crucial for attaining trustworthy AI: (i)

Safety & Robustness, (ii) Non-discrimination & Fairness, (iii) Explainability, (iv) Privacy, (v) Accountability & Auditability, and (vi) Environmental Well-being. For each dimension, they assess the associated technologies, outline their real-world applications, and explore the corresponding and conflicting interactions among these various dimensions. On the other hand, Kaur et al. (2022) analyse trustworthiness requirements (fairness, explainability, accountability, reliability, and acceptance), adopting a human-centred approach by examining different levels of human involvement in making AI systems trustworthy.

Table 6. Technologies and tools for trustworthy AI

Organisation	Software	Description	Online references
AI Verify Foundation	https://github.com/IMDA-BTG/aiverify	Single integrated toolkit that can perform technical tests on common supervised learning classification and regression models for most tabular and image datasets.	https://aiverifyfoundation.sg
IBM research trustworthy AI	https://github.com/Trusted-AI	Various projects available on Linux Foundation AI Trusted AI organisation e.g. AI Fairness 360, AI Explainability 360, etc.	https://research.ibm.com/topics/trustworthy-ai https://aif360.res.ibm.com
IBM product watsonx.governance	Not openly available	watsonx.governance™ employs software automation to strengthen an organisation's ability to mitigate risks, manage regulatory requirements and address ethical concerns.	https://www.ibm.com/products/watsonx-governance
Google TensorFlow Responsible AI	https://github.com/tensorflow	Responsible AI practices (e.g. fairness, privacy, interpretability) integrated in the ML workflow using TensorFlow.	https://www.tensorflow.org/responsible_ai
Microsoft Responsible AI	https://github.com/microsoft/responsible-ai-toolbox tutorials & walkthroughs https://github.com/microsoft/responsible-ai-workshop	Suite of tools providing a collection of model and data exploration and assessment user interfaces and libraries that empower developers and stakeholders of AI systems to develop and monitor AI more responsibly. Microsoft also provides a series of hands-on tutorials for developers and data scientists.	https://www.microsoft.com/en-us/ai/ai-lab-responsible-ai-dashboard (link require login)
SAS product Viya Platform	Not openly available	SAS Viya AI and data analytics platform with AI-based automation produces outcomes that are repeatable, reliable, explainable and compliant.	https://www.sas.com/sas/whitepapers/a-comprehensive-approach-to-trustworthy-ai-governance.html https://www.sas.com/en_us/company-information/innovation/responsible-innovation.html
Data Robot	Not openly available	DataRobot's enterprise AI platform incorporates features and tools that make trustworthy AI accessible and standardized.	https://www.datarobot.com/trusted-ai-101/
DataIku Govern	https://github.com/dataiku	Dataiku governance framework features a centralized monitoring capability and integrated MLOps to close the governance loop after models are deployed into production.	https://blog.dataiku.com/a-lifecycle-approach-for-responsible-ai https://knowledge.dataiku.com/latest/mlops-o16n/govern/index.html

Captum	https://github.com/pytorch/captum	Captum provides algorithms that allow developers to understand which features are contributing to a model's output.	https://captum.ai
Alexandra Institute	https://github.com/alexandrainst/responsible-ai	Knowledge base for responsible AI	https://alexandra.dk/about-the-alexandra-institute/
Holistic AI library	https://github.com/holistic-ai/holisticai Documentation https://holisticai.readthedocs.io/en/latest/	The Holistic AI library is an open-source tool to assess and improve the trustworthiness of AI systems. Currently, it offers a set of techniques to easily measure and mitigate bias, and in the future, it will be extended to include tools for efficacy, robustness, privacy and explainability.	platform https://www.holisticai.com/ai-governance-platform
Trustible solution	Not openly available	Responsible AI Governance platform, a turnkey solution to maximize trust and facilitate AI governance.	https://www.trustible.ai/future-why-trustible

In addition to academically available research efforts, major technology companies have started providing technologies and toolkits to support trustworthy AI. **Table *6*** outlines the efforts of major well-known corporations, some of which (like IBM and Microsoft) provide open-source versions of their toolkits. Noteworthy is the fact that two foundations supported by many companies provide open-source toolkits. The first is the AI Verify Foundation [6], a not-for-profit subsidiary of IMDA, the Infocommunications Media Development Authority of Singapore. This initiative seeks to leverage the collective expertise and efforts of the global open-source community to create AI testing tools that promote responsible AI practices. Key members of this foundation include industry giants such as Google, IBM, Microsoft, RedHat, Aicadium, Salesforce, among others. The foundation is responsible for the development of AI Verify[8], a framework and software toolkit designed for AI governance testing. AI Verify validates the performance of AI systems based on a set of principles and aligns with AI governance frameworks such as those established by the European Union, OECD, and Singapore. The second is the LF AI and Data Foundation, a Linux Foundation project that supports and sustains open-source projects within AI and the data space[9]. Of relevance is the Trusted-AI that hosts LF AI Foundation projects in the category of Trusted and Responsible AI. Among them are IBM's toolkits[10] such as AI Fairness 360, AI Explainability 360, Adversarial Robustness 360, AI Privacy 360, etc. While all these tools primarily concentrate on evaluating the trustworthiness of the AI system itself, there are recent research endeavours that redirect attention towards assessing the perceived trustworthiness of the development process. The rationale behind this shift is that while trustworthy AI is defined by system requirements, its practical implementation necessitates an understanding of its connection to specific measures throughout the development process. For example, Hohma and Lütge (2023) present a concept for establishing a trustworthy development process for AI systems, introducing a framework derived from a semi-systematic analysis of AI governance activities. This framework aims to identify the obligations and measures necessary to meet established AI ethics requirements and align them with the AI development lifecycle. Another effort, by Ronanki et al. (2023) focuses on requirements engineering and examines the applicability of ethical AI development frameworks for performing effective requirements engineering during the development of trustworthy AI systems.

As it has become evident from the overview above, numerous technologies and tools already exist to support trustworthy AI development and deployment. However, most of these tools focus on specific technical aspects of AI models and fall short of providing a comprehensive framework. A more holistic approach should involve human domain experts and users while also considering the broader application environment during the assessment and optimization of trustworthiness. THEMIS 5.0 aims to bridge this gap by leveraging technical approaches from various tools and taking into account the different AI stakeholders who interact with the AI system in order to support them through the trustworthiness process.

[8] AI Verify Foundation, https://aiverifyfoundation.sg
[9] Linux Foundation LF AI & Data Foundation https://lfaidata.foundation
[10] IBM. AI fairness 360. https://aif360.res.ibm.com

2.7. Ethical and Legal Requirements for Trustworthy AI

2.7.1. Introduction

THEMIS 5.0 aims for a personalised Trustworthiness Optimisation Ecosystem that is, per se, AI-driven. As such, the THEMIS 5.0 technology needs to be designed in adherence to the ethical and legal requirements of Trustworthy AI, applicable in the EU, and can then be meaningfully used to evaluate and optimize the trustworthiness of third-party AI systems.

AI governance implies regulation in a very broad sense, not only by law but rather by general steering of human behaviours and practices. In recent years, due to the fast advancement and widespread availability of AI deployments, AI governance gained exceptional momentum. However, at the present moment, we are still in an early stage of the emergence of AI rules and practices, characterized by changing and blurred frameworks.

In particular, so far, the role of ethics and soft law is much more prominent than the role of hard law, which is in the course of development, with the AI Act as the first horizontal regulation specifically tailored for AI in the world. Ethics strongly contributes to "making, shaping or changing the law" (so-called hard ethics), while in the foreseeable future ethics may acquire a more complementary role to law "by considering what ought and ought not to be done over and above existing regulation" (so-called soft ethics) (Floridi, 2023 p. 82).

With this premise, building on the evermoving ethical and legal landscape for Trustworthy AI this section[11] focuses on the requirements that the design and development of THEMIS 5.0. technology needs to take into consideration.

2.7.2. Overview of the philosophical ethics landscape of Trustworthy AI

In this section we provide an overview of the ethical basis for Trustworthy AI with particular relevance for THEMIS 5.0.

The recent development of AI technologies has necessitated a thorough ethical evaluation of their implications within the field of AI ethics (Floridi, 2023; Coeckelbergh, 2020) and within various ethical AI frameworks (Prem, 2023). This is, for example, seen in the substantial development of, and interest in, ethical guidelines for AI systems developed by industry, government and interest organizations. Hagendorff (2020) provided an overview of more than 20 such guidelines intended for an international context, examples of which include the European AI HLEG (2019) and OECD Recommendation of the Council on Artificial Intelligence (OECD, 2024). AI ethics guidelines suggest that trustworthiness may be compromised by various factors such as opacity, misuse and malpractices, risk biases and issues concerning data sharing and surveillance, and in response recommends adherence to ethical requirements including, e.g., transparency, fairness, non-discrimination, diversity, data security and privacy, reliability, robustness, accuracy, traceability and verifiability, understandability, interpretability, reasonable explainability, predictability, human oversight, liability, avoiding harm and compliance with norms and standards. The connection between trustworthiness and human values is also underlined in relation to autonomy as a human right, e.g. with regard to non-manipulation and human consent (Reinhardt ,2023; Buruk, 2020).

A thorough critical examination of existing AI ethics guidelines has been undertaken (Buruk, 2020) with an emphasis on trust and trustworthiness within the field of AI and the need for a practical philosophy approach. Reinhardt (2023) highlights issues of clarity and ambiguity, drawing attention to the potential danger of overloading the notion of trustworthiness in AI research. Reinhardt further emphasizes that design and development factors are rarely mentioned in the guidelines. The analysis proposes a practical philosophy perspective to address these concerns and various principles related to trustworthiness are identified. For the needs of THEMIS 5.0, emphasis is placed on fairness, robustness, and accuracy. While fairness is explored independently, robustness and accuracy are typically considered together, or as closely related, in the ethics literature.

On this background, the ethical balance between human and artificial autonomy is essential in the development of Trustworthy AI. AI ethics here serves as a foundational step to inform legislation and promote autonomy in moral agency (Floridi, 2019). While AI trustworthiness involves meeting its function-based obligations, philosophical discussions bring several criticisms (Simion, 2023), including the absence of responsible ethical AI leaders, a lack of ethics audits, moral deskilling, inadequate inclusivity in AI multistakeholder governance, and a deficiency in scalable training programs for workforce sensitivity to ethical issues (Akbar et al., 2024).

Recognizing AI systems as socio-technical systems, the roles of individuals in design, development, deployment, and usage become crucial in determining trustworthiness. If we ignore these details, it has been argued that terms like "trust

[11] This section is complemented by the detailed analysis on EU and non-EU regulatory instruments for Trustworthy AI included in Annex 2.

in AI" and "trustworthy AI" might lose their clear meaning and become vague labels for any positive feature in AI systems (Duenser and Douglas, 2023). Moreover, a skeptical attitude towards the capacity of AI to be trusted in an anthropic manner has been presented, and it is proposed that trust in AI should be viewed as "a form of reliance" (Ryan, 2020). Freiman (2023) suggests shifting from the concept of "Trustworthy AI" to "Reliable AI", but critically argues that the former will persist. It is noted that AI ethics should extend beyond technical considerations and encompass power dynamics, social justice, and scholarly activism. Additionally, it is proposed that, to prioritize citizens' interests in AI, ethics and trust scholars focusing on community and social justice should (a) explore democratic aspects of trust formation and (b) shed light on critical social dimensions revealed by instances of distrust.

The importance of human values and virtues in digital ethics and AI has been discussed (Vallor, 2016 and 2024; Coeckelbergh, 2020; Stamatellos, 2007) and with an emphasis to the need for the design of value-sensitive AI and a meaningful human-centered control in the development of AI systems (Sadek et al., 2023; Cavalcante Siebert et al., 2023). In the emerging field of Human-Centred AI, as discussed in Section 2.4.1, comprehensive insights into the governance and ethical considerations of trustworthy AI have been provided, with an emphasis on a holistic perspective (Shneiderman, 2020; Kaur et al., 2022; Díaz-Rodríguez et al., 2023). Shneiderman (2020) argues that the gap between the ethical principles of Human-centered AI (HCAI) and practical governance should be bridged (See also Lin 2021) and proposes concrete recommendations towards this, across the governance levels of teams, organizations, and industries. At team-level, recommendations include implementing sound software engineering practices, such as audit trails, workflows, verification, validation testing, and bias testing. At organization-level recommendations emphasize creating a safety culture through leadership commitment, safety-oriented hiring and training, extensive reporting, internal review boards, and alignment with industry standards. Trustworthiness certification at industry-level involves government intervention, external audits by accounting firms, compensation by insurance companies for failures, advocacy by non-governmental organizations, civil society, and standards development by professional organizations and research institutes. The overarching goal is to mitigate risks, maximize benefits, and ensure the well-being of individuals, organizations, and society through HCAI (Shneiderman, 2020).

Díaz-Rodríguez et al. (2023) suggest a more holistic approach that encompasses four essential axes: the global principles for ethical use and development of AI-based systems, a philosophical take on AI ethics, a risk-based approach to AI regulation, and the aforementioned pillars and requirements. The seven requirements—human agency and oversight, robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability—are analyzed from a triple perspective: what each requirement for trustworthy AI is, why it is needed, and how each requirement can be implemented in practice. However, a practical approach to implement trustworthy AI systems involves defining the concept of responsibility for AI-based systems facing the law through a given auditing process. Therefore, a "responsible AI" system is the resulting notion introduced in this work, which can be realized through auditing processes. It is suggested that regulation is key for reaching a consensus among these views and that trustworthy and responsible AI systems will be crucial for the present and future of our society. Coeckelbergh (2020) emphasizes the importance of responsibility and explainability for ethical AI.

2.7.3. Overview of key legal and ethical requirements for THEMIS 5.0 results based on the EU regulatory instruments for Trustworthy AI

In response to the need for regulation of AI technology, more firm ethical requirements and regulations are emerging. These are grounded in the evolving ethics research and guidelines of relevance for AI systems development and deployment and formalized as policy or regulation by governments or international organizations.

Of particular relevance for AI research and development in Europe is the EU legislation on AI, but given the broad international scope of AI technology research and development, related efforts in Europe and beyond may also be of relevance for THEMIS 5.0. In this groundwork, we have therefore provided an overview of (a) relevant regulatory instruments in the EU as well as (b) related organizations and collaborating countries, specifically international and regional fora where the European Union operates, such as OECD, G7, G20, and UN, and (c) countries with which EU has strong cooperation on AI, including the US and UK. This overview is provided in Annex 2.

Since THEMIS 5.0. is an EU-based project, among the increasing number of global, regional and national initiatives to govern AI, we focus on official EU sources.

In this section, we provide the key legal and ethical requirements for THEMIS 5.0. results based on the EU regulatory instruments for Trustworthy AI, particularly the AI Act. While a more detailed overview of the AI Act is provided in Annex 2, we provide an overview of its background, its stated scope of AI systems, and its considerations of AI system risk before presenting the legal and ethical requirements.

2.7.3.1 Background on the AI Act

Ensuring the needed ethical and legal framework for AI development and deployment – so as to ensure Trustworthy AI – has been a key priority in the EU for years, as summarized in the 2018 EC Communication on Artificial Intelligence for Europe (EC, 2018).

Following this, AI ethics guidelines were developed by the EC HLEG (AI HLEG, 2019). Here, key ethical requirements for Trustworthy AI are grounded in basic ethical principles of respect for autonomy, prevention of harm, fairness, and explicability. The guidelines have subsequently been operationalized to support AI developers, specifically the Assessment List for Trustworthy AI – ALTAI (EC, 2020).

A further shift towards a law approach to AI was taken in the EU Commission White Paper on AI (EC, 2020), where legal initiatives specifically for AI were promoted. The first draft of the AI Act was provided in 2021 (EC, 2021). The final version was approved in 2024 and will be applicable from 2026.

2.7.3.2 The scope of AI systems

The AI Act adopts a legal definition of AI systems which delimits the scope of its provisions. The final version of the legal text opts for a definition of AI systems more oriented to technology neutrality. It states that 'AI system' means a "machine-based system designed to operate with varying levels of autonomy, 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)). Here, autonomy and adaptiveness are key concepts.

2.7.3.3 A risk-based approach to AI

The AI Act adopts a risk-based approach to AI. That is, the obligations of AI developers and deployers depend on the risk posed to fundamental rights, health and safety. On this basis, AI systems are categorized as (a) unacceptable risk, (b) high risk, (c) limited risk, and (d) minimal risk.

High-risk AI systems are systems that may pose significant potential harm to health, safety, fundamental rights, environment, democracy and the rule of law. High-risk AI systems have specific obligations and requirements.for THEMIS 5.0, a risk management system must be established, transparency and explainability requirements have to be implemented during the design and development of the system, human oversight shall be ensured, and an appropriate level of accuracy, robustness and cybersecurity shall be guaranteed (AI Act, art. 8-15).

Many of the requirements legally enforce the ethical requirements identified by the HLEG AI Guidelines. For instance, art. 10 (2) (g) demanding data governance and management practices with a view to identify possible biases relates to the requirement of diversity, non-discrimination and fairness; art. 13 on transparency measures; art. 14 on human oversight; art. 15 on accuracy, robustness and cybersecurity.

For limited-risk AI systems, requirements concern minimum transparency obligations, including to inform users that they are interacting with an AI system. For minimal-risk AI systems, no new legal obligations are added through the AI Act. However, they are encouraged to voluntarily adopt some of the legal requirements for high-risk AI systems or to apply additional requirements such as those of the EC AI HLEG (2019).

2.7.3.4 Key legal and ethical requirements for THEMIS 5.0

In the following, we have identified key legal and ethical requirements for THEMIS 5.0. The requirements have been mapped based on our review of the published version of the AI Act, as well as a review of relevant legal regulation and ethical requirements such as GDPR, regulation concerning Cybersecurity, and the EC HLEG ethical guidelines (AI HLEG, 2019).

The requirements are presented in a tabular format, to facilitate the access and consultation by the Partners who will design and develop THEMIS AI System(s) and components.

The ethical and legal requirements have been drawn out with an assumption that it is crucial that these are built into the technology from the design phase. Considering them only after the technical design choices have been made may be ineffective and time-consuming.

KEY LEGAL AND ETHICAL REQUIREMENTS FOR THEMIS 5.0. RESULTS BASED ON EXISTING AND UPCOMING EU REGULATORY INSTRUMENTS AND INTERPRETATIONS[12]	
LEGAL REQUIREMENTS	
AI *ad hoc* Regulation **AI ACT - REGULATION (EU) 2024/1689**	
Disclaimer	*The legal requirements of the AI Act will not be applicable to THEMIS 5.0. AI system(s) in the course of the project, because THEMIS 5.0. AI system(s) are covered by the research exemption for which the AI Act is not applicable to "AI systems or AI models, including their output, specifically developed and put into service for the sole purpose of scientific research and development." (AI Act, art. 2(6)).* *However, in order to design and develop a future-proof, human-centric and marketable AI technology, the legal requirements of the AI Act must be taken into consideration, even if not directly prescriptive in the course of the project.*
Premise	*According to the current initial and provisional understanding and assessment of THEMIS 5.0. trustworthiness optimization ecosystem[13], the technology does not seem to fall in the categories of prohibited practices and of high- risk AI systems.* *THEMIS 5.0. seems to fall under the category of limited risk, subject to transparency obligations, since i) it is intended to interact directly with natural persons, ii) it may generate text content (AI Act, art. 50 (1)(2)).*

Requirements from AI ACT - REGULATION (EU) 2024/1689: **Providers[14] of THEMIS AI system(s) and/or components SHOULD**	**Legislative references**
Inform any person interacting with THEMIS, that they are interacting with an AI system in a clear manner and at the time of the first interaction with the system (so-called detection obligation).	AI Act, art. 50 (1)
Ensure that the modalities in which this information is given is appropriate to the characteristics of the intended final users, for instance belonging to vulnerable groups due to their age or disability.	AI Act, rec. 132
Ensure, through the adoption of technical solutions, that the output generated by THEMIS (i.e. text content) is machine readable and detectable as artificially generated (so-called labelling obligation).	AI Act, art. 50(2))
Choose technical solutions to mark the AI output as artificially generated, by considering the State-of-the-Art in terms of reliability, interoperability, effectiveness and robustness. This can be achieved by combining different techniques, including watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods, fingerprints. Techniques can be implemented at the level of the system or at the level of the model.	AI Act, rec. 133
Take measures to ensure that the deployers dealing with the operations and use of THEMIS have the means, skills, knowledge and understanding ("AI literacy") to make an informed use of the AI systems and are aware of the opportunities and risks of AI and possible harms.	AI Act, art. 3(56) and art. 4
Adopt measures (technical in the design of the systems and organizational/contractual in the relationship with the deployers) to make sure that affected persons can access the knowledge	AI Act, art. 3(56), art. 4 and rec. 20

[12] This list includes the key legal and ethical requirements for the design and development of THEMIS 5.0. AI technologies in line with the principles of AI trustworthiness for the EU. The respect of this list does not ensure automatic compliance with all the legal requirements to which THEMIS 5.0. AI system(s) will subject, which will depend on the actual features of all the components of the AI systems under development, their use cases and the applicable national legislations in combination with the relevant European frameworks. The compilation of this list was inspired by Rosenthal's (2024) summary of compliance requirements for AI.
[13] A thorough ethical and legal impact assessment of the technology will be conducted by following the methodologies put forward in D1.3., including a re-assessment of the provisional initial understanding of the AI system(s) and its risk category.
[14] Provider is anyone – person, public authority, company etc. – who either develops, places on the market or puts into service an AI system or a general-purpose AI model whether for payment or free of charge.

necessary to understand how decisions taken with the assistance of AI will have an impact on them.	

Data Protection **GDPR - REGULATION (EU) 2016/679**	
Disclaimer	*Additional national data protection obligations must be considered.*
Premise	*The GDPR applies to any operation or set of operations ("processing") performed on personal data or on sets of personal data, either or not by automated means.* *Personal data are not only those that directly identify a person - such as the name or an identification number - but also any information that can be used to trace back the person to which it pertains, including factors related to the social or cultural identities.* *Pseudonymised data are still "personal data", while fully anonymised personal data are not anymore personal data and therefore not subject to the GDPR.* *The following requirements apply only if and when THEMIS 5.0. AI system(s) process personal data.*

Requirements from GDPR - REGULATION (EU) 2016/679: **Providers of THEMIS AI system(s) and/or components HAVE TO**	**Legislative references**
Identify and map all the operations in which THEMIS trustworthiness optimization ecosystem may process personal data, by distinguishing "ordinary" personal data, from special categories of personal data concerning for example political opinions, religious or philosophical beliefs and health data.	GDPR, art. 9(1)
Design THEMIS by minimising the processing of personal data. Anonymize personal data to a possible extent/an extent possible, and where not possible put in place pseudonymization techniques or/and other measures that implement the data minimisation principle.	GDPR, art. 5(1) (c) and art. 89(1) [on data minimisation and research purposes]
Identify for each processing operation the Partner(s) who play the role of the "data controller", by determining the purposes and the means of the processing or "data processor", by processing personal data on behalf of the controller. If controllers and processors are identified, a data processing agreement among them must be put in place. If more Partners jointly determine the purposes and means of processing, they shall be joint controllers**,** and a joint controllership arrangement must be put in place.	GDPR, art. 4 (7)(8) [on the definitions of data controller and data processor] GDPR, art. 28 [on processors and data processing agreements] art. 26 [on joint controllers]
Identify what is the lawful basis for each processing operation, by consulting your DPO and the DPO of the project. The controller shall implement it and inform the data subject. For instance, if consent is identified as the appropriate lawful basis, the controller(s) acquires consent from the data subjects in a manner that is freely given, specific, informed and unambiguous.	GDPR, art. 5(1)(a), art. 6, art. 9 and art. 13 (1)(c)) [on lawfulness of processing and transparency to the data subject] GDPR, art. 4(11) and art. 7 [on consent]
Communicate to the data subject what data processing operations are performed trough AI systems, what is the purpose of each data processing operation and use the data only for the original declared purpose.	GDPR, art. 5 (1)(b), art. 13(1)(c) [on purpose limitation and transparency to the data subject]
Map with particular attention, the processing operations aimed at "profiling", which occurs when automated means, including AI, are used to evaluate, analyse or predict personal aspects related, for instance, to personal preferences, interests and behaviour.	GDPR, art. 4 (4)
For each profiling activity, clearly inform the data subject about the logic of the profiling and the possible consequences of such processing.	GDPR art. 13 (2)(f)
Ensure that anytime THEMIS produces a decision that is solely based on automated processing, including profiling, the data subjects i) are informed about the existence and the	GDPR art. 13 (2)(f) and art. 22

nature of the processing and can: ii) obtain human intervention from the controller iii) express their point of view iv) contest the decision.	[on transparency to the data subjects and automated individual decision making]
Ensure that possible inaccuracies of personal data processed can be remedied with timely erasure and rectification.	GDPR, art. 5(1) (d) and Art. 16, 17 [on accuracy principle and related rights of the data subjects]
Store personal data only for as long as necessary in accordance with the purpose of the processing operation, and clearly communicate the storage period to the data subject.	GDPR, art. 5(1)(e) and art. 13(2)(a)) [on storage limitation and transparency to the data subject]
Put in place measures that protect the integrity, the confidentiality and the security of the personal data processed. Measures could also concern the exclusion from use of THEMIS 5.0. AI systems by downstream deployers[15] who do not offer sufficient guarantees in terms of integrity, confidentiality and security.	GDPR, art. 5(1)(f)) and art. 32 [on integrity, confidentiality and security]
Be able to comply with data subject requests (e.g., access, objection, deletion)	GDPR, art. 15/22 [on data subjects rights]
Map for each processing operation and communicate to the data subject, all the entities to which personal data are disclosed ("recipients"), either Partners or third parties	GDPR, art. 4(9) and art. 13(1)(e) [on recipients and transparency to the data subject]
Prior to the beginning of the processing operation, evaluate if a data protection impact assessment (DPIA) is needed and carry it out. Guidance on the conditions that trigger a DPIA is included in D1.3. and its Annex I.	GDPR, art. 35
Keep records of the processing activities carried out by each controller and processor, in particular in case of processing activities that require a DPIA.	GDPR, art. 30
CYBERSECURITY	
Requirements from Cybersecurity laws: **Providers of THEMIS AI system(s) and/or components HAVE TO**	**Legislative references**
Put in place technical and organizational measures to guarantee an appropriate level of cybersecurity, encompassing the system(s) *per se*, the direct users and all the persons that can be affected by cyber threats. Such endeavour can be facilitated by the adoption of a (voluntary) European cybersecurity certification scheme, following the Cybersecurity Act, Regulation (EU) 2019/881.	Cybersecurity Act, art. 46 and ff.
Consider that THEMIS falls within the definition of "products with digital elements" of the possible upcoming Cyber Resilience Act[16] (CRA), for which manufacturers have to analyse cybersecurity risks and implement cybersecurity requirements.	Cyber Resilience Act, art. 13 and Annex I (text adopted by EU Parliament 12 March 2024)[17]
Distinguish between the cybersecurity obligations of THEMIS AI system(s) and/or components per se and those of the downstream professional deployers, which have to take into account *ad hoc* obligations, such as the NIS (Network and Information System Security) Directives, considering the peculiarities, risk of attacks and impairment of the different use cases. Guidance on cybersecurity risk management is included in D1.3. and in the Annex I.	Directive (EU) 2016/1148 Directive (EU) 2022/2555

[15] Deployer is anyone using an AI system under its authority, excluding the use in the course of personal non-professional activities. For instance, Pilot Partners in Themis are deployers.
[16] After the European Parliament legislative resolution of 12 March 2024, the Council's 1st reading position is awaited for the final approval of the Act (European Parliament, 2024) Carriages preview | Legislative Train Schedule (europa.eu).
[17] Reference to the CRA is to the text adopted by the European Parliament on 12 March 2024 (European Parliament, 2024 TA (europa.eu).

THIRD PARTY RIGHTS AND OWNERSHIP		
Premise	*Third Party rights and ownership rights are diversified and heavily regulated at national level. For this reason, specific legislative references are not included, but the requirements are expressions of established legal principles.*	
Requirements concerning third party rights and ownership: **Providers of THEMIS AI system(s) and/or components HAVE TO**		
Prevent, as far as possible, through, for instance, filtering or blocking measures, that the AI input and/or output is detrimental to third parties' rights, such as copyright and other protected content, as trade secrets.		
Train and fine-tune the AI model(s) used by THEMIS on content which can be lawfully used for this purpose		
Agree with the downstream deployers on indemnification obligations in the event of third parties claiming that THEMIS AI system(s) and or/components have infringed their IP rights		
Put in place agreement (within and outside the Consortium) that clearly identify the owners of THEMIS outputs		
CONTRACTUAL COMMITMENTS BETWEEN PROVIDERS AND DOWNSTREAM PROFESSIONAL DEPLOYERS		
Premise	*Contract law and confidentiality obligations are regulated at national level, therefore specific legislative references are not included here, but must be taken into consideration for compliance.*	
Requirements concerning contractual commitments between providers and downstream professional deployers: Providers of THEMIS AI system(s) and/or components HAVE TO		
Put in place contracts between the provider(s) of THEMIS AI system(s) and/ or components and the downstream professional deployers in the context of the use cases to ensure respect for the confidentiality obligations to which the deployers are subject. For instance, data acquired from the downstream deployers must be used by the providers only for the provision of the AI services; the providers (including all their personnel) must access deployers' data only when expressly allowed and keep them confidential; disclosure or access by third parties to deployer's data accessed by the providers must not be permitted. It is relevant to consider that the contractual commitments related to confidentiality are not restricted to personal data.		
Delete any data acquired from the downstream deployers by the providers of THEMIS AI system(s) and/or components after the termination of the contract/provision of the AI service, except when the law prescribes retention periods or when the retention is necessary for motivated security reasons (e.g. back-up).		
Put in place terms and conditions of use of THEMIS AI system(s) and/or components which regulate how and under which conditions THEMIS and its outputs can be legitimately used by the professional downstream deployers in the use cases context.		
Monitor the use, potential misuse, and abuse of THEMIS AI system(s) and its components. Require downstream deployers to report any incident, malfunctioning or performance issues.		
ETHICAL REQUIREMENTS		
Premise	*Since THEMIS 5.0. appears to fall under the category of limited risk under the AI Act, the legal requirements prescribed by the AI Act for high-risk AI systems are not directly applicable as such but should be considered from an ethical perspective. The AI Act strongly promotes voluntary compliance with high-risk AI legal requirements - already drawing from the EU Commission Ethics Guidelines for Trustworthy AI (2019)- to design, develop and deploy ethical AI systems. Some of the listed ethical requirements are elaborated in view of topics that emerged in the AI ethics debate*[18].	
Requirements from ethical guidelines: **Providers of THEMIS AI system(s) and/or components SHOULD**		**References**
Implement the following ethical requirements by incorporating available technical solutions and adhering to industry best practices such as using model and data cards.		AI Act, rec. 165
If THEMIS AI System(s) are based on machine learning, carefully build or select the data sets for training, testing and validation of the models so that they are relevant, representative and		AI Act, art. 10, rec. 67

[18] For an overview on AI ethics see par. 2.7.2.

to the possible extent free of errors considering the intended purpose of the system and the prospected persons or group of persons that will be impacted by the AI system(s).	
Put in place measures taking into consideration each AI system(s) component and the relevant socio-technical environment (e.g. model, datasets, environments of use, targeted deployers and affected persons) to avoid discriminatory outputs.	EU Commission Ethics Guidelines for Trustworthy AI (2019)[19]
Ensure that measures for the automatic recording of events are put in place to trace the system's functioning and facilitate accountability.	AI Act, art. 12
Inform the downstream deployers about the characteristics, capabilities, logic and limitations of the AI system(s) and the foreseeable risks connected to its use and misuse.	AI Act, art. 13 and 14
Inform the deployers on how to use the systems and how to interpret the output, as well as how to use it.	AI Act, art. 13 and 14
Document and inform the deployers about the estimated energy consumption related to the operation of the system(s).	EU Commission Ethics Guidelines for Trustworthy AI (2019); AI Act rec. 165 and art. 13
Design and develop THEMIS with features that ensure meaningful control by the humans operating the system(s), including the detection of possible malfunctioning. Oversight measures can also be identified by the providers and implemented by the downstream deployers, depending on the characteristics of the use cases.	AI Act, art. 14
Make sure that the deployers are able to opt-out from the THEMIS decision, in particular when the decision has significant effects (for instance affects the right to health or freedom of expression).	AI Act, art. 14
Adopt appropriate safeguards to guarantee that THEMIS AI system(s) is accurate, by delivering correct output. You should choose appropriate accuracy metrics, based on the State-of-the-art and communicate in a clear manner to the deployer.	AI Act, art. 15
Equip THEMIS AI system(s) with backup and fail-safe plans and other redundancy solutions to guarantee the system's robustness and consistent performance through time.	AI Act, art. 15
Put in place an inclusive, diverse and participatory design of the AI system(s), by fostering the involvement of prospected deployers and people outside the design team (e.g. people working in the same sector and performing the same functions as the prospected deployers).	AI Act, rec. 165; EU Commission Ethics Guidelines for Trustworthy AI (2019)
Define accountability with regard to THEMIS AI system(s) and/or components. This implies i) identifying who is accountable for the functioning of THEMIS AI services; ii) identifying who is accountable for the downstream use of THEMIS AI services; iii) establishing redress mechanisms to report incidents and malfunctioning.	EU Commission Ethics Guidelines for Trustworthy AI (2019)
Enable the deployers to explain to affected persons the role (if any) that THEMIS AI system(s) plays in the decision-making process.	AI Act, art. 84

2.8. Trustworthiness optimization

The field of AI Trustworthiness only recently received great attention from academia, industry, and government bodies, mainly due to the unprecedented and rapid growth of AI which hinders the trust of humans in such AI Systems. Despite the massive effort and progress around AI Trustworthiness, the literature does not provide a clear definition for its optimization or even a clear approach or framework to resolve the issues related to trust.

Although a formal definition for Trustworthiness optimization is absent, academia and government bodies give directions on how to proceed. The optimization in this context seeks to find a balance between the high performance of AI tasks while also ensuring the minimization and mitigation of trustworthiness issues to foster trust among users, stakeholders,

[19] Discriminatory outputs can be relevant also from a legal point of view, taking into consideration the EU non-discrimination legal framework and the national laws. The principles of fairness and non-discrimination are detailed from an ethical and legal perspective in par. 2.3.7.2. and a Guide for the impact assessment of THEMIS technology vis-à-vis fairness and non-discrimination is included in D1.3.

and the broader society. Also, Trustworthiness optimization should be perceived as the systematic and intentional improvement of AI Systems to increase their trust by enhancing aspects such as reliability, transparency, fairness, accountability, and security.

In the literature, two articles were identified in the field of AI Trustworthiness that also included the word "optimization". One by Mannion et al. (2021), where an explicitly multi-objective decision making approach was used, and another from Mattioli et al. (2023) proposing a multi-criteria decision analysis as a solution based on the Confiance.ai program[1]. Many scholars consider the Trustworthy AI problem as a set of risks that need to be dealt with, that depend on many aspects of the AI system under examination. For this reason, many approaches default to risk management as a tool for improving AI Trustworthiness.

Hence, **risk management** can be considered a human-centred approach for optimizing Trustworthy AI, ensuring human involvement and applicability throughout the life cycle of the AI system. In the NIST AI Risk Management Framework (NIST, 2023), Trustworthy AI is established and maintained by organizations actively relating to measures for mapping, measuring, and managing risks towards trustworthiness.

Achieving trustworthiness optimization is not a straightforward task. It requires thorough analysis of the risks, challenges and issues an AI system faces, insightful search for possible mitigation methods and tools, and careful planning of testing and implementing mitigating actions as well as verifying their effects. Hence, understanding the lifecycle of trustworthiness of an AI system can be considered as a steppingstone for Trustworthiness optimization.

2.8.1. Trustworthy AI System Lifecycle

To provide a comprehensive description of the Trustworthy AI lifecycle, we delved into the literature for research papers and frameworks related to Trustworthy AI that also define or describe a lifecycle. We searched the literature using combinations of the keywords: AI System lifecycle; Trustworthy AI lifecycle; and Machine Learning lifecycle. We identified 10 references that were strongly related to the Trustworthy AI lifecycle.

Table 7: Surveyed literature on lifecycle perspective of trustworthy AI

OECD	NIST	ENISA	KAUR et. al.	Calegari et. al.	CDAC	Schlegel, et. al. (ML lifecycle)	Toreini et. al (ML lifecycle)	Suresh et. al. (ML system)	Li et. al. (AI product)	
Plan and Design	Plan and Design	Business Goals	Planning	Planning	Design	Requirements Stage			Requirement Analysis & Problem Modeling	
		Data Ingestion								
Collect and process data	Collect and process data	Data Exploration	Data	Data		Data-oriented Stage	Data Collection	Data Collection	Data Management	Management
		Data processing					Data Preparation	Data Preparation	Algorithm Design	
Build and use model	Build and use model	Feature Selection	Modeling	Development	Develop	Model-oriented Stage	Feature Engineering	Model Development	Development	
		Model Training								
Verify and validate	Verify and validate	Model Tuning					Training/ Testing	Model Evaluation		
		Model Adaptation								
Deploy	Deploy and Use	Model Deployment	Deployment	Deployment	Deploy	Operation stage	Inference	Model Postprocessing	Deployment	
		Model Maintenance						Model Deployment		
Operate & Monitor	Operate and Monitor	Business Understanding	Oversight	Planning					Product delivery	

Data-centric
Model-centric

Table 7 groups the surveyed approaches. Starting from the OECD (Framework for the Classification of AI Systems), the approach first identifies the key dimensions involved in AI Systems (Data & Input, AI Model, Task & Output, Economic context, and People & Planet) based on which they map six phases of the AI system lifecycle as complementary for the technical characteristics of the System: plan and design; collect and process data; build and use model; verify and validate; deploy; operate and monitor.

Building on the aforementioned approach, the NIST Artificial Intelligence Risk Management Framework (AI RMF) modified the lifecycle from OECD, to include five dimensions (Data and Input, AI Model, Tosk and Output, Application Context, and People and Planet) and only changed the phase of deploy to "Deploy and Use". The modifications highlight the importance of test, evaluation, verification, and validation (TEVV) processes and also make the operational context of the AI system more generalizable.

Similarly, ENISA (ENISA Report – Artificial Intelligence Cybersecurity Challenges) introduces a generic AI lifecycle as a reference model that includes several stages but also notes that there are several interdependent phases of the AI system and the lifecycle focuses on the ML subfield of AI. The lifecycle enumerates a total of thirteen stages (Business Goals, Data Ingestion, Data Exploration, Data Processing, Feature Selection, Model Training {Model Selection/Building, Model Training and Testing}, Model Tunning {Model Validation, Model evaluation}, Model Adaptation, Model Deployment, Model maintenance, Business Understanding). The connections for this lifecycle follow a sequential flow but with multiple feedback loops.

Kaur et. al. (2022) and Calegari et al. (2023) propose a similar lifecycle with some minor differences in the terminology. The first one focuses on Trustworthy AI and defines five stages (Planning, Data Modeling, Deployment, and Oversight), while the second one focuses on Fairness and changes from Modeling to Development and from Oversight to Planning while also indicating multiple sub-processes for each stage. De Silva and Alahakoon (2022) adopt the AI system lifecycle from the Centre for Data Analytics and Cognition (CDAC) which has three main phases (design, develop, deploy) and 19 stages as sub-processes.

Another approach that strongly focuses on Trustworthy AI is from Li et. al. (2022), who identify five stages for the AI System lifecycle (Data Management, Algorithms Design, Development, Deployment, and Management) that are linked to multiple aspects of AI Trustworthiness to map existing methods to improve trustworthiness through the AI System lifecycle. In addition to the five main stages, two more stages supplement the lifecycle: one at the very start (Requirement Analysis & Problem Modeling) and one at the very end (Product Delivery).

The rest of the literature covered in this section covers the lifecycle from the perspective of ML, which is a subset of AI, but it covers the majority of AI systems (ENISA). Schlegel and Sattler (2022) use a four-stage ML lifecycle (Requirements, Data-oriented, Model-oriented, Operations) with sub-processes to examine the management of ML systems artifacts. Toreini et al. (2022) indicate a six-stage (Data Collection, Data Preparation, Feature Engineering, Training, Testing, Inference) generic lifecycle to frame the discussion of trustworthy ML across several AI-based systems. Suresh and Guttag (2021) describe an ML lifecycle with six stages (Data Collection, Data Preparation, Model Development, Model Evaluation, Model Postprocessing, Model Deployment) used to understand and indicate harm sources.

From the reviewed literature one can observe that almost all approaches adhere to a similar AI System lifecycle structure, with some differences in terminology. The high-level stages of these lifecycles can serve as a foundation to be followed by the optimization process, which should enrich them with all the various methods, entities, characteristics, and considerations related to AI Trustworthiness.

2.8.2. Trustworthiness optimization building blocks

The direction for optimizing Trustworthiness is built upon all the SOTA approaches and methods that unilaterally aim to mitigate the limitations of some Trustworthiness issues and aspects. In the literature review of Kaur et al. (2022) on Trustworthy AI, three main pillars are identified to establish Trustworthy AI:

- **Approaches and methods at different stages of the AI life cycle**, including system design, system modeling, implementation, and oversight, as well as data gathering and management.
- **Human involvement**, as detailed by the HLEG on Trustworthy AI (AI HLEG, 2019) – and also in the NIST AI Risk Management Framework (NIST, 2023)
- **Collaborative decision making**, involving both humans and AI in the decision-making process.

Already methods exist in the literature that are exhaustively analyzed but only focus on specific Trustworthiness aspects or stages of the AI lifecycle. Li (Li et. al. 2022) navigates through the stages of the AI lifecycle while recording methods related to multiple aspects of AI Trustworthiness, such as Robustness, Generalization, Explainability, Transparency, Fairness, Privacy Protection, Value Alignment, and Accountability. Also, some of those methods are applicable to more than one aspect. Liu et al. (2021) go through each of the Trustworthiness aspects in the scope of the article, describing concepts around the aspect, defines methods against issues, domains that can be applied, and tools to implement them. Tghere are also papers that focus only on one aspect of Trustworthiness, such as the one from Calegari et al. (2023) that focuses on Fairness and provides multiple methods and approaches from the literature for the stages of the AI lifecycle, while also grouping them in the context of pre/in/post-processing. Diaz-Rodriguez et al. (2023) provide a holistic vision of the multifaced Trustworthy AI, primarily based on the principles of the EU, and give high-level descriptions and directions on how to address the aspects of Trustworthiness. A more comprehensive approach is the survey from Wu (Wu et. al., 2023) that introduces a novel taxonomy or framework for Trustworthy AI based on the perspective of strategic decision-making and provides a large pool of methods and tools that are related to multiple aspects of Trustworthiness.

Based on all the above, we foresee that the optimization and execution of trustworthiness can be achieved through a process strongly related to the AI System lifecycle, while also requiring human input and supervision. As depicted in Figure

14, the Trustworthiness Optimization Process consisting of four correlated stages can enable the effective implementation of trustworthiness and the incorporation of its related aspects and considerations.

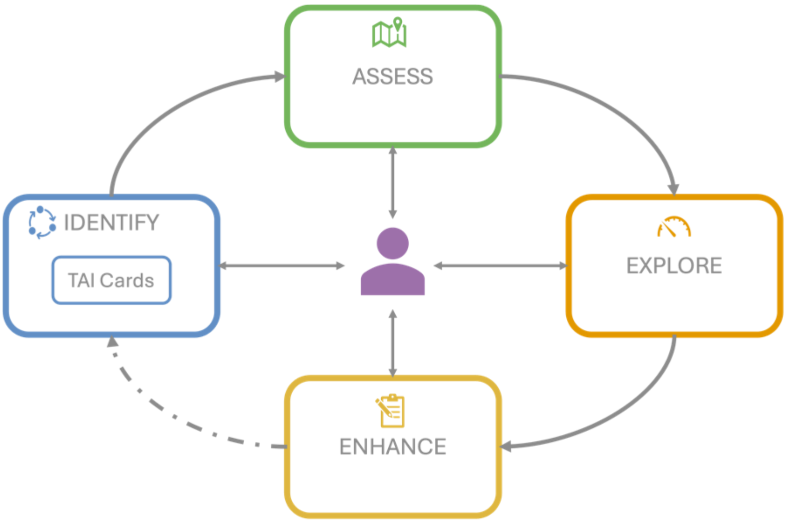

Figure 14: Proposed Trustworthiness Optimization Process overview

Although the proposed *Trustworthiness Optimization Process* is developing under THEMIS T2.3 and it is a subject of THEMIS D2.2, a short description of the stages is provided:

- **Identify:** The initial phase of the lifecycle requires human involvement to gather all the related information around the socio-technical specifications of the AI System. This data is stored as a structured set of cards, referred to as "Trustworthy AI Cards" (TAI Cards).
- **Assess:** After the necessary information is collected a comprehensive assessment begins, generating metrics and risks tailored to the use case.
- **Explore:** Based on the initial assessment of the system the lifecycle aims to find optimal solutions by testing several methods and comparing changes in the metrics and risks.
- **Enhance:** Finally, when an acceptable solution is found, the respective changes to the AI system are applied while also monitoring the system for potential shortcomings.

2.8.3. The ethical and legal basis for Trustworthiness optimization

In this section, we provide a summary of relevant ethical and legal background for trustworthiness optimization. A more detailed overview is provided in Annex 1.4.

Prominent ethical documents point out that ethical requirements of Trustworthy AI cannot be dealt with individually, but need to be balanced against each other, depending on the context of use (e.g. AI HLEG, 2019). Hence, the most precise meaning of each requirement in a certain domain also depends on the interaction with the other requirements and on the balanced trade-off that can be achieved in practice. Requirements can work synergically or, on the contrary, there can be tensions among them.

Synergies between the legal requirements are reflected in the AI Act. For example, the AI Act (art. 13) mandates transparency and provision of information to deployers, which in turn may contribute to enhancing other requirements such as accuracy, robustness, and fairness.

Tensions are also possible, such as the balancing of a need for security and privacy which may work against active measures to reduce bias and improve fairness (NIST, 2023).

Hence, trade-offs between competing interests and, by extension ethical requirements, are needed. Drawing on the literature, particularly the work of EC AI HLEG (2019), some key principles for trade-offs may be identified.

- **Case-by case:** Trade-offs between trustworthiness requirements needs to be done on a case-by-case basis, reflecting a system's context of use as well as relevant values and legal requirements which may depend on region or jurisdiction. Such a case-by-case approach is grounded in case law that guides the balancing exercise in specific domains.
- **Concern for fundamental rights:** When there are more means to reach a certain end, the least adverse to fundamental rights must be preferred in the application of the principle of proportionality (HLEG on AI, 2019).

- **Traceability of the process:** Trade-offs among requirements need to be addressed rationally and methodologically, with full traceability of the decision-making process and accountability of the decision-maker (HLEG on AI, 2019).
- **Overall benefits should exceed individual risks:** Tensions must be balanced in a way that the AI systems' overall benefits substantially exceed the foreseeable individual risks (AI HLEG, 2019).

These principles imply that it is not always possible to achieve acceptable trade-offs because there are fundamental rights which are absolute and that cannot be subject to balancing exercise, such as human dignity. When the risk is not acceptable, the system is not trustworthy (Laux et al., 2024). Acceptability of the risks is the threshold to operationalize the requirements proposed by the HLEG AI Guidelines and by the self-assessment List for Trustworthy Artificial Intelligence (ALTAI) and at the basis of the AI Act.

Takeaway: *The trustworthiness optimization that THEMIS 5.0. intends to achieve needs to take into account (e.g. by providing appropriate guidance to the users) the interactions between the requirements that will be detected, analyzed and optimized by Themis, i.e. fairness, robustness and accuracy and the other requirements for Trustworthy AI, which are not directly optimized by Themis e.g. accountability, privacy.*

3. T2.2: HUMAN-CENTRIC REQUIREMENTS ANALYSIS FOR THE THEMIS 5.0 TRUSTWORTHINESS OPTIMISATION ECOSYSTEM

To ensure a human-centric approach for the THEMIS 5.0 solutions, we have provided the material needed for the execution of co-creation A, processed its results, and also collected user requirements from the user-case partners. This work was conducted within Task 2.2. In this chapter, we present our approach, and a summary of the user requirements and outcomes.

3.1. Approach

An essential step to reach the objective of THEMIS 5.0 for a personalised Trustworthiness Optimisation Ecosystem is to involve citizens' values, hopes and concerns. To achieve this, we leverage two channels of information: the co-creation workshops and the three use cases of the THEMIS 5.0 project. Hence, to collect the human-centric requirements we: a) prepared the material needed for the execution of the workshops of Co-creation Phase A and processed their output, and b) to further enrich this knowledge, we collected input from the use case partners. For the development of the 'user journeys' and, eventually, of mock-ups, we leveraged input from both channels.

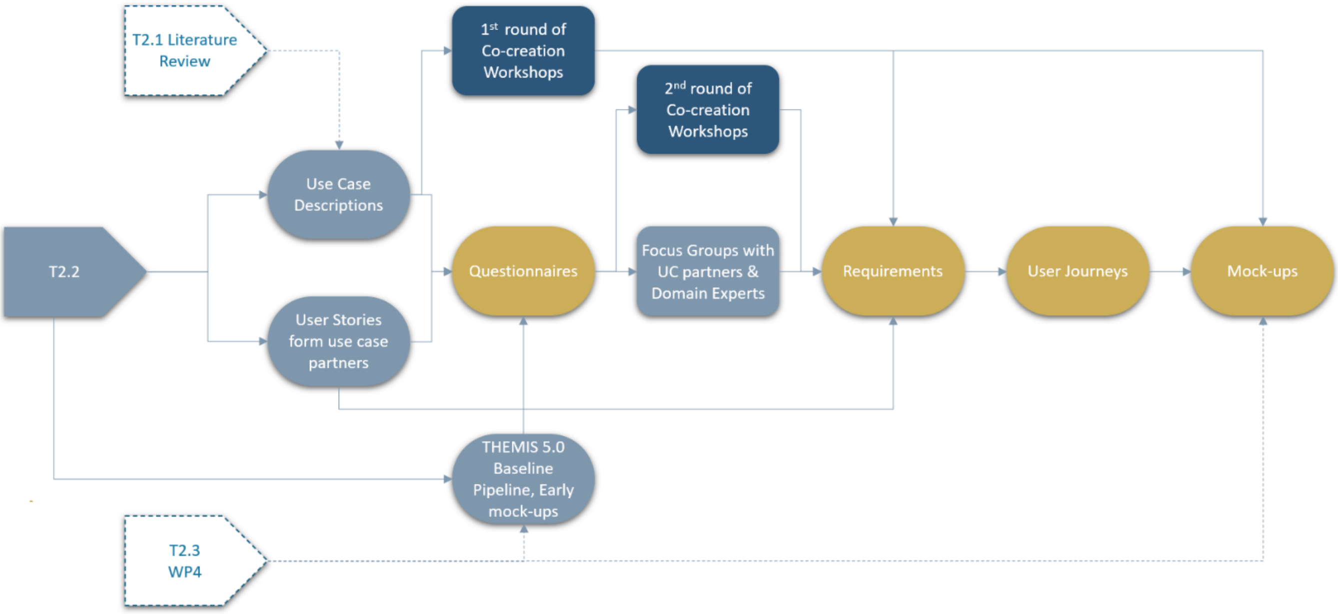

Figure 15: Workflow of the collection of user requirements and mock-ups development.

The workflow that we followed is illustrated in Figure 15, where it is shown that a variety of tools were used in a complementary manner:

- **Literature Review:** We conducted an extensive review of existing literature, including scholarly articles, research papers, and studies that explored the trustworthiness characteristics studied in THEMIS 5.0. This is detailed in Section 2. We leveraged this knowledge to formulate suitable material for the collection of the use case descriptions and to contribute to the preparation material for the 1st round of co-creation workshops.
- **Use Cases Descriptions**: We collected information for each use case regarding the AI system used, its use, the types of end users and its usage scenarios (presented in brief in Section 3.2). Additionally, for each scenario, we collected the decision points along with relevant business, ethical and legal risks and the associated trustworthiness vulnerabilities (presented in brief in Section 3.3).
- **1st round of co-creation workshops:** These workshops were implemented in WP3 using material prepared in T2.2 also leveraging the aforementioned Use Case Descriptions. The workshop methodology explored two themes: 1) end-users' attitudes towards AI tools in their work, and 2) contextual understanding of trustworthiness parameters. The results of the co-creation workshops (detailed in D3.1) were exploited for the definition of end-users' expectations from AI tools (Section 3.3.1) and for the identification of the user personas that will be used for the estimation of the users' trustworthiness preferences.

- **User Stories and Pilot Partners**: To gather user needs and understand their perspectives on how the THEMIS 5.0 AI platform should be designed as a user-friendly tool that can adapt to the needs of their respective sectors, we collected user stories from the pilot partners and their technical counterparts. The user stories had the form "As person role X, I want functionality Y so I get business benefit Z", were translated into user requirements and have been included alongside the user requirements that were collected via other means in Section 3.5. The list of all user stories can be found in Appendix A.1.
- **Questionnaires**: To further enrich the set of user requirements, we formulated three questionnaires (one for each use case) based on the input from the user stories and the use cases descriptions. The purpose of this questionnaire was to collect from the use case partners, their respective technical counterparts and the 2nd co-creation workshop further information about (i) the trustworthiness parameters definitions, (ii) business, ethical and legal risks stemming from trustworthiness vulnerabilities, (iii) user preferences regarding the importance and ranking of the trustworthiness characteristics, (iv) user requirements, and (v) user stories. To achieve this, we exploited the evolvements in the Conceptual Modelling of AI THEMIS 5.0 Trustworthiness Optimisation Ecosystem (T2.3) as described in Section 2.8.1, the system requirements -as defined in DoA (summarized in Section 0), and input from previously collected material, to formulate a baseline pipeline of THEMIS 5.0 ecosystem as well as early mock-up dialogues. This helped the respondents to comprehend the vision and the foreseen user interaction with the THEMIS 5.0 framework and provide meaningful input. The questionnaires are publicly available[20].
- **2nd round of co-creation workshops:** The questionnaires alongside the aforementioned explanatory material (THEMIS 5.0 baseline pipeline, early mock-up dialogues and THEMIS usage scenarios for each use case) were also provided to the 2nd round of co-creation workshops. These workshops were focused on gathering user stories from AI-users and business stakeholders related to the THEMIS 5.0 conversational agent and the prediction of trustworthiness preferences, as well as the presentation of the trustworthiness assessment and optimisation of the AI system. The user requirements that were translated from the user stories collected from the 2nd round of co-creation workshop are presented in Table 26 (and the user stories in Table 22).
- **Focus groups with partners:** The questionnaires were also provided to the use case partners and their respective technical counterparts and, when further support was needed (e.g., media use case), we organized focus groups. Through these expert interviews, discussions, and consultations, we gained valuable insights into the practical considerations to refine the user requirements, ensuring that the THEMIS 5.0 platform addressed user needs and trust. The user requirements from this process are presented in Table 20 and Table 27.

The user requirements stemming from both channels (co-creation workshops and use case partners) were analysed by identifying the various commonalities and their relevance to the scope of THEMIS 5.0. The summary of results is presented in Section 3.5. The outcome of this process is the THEMIS 5.0 *user journeys* and the *mock-ups* that reflect the broad user requirements for how the THEMIS 5.0 AI ecosystem should be designed and developed in a societally responsive and ethically sound way in accordance with citizen's values, hopes and concerns.

3.2. Description of Use Cases

In this section, we will define the three key use cases of THEMIS 5.0, each addressing a specific high-priority and critical application and industrial sectors, (i.e., in the healthcare, transportation and media sectors), engaging diverse communities of users of relevant AI systems from Greece, Bulgaria and Spain to test and evaluate THEMIS 5.0 results. We briefly present the AI system used, its use, the types of end users and its usage scenarios of the three use cases of THEMIS 5.0 project: AI in Healthcare, AI in Port and AI for fighting disinformation. For the collection of the use case descriptions, we prepared and distributed a template[21] to the use case partners and technical counterparts. The completed templates can be found in the shared drive[22] of the THEMIS 5.0 project.

[20] Questionnaires for URs
[21] THEMIS Use Case Description templates
[22] Use case descriptions

3.2.1. AI in Healthcare

AI System: Pancreatic Cancer Risk Predictor

User types: General Practitioners, Specialists in Internal Medicine, Endocrinology, Emergency Medicine, General and Abdomen surgery; Patients.

Scenario description: Recently created as a result of HORIZON 2020 iHelp project[23], and implemented into medical practice, the platform uses AI to triage the patients at risk of developing Pancreatic cancer (PC), by utilizing the collected Holistic Health Records (HHR). The platform provides Healthcare professionals (HCPs) with a powerful tool to evaluate the patients' health risk level and build personalised preventive programs, by constantly monitoring patients' health status and utilizing a decision support system (DSS).

Notwithstanding the excellent results achieved in raising awareness within HCPs about the plausibility of AI use in medical practice, a significant number of medics and patients have raised questions about the level of AI risk assessment accuracy and how it is calculated. Another raised concern is related to the fairness of the recommendation and advice provision to every patient at risk and what is the level of protection of the HHR data especially during the health status monitoring and secondary data collection.

Usage Scenarios		
Scenario Name	End Users	Scenario Description
Health-risk assessment, patients' triage, early diagnosis and patients' monitoring	General practitioners, Specialists in Internal Medicine, Endocrinology, Emergency Medicine, General and Abdomen surgery	Step 1: identify the patients to be enrolled in the AI system Step 2: Insert patient's data into the AI system Step 3: Launch the Risk predictor Step 4: Receive the Risk level. Step 5: Receive the AI recommendations. Step 6: Create a Risk mitigation plan. Step 7: Set the monitoring rules and objectives. Step 8: Set the Machine–patient automated dialogues. Step 9: Receive the DSS and patient's monitoring data. Step 10: Adjust the risk mitigation plan and set the new monitoring rules and objectives, when required.
Patient monitoring and coaching	Patients, HCPs	Step 1: Validate the AI platform's recommendations. Step 2: Set monitoring rules and objectives. Step 3: Launch the Patient's monitoring and DSS. Step 4: Receive the AI platform recommendations. Step 5: Receive and fill in the questionnaires. Step 6: Insert required data or use the recommended wearable for secondary data collection. Step 7: Receive the AI-driven alerts and virtual coaching. Step 8: Follow recommendations and advice.

3.2.2. AI in Port

AI System: ETA prediction system

User types: Port Traffic Control; Port Terminal; Towing Services, Pilot; Mooring Services; Transport Companies; Shipping Agencies; Port Authority; Freight Forwarder.

Scenario description: Port logistics entails the involvement of a multitude of agents that must coordinate efficiently. By predicting the Estimated Time of vessel Arrival (ETA) and its adoption by the different agents that make up the port-logistic chain, we can obtain a series of benefits such as: improving the efficiency of the different actors, reducing traffic congestion in the port, facilitating port traffic control operations, etc. Port of Valencia believes that using THEMIS 5.0 will

[23] GA No 101017441, https://ihelp-project.eu/

benefit us greatly by facilitating user adoption and ensuring that the systems are not biased and do not harm or benefit any agent.

Errors in the accuracy of the prediction of ETA can have significant and unfavourable consequences. For example, they could cause errors in the planning of resources in container terminals, resulting in delays in operations and economic losses. The ETA of vessels can be used to communicate to vessels en route to the Port of Valencia to delay arrival to avoid waiting in the anchorage area. An error in the prediction could result in excessive arrival delays but could also cause prolonged stays of the vessels in the anchorage area, which would have an impact on greenhouse emissions to the city. Additionally, this situation could result in trucks arriving at the port prematurely for container delivery or removal, leading to unwarranted congestion on various port roads. This congestion poses challenges for trucks that genuinely require timely access for import and export operations. Furthermore, erroneous ETA predictions can cause controversy among the involved agents, such as maritime agencies, which generally provide an ETA established by the captain of the vessel. As a parameter to be considered for the planning of nautical services (e.g. pilotage and towing) to attend a specific vessel, the ETA corrected by the AI system may cause the decision to attend another vessel from another shipping company different from the one initially pre-established by the original ETA, causing discontent among shipping companies and the port.

This use case will enhance the AI-based ETA prediction systems at the Port of Valencia with explainability and trustworthiness features to support its increased use and adoption by the port authorities, port terminals, shipping agencies, and transport carriers. The innovations developed in THEMIS 5.0 will help Port of Valencia achieve short-, medium-, and long-term trust in the AI-driven decision support and better use the AI recommendations to address dynamic situations, resulting in potential gains in terms of time, cost, environmental impact, and user satisfaction. Through the enhancements in THEMIS 5.0, the explainable AI system will be able to provide more relevant and understandable decision support, resulting in reduced risks, costs, and effort by all the agents involved in the port-logistic chain.

Usage Scenarios		
Scenario Name	End Users	Scenario Description
Scheduling Vessel Arrival and Exit	Port Control, Ship Captain, Pilot, Port Terminal, Towing Services, Mooring Services	Port Control must schedule vessel arrivals and exits regarding efficiency and safety. Port Control proposes a schedule to the pilot/s, which gets accepted or discussed. Step 1: Port Control and Ship Captains communicate with each other. Captains give their position, direction and speed when approaching the Port. Step 2: Port Control consults the AI system and schedules the order of entrance/exit and communicates with the pilot/s Step 3: Pilots also consult the AI system and accept or discuss the proposed scheduling with Port Control. Step 4: Port Control communicates with the Captains and gives them instructions (such as accelerate, lower your speed, go to the anchoring area, etc…) When the scheduled vessel arrives at the port entrance, it is attended by the necessary services, and the maneuver is executed. For exiting vessels, they wait at the port until they get attended.
Coordination of Terrestrial Transportation Schedule	Transport Companies, Shipping Agencies, Freight Forwarder, Port Terminal	Shipping Agencies contract Transport Companies to move the cargo from or to the port. Both parties agree on a promised date to deliver the cargo to its destination. Step 1: The Freight Forwarders/Shipping Companies must schedule and assign the travel orders to the Transport Companies. Step 2: Using the AI system, the Transport Companies know the ETA of the involved vessels, and can adjust the start of the travel schedule, reducing waiting times and traffic congestion at the port.
Scheduling Towing Ship's Maintenance	Towing Services	The maintenance operations of a Towing Ship require a specialized technician and a window of several hours where the Towing Services can carry out their functions with one less ship. Step1: Consult the data of future arrivals and exits. Step 2: Schedule a maintenance window.

3.2.3. *AI for fighting disinformation*

AI System: Hate speech detection system; Fake news detections system.

User types: Journalists, Fact-checkers.

Scenario description: News institutions and journalists increasingly use AI-based tools to help them avoid spreading disinformation and publishing unchecked information.

ANA media organisation has set up an AI-driven quality control system that ensures systematic monitoring, investigating, and reporting on disinformation. The core of the quality control system is the Truly Media tool that supports journalists with the fact-checking and verification tasks on the information they receive and propagate through their journalism. TrulyMedia was co-developed by ATC and Deutsche Welle (www.truly.media) and it is used by several major media organisations in Europe and beyond, including the European Digital Media Observatory, EDMO.

Journalists and/or fact checkers of ANA work every day to deliver accurate news to a wide range of outlets around Greece, but they are also investing part of their time in running a fact-checking service. They want to be able to monitor disinformation narratives and come up with an assessment, in a fast and effective manner. By using the AI features of Truly Media, ATC's collaborative platform, they are equipped with the tools and the steps necessary to reach a solid conclusion. Disinformation is one of the primary dangers to the world's prosperity, as it directly affects the ability of the public to make crucial decisions based on facts. Subsequently, it's in the interest of any highly acclaimed media outlet to fact-check their content or to inform their audience about narratives that appear widely online and might emerge as a direct danger to their readers' state of knowledge.

This use-case will focus on a dominant disinformation narrative, e.g., COVID-19, to pilot the results of THEMIS 5.0. In this context, the use case will allow journalists tasked with news verification to optimise the trustworthiness of the Truly Media AI services. Moreover, THEMIS 5.0 will ensure that Truly Media AI services balance the socio-technical organizational constraints for news production, resulting in the societal benefits of wider understanding and mitigation of the threat of disinformation. Partner ATC will lead the technical support to this Use Case.

Usage Scenarios		
Scenario Name	End Users	Scenario Description
Debunking Disinformation Item in Truly Media with the use of AI	Journalists, Fact Checkers	Truly Media is a collaborative fact-checking platform that allows journalists and researchers to work together to verify claims that are seen online. Gradually, the platform is being strengthened, with AI-based tools that are integrated into its 2-step solution, to make fact-checking easier and more efficient. Step 1: The user in our scenario starts by logging into Truly Media to find stories that are potentially flawed around Covid-19. For this reason, we have initiated an aggregation based on a specific set of keywords, web sources, RSS feeds and social media accounts. The AI-based system chooses through a fake news detector every hour stories from this pool of sources and feeds them automatically to the dashboard, so the user gets to pick the one he wants to work with. Step 2: In such a scenario, the user proceeds to the final step, which is to consider the fake news detector's probability, verify it, and start writing a fact-checking report on the story, by clicking on the verification button.
Identifying Potentially Misleading Narratives in TM with the use of AI	Journalists, Fact Checkers	Step 1: In this scenario, the user logs into Truly Media and gets fed with stories, especially posts that are considered to include some sort of hate speech. This means that their content isn't balanced and that the motive behind their spread is to increase a specific flawed narrative on the web. The labelling of the items as ones that contain hate speech is being done through AI. Step 2: Same process as in scenario UCS1. The user confirms the calculation of the previous step and starts writing a fact-checking report.
Manually uploading content to be verified	Journalists, Fact Checkers	Step 1: In the context of UCS 3, the user is not just receiving the aggregated feed of potentially false stories, but on the contrary, is the one who chooses the type of content that he/she wants to verify. Any media type is acceptable (picture, video, or text). Step 2: Verification of any given media type is partly possible, by existing technological solutions within TM. Step 3: The same process as in UCS1 and UCS2 is followed.

3.3. Trustworthiness Aspects of the Systems Under Test

In this section, the trustworthiness aspects of the Systems Under Test that relate to the project's Use Cases, are presented. First off, the end-users' expectations from the AI tools are listed, followed by end-users' expectations regarding the trustworthiness categories. The importance of each trustworthiness characteristic for each use case is analysed and considerations for measuring an AI system's trustworthiness parameters in the framework of each use case are showcased. Finally, the business, ethical and legal risks that, according to the end-users, are connected to each of the trustworthiness categories are listed.

This section aims to serve as a pointer to the direction that the THEMIS 5.0 framework should be designed and developed to ensure personalized assessment and optimization of the Systems Under Test, taking into account the specific needs, expectations, limitations, preferences and risks for each Use Case.

3.3.1. End-users' expectations from AI Tools

In this section, the co-created attention points that formulate the AI users' anticipation for the implementation of AI tools in their respective domains and determine the users' preferences affecting the perceived trustworthiness of the AI Systems Under Test, are presented. The following

Table **8** lists these attention points that were collected from the 1st round of creation workshops, per use case, as they were documented in the deliverable D3.1[24], alongside the related Users' Expectations that have been extracted. The relevance of these expectations to the context of THEMIS 5.0, is indicated in the last column of the table. THEMIS 5.0 should consider these expectations while designing and implementing the THEMIS 5.0 framework, whose ultimate goal is to help AI-users enhance an AI-system's trustworthiness. As expected, input from the co-creation workshops could in some instances exceed the scope of the project; Being the result of a live and open discussion amongst participants, there is always the possibility that the discussion touches upon issues beyond the use case descriptions and objectives. This input, while being indicated as not relevant to THEMIS 5.0, is still documented here, as an indicator of the participants' expectations.

Based on input from all three sectors (i.e. healthcare, port management, and media), there is the common notion that AI systems should serve as decision-support tools and not function independently.

In the **healthcare sector**, AI users need to be informed about how the AI-produced recommendations are made, the quality of the training dataset, any potential bias as well as any system limitations. The AI tools should be adaptable; being able to address the healthcare sector's evolving challenges and support healthcare professionals in their work without hindering their potential for skill development. They should also help professionals minimize environmental impact as traditional workflows integrate new technologies; climate impact should be considered in the cost vs. efficiency balance.

In **the port management sector**, AI users want the AI systems they are using to be always available, user-friendly and easy to use, even for those with limited experience in AI tools, as well as to be transparent in their decision-making. They should be able to adapt to the specific needs of each port by being able to model port-specific structures and commercial interests. The integration of AI systems should benefit all stakeholders proportionally to their needs.

In the **media sector**, AI users want the AI systems they are using to be transparent in their decision-making and accurate in their outputs. AI users want to be informed about the parties involved in the tools' development, the training datasets used, and the overall purpose of the system. They should be able to adapt to ever-changing contexts while remaining objective and adhering to sector-specific ethical standards. They should be able to support the work of journalists and fact-checkers while optimizing time without sacrificing quality.

[24] THEMIS 5.0: D3.1 Report on citizen co-created design principles for THEMIS 5.0 Ecosystem of services

Table 8: Users' Expectations from AI systems

Sector	ID	Attention Points	Users' Expectations from AI systems	Relevant to THEMIS (Y/N)
Healthcare	UR_W1	***Responsibility**: AI developers of healthcare-related AI tools should consider the way their tools deliver output to the healthcare professional to ensure that it is functioning as a support tool for human decision rather than the tool taking decision for the user.*	AI systems for the Healthcare sector should not function independently and should work as decision support tools assisting users to reach a decision.	Y
	UR_W2	***Responsibility**: AI-generated recommendations should be transparent, ensuring that healthcare professionals understand the basis and limitations of AI suggestions*	The user should be aware of how an AI system has produced a recommendation.	Y
	UR_W3	***Responsibility**: AI developers of healthcare related AI tools should develop their tools with a patient-centric approach*	AI systems for the Healthcare sector should be developed based on a patient-centric approach.	N (Overly generic)
	UR_W4	***Transparency & Accuracy:** AI developers of healthcare related AI tools should communicate clearly about the data used to train their AI tools and use evidence-based data*	The users of AI systems for the Healthcare sector need to be informed about the quality of the training datasets that have been used.	Y
	UR_W5	***Transparency & Accuracy:** AI developers of healthcare related AI tools should investigate potential biases in their training data*	Potential bias in the training datasets used for healthcare related AI tools should be investigated.	Y
	UR_W6	***Transparency & Accuracy:** AI developers of healthcare related AI tools should clearly define the limitations of their tools, and enable ongoing improvements of AI tools to address contextual and evolving healthcare challenges*	Healthcare-related AI tools' limitations should be clearly communicated to the user.	Y
	UR_W7		AI tools need to be adaptable to address contextual and evolving healthcare challenges.	Y
	UR_W8	***Transparency & Accuracy:** AI developers of healthcare related AI tools should find a succinct way to present transparency due to the limited time of healthcare professionals*	The AI system should ensure the explainability and transparency of results, while respecting the users' time constraints. Explanations and details should be brief and to the point.	Y
	UR_W9	***Cost vs. Efficiency:** AI developers of healthcare related AI tools should consider the cost vs. efficiency of their tools*	The development of AI systems for the Healthcare sector should consider cost vs. efficiency.	N (Overly generic)
	UR_W10	***Cost vs. Efficiency:** AI developers of healthcare-related AI tools should consider and minimise the climate impact of their tools*	The development of AI systems for the Healthcare sector should minimise climate impact.	Y
	UR_W11	***Cost vs. Efficiency**: AI developers of healthcare-related AI tools should align the development of AI tools with what the healthcare sectors and professionals' needs*	The development of AI systems for the Healthcare sector should be based on the needs of the healthcare sector and professionals.	Y
	UR_W12	***Attitudes towards AI:** AI developers of healthcare-related AI tools should ensure that AI tools are beneficial for healthcare professionals and do not become more time-consuming than not using AI*	AI systems for the Healthcare sector should facilitate and optimise the work of healthcare professionals.	Y
	UR_W13	***Attitudes towards AI:** AI developers of healthcare related AI tools should ensure that the use of AI tools does not negatively impact healthcare professionals' capabilities and the possibility to develop competencies*	AI systems for the Healthcare sector should not hamper healthcare professionals from acquiring and developing competences.	Y (see also UR_W1 above)
Port	UR_W14	***Responsibility**: AI developers of port management related AI tools should consider the way their tools deliver output to the port management professional to ensure it is functioning as a support tool for human decision making rather than the tool making decisions for the user.*	AI systems for the port management sector should not function independently and should work as decision support tools assisting users to reach a decision.	Y
	UR_W15	***Responsibility**: AI developers of port management related AI tools should foster a collaborative relationship between their tools and their users.*	AI systems for the port management sector should foster a collaborative way of working, so that users are involved in the AI-supported decision-making process.	N (Not aligned with UC2 SUT)

	UR_W16	***Responsibility**: AI developers of port management related AI tools should be transparent about who has the responsibility of AI tools contextual output, considering existing laws and limitations of the law.*	AI systems for the port management sector should be able to communicate the roles that are accountable for the AI tools' output according to the legal and regulatory framework.	N (Not aligned with UC2 SUT)
	UR_W17	***Transparency & Accuracy**: AI developers of port management related AI tools should consider how the AI tools will provide explanations for their outputs to foster trust*	The AI system should ensure explainability and transparency of results to foster trust amongst users.	Y
	UR_W18	***Transparency & Accuracy**: AI developers of port management related AI tools should consider how to strike a balance between the need for objectivity and subjectivity in port management*	AI systems for the port management sector should be developed balancing the need for objectivity and subjectivity in the decision-making process.	N (Overly generic)
	UR_W19	***Contextual Nuances:** AI developers of port management related AI tools should ensure the constant availability of their tools*	AI systems for the port management sector should be constantly available.	Y
	UR_W20	***Contextual Nuances:** AI developers of port management related AI tools should be able to take into account the complexity and different protocols and politics of individual ports*	AI systems for the port management sector should be able to model the structure, protocols and policies of each port.	Y
	UR_W21	***Contextual Nuances:** AI developers of port management related AI tools should ensure that their tools benefit the many, not the few*	AI systems for the port management sector should ensure that all stakeholders are equally (or proportionally) benefited.	Y
	UR_W22	***Contextual Nuances:** AI developers of port management related AI tools should consider how commercial interests can affect how ports function*	AI systems for the port management sector should model how commercial interests interact with the port's operation.	Y
	UR_W23	***Attitudes towards AI:** AI developers of port management related AI tools should consider how the work between the tool and the users becomes a collaborative effort to ensure users do not become too dependent on AI*	AI systems for the port management sector should foster a collaborative way of working, so that users are involved in the AI-supported decision-making process, as well as ensuring that users can still operate without the use of AI.	N (Not aligned with UC2 SUT)
	UR_W24	***Attitudes towards AI:** AI developers of port management related AI tools should take into consideration how their tools can interfere with the current cultural norms and established hierarchy in ports.*	AI systems for the port management sector should be able to adapt to the cultural norms and established hierarchy of each port.	N (Not aligned with UC2 SUT)
	UR_W25	***Attitudes towards AI:** AI developers of port management related AI tools should consider ways to encourage AI uptake from end-users not familiar with AI technologies.*	AI systems for the port management sector should facilitate users with limited or no experience with AI tools to use them in their line of work.	Y
Media	UR_W26	***Responsibility**: AI developers of media related AI tools should consider the way their tools deliver output to the journalists and fact-checkers to ensure that it is functioning as a support tool for human decision making rather than the tool making its own decisions.*	AI systems for the media sector should not function independently and should work as decision support tools assisting users to reach a decision.	Y
	UR_W27	***Responsibility**: AI developers of media related AI tools should be transparent about who is involved in the development of AI tools and what data is being used.*	The users of AI systems for the media sector need to be informed about the parties that have been involved in their development and the training datasets that have been used.	Y
	UR_W28	***Responsibility**: AI developers of media related AI tools should be clear about what the intended purpose and use of their AI tool is.*	The user should be aware about the purpose and use of the AI system.	Y
	UR_W29	***Transparency & Accuracy:** AI developers of media related AI tools should make it possible for users to examine the 'line of thought' going on in the AI tool to ensure trust.*	The AI system should ensure explainability and transparency of results, informing the user about how a recommendation has been produced, to foster trust amongst users.	Y

	UR_W30	***Transparency & Accuracy:*** *AI developers of media related AI tools should understand how transparency, accuracy and trust are closely connected in the work of journalists and fact-checkers.*	AI systems for the media sector should emphasise accuracy and transparency of results to foster trust.	Y
	UR_W31	***Transparency & Accuracy:*** *AI developers of media related AI tools should inform the end-user about the data used for training their AI tools.*	The users of AI systems for the media sector need to be informed about the training datasets that have been used.	Y
	UR_W32	***Contextual Nuances:*** *AI developers of media related AI tools should consider that most of work in media in heavily dependent on the context and real-time state of the world.*	AI systems for the media sector should be able to adapt their results to the everchanging context and state of the world.	Y
	UR_W33	***Contextual Nuances:*** *AI developers of media related AI tools should make it possible to still adhere to and support working with industry ethical standards and norms.*	AI systems for the media sector should adhere and promote the industry's ethical standards and norms.	Y
	UR_W34	***Contextual Nuances:*** *AI developers of media related AI tools should not make AI tools that are subjective and make moral judgements but rather focus on tasks were objectivity is the goal.*	AI systems for the media sector (i) should focus on tasks where objectivity is the goal and (ii) should refrain from making moral judgements and taking subjective decisions.	Y
	UR_W35	***Attitudes towards AI****: AI developers of media related AI tools should make their AI tools supportive in their functionality for journalists and fact-checkers.*	AI systems for the media sector should facilitate and optimise the work of journalists and fact-checkers.	Y
	UR_W36	***Attitudes towards AI:*** *AI developers of media related AI tools should consider how to balance time optimisation by enabling use of AI tools without impacting the quality of work.*	AI systems for the media sector should support time-optimisation while not sacrificing the quality of work.	Y

3.3.2. End-users' expectations regarding fairness, accuracy, and robustness in AI functionality.

In this section the end-users' expectations and concerns regarding fairness, accuracy, and robustness in AI systems used in their professional context are presented.

As documented in the following Table 9, the trustworthiness expectations (per Use Case and per Trustworthiness category) combine input stemming from (i) the 1st round of co-creation workshops as documented in the deliverable D3.1[25], (ii) the Use Case Description templates that were filled in by end-users engaged through the Use Case partners and (iii) user requirements from this document from Section 3.5. For each entry, the respective source, the categories that they fall within and their relevance to THEMIS 5.0 are listed. The assigned categories can be used as keywords to swiftly sort and filter the expectations for a specific category (e.g. Training datasets, Adaptation to the needs of end-users, Explainability, End-users' engagement, Bias, etc.). THEMIS 5.0 should consider these expectations while designing and implementing the THEMIS framework, whose ultimate goal is to help AI-users enhance an AI-system's trustworthiness in terms of their actual as well as perceived fairness, accuracy, and robustness.

The key points per trustworthiness category and Use case are listed below:

- **Fairness:**
  - Healthcare:
    - Consider individual patient needs
    - Equal treatment
    - Fair access for all users
  - Ports:
    - Account for diverse user needs
    - Transparency and bias avoidance
    - Explainability
    - Ensure user engagement and onboarding
  - Media:
    - Transparent about data sources

[25] THEMIS 5.0: D3.1 Report on citizen co-created design principles for THEMIS 5.0 Ecosystem of services

- Recognize biases, and present balanced perspectives
- Human-in-the-loop
- Diverse dataset

➢ **Accuracy:**

- Healthcare:
  - Training data quality
  - Improvement over time
  - Contextual factors affecting accuracy
  - Transparency allows for identifying errors and achieving precise support
- Media:
  - AI-user interactive collaboration and guidance
  - Explainability of results
  - Quality of training datasets

➢ **Robustness:**

- Healthcare:
  - Security measures
  - Adaptability to different environments
  - Compliance with existing healthcare standards
- Ports:
  - Data Security
  - Adaptability for error mitigation
  - 24/7 availability across devices
- Media:
  - Regulation compliance
  - Transparency
  - Smooth functionality
  - Accessible in various conditions
  - Continuous, reliable information delivery

Table 9: Users' Trustworthiness Expectations

T/W parameter	Sector	ID	Users' Trustworthiness Expectations	Source	Category	Relevant to THEMIS (Y/N)
Fairness	Healthcare	TE_1	Utilize applicable data in AI training and in AI use to accommodate the individual patient and the exact situations that AI is used in, to achieve fairness in treatment of all patients.	Contextual indicators D3.1	Training datasets	Y
		TE_2	Consider who are the end users (hospitals, professionals, patients), and how to adapt to their needs to ensure equal and fair access.	Contextual indicators D3.1	Adaptation to the needs of end-users	Y
	Port	TE_3	Considering the large number of different types of users in the port management sector is key to fairness. AI needs to consider the different tasks, commercial interests, the complexity of vessel prioritization, and existing principles like FIFO.	Contextual indicators D3.1	Adaptation to the needs of end-users	Y
		TE_4	Transparency in data used will enhance users' comprehension of AI and trust in AI, allowing them to identify biased data and avoid preferential treatment of certain companies. Enabling a fairer service to all customers, striving towards more neutrality in the sector.	Contextual indicators D3.1	Training datasets suitability, Explainability	Y
		TE_5	A gradual implementation of new AI tools is needed to give a fair chance to the different types of users to adapt to the use of new technologies and avoid alienating specific types of users or creating resistance towards new technology.	Contextual indicators D3.1	End-users' engagement	Y

		TE_6	The system should not have biases and should be able to explain its predictions. In a user-friendly manner, elucidate the key input variables of the model that exerted significant influence on the prediction's outcome.	UC2T_5 Perceived T/W aspects	Bias, Explainability	Y
		TE_7	The user should be presented with analytical results about bias: which are the favoured and which the discriminated groups.	UR_S8	Bias, Explainability	Y
	Media	TE_8	The disinformation in the media sector has constantly flowing and changing information which requires updates and continuous review, combined with AI built on a foundation of diverse data, media ethics, and legal compliance. To achieve fairness based on these requirements, the user needs to have insights into the foundation of the AI in use, explained to them in an understandable way.	Contextual indicators D3.1	Training datasets, Ethics, Legal Compliance Explainability	Y
		TE_9	Accessibility demands cover a wide set of features and ways of communicating from AI to the user to achieve fairness. There is both actions to be done in the AI development, but also in the education and guidance of AI users, to achieve fair accessibility.	Contextual indicators D3.1	Accessibility, User Engagement	Y
		TE_10	There is a demand for clarity of what information is used in an AI outcome and that the AI can both identify stereotypes and political influence and moderate its output to not reproduce bias and deliver fair results.	Contextual indicators D3.1	Explainability, Bias, Training Datasets	Y
		TE_11	Human guidance and intervention are needed to have a fair AI tool.	Contextual indicators D3.1	Human-in-the-loop	Y
		TE_12	AI-generated reports should also include information that contradicts the final verdict.	UC3T_5 Perceived T/W aspects	Explainability, Human-in-the-loop	Y
		TE_13	The sources must be balanced and from a variety of outlets, covering most of the political spectrum (from the Ethics Aspects column).	UC3T_5 Perceived T/W aspects	Bias, Training datasets	Y
Accuracy	Healthcare	TE_14	Transparency is needed for users to identify errors in AI tools and to be able to take responsibility for accuracy in AI output and support.	Contextual indicators D3.1	Explainability, Error correction, User Responsibility	Y
		TE_15	Accuracy in data is influenced by data generated by patients and healthcare professionals, research data, data used for training, and it depends on whether AI can improve data over time.	Contextual indicators D3.1	Training Datasets	Y
		TE_16	Expected accuracy will depend on the specific situation in which AI is used.	Contextual indicators D3.1	Adaptability	Y
		TE_17	AI needs to consider multiple patient factors to assist accurately in the decisions of healthcare professionals.	Contextual indicators D3.1	Input criteria	Y
	Port	TE_18	The complex and globally connected activities of ports require accuracy across many factors, from vessel information, ETA, and port operations like loading and unloading, and at the same time, be able to navigate data discrepancies e.g. weather forecasts.	Contextual indicators D3.1	Adaptability, Training Datasets, Human-in-the-loop	Y
		TE_19	Automatic retraining will allow AI to improve data accuracy over time and learn from commonly made errors caused by users and from improved data quality within the sector. This will not only increase accuracy over time but also efficiency, and reliability in their environment.	Contextual indicators D3.1	Training Datasets, Error correction, Retraining for Improvement	N (Automatic retraining is not foreseen)
		TE_20	ETA accuracy is key to the entire port operation, and therefore, there needs to be a high level of accuracy for ETA.	Contextual indicators D3.1	Accuracy	Y

		TE_21	The system needs to have a high precision. The AI system should estimate its prediction accuracy by selecting the actual accuracy of models under comparable conditions.	UC2T_5 Perceived T/W aspects	Accuracy	Y
	Media	TE_22	To achieve accuracy in the output from AI, AI should be able to guide the user in how to prompt in the right way and accommodate a collaborative interaction between user and AI to make sure requests from users are understood accurately, hence AI delivering more accurate results.	Contextual indicators D3.1	User engagement, Human-in-the-loop.	Y
		TE_23	Regardless of the accuracy of AI output, the tool should be able to present and explain the level of accuracy.	Contextual indicators D3.1	Explainability	Y
		TE_24	AI needs to work with standards for accuracy, such as thresholds of the amount and quality of data, protocols for source reputation, periodic reviews of accuracy of output, and media ethics.	Contextual indicators D3.1	Training datasets, Testing and Validation, Ethics	Y
Robustness	Healthcare	TE_25	Security measures are closely linked to robustness and cover the wish to be protected against hacker attacks, performance issues, human errors, natural disasters, and political conflicts.	Contextual indicators D3.1	Security, Resilience, Error handling	Y
		TE_26	AI is robust if its performance considers the environment and external factors in which the healthcare professional is using the tool.	Contextual indicators D3.1	Adaptability	Y
		TE_27	If AI tools use existing legislations and standards from the healthcare sector, it will create trust and quality control would provide robustness.	Contextual indicators D3.1	Standard procedures, Quality control	Y
		TE_28	AI robustness in the healthcare sector requires a broad view of users, operating AI in their homes and in various professional settings.	Contextual indicators D3.1	Adaptability, accessibility	Y
	Port	TE_29	Applying safety and security measures to AI is essential for robustness. In the port management sector safety measures are needed to detect anomalies, have secure hosting, avoid hacker attacks, provide data backup, and data protection, and have a plan B in case of security breaches.	Contextual indicators D3.1	Safety, Security, Resilience, Data safety, Data protection	Y
		TE_30	AI needs to apply some kind of flexibility and ability to mitigate errors, ensuring a level of adaptability is needed for achieving robustness in a work environment with many changes.	Contextual indicators D3.1	Flexibility, Adaptability, Error handling	Y
		TE_31	Availability to operate 24/7 and on various devices is key for AI to be robust when used in port management.	Contextual indicators D3.1	Operational availability, Accessibility	Y
		TE_32	Robustness is understood as delivering a solid data foundation with a variety of algorithms in use, combined with data transparency towards the user and reliability in the accuracy of the information delivered to the user.	Contextual indicators D3.1	Training datasets, Multiple solutions, Reliability, Transparency	Y
		TE_33	The system needs to have high reliability. The system needs to function in real-time 24/7.	UC2T_5 Perceived T/W aspects	Operational availability	Y
	Media	TE_34	Robustness can be achieved by applying existing regulations and legislation, combining it with additional attempts at of regulations, e.g. context-specific code of conduct.	Contextual indicators D3.1	Regulatory framework, Legal Framework, Code of conduct	Y
		TE_35	Robust operation is a twofold approach of training and implementing appropriate AI functionalities which guide the user to appropriate and safe use.	Contextual indicators D3.1	Training, User engagement, User guidance	Y
		TE_36	Enabling user and AI collaboration requires transparency and user-friendliness, which will allow the user to trust the robustness of the system.	Contextual indicators D3.1	Transparency, User engagement	Y

		TE_37	Accessing AI functionality under various conditions, e.g. access in offline mode, is also perceived as robustness.	Contextual indicators D3.1	Accessibility, Resilience	Y
		TE_38	An aggregated stream that runs smoothly, provides information regularly and ensures relevancy of the suggested items is an essential part of any process that has technical solidity.	UC3T_5 Perceived T/W aspects	Operational availability, Reliability, Accuracy	Y
		TE_39	Tools that work well all the time	UC3T_5 Perceived T/W aspects	Resilience	Y

3.3.3. End-users' AI-Trustworthiness Preferences

3.3.3.1 Rank and importance of trustworthiness parameters

In this section, an initial analysis is presented of the end-users' preferences against the three major AI trustworthiness categories that THEMIS 5.0 addresses, namely Fairness, Accuracy and Robustness. The end-users' preferences were collected via interviews and questionnaires[26] from a small sample of domain experts/end-users (four from each use case) engaged through the Use Case partners, who were asked to (i) rank the trustworthiness parameters according to the needs of their sectors, (ii) to assign a number from 1-10 indicating their importance, and (iii) to provide the minimum level of each characteristic that they could tolerate using H (High level), M (Medium level), and L (Low level). The results are presented in the following table (Table 10). These preferences against trustworthiness aspects will be used as pointers for the initial steps of the development of the relevant personalization components and will be used as a basis for the collection of preferences from a substantial sample of end-users/experts during the co-creation phases B (Living Labs) and C (Pilots).

Table 10: End-users' trustworthiness preferences

Sector	Healthcare			Port Management												Media					
Role	Healthcare professionals[27]			Vessel Traffic Service Supervisor			Port Authority Manager			Port Authority AI developer			Transportation Traffic Manager			Business Development Manager			Fact Checkers[28]		
Preference	Rank	Importance	Min. Level	Rank	Importance	Min. Level	Rank	Importance	Min. Level	Rank	Importance	Min. Level	Rank	Importance	Min. Level	Rank	Importance	Min. Level	Rank	Importance	Min. Level
Fairness	3	8	-	2	9	H	3	6	H	3	7	M	3	5	M	1	10	H	1	10	H
Accuracy	2	9	H	1	10	H	1	10	H	1	10	H	1	9	H	2	8	-	2	9	H
Robustness	1	10	H	3	9	M	2	8	H	2	9	H	2	7	M	3	6	-	3	7	M

Based on this initial sample, it seems that for the Healthcare sector **robustness** comes first, although it's almost as important as **accuracy** that comes close second (10 to 9). For the Port management sector **accuracy** is the most sought-after trustworthiness characteristic, while **robustness** and **fairness** come second and third, respectively. As for the News media sector, **fairness** is the most critical characteristic, followed by **accuracy** and **robustness**. Based on this initial sample, it seems that the three use-cases have a different prioritization of the trustworthiness characteristics of AI systems, although all aspects are graded as important, having values that range from 5-10/10.

These findings are for the most part aligned with the findings reported in deliverable D3.1[29] where, based on the input collected from the co-creation Phase A workshops, it appears that the most important principle for trustworthy AI (i) in

[26] Questionnaires for URs
[27] Collective input from 4 UC1 representatives (general practitioner, internal medicine specialist, Technical Lead, AI R&D)
[28] Collective input from 3 Fact Checkers
[29] THEMIS 5.0 D3.1 Report on citizen co-created design principles for THEMIS 5.0 Ecosystem of services

the health sector is robustness (49%), followed by accuracy (32%) and fairness (25%), (ii) in the port management sector is accuracy (59%) while the least important principle is fairness (15%) - making robustness (27%) fall right between the two, and (iii) in the media sector, accuracy stands out as the most crucial parameters for trustworthy AI, with 53% prioritizing it. Fairness and Robustness follow closely with 25% and 22% of the participants respectively. These findings differ only in the order of the 1st and 2nd trustworthiness characteristics for the media sector as in the co-creation workshops accuracy was selected by most participants as being the most important. This discrepancy can be partially attributed to the fact that the interviews took place later in the development of the project, than the relevant co-creation workshop, and interviewees were provided with more elaborated explanatory material and contextual information. In addition, the interviewees were domain experts with previous experience in using AI tools to identify fake news, whose perspective towards trustworthy AI, may be slightly different from that of the participants in the co-creation workshops.

3.3.3.2 Consideration for AI systems' Trustworthiness Assessment

For the collection of the user requirements and the end-users' considerations, expectations and preferences with regards to the optimisation of AI systems in terms of fairness, accuracy and robustness, the definitions and examples for each use case listed in Table 11, presented next, were communicated to the end users.

Table 11: Working definitions for fairness, accuracy and robustness

What is "**FAIRNESS**" for an AI System? *[The following definition of an AI system's fairness was presented to end-users as contextual information.]*	The principle of justice in trustworthy AI also underlines fairness and the prevention of discrimination (Thiebes et al., 2021). For Kaur et. al (2022), fairness of an AI system "ensures that there is an absence of any discrimination or favoritism toward an individual or a group based on any inherent or acquired characteristics that are irrelevant in the context of decision making". (see section 2.6.3.2 above)
Health: What does it mean for an AI-based personalised risk prediction system for pancreatic cancer to have low fairness?	It can mean that: a) the system may make correct predictions (or tends to identify as high/medium/low risk) only for people from particular groups (based on factors irrelevant to pancreatic cancer, e.g., based on religious belief), b) healthcare professionals with inabilities may not be able to operate the risk prediction system.
Port: What does it mean for an ETA system to have low fairness?	It can mean that: a) the system has biases against Shipping Companies or types of cargo, etc. ii) Also, it may mean that the system cannot be accessed by certain types of end-users (e.g., persons with inabilities).
Media: What does it mean for a fake news and hate speech detection system to have low fairness?	It can mean that the fake news detection tool may flag fake news only from particular agendas (political/religious/ etc) and omit the rest. Also, it may mean that the tool cannot be accessed by journalists with inabilities.
What is "**ACCURACY**" for an AI System? *[The following definition of an AI system's accuracy was presented to end-users as contextual information.]*	Accuracy may be understood as the "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true" (NIST, 2023). For AI-systems it may be important to note that this particularly concerns how well the AI system "do on new (unseen) data compared to data on which it was trained and tested?" (Wing, 2021). (see section 2.3.4.1 above)
Health: What does it mean for an AI-based personalised risk prediction system for pancreatic cancer to have low accuracy?	A low accuracy for a such system may mean that: a) a patient may be wrongly predicted as being in low (or medium) risk for pancreatic cancer while s/he is in high risk, b) a patient may be wrongly predicted as being in high (or medium) risk for pancreatic cancer while s/he is in low risk.
Port: What does it mean for an ETA system to have low accuracy?	It can mean that the estimated time of arrival of a vessel is significantly different from its actual time of arrival.
Media: What does it mean for a fake news and hate speech detection system to have low accuracy?	It can mean that: a) Content in news articles that is flagged as fake (or that contain hate speech) may not be indeed fake (or that contain hate speech), b) articles that contain fake news (or hate speech) may not be flagged as such.
What is "**ROBUSTNESS**" for an AI System? *[The following definition of an AI system's robustness was presented to end-users as contextual information.]*	Robustness concerns the AI-systems' ability to perform as expected under varying conditions. For example, Wing (2021) notes that robustness concerns the "sensitivity of the system's outcome to a change in the input". NIST (2023) accentuate that robustness "is a goal for appropriate system functionality in a broad set of conditions and circumstances, including uses of AI systems not initially anticipated". Kaur et al. (2022) also specifically note that robustness concerns system's ability to deal with error at any point in the lifecycle and that the system is resilient to attacks. (see section 2.3.4.2 above)

Health: What does this mean for an AI-based personalised risk prediction system for pancreatic cancer?	It can mean that: a) the risk prediction system works properly besides any malicious attacks (e.g., deliberately end-user gives wrong feedback). 2) When exposed to different types of input, the system can still work properly. For instance, the system responds correctly for 43-year-old patients even though it is not trained with data specifically from 43-year-old patients, but from patients of similar ages.
Port: What does it mean for an ETA system to have low robustness?	It can mean that: a) it is vulnerable to malicious attacks (e.g., deliberately end-user gives wrong feedback), b) it does not function 24/7, c) it does not behave as expected in input different from the one that has been trained on. For instance, it does behave properly for a specific vessel type, although it has been trained on similar vessel types.
Media: What does this mean for a fake news and hate speech detection system?	1) The fake news/hate speech detection tool works properly besides any malicious attacks (e.g., deliberately end-user gives wrong feedback). 2) When exposed to different types of input (e.g., website articles written differently or in different topics) from the ones that has been trained on, the detection system can still works properly.

Moreover, in Table 12 below, additional end-users' considerations for measuring an AI system's trustworthiness in terms of fairness, accuracy and robustness are listed. These considerations originate from several sources including the interviews and the questionnaires and the users' trustworthiness expectations (Table 12). For each consideration, the table lists the respective sources.

Table 12: Trustworthiness measuring considerations

T/W characteristic	Use Case	Trustworthiness measuring considerations	Source	Role
Fairness	Media	To ensure unbiased and equitable treatment of all content and users, regardless of their background or characteristics. It requires transparency, accessibility, regular auditing, diverse training data, and robust accountability mechanisms to prevent discrimination and promote trust in the system.	Questionnaires for Collection of User Requirements	Fact Checkers
		The sources must be balanced and from a variety of outlets, covering most of the political spectrum.	Users' Trustworthiness Expectations, TE_13, Table 9	-
Accuracy	Healthcare	Accuracy calculation should apply to different weights to different types of errors, emphasising the correct identification of the critical risk categories ("high risk" and "very high risk"). Accuracy should also be based on other analyses (e.g. confusion matrix, mean Average Precision (mAP), precision-recall and F1 score, etc.)	Questionnaires For Collection of User Requirements	Health care professionals
	Port	The accuracy measurement should be based on the AI system's previous predictions vs observations. The comparison period and conditions should be customisable by the user	UR_S6	-
		The AI system should estimate its prediction (of) accuracy by selecting the actual accuracy of models under comparable conditions.	Users' Trustworthiness Expectations, TE_21, Table 9	-
	Media	Accuracy measurement and tolerance varies between false positives and false negative	UR_Q13	
		False Positives: Content is incorrectly flagged as fake news or hate speech. False Negatives: Fake news or hate speech content is not flagged	Questionnaires For Collection of User Requirements	Business Development Manager, Fact Checkers
Robustness	Healthcare	Robustness is the ability of the model to generalize and to provide similar suggestions for similar cases.	Questionnaires For Collection of User Requirements	Health care professionals
	Port	Different ETA calculation for vessels with similar characteristics	Questionnaires For Collection of User Requirements	Vessel Traffic Service Supervisor

3.3.4. *Trustworthiness-related Business, Ethical and Legal Risks*

In this section, the trustworthiness-related business and ethical risks are documented, also based input collected via interviews and questionnaires[30] filled in by domain experts/end-users engaged through the Use Case partners. THEMIS should consider the domain-specific risks that are associated with each trustworthiness characteristic as a pointer for the design and development of the solutions that (i) will help users to build an AI system's socio-technical model and (ii) will support users to take an informed decision also considering the impact of each potential trustworthiness optimization measure on the business targets and objectives. Hereinafter, Table 13, Table 14 and Table 15, present the business, ethical and legal risk that according to the end-users are related to the AI system's fairness, accuracy and robustness.

Table 13: Business, Ethical and Legal Risks connected to the fairness of the AI system.

Fairness				
Sector	**Role**	**Business Risks**	**Ethical Risks**	**Legal Risks**
Healthcare	Health care professionals	-	Unequal distribution of accurate and non-accurate results between patient groups (age, race, gender, obesity, type 2 diabetes, smoking and drinking habits, diet, etc.).	-
Port	Vessel Traffic Service Supervisor	-	Unjustified and unjust vessel prioritisation could cause increased costs to achieve the deserved priority.	-
	Port Authority Manager	■ Slow adoption of the ETA prediction by stakeholders ■ Economic Losses ■ Reputation damage ■ Legal issues	Unjust treatment of stakeholders (company, terminal destination, etc…)	-
	Port Authority AI developer	■ Slow adoption of the ETA prediction by stakeholders ■ Economic Losses ■ Reputation damage ■ Legal issues	Unjust treatment of stakeholders (company, terminal destination, etc…)	
	Transportation Traffic Manager		Unfair treatment of specific group of port calls ■ Specific group of transport orders linked to these port calls that a certain transport company is mostly covering.	
	Use Case Partners	■ Shipping companies loose trust in the system. ■ Shipping agencies risk business with Transport companies ■ Decrease of the resource usage in Transport Companies.	■ Discrimination against specific shipping companies ■ Delays in the delivery of essential cargo ■ Suboptimum port management causes increased vessels' and trucks' fuel consumption.	
Media	Business Development Manager, Fact Checkers	■ Reputational damage ■ Loss of user trust ■ Legal challenges ■ Operational inefficiencies ■ Low competitive advantage ■ Low advertiser confidence ■ Low employee morale ■ Troubles with stakeholder relations	■ Discrimination and bias ■ Violation of human rights ■ Erosion of trust ■ Negative social impacts ■ Breaches of moral responsibility ■ Threats to democratic processes ■ Psychological harm	■ Violation of anti-discrimination laws ■ Breaches of data protection and privacy regulations ■ Consumer protection issues ■ Contractual breaches ■ Defamation claims ■ Increased regulatory scrutiny ■ Litigation

[30] Questionnaires for URs

		■ Decreased diversity and inclusion efforts.	■ Misinformation ■ Lack of accessibility.	■ Intellectual property disputes ■ Complications in mergers and acquisitions ■ Employment law violations.
	Use Case partners	■ Increased cost of infrastructure to make the system more accurate. ■ Increased operational costs.		

Table 14: Business, Ethical and Legal Risks connected to the accuracy of the AI system.

ACCURACY				
Sector	**Role**	**Business Risks**	**Ethical Risks**	**Legal Risks**
Healthcare	Health care professionals	■ Affect the quality of the service ■ Impact the business and the hospital image ■ AI users (doctors) opt for a hospital with better AI tools.	■ Conflict and stress in case of disagreement between the doctor and the AI. ■ Encourage transfer of decision to the AI tool.	■ Legal responsibility for medical decisions lies with the doctor ■ Doctors can sue the hospitals for offering a misleading tool.
	Use Case Partners	■ Affect the quality of the service ■ Medics are not confided to integrate the system's suggestions in their work	■ Patients are not following the recommended preventive program	
Port	Vessel Traffic Service Supervisor	■ Berthing schedule disruption ■ Inefficient utilization of port resources ■ Impact other port calls ■ Unnecessary waiting and idle times.	*(moved to Robustness)*	
	Port Authority Manager	■ Lack of trust by the end users of the information provided by the PCS in the prediction provided by the model ■ Confusion among PCS stakeholders ■ Negatively influence decisions based on the ETA ■ Traffic Disruptions and congestion.	■ Suboptimum traffic management ■ Port operations' disruption ■ Negative impact on the environment.	
	Port Authority AI developer	■ Lack of trust of the end users in the prediction provided by the model ■ Stakeholders ignore or stop using the model.	Low accuracy could be biased against specific shipping companies or type of vessel.	
	Transportation Traffic Manager	■ Inefficient use of companies' own resources ■ Assigned drivers endure long waits and cannot be used in other trips ■ Increased costs due to subcontracting other trips.	■ Increased fatigue and stress compromising the drivers' safety and well-being. ■ Increased operational costs due to inefficient use of companies' resources (time and fuel). ■ Increased environmental impact. ■ Negative work environment and labour conflicts if drivers feel they are not adequately valued.	Longer working hours conflicting with labour laws.
	Use Case Partners	■ Shipping agencies risk business with Transport companies ■ Decrease of the resource usage in Transport Companies.		

Media	Business Development Manager	■ Reputational Damage (False Positives / False Negatives) ■ User Trust and Retention (Decreased Trust / User Exodus) ■ Operational Costs (Increased Moderation Costs / Resource Allocation) ■ Advertiser Confidence and Revenue (Advertiser Trust / Revenue Loss) ■ Competitive Disadvantage (Market Position / Innovation Lag) ■ Brand Image (Public Perception / Crisis Management).	■ Censorship and Freedom of Expression ■ Misinformation ■ Harm to Individuals and Groups ■ Bias and Fairness ■ Trust and Transparency ■ Responsibility and Accountability ■ Social Cohesion ■ Informed Decision-Making ■ Psychological Well-being ■ Long-term Ethical Impact.	■ Non-Compliance with Regulations ■ Defamation and Libel ■ Failure to Protect Users ■ Discrimination and Bias ■ Consumer Protection Laws ■ Intellectual Property ■ Regulatory Fines and Sanctions ■ User Lawsuits and Class Actions ■ Contractual Obligations.
	Fact Checkers	■ Overhead of work for journalists/fact checkers to recheck each output. ■ Reliability issues ■ Financial issues.	■ Spreading of fake news ■ An undetected fake news item may cause a propagation of multiple other fake news.	
	Use Case partners	■ Increased cost of infrastructure to make the system more accurate ■ Increased operational costs.		

Table 15: Business, Ethical and Legal Risks connected to the robustness of the AI system.

Robustness				
Sector	**Role**	**Business Risks**	**Ethical Risks**	**Legal Risks**
Healthcare	Health care professionals	■ Reduced trust in using the AI system could lead to ignoring the model output ■ AI system users spend more time to assess the risk.	■ Increased stress and uncertainty for AI users due to unstable results and unpredictable behaviour ■ Reduced quality of service for patients.	
Port	Vessel Traffic Service Supervisor	■ Economic Losses ■ Reputation damage ■ Legal issues	■ Tool vulnerable to inappropriate use. ■ Optional ETA update can be exploited by stakeholders to maintain previously acquired slot.	-
	Port Authority Manager	■ Reputation damage due to compromised system ■ Operational disruptions at all levels in the logistic chain.	Unfavourable or favourable treatment of specific shipping agencies due to attacks.	-
	Port Authority AI developer	■ Economic losses due to an outage or a system failure ■ Loss of efficiency in vessel traffic control.	Unfavourable or favourable treatment of specific shipping agencies due to attacks.	-
	Transportation Traffic Manager	■ Implementation of traditional traffic management due to interruptions in the ETA prediction service ■ Economic loss ■ Suboptimum usage of companies' resources ■ Reduced clients' satisfaction.	■ Inefficient allocation of resources ■ Favourable for specific type of drivers ■ Unnecessary traffic jams within the port. ■ Negative working environment	Longer working hours conflicting with labour laws.
	Use Case Partners	■ Shipping agencies risk business with Transport companies		

		■ Decrease of the resource usage in Transport Companies.		
Media	Business Development Manager, Fact checkers	■ Operational disruptions ■ Loss of user trust ■ Reputational damage ■ Legal and regulatory challenges ■ Financial implications ■ Competitive disadvantages ■ Security risks ■ User experience issues ■ Increased operational complexity ■ Impact on partnerships.	■ Inconsistent moderation ■ Unfair treatment ■ Erosion of trust and accountability ■ Freedom of expression violation ■ Vulnerability to exploitation ■ Social and psychological harm ■ Lack of inclusivity ■ Compromised ethical integrity ■ Harm democratic processes ■ Injustice and inequality.	■ Regulatory non-compliance ■ Liability for harmful content ■ Infringement of rights ■ Consumer protection issues ■ Regulatory scrutiny ■ Intellectual property disputes ■ Impacts on partnerships ■ Employment law concerns.

3.4. System Requirements from THEMIS 5.0 Project Objectives

In this section, we describe briefly the system requirements as defined in the THEMIS 5.0 project objectives. The reader is referred to Table 23 in Appendix A.2 for a more detailed view of the system requirements translated from THEMIS 5.0 DoA. In the following, each system requirement is accompanied by the respective references (indicated with IDs SR_X) to Table 23.

First, we describe the system requirements that concern the overall THEMIS 5.0 platform.

THEMIS 5.0 must:

1. Be a micro-service-based platform supporting cycles of design, deployment, evaluation, and tuning of experiment variants of AI systems. These services will be cloud-based [SR_12].
2. Interact with the user via an AI-driven conversational agent [SR_01].
3. Facilitate dialogues adjusted to users' preferences and traits [SR_07].
4. Comply with the European legal and ethical framework [SR_03].

For readability, we describe the system requirements per stage of the Trustworthiness Optimization Process (proposed in Section 2.8.2). However, the distribution of requirements among the stages is subject to changes as the conceptual modelling of the THEMIS 5.0 ecosystem (T2.3) is still under development.

In the IDENTIFY stage, THEMIS 5.0 must:

1. Capture users' role (job role) and *persona*, i.e., the preferences, requirements, objectives capabilities, motives and behavioural patterns, decision support needs, and legal, moral and ethical principles [SR_14, SR_21, SR_22].
2. Capture business objectives and KPIs that the optimised AI system will need to support [SR_22].
3. Enable the user via a GUI and a chatbot to create a qualitative model of the socio-technical environment using a combination of high-level knowledge specification and machine learning. In this GUI the user will insert [SR_18, SR_19, SR 24, SR_47]:
    - all the possible actions that the AI system can recommend,
    - all the KPIs that might be affected by any of the possible actions,
    - any external factors that might influence the actions or the actions' effect on the KPIs, and
    - how the actions and the external factors affect the KPIs through pairwise relations
4. Generate a qualitative model based on user input [SR_24].
5. Generate a quantitative model or a simulation model from the qualitative model [SR_24].
6. Receive feedback from the users concerning their trust in the socio-technical model [SR_47].
7. Update either the qualitative or the quantitative model of the socio-technical environment based on user feedback [SR_47].
8. Enable the user to model an AI system in the context of a socio-technical environment in accordance with decision support needs and moral values [SR_56].

In the ASSESS stage, THEMIS 5.0 must:

1. Perform a personalised AI Trustworthiness assessment of the AI system based on the knowledge captured in the IDENTIFY stage (1., 2.) [SR_13, SR_14, SR_54, SR_20]

2. Consider vulnerabilities related to AI system fairness and technical accuracy and robustness, as well as to the EU legal framework for trusted AI.
3. Perform a personalised AI Trustworthiness assessment based on at least 50 anomaly detection indicators and metrics [SR_25, SR_26]:
    - The fairness bias indicators will be defined according to NIST, including (a) systemic bias, (b) computational bias and (c) human bias [SR_29].
    - Technical accuracy and robustness metrics will be able to adapt to the severity of risks in the business environment as well as to the users' needs and preferences [SR_30].
    - The indicators must be related to the embedded socio-technical environment.
4. Forecast risks and vulnerabilities based on users' persona and preferences [SR_03, SR_39].
5. Be dynamic and be updated based on the implementation (or the projection) of trustworthiness optimisation measures [SR_28].

In the EXPLORE stage, THEMIS 5.0 must:

1. Provide solutions that will improve the trustworthiness of the AI System
2. The provided solutions will be explainable and transparent by providing the users with [SR_07, SR_36, SR_41, SR_42]:
    - input criteria used in the decision support process (training data sets, AI models, algorithms);
    - the output of that process; and
    - the perceived causal relationship between input and output, taking into consideration (i) the characteristics of the human recipients of decision support, (ii) the context and circumstances that triggered the decision support, and (iii) the user-defined objectives.
3. Identify where human-in-the-loop and human-in-command are needed and ensure human involvement in the AI-based decision support [SR_35].
4. Analyse alternative algorithmic approaches and metrics for evaluating the solutions produced by multi-actor, multi-objective decision-making AI systems [SR_27].
5. Explore the adoption of negotiation strategies for the selection between multiple potential solutions, resulting from the different actors and objectives there may be [SR_37].
6. Inform the user beforehand of the impact of the optimisation measures on the established business-related KPIs [SR_38].
7. Allow users to choose when and if a trustworthiness optimisation measure will be applied to the AI system under test [SR_02, SR_48].
8. Feature a fully equipped sandbox to design and deploy experiment variants and evaluate them against defined sets of criteria, through a no-code Big Data Analytics as a Service cloud-based designer (Interface). This sandbox must provide access to datasets for experimentation, Open APIs for experiment development, ML/DL algorithms models, advanced analytics, visualization configurations, human/machine interactions schemas, as well as the legal framework for trusted AI [SR_33, SR_34]. A ModelOps approach should be followed for the operationalisation of the produced experimental variables [SR_49, SR_51].

In the ENHANCE stage, THEMIS 5.0 must:

1. Implement enhancement in improvement continuous cycles. THEMIS 5.0 must monitor AI systems' trustworthiness optimisation at each cycle by calculating quantitative and qualitative KPIs, benchmarking and measures on progress monitoring based on a defined set of criteria on the various process assets (AI models, algorithms, datasets, as well as specific configurations).
2. Use feedback from users on the produced results to fine-tune the implemented AI models [SR_02].
3. Ensure that the accuracy of the AI systems that have gone through one (or several) trustworthiness optimisation cycles will reach at least 70%.

3.5. User Requirements

In this section, we present the broad user requirements collected from both channels: the co-creation workshops and the use case partners. In particular, Table 16 includes the user requirements collected from:

i) the 2nd co-creation workshop (held in June with AI-users and business stakeholders – ID code UR_C[X]),
ii) the user stories distributed to the use case partners and technical counterparts – ID code UR_S[X]),
iii) the questionnaires distributed to the use case partners and technical counterparts -ID code UR_Q[X]).

After merging the similar ones, we concluded with 81 unique user requirements. As in Section 0, we have distributed the requirements to the four stages of the Trustworthiness Optimization Process (proposed in Section 2.8.2). However, the distribution of requirements among the stages is subject to changes as the conceptual modelling of the THEMIS 5.0 ecosystem (T2.3) is still under development.

The ID codes in Table 16 indicate the corresponding requirement as recorded in the respective table in APPENDIX A (Table 24, Table 26, Table 27). For readability, we simplified the various roles to [Use Case] DE, for domain experts and [Use Case] AI, for AI developers. Also, we have identified the requirements that have already been mentioned in DoW.

For the prioritization, we utilized the MoSCoW method, as it is one of the most common methods for software requirements prioritization (Achimugu et al, 2014). The term MoSCoW is an acronym for "Must have" (Mo), "Should have" (S), "Could have" (Co), and "Won't have" (W), each denoting a level of priority:

i) "Must have" defines the requirements that must be included in the final product. In this case, all requirements mentioned in DoA are indicated as such.
ii) "Should have" defines high-priority requirements that should be included, if possible, within the delivery time frame.
iii) "Could have" requirements are desirable or nice to have requirements and could be included without incurring too much effort or cost.
iv) "Won't have" requirements are those recorded requirements, which, however, are out of the scope of the THEMIS 5.0 project.

The requirements have been prioritized by WP2 and WP4 partners, but this process will be finalised within T4.1.

Finally, the prioritization of requirements referring to duration is denoted as "NA" (not applicable) as it is too early to estimate the duration of the various stages.

Table 16 Summary of User Requirements

ID	User Requirement	User Req. ID	Role	System Requirement from DoA	Prioriti-zation
			OVERALL		
UR_1	THEMIS 5.0 interactions to be based on user's familiarity with AI.	UR_S1, UR_C13, UR_C38, UR_C39 UR_Q11, UR_Q23 UR_Q10	Health DE, Media DE	SR_07. THEMIS conversational interface must (i) adjust to the users' preferences and traits (....) SR_20. THEMIS personalised trustworthiness assessment will be based (i) on the users' (...) Technical awareness (...)	MUST HAVE
UR_2	Provide to the user a user-manual with explanations of the various functionalities of THEMIS 5.0	UR_Q1, UR_Q12	Media DE	-	SHOULD HAVE
UR_3	Provide to the user a user-manual with explanations of the terminology used by THEMIS 5.0.	UR_Q1, UR_Q12	Media DE	-	COULD HAVE
	THEMIS 5.0 interactions to be based on user's:			-	
UR_4a	-Attitude towards AI (to provide appropriate arguments)	UR_C22, UR_C44	Port AI, Health DE	-	COULD HAVE
UR_4b	- Role within organization (to ensure alignment with organization guidelines).	UR_C3, UR_C4, UR_C21	Media DE, Health DE	-	COULD HAVE
UR_4c	-For clarity, it is advised that the user will select their role from predefined list	UR_C4	Health DE	-	COULD-HAVE
UR_5	The interaction with the chatbot to be text-based	UR_C1	Health DE	-	MUST HAVE
UR_6	The chatbot to ask the same questions with multiple different	UR_C2	Health DE	-	COULD-HAVE

	ways (with paraphrases) to ensure clarity.				
UR_7	THEMIS 5.0 users can verify that functionalities are in line with organizational security guidelines	UR_C39 UR_C42 UR_Q9	Media DE	-	SHOULD HAVE
UR_8	THEMIS 5.0 users can verify that functionalities are in line with organizational ethical guidelines	UR_C41	Media DE	-	SHOULD HAVE
UR_9	THEMIS 5.0 platform can be reached and used by laptops, computers and smartphones	UR_C43	Health DE Media DE	-	SHOULD HAVE
IDENTIFY					
	The chatbot to collect information about:			-	
UR_10a	-organizational ethical guidelines that the end-user must comply with.	UR_C7	Media DE	SR_22. THEMIS conversational interface must be able to capture (…) the users' ethical values	MUST HAVE
UR_10b	-organizational security measures	UR_C8, UR_C24	Media DE Port AI	-	COULD HAVE
UR_10c	-end-user's requirements with regards to trustworthiness characteristics	UR_C9	Media AI Port AI	SR_39. Taking into account the users' profile and preferences, THEMIS will (i) forecast risks and vulnerabilities	MUST HAVE
UR_10d	-end-user's climate concerns	UR_C10	Media DE	-	COULD HAVE
UR_10e	-where in the work process is the end-user using the SUT and the purpose of use.	UR_C11 UR_C12	Media DE	SR_22. THEMIS conversational interface must be able to capture (i) the users' decision support needs, (…) objectives and KPIs that the optimised AI system will need to support.	MUST HAVE
UR_10f	-organisation's rules and established procedures	UR_C23	Port AI	-	COULD HAVE
TRUSTWORTHINESS PREFERENCE PREDICTION					
UR_11	To first make the 20 most crucial questions for a quick start, and later refine the persona if necessary	UR_C14	Health DE	-	COULD HAVE
UR_12	Interview the end-users only once, in the initialization of the platform and then update only if necessary (e.g., change of role)	UR_C17 UR_C18	Health DE	-	SHOULD HAVE
UR_13	Display the prediction of trustworthiness preferences to the user	UR_Q2	Health DE, Port DE, Media DE	-	MUST HAVE
UR_14	Enable the user to create various personas, based on role.	UR_C19	Health DE	-	COULD HAVE
UR_15	Enable the organization to define the same trustworthiness preferences among all employees (for the same SUT).	UR_C20 UR_Q14	Port DE, Media DE	-	COULD HAVE
UR_16	Enable users to give feedback on the displayed predictions of preferences	UR_Q3	Health AI, Media DE	-	COULD HAVE
UR_17	Provide users with explanations about the prediction of trustworthiness preferences, upon request.	UR_Q5	Health AI Port DE Port AI Media AI	SR_01. THEMIS (..) should interact with the user via an AI-driven conversational interface that will be able to explain in the necessary level of detail the outputs of AI systems. *#Comment: Although SR_01 refers to the system under test, as THEMIS 5.0 platform is also an AI System that the user needs to trust we conclude that it*	MUST HAVE

				refers to the THEMIS 5.0 platform as well.	
UR_18	Enable users to adjust the estimated trustworthiness preferences.	UR_Q6	Port DE Port AI	-	SHOULD HAVE
UR_19	Enable users to copy the preferences of users from the same organization.	UR_Q7	Health AI	-	COULD-HAVE
UR_20	Enable users to review their historical trustworthiness preferences.	UR_Q8	Health DE, Port DE, Port AI, Media AI	-	SHOULD HAVE
UR_21	The identify stage should not last more than 5 mins	UR_Q16	Fact checkers	-	NA
TRUSTWORTHINESS ASSESSMENT					
UR_22	Trustworthiness assessment must take into account end-user's trustworthiness requirements.	UR_C25	Media DE	SR_39. Taking into account the users' profile and preferences, THEMIS will (i) forecast risks and vulnerabilities	MUST HAVE
UR_23	Consider for the trustworthiness assessment the purpose of use of the SUT & where in the work process the SUT is used.	UR_C26 UR_C27	Health DE Media DE	SR_25. (...) trustworthiness assessment which will be based on (…) (iii) the embedding socio-technical environment. SR_24. THEMIS should (i) feature a GUI that will allow the users to create a qualitative model of the socio-technical environment by inserting (i.a) all the possible actions that the AI system can recommend , (i.b) all the KPIs that might be affected by any of the possible actions, (i.c) any external factors that might influence the actions or the actions' effect on the KPIs, and (i.d) how the actions and the external factors affect the KPIs through pairwise relations; THEMIS should be able to (ii) generate a qualitative model.	MUST HAVE
UR_24	Present AI system's trustworthiness assessment to the users via a chatbot-based interface, that is initiated by them and is tailored to their preferred level of detail.	UR_S2 UR_S3	Port DE	SR_01. THEMIS cloud-based services must interact with the user via an AI-driven conversational interface that will be able to explain in the necessary level of detail the outputs of AI systems. *#Comment: Although SR_01 refers to the system under test, as THEMIS 5.0 platform is also an AI System that the user needs to trust we conclude that it refers to the THEMIS 5.0 platform as well.*	MUST-HAVE
UR_25	Present initially to the user the trustworthiness assessment results (chatbot or document) in the form of a high-level report with the most important results.	UR_S4 UR_Q22	Port DE Media DE	-	SHOULD HAVE
UR_26	Present to the user more details on specific results from the trustworthiness assessment overview on demand (chatbot or hyperlink in the document)	UR_S5	Port DE	-	SHOULD HAVE
UR_27	Present to the user the trustworthiness assessment results in a data-visualisation like	UR_Q25	Health AI	-	COULD-HAVE

	dashboard (e.g. PowerBI) with descriptions				
UR_28	Present to the user the trustworthiness assessment results with quantitative metrics.	UR_Q26	Media AI, Port AI	-	SHOULD HAVE
UR_29	Display to the user the trustworthiness assessment in context-specific way (General error track record, Correlation of error with conditions and events, Training datasets details, e.g., 20% of the patients may be wrongly identified as of high risk)	UR_Q27	Health AI, Port DE Media DE	-	SHOULD HAVE
UR_30	Provide to the user the trustworthiness assessment report in a user-friendly way to save in a digital format.	UR_C27	Media DE	-	SHOULD-HAVE
UR_31	Present to the user which parts of the SUT suffer (e.g., model, deployment, datasets), but the level of detail will be based on user preferences.	UR_S7 UR_Q28 UR_Q29	Health AI, Port AI, Media AI, Media DE	-	COULD-HAVE
UR_32	Present to the user explanations on SUT trustworthiness assessment on demand, but the level of detail will be based on user preferences.	UR_Q30 UR_Q31 UR_Q35 UR_Q37	ALL	SR_25. THEMIS must provide explanations for the trustworthiness assessment (…)	MUST-HAVE
UR_33	Present to the user the factors that have mostly influenced the trustworthiness assessment.	UR_Q37	ALL	-	SHOULD HAVE
UR_34	Display to the user system's current trustworthiness assessment against previous trustworthiness assessments. Access to previous trustworthiness assessments along with the corresponding state of the AI to be accessible	UR_Q5 UR_S13 UR_Q32	Health DE, Health AI, Port DE, Port AI, Media DE	-	COULD-HAVE
UR_35	Enable the user to obtain a comparative analysis of multiple AI systems and services for the same task (selected by the end-user).	UR_Q34	ALL	-	COULD-HAVE
UR_36	Enable the user to choose which trustworthiness characteristics would like to be assessed	UR_Q36	Health DE, Health AI, Media DE,	-	COULD HAVE
UR_37	Every time the SUT is updated, the system signal to rerun the trustworthiness assessment			SR_28. THEMIS trustworthiness assessment will be dynamic and will be updated based on the implementation (or the projection) of trustworthiness optimisation measures.	MUST HAVE
UR_38	The Trustworthiness Assessment should not last more than an hour			-	NA
UR_39	Enable the user to perform experiments both for different scenarios (e.g., articles when trained only in posts), and on own datasets (e.g., sets of "hard"articles).	UR_Q43 UR_Q44 UR_S12	Media DE, Media AI, Port DE	SR_33. THEMIS will provide access to datasets for experimentation, Open APIs for experiment development, ML/DL algorithms models, advanced analytics, visualization configurations, human/machine interactions schemas,	MUST HAVE

				as well as the legal framework for trusted AI. SR_34. THEMIS will feature a fully equipped sandbox to design and deploy experiment variants and evaluate them against defined sets of criteria.	
RISK ASSESSMENT					
UR_40	Present to the user the related risks stemming from trustworthiness vulnerabilities.	UR_S9	Port DE	SR_31. THEMIS will implement trustworthiness assessment based on a risk assessment approach. SR_32. THEMIS (..) will be able to assess the AI system's accuracy and fairness their impact on the business-related risks and KPIs.	MUST HAVE
UR_41	The system could have a BI tool to analyse risk assessment results	UR_Q46	Health AI	-	COULD HAVE
UR_42	Provide on demand explanations to the user on how the presented risks have been calculated.	UR_Q47	ALL	SR_01. THEMIS (..) should interact with the user via an AI-driven conversational interface that will be able to explain in the necessary level of detail the outputs of AI systems. *#Comment: Although SR_01 refers to the system under test, as THEMIS 5.0 platform is also an AI System that the user needs to trust we conclude that it refers to the THEMIS 5.0 platform as well.*	MUST HAVE
UR_43	Inform the user about the factors which have mostly influenced the risk assessment	UR_Q48	Health AI Health DE Media DE Media AI	-	SHOULD HAVE
UR_44	Present the risk assessment results at once (chatbot or document) with a high-level report with the most important results and present more details on demand.	UR_S10 UR_S11	Port DE	-	SHOULD HAVE
UR_45	The Risk Assessment should not last more than a few seconds			-	NA
EXPLORE					
UR_46	Inform the user about positive and negative business impacts for each enhancement measure	UR_S14, UR_C29, UR_Q54	Media DE, Health DE, Port AI	SR_38. THEMIS will inform the user beforehand for the impact of the optimisation measures on the established business related KPIs.	MUST HAVE
UR_47	Inform the user about positive and negative impacts in SUT's trustworthiness for each enhancement measure.	UR_S16, UR_Q59	Media AI, Media DE	-	SHOULD HAVE
UR_48	Inform the user about the required time, and cost for the implementation of each enhancement measure.	UR_Q58	Media DE	-	WON'T HAVE (beyond scope)
UR_49	Inform the user about the estimated changes in the usability of the tool (e.g., certain functionalities become slower by 10%).	UR_Q60	Media DE	-	WON'T HAVE (NA)
UR_50	For each enhancement suggestion inform the user about	UR_C30	Media DE	-	COULD HAVE

	the respective climate impact rating				
UR_51	For each enhancement suggestion provide explanations confirming that ethical/legal/security/climate organization's restrictions are respected.	UR_C31 UR_Q6 UR_Q67	Health DE, Port AI Media DE	-	WON'T HAVE (out of scope)
UR_52	Present enhancement suggestions in tabular form with the pros and cons	UR_Q55	Health AI	-	COULD HAVE
UR_53	Present enhancement suggestions in using a rating system	UR_C35	Health DE	-	COULD HAVE
UR_54	Provide the user with a visual representation of the trade-offs (when increasing fairness-accuracy drops, etc) for each suggestion	UR_Q56	Media DE	-	COULD HAVE
UR_55	All different types of roles that are using the SUT within the organization are considered for calculating enhancement suggestions, to ensure that no contradictory suggestions are made.	UR_C34	Port DE	SR_37. THEMIS will explore the applicability implement negotiation strategies for the selection between multiple potential solutions, resulting from the different actors and objectives there may be	COULD HAVE
UR_56	User's years of experience (to ensure motivation of decision making) are considered for calculating enhancement suggestions.	UR_C5, UR_C32	Health DE	-	WON'T HAVE (Unclear how years of experienc e is relevant)
UR_57	User's professional interests (to ensure motivation of decision making) are considered for calculating enhancement suggestions.	UR_C6	Health DE	SR_20. THEMIS personalised trustworthiness assessment will be based on (…) Motivation (…)	MUST HAVE
UR_58	Exploration of solutions should be accessible only by specific roles	UR_Q57	Media DE	-	SHOULD HAVE
UR_59	Provide the user with a testbed environment to allow the application of the provided solutions in a pipeline.	UR_S17	Media AI	-	SHOULD HAVE
UR_60	Enable users to add new potential risks of trustworthiness vulnerabilities.	UR_Q61	Health AI, Port DE, Port AI, Media DE	-	SHOULD HAVE
UR_61	Enable only specific roles to add new potential risks of trustworthiness vulnerabilities.	UR_Q62	Media DE	-	SHOULD HAVE
UR_62	Enable the user to select when and if a trustworthiness optimisation measure will be applied to the AI system under test.	UR_S15	Media DE	SR_48. THEMIS users will be able to choose when and if a trustworthiness optimisation measure will be applied to the AI system under test.	MUST HAVE

UR_63	The optimal solution should be directly implemented without asking the user.	UR_Q64	Health DE	-	WON'T HAVE (SR_48)
UR_64	Enable the user to choose between allowing THEMIS 5.0 to perform the optimal solution automatically and allowing the user the one that prefers	UR_Q65	Media DE	-	COULD HAVE
UR_65	Present to the user how the changes made to the system will affect the risk assessment against previous assessments.	UR_Q45	Health AI, Health DE, Port AI, Port DE Media AI, Media DE	-	MUST HAVE
UR_66	The explore stage should not last more than a few days	UR_Q71	Port DE	-	NA
ENHANCE					
UR_67	Present to the user the new trustworthiness assessment results.	UR_S19	Media DE	-	SHOULD HAVE
UR_68	Present to the user an overview of the updated risk assessment of the AI system against previous values.	UR_Q72	Health DE Health AI Port DE Port AI Media DE Media AI	-	SHOULD HAVE
UR_69	Provide the user with a visual representation of trustworthiness parameters changes through time (e.g. in tabular form along with timestamps), given a predefined timeframe.	UR_Q73 UR_Q74	Media DE Health AI	-	COULD HAVE
UR_70	Present to the user upon request analytical details on what has changed to the SUT (e.g., fine-tuned in dataset X)	UR_Q75	Port DE	-	SHOULD HAVE
UR_71	Present to the user upon request the sources used for generating the enhanced results.	UR_Q76	Port DE	-	COULD HAVE
UR_72	The enhancement stage should not last more than 24 hours	UR_Q77		-	NA

3.6. Co-created User Journeys

In this section, we present the co-created user journeys. They are based on both the system requirements (Section 0) and on the user requirements (Section 3.5), and specifically on those prioritized as "SHOULD-HAVE". However, as T2.3 and T4.1 progress, this prioritization is subject to change and therefore the user journeys may change accordingly.

The User Journey, or user journey map, helps document and visualize the step-by-step user experience with a product or service from beginning to end and lists the different actions users take to accomplish a goal (Walter, 2022). These actions are arranged in chronological order, often presented as a timeline. During the ideation and system specification process, the User Journey is used as a tool to describe each step of the "journey" and, to keep it systematic, the description is ordered according to generic steps. In this way, service designers and stakeholders can obtain a global overview of all possible journeys. According to Walter (2022), there is no "one size fits all" for building a journey map. However, often a journey map contains the following fields:

- **User goal:** What the user is trying to accomplish
- **Journey stages:** the big stages that a user has to go through to accomplish a goal. Here these stages are based on the four stages of the Trustworthiness Optimization Process described in Section 2.8.2 that is Identify-Assess-Explore-Enhance.

- ○ **Actions:** the actions that a user has to do at each stage,
- ○ **Pain points:** the difficulties that the user may experience at each stage,
- ○ **Opportunities:** ideas for overcoming the pain points.

Additionally, as THEMIS 5.0 supports various types of users with different capabilities and authorization levels, we also include the "User type" attribute. Based on the broad user requirements, we concluded that for the design of the user journeys we should consider three main types of users:

- **Type A:** Domain expert in a decision-making position for the organization (e.g., editor-in-chief/business product manager for the media use case/hospital manager for the health care use case/port control manager for the port management use case). This type of user has full accessibility to the THEMIS 5.0 platform but is not interested in technical information (e.g., trustworthiness assessment results)
- **Type B:** Domain expert not in a decision-making position (e.g., journalist/general practitioner/ship captain). This type of user has limited accessibility to the THEMIS 5.0 platform and can only perform risk assessment.
- **Type C:** AI-developer. This type of user has full accessibility to the THEMIS 5.0 platform and is interested in technical information (e.g., trustworthiness assessment results)

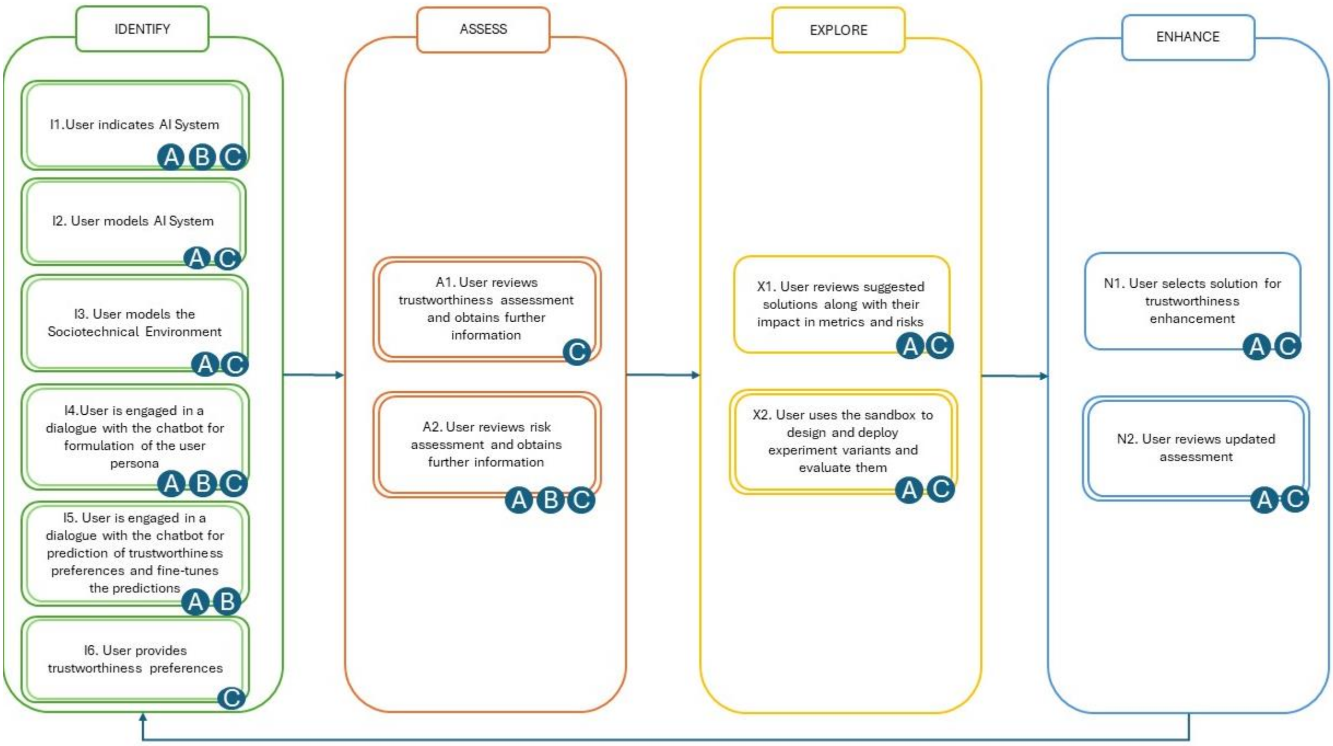

Figure 16: The main stages and actions considered in the generation of the user journeys. With A, B, C we represent the type of user that performs an action.

The stages and the various *main* actions (in the user journeys, these actions are decomposed into more detailed ones) are illustrated in Figure 16. The actions that can be final in a user journey are represented with a double frame. For instance, a user journey may include only the action I3 (and the sub-actions that entails), or the actions <I1, I3, I4, I6, A1>, or the actions <I1, I3, I4, I6, A1, X2>, etc.

To make the user journeys more tangible, we provide three user journeys, one for each type of user, for the three use cases (healthcare, port, media) and we also provide a user journey for the case that a user wants to generate a new AI system from scratch (UJ10). Hence, in total, we provide 10 user journeys. It is important to highlight that, as illustrated in Figure 16, these user journeys could be divided into smaller ones. For example, a hospital director could be interested only in modelling the socio-technical environment, or only in performing risk assessment of an AI system that has been already assessed.

As in the THEMIS 5.0 project, the user journeys are used as a basis for the development of the mock-ups which will be presented in the Co-creation Phase B to collect decision-making specifications for THEMIS 5.0 AI ecosystem services, it suffices to include only the “User goal”, the “Journey stages” and “Actions”. The “Pain points” and “Opportunities” will be explored during the Co-creation Phase B.

<table>
<tr><th>User Journey 1</th><th>User type:
A: Unit Director

Use case:
Healthcare</th><th>Goal:
The user wants to perform trustworthiness enhancement of an existing pancreatic disease risk prediction system.</th><th colspan="2">Expectations:
-Deep understanding of the risk assessment results.
-Enhancement is tailored to the KPI objectives of all actors involved (hospital, AI tech company)</th></tr>
<tr><td rowspan="2">Stage</td><td colspan="2" rowspan="2">Stage 1: IDENTIFY</td><td colspan="2">Stage 2: ASSESS</td></tr>
<tr><td>Trustworthiness Assessment (Optional)</td><td>Risk Assessment</td></tr>
<tr><td>Action:</td><td colspan="2">a. The user logs in.
b. Provides legal, ethical, and regulatory frameworks that the hospital should comply with (if not provided already) [SR_43].
c. The user is engaged in personalized dialogues through a chatbot and if they have not already used THEMIS 5.0, they provide their role, technical awareness [UR_1], attitude towards AI [UR_4a], Unit Director role requirements [UR_4b], objectives, capabilities, motives, decision support needs, targets, objectives and KPIs [SR_14, SR_21, SR_22].
Otherwise, they can proceed with one of the following: i) review & update the existing persona [UR_12], ii) add a new persona (e.g., if they have multiple roles – Unit Director and General Practitioner) [UR_14], iii) insert the user preferences of other users from the same organization [UR_19], iv) to next action.
d. The user reviews the generic trustworthiness preference prediction results as calculated by THEMIS 5.0, asks for explanations if necessary, and if needed, fine-tunes the results [UR_14, UR_17, UR_18].
e. The user indicates the type of process (risk prediction for pancreatic cancer) that the AI system is used for and through a high-level discussion with the chatbot, provides their specific trustworthiness preferences for such processes.
f. The user reviews and fine-tunes the process-specific trustworthiness preferences [UR_18].
g. The user indicates the risk prediction system that wants to assess and enhance its trustworthiness.
h. If someone in the organization has already assessed the indicated pancreatic disease risk prediction system, then the user reviews and refines if necessary the qualitative/quantitative model of the sociotechnical environment.
Otherwise, the user creates a qualitative model for the hospital socio-technical environment with regards to the pancreatic disease risk prediction system via the chatbot and with the help of a GUI, by inserting all the ethical/business/legal risks and implications on the organizational KPIs stemming from all potential trustworthiness vulnerabilities of the AI</td><td>a. The user selects the trustworthiness characteristics they want to be assessed (they can select also all trustworthiness characteristics) [UR_36].
b. Reviews a high-level report of the trustworthiness assessment results [UR_25].
c. Can obtain further information on the trustworthiness assessment results (e.g., timeline of trustworthiness assessments of the pancreatic disease risk prediction system over the various trustworthiness enhancement cycles) [UR_26].
d. Can obtain high-level explanations on the trustworthiness assessment results (e.g., why the accuracy is low, how it was calculated, sources used) [UR_32].
e. Can observe the performance of the system in various scenarios (e.g., what is the probability of a prediction in young patients) [UR_39]
f. Stores locally the trustworthiness assessment report [UR_30].</td><td>a. The user selects to perform risk assessment.
b. Reviews the business-related risks displayed on their screen [UR_40].
c. The user is notified about potential ethical and/or legal risks that need to be checked [UR_7, UR_8].
d. Requests from the chatbot for more detailed information on the risk assessment results (e.g., a timeline of risk assessments of the pancreatic disease risk prediction system over the various trustworthiness enhancement cycles) [UR_44].
e. Asks the chatbot for explanations on the risk assessment results (e.g., which socio-technical rules and which conditions of the AI system resulted in there being a high risk of lawsuits against the hospital) [UR_42].
f. The user may add through the chatbot missing business/ethical/legal risks for the Hospital from certain trustworthiness vulnerabilities [UR_60].
g. Asks the chatbot for further clarifications on terms that they do not understand [UR_2].
h. The user can store locally the final risk assessment report [UR_30].</td></tr>
</table>

	system [SR_18, SR_19, SR 24, SR_47]. For instance, by indicating that a low accuracy of the system may result in lawsuits against the hospital by the doctors for offering a misleading system (see Table 14). i. The user reviews the qualitative/quantitative socio-technical model generated by THEMIS 5.0 asks the chatbot for explanations if necessary, and, if needed, fine-tunes the results through the chatbot [SR_47].		

User Journey 1 (cntd)	**User type: A: Unit Director** **Use case: Healthcare**	**Goal: The user wants to perform trustworthiness enhancement of an existing pancreatic disease risk prediction system.**	**Expectations:** **-Deep understanding of the risk assessment results.** **-Enhancement is tailored to the KPI objectives of all actors involved (hospital, AI tech company)**
Stage	Stage 3: EXPLORE		Stage 4: ENHANCE
Action:	a. The user reviews the THEMIS 5.0 solutions to mitigate vulnerabilities along with the estimated business/ legal/ethical benefits and risks for the hospital that the implementation of each solution entails [UR_46]. b. The user asks for explanations (e.g., input criteria used in the decision support process) in this process [SR_07, SR_36, SR_41, SR_42]. c. The user uses the sandbox to design and deploy experiment variants and evaluate them [UR_39]. d. The user identifies the optimal solution *for the hospital and the actors involved* with the support of the conversational interface.		a. The user selects a solution [UR_62]. b. Upon request, the user reviews the new trustworthiness & risk assessment results, after the implementation of the solution [UR_67, UR_68] c. Upon request, the user is presented with analytical details on what has changed to the pancreatic disease risk prediction system (e.g., fine-tuned in dataset X) [UR_70]

User Journey 2	**User type: B: Doctor** **Use Case: Healthcare**	**Goal: The user wants to perform risk assessment of an existing pancreatic disease risk prediction system.** ***As the user is of type B, they cannot perform Stages 3, 4, which affect the AI system.***	**Expectations:** **-Deep understanding of the risk assessment results.** **-Short interaction with the tool**	
Stage	Stage 1: IDENTIFY		Stage 2: ASSESS	
			Trustworthiness Assessment (Optional)	Risk Assessment
Action:	a. The user logs in. b. The user is engaged in personalized dialogues through a chatbot and if they have not already used THEMIS 5.0, they provide their role, technical awareness [UR_1], attitude towards AI [UR_4a], requirements stemming from the specific role (Doctor)[UR_4b], personal and role capabilities, motives, objectives and KPIs [SR_14, SR_21, SR_22]. Otherwise, they can proceed with one of the following: i) review & update the existing persona [UR_12], ii) add a new persona (e.g., if they have multiple roles -Doctor and Unit Director) [UR_14], iii) insert the user preferences of other users from the same organization [UR_19], iv) to next action. c. The user reviews the preference prediction results as calculated by THEMIS 5.0, asks for explanations if necessary, and if needed, fine-tunes the results [UR_14, UR_17, UR_18].		a. The user selects the trustworthiness characteristics they want to be assessed (they can select also all trustworthiness characteristics) [UR_36]. b. Reviews a high-level report of the trustworthiness assessment results [UR_25]. c. Stores locally the trustworthiness assessment report [UR_30].	a. The user selects to perform risk assessment. b. Reviews the business risks related to the provided KPIs [UR_40]. c. The user is notified about potential ethical and/or legal risks that need to be checked [UR_7, UR_8]. d. Requests from the chatbot for more detailed information on the risk assessment results (e.g., which types of patients are underrepresented, what are the underpinning risks from this?) [UR_44]. e. Asks the chatbot for further clarifications on terms that they do not understand [UR_2].

<table>
<tr><td></td><td>d. The user indicates the type of process (risk prediction for pancreatic cancer) that the AI system is used for and through a high-level discussion with the chatbot, provides their specific trustworthiness preferences for such processes.
e. The user reviews and fine-tunes the process-specific trustworthiness preferences [UR_18].
f. The user indicates the pancreatic cancer risk prediction for system that wants to assess its trustworthiness.</td><td></td><td>f. The user can store locally the final risk assessment report [UR_30].</td></tr>
</table>

<table>
<tr><th>User Journey 3</th><th>User type:
C: AI tech support for hospital (either as hospital employee or as an employee of a collaborating tech company)

Use Case: Healthcare</th><th>Goal:
The user wants to perform trustworthiness enhancement of an existing pancreatic disease risk prediction system.</th><th colspan="2">Expectations:
-Deep understanding of the trustworthiness & risk assessment results.
-Trustworthiness enhancement is tailored to the KPI objectives of the organization the AI-tech is working on
-Short interaction with the tool</th></tr>
<tr><td rowspan="2">Stage</td><td colspan="2" rowspan="2">Stage 1: IDENTIFY</td><td colspan="2">Stage 2: ASSESS</td></tr>
<tr><td>Trustworthiness Assessment</td><td>Risk Assessment</td></tr>
<tr><td>Action:</td><td colspan="2">a. The user logs in.
b. The user provides legal, ethical, and regulatory frameworks their organization should comply with (if not provided already) [SR_43].
c. The user is engaged in personalized dialogues through a chatbot and if they have not already used THEMIS 5.0, they provide their role, their generic trustworthiness preferences [UR_10c], AI tech support for hospital role requirements, objectives, capabilities, motives, decision support needs, targets, objectives and KPIs [SR_14, SR_21, SR_22].
Otherwise, they can proceed with one of the following actions: i) review & update the existing persona [UR_12], ii) add a new persona (e.g., if they have multiple roles -AI tech aware& General Practitioner) [UR_14], iii) insert the user preferences of other AI-developers from the same organization [UR_19], iv) to next action.
d. The user indicates the pancreatic disease risk prediction system that wants to assess and enhance its trustworthiness. If the system has not been assessed before, they will upload the respective AI model & data cards.
e. The user provides their technical trustworthiness preferences for the pancreatic disease risk prediction system.
f. If someone in the organization has already assessed the indicated system then the user reviews and refines if</td><td>a. The user selects the trustworthiness characteristics they want to be assessed (they can select also all trustworthiness characteristics) [UR_36]
b. The user reviews the technical (quantitative) trustworthiness assessment results presented by THEMIS 5.0 [UR_28].
c. The user can obtain further information on the trustworthiness assessment results (e.g., timeline of trustworthiness assessments of the disease risk prediction system over the various trustworthiness enhancement cycles) [UR_26]
d. The user can obtain detailed technical explanations on the trustworthiness assessment results (e.g., why the accuracy is low, how it was calculated, sources used) [UR_31, UR_32, UR_33].
e. The user obtains deeper understanding of the trustworthiness of the pancreatic disease risk prediction system by performing focused experiments, e.g., testing its performance in “hard” risk prediction datasets [UR_39, SR_33, SR_34, SR_49, SR_51].
f. Stores locally the trustworthiness assessment report [UR_30].</td><td>a. The user selects to perform risk assessment.
b. Reviews the business risks related to the provided KPIs [UR_40].
c. The user is notified about potential ethical and/or legal risks that need to be checked [UR_7, UR_8].
d. The user requests from the chatbot more detailed information on the risk assessment results (e.g., timeline of risk assessments over the various trustworthiness enhancement cycles) [UR_44].
e. Asks the chatbot for explanations on the risk assessment results (e.g., which rules triggered a specific risk) [UR_2].
f. The user stores locally the final risk assessment report [UR_30].</td></tr>
</table>

	necessary the qualitative/quantitative model sociotechnical environment. Otherwise, the user creates a qualitative model for the socio-technical environment hospital/tech company that works for via the chatbot and with the help of a GUI, by inserting all the ethical/business/legal risks and implications on the organization (hospital or tech company) stemming from all potential trustworthiness vulnerabilities of the prediction system [SR_18, SR_19, SR 24, SR_47]. For instance, the lawsuits from the hospital to the tech company for offering a misleading tool. g. The user reviews the qualitative socio-technical model generated by THEMIS 5.0 asks the chatbot for explanations if necessary, and, if needed, fine-tunes the results through the chatbot [SR_47].		

User Journey 3 (cntd)	**User type: C: AI tech support for hospital (either as hospital employee or as an employee of a collaborating tech company)** **Use Case: Healthcare**	**Goal:** **The user wants to perform trustworthiness enhancement of an existing pancreatic disease risk prediction system.**	**Expectations:** **-Deep understanding of the trustworthiness & risk assessment results.** **-Trustworthiness enhancement is tailored to the KPI objectives of the organization the AI-tech is working on** **-Short interaction with the tool**
Stage	Stage 3: EXPLORE		Stage 4: ENHANCE
Action:	a. The user reviews the THEMIS 5.0 solutions to mitigate vulnerabilities along with the estimated business/ legal/ethical benefits and risks for the hospital that the implementation of each solution entails [UR_46]. b. The user asks for explanations (e.g., input criteria used in the decision support process) in this process [SR_07, SR_36, SR_41, SR_42]. c. The user can experiment with alternative solutions to see their impact [SR_34]. d. The user identifies the optimal solution for the hospital and the actors involved based on the provided metrics and with the support of the conversational agent.		a. The user selects a solution [UR_62]. b. Upon request, the user reviews the new trustworthiness & risk assessment results, after the implementation of the solution [UR_67, UR_68] c. Upon request, the user is presented with analytical details on what has changed to the prediction system (e.g., fine-tuned in dataset X) [UR_70]

<table>
<tr>
<td>User Journey 4</td>
<td>User type:
A: Port Authority Manager

Use case: Port Management</td>
<td>Goal:
The user wants to perform trustworthiness enhancement of an existing ETA system.</td>
<td colspan="2">Expectations:
-Deep understanding of the risk assessment results.
- Enhancement is tailored to the objectives of all actors involved (i.e., Transportation Companies, Pilots, etc)
-Short interaction with the tool</td>
</tr>
<tr>
<td rowspan="2">Stage</td>
<td colspan="2" rowspan="2">Stage 1: IDENTIFY</td>
<td colspan="2">Stage 2: ASSESS</td>
</tr>
<tr>
<td>Trustworthiness Assessment (Optional)</td>
<td>Risk Assessment</td>
</tr>
<tr>
<td>Action:</td>
<td colspan="2">a. The user logs in.
b. The user provides legal, ethical, and regulatory frameworks that the hospital should comply with (if not provided already) [SR_43].
c. The user is engaged in personalized dialogues through a chatbot and if they have not already used THEMIS 5.0, they provide their role, technical awareness [UR_1], attitude towards AI [UR_4a], Port Authority role requirements [UR_4b], objectives, capabilities, motives, decision support needs, targets, objectives and KPIs [SR_14, SR_21, SR_22].
Otherwise, they can proceed with one of the following: i) review & update the existing persona [UR_12], ii) add a new persona (e.g., if they have multiple roles – Port Authority Manager and AI developer) [UR_14], iii) insert the user preferences of other users from Port Authority [UR_19], iv) to next.
d. The user reviews the generic preference prediction results as calculated by THEMIS 5.0, asks for explanations if necessary, and if needed, fine-tunes the results [UR_14, UR_17, UR_18].
e. The user indicates the type of process (ETA prediction) that the AI system is used for and through a high-level discussion with the chatbot, provides their specific trustworthiness preferences for such processes.
f. The user reviews and fine-tunes the process-specific trustworthiness preferences [UR_18].
g. The user indicates the ETA system that wants to assess and enhance its trustworthiness.
h. If the indicated system has been previously assessed by someone in the Port Authority organization, the user reviews and refines if necessary, the qualitative/quantitative model of the sociotechnical environment.
Otherwise, the user creates a qualitative model for the port socio-technical environment via the chatbot and with the help of a GUI, by inserting all the ethical/business/legal risks and implications on the Port Authority KPIs stemming from all potential ETA trustworthiness vulnerabilities [SR_18, SR_19, SR 24, SR_47]. For instance, the level of traffic disruption and congestion that may be caused by the low accuracy of the ETA (see Table 14).
i. The user reviews the qualitative/quantitative socio-technical model generated by THEMIS 5.0 asks the chatbot for explanations if necessary, and, if needed, fine-tunes the results through the chatbot [SR_47].</td>
<td>a. The user selects the trustworthiness characteristics they want to be assessed (they can select also all trustworthiness characteristics) [UR_36].
b. Reviews a high-level report of the trustworthiness assessment results [UR_25].
c. Can obtain further information on the trustworthiness assessment results (e.g., timeline of trustworthiness assessments of the ETA over the various trustworthiness enhancement cycles) [UR_26].
d. Can obtain high-level explanations on the trustworthiness assessment results (e.g., why the accuracy is low, how it was calculated, sources used) [UR_32].
e. Can observe the performance of the system in various scenarios (e.g., the accuracy of ETA made by the AI System in comparison to different scenarios -vessels of similar type, weekday, shipping agency, etc.) [UR_39]
f. Stores locally the trustworthiness assessment report [UR_30].</td>
<td>a. The user selects to perform risk assessment.
b. Reviews the business-related risks displayed on their screen [UR_40].
c. The user is notified about potential ethical and/or, legal risks that need to be checked [UR_7, UR_8].
d. Requests from the chatbot for more detailed information on the risk assessment results (e.g., which shipping companies are being favoured with the current version of the system) [UR_44].
e. Asks the chatbot for explanations on the risk assessment results (e.g., what is the cause of a business risk) [UR_42].
f. Asks the chatbot for further clarifications on terms that they do not understand [UR_2].
g. The user may add through the chatbot missing risks related to AI Port Authority from certain trustworthiness vulnerabilities [UR_60].
h. The user can store locally the final risk assessment report [UR_30].</td>
</tr>
</table>

<table>
<tr><th>User Journey 4 (cntd)</th><th>User type:
A:
Port Authority Manager

Use Case: Port Management</th><th>Goal:
The user wants to perform trustworthiness enhancement of an existing ETA system.</th><th>Expectations:
-Deep understanding of the risk assessment results.
- Enhancement is tailored to the objectives of all actors involved (i.e., Transportation Companies, Pilots, etc)
-Short interaction with the tool</th></tr>
<tr><td>Stage</td><td colspan="2">Stage 3: EXPLORE</td><td>Stage 4: ENHANCE</td></tr>
<tr><td>Action:</td><td colspan="2">a. The user reviews the THEMIS 5.0 solutions to mitigate vulnerabilities along with the estimated business/ legal/ethical benefits and risks for the Port Authority that the implementation of each solution entails [UR_46].
b. The user asks for explanations (e.g., input criteria used in the decision support process) in this process [SR_07, SR_36, SR_41, SR_42].
c. If the user is familiar with AI they can use the sandbox to design and deploy experiment variants and evaluate them [UR_39].
d. The user identifies the optimal solution for all stakeholders (e.g., all shipping agencies) with the support of the conversational interface.</td><td>a. Selects and implements the optimal solution [UR_62].
b. Reviews the new trustworthiness & risk assessment results, after the implementation of the solution [UR_67, UR_68].
c. Upon request, the user is presented with analytical details on what has changed to the ETA prediction system (e.g., fine-tuned in dataset X) [UR_70]</td></tr>
</table>

<table>
<tr><th>User Journey 5</th><th>User type:
B: Port Authority employee/ship captain with limited authorization rights

Use Case: Port Management</th><th>Goal:
The user wants to perform risk assessment of an existing ETA system.

*As the user is of type B, they cannot perform Stages 3, 4, which affect the AI system.</th><th colspan="2">Expectations:
-Deep understanding of the risk assessment results.
-Short interaction with the tool</th></tr>
<tr><td rowspan="2">Stage</td><td colspan="2" rowspan="2">Stage 1: IDENTIFY</td><td colspan="2">Stage 2: ASSESS</td></tr>
<tr><td>Trustworthiness Assessment (Optional)</td><td>Risk Assessment</td></tr>
<tr><td>Action:</td><td colspan="2">a. The user logs in.
b. The user is engaged in personalized dialogues through a chatbot and:
1. If the user has not already used THEMIS 5.0, they provide their role, technical awareness [UR_1], attitude towards AI [UR_4a], role requirements related to their position as employees in port authority/ship captain [UR_4b], objectives, capabilities, motives, decision support needs, targets, objectives and KPIs relevant to their role [SR_14, SR_21, SR_22].
Otherwise, they can proceed with one of the following: i) review & update the existing persona [UR_12], ii) add a new persona (e.g., if they have multiple roles – Port Authority employee and AI developer) [UR_14], iii) insert the user preferences of other users from Port Authority [UR_19], iv) to next action.
2. The user reviews the trustworthiness preference prediction results as calculated by THEMIS 5.0, asks for explanations if necessary, and if needed, fine-tunes the results [UR_14, UR_17, UR_18].
c. The user indicates the type of process (ETA prediction) that the AI system is used for and through a high-level discussion with the</td><td>a. The user selects the trustworthiness characteristics they want to be assessed (they can select also all trustworthiness characteristics) [UR_36].
b. Reviews a high-level report of the trustworthiness assessment results [UR_25].
c. Stores locally the trustworthiness assessment report [UR_30].</td><td>a. The user selects to perform risk assessment.
b. Reviews the business-related risks displayed on their screen [UR_40].
c. The user is notified about potential ethical and/or legal risks that need to be checked [UR_7, UR_8].
d. Requests from the chatbot for more detailed information on the risk assessment results (e.g., which shipping companies are being favoured with the current version of the system and what is the business risk from such behaviour of the system) [UR_44].
e. Asks the chatbot for further clarifications on terms that they do not understand [UR_2].
f. The user can store locally the final risk assessment report [UR_30].</td></tr>
</table>

	chatbot, provides their trustworthiness preferences for such processes. d. Reviews and fine-tunes the process-specific trustworthiness preferences [UR_18]. e. Indicates the ETA system that wants to assess its trustworthiness.		

<table>
<tr><td>User Journey 6</td><td>User type:
C: Port Authority-AI developer

Use Case: Port Management</td><td>Goal:
The user wants to perform trustworthiness enhancement of an existing ETA system.</td><td colspan="2">Expectations:
-Deep understanding of the trustworthiness & risk assessment results.
- Trustworthiness enhancement is tailored to the KPI objectives of the organization the AI-developer is working on
-Short interaction with the tool</td></tr>
<tr><td rowspan="2">Stage</td><td colspan="2" rowspan="2">Stage 1: IDENTIFY</td><td colspan="2">Stage 2: ASSESS</td></tr>
<tr><td>Trustworthiness Assessment</td><td>Risk Assessment</td></tr>
<tr><td>Action:</td><td colspan="2">a. The user logs in.
b. The user provides legal, ethical, and regulatory frameworks their organization should comply with (if not provided already) [SR_43].
c. The user is engaged in personalized dialogues through a chatbot and if they have not already used THEMIS 5.0, they provide their role, trustworthiness preferences [UR_10c], the role requirements for AI tech support for Port Authority, objectives, capabilities, motives, decision support needs, targets, objectives and KPIs [SR_14, SR_21, SR_22].
Otherwise, they can proceed with one of the following actions: i) review & update the existing persona [UR_12], ii) add a new persona (e.g., if they have multiple roles -AI tech aware& Porth Authority manager) [UR_14], iii) insert the user preferences of other AI-developers from Port Authority [UR_19], iv) to the next action.
d. Indicates the ETA system that wants to assess and enhance its trustworthiness and, if it has not been assessed before, they will upload the respective AI model & data cards.
e. The user provides their specific technical trustworthiness preferences for ETA systems.
f. If someone in the organization has already assessed the ETA system, then the user reviews and refines if necessary, the qualitative/quantitative model sociotechnical environment.
Otherwise, the user creates a qualitative model for the socio-technical environment of the Port Authority via the chatbot and with the help of a GUI, by inserting all the ethical/business/legal risks and implications on the Port Authority KPIs stemming from all potential trustworthiness vulnerabilities of the ETA [SR_18, SR_19, SR 24, SR_47]. For instance, the level of traffic disruption and congestion that may be caused by low accuracy of the ETA.
g. The user reviews the qualitative socio-technical model generated by THEMIS 5.0 asks the chatbot for explanations if</td><td>a. The user selects the trustworthiness characteristics they want to be assessed (they can select also all trustworthiness characteristics) [UR_36]
b. The user reviews the technical (quantitative) trustworthiness assessment results presented by THEMIS 5.0 [UR_28].
c. The user can obtain further information on the trustworthiness assessment results (e.g., timeline of trustworthiness assessments of the ETA over the various trustworthiness enhancement cycles) [UR_26]
d. The user can obtain detailed technical explanations on the trustworthiness assessment results (e.g., why the accuracy is low, how it was calculated, sources used) [UR_31, UR_32, UR_33].
e. The user obtains deeper understanding of the trustworthiness of the AI system by performing focused experiments, e.g., calculating the just-in-time arrival, or testing its performance in specific scenarios [UR_39, SR_33, SR_34, SR_49, SR_51].
f. Stores locally the trustworthiness assessment report [UR_30].</td><td>a. The user reviews the business-related risks displayed on their screen [UR_40].
b. Reviews the business risks related to the provided KPIs [UR_40].
c. The user is notified about potential ethical and/or legal risks that need to be checked [UR_7, UR_8].
d. The user requests from the chatbot for more detailed information on the risk assessment results (e.g., timeline of risk assessments over the various trustworthiness enhancement cycles) [UR_44].
e. Asks the chatbot for explanations on the risk assessment results (e.g., which rules triggered a specific risk) [UR_2].
f. The user can indicate other ETAs for comparative results with respect to risk assessment [UR_35].
g. The user stores locally the final risk assessment report [UR_30].</td></tr>
</table>

	necessary, and, if needed, fine-tunes the results through the chatbot [SR_47].		

User Journey 6 (cntd)	**User type:** **C: Port Authority-AI developer** **Use Case: Port Management**	**Goal:** **The user wants to perform trustworthiness enhancement of an existing ETA system.**	**Expectations:** **-Deep understanding of the trustworthiness & risk assessment results.** **- Trustworthiness enhancement is tailored to the KPI objectives of the organization the AI-tech is working on** **-Short interaction with the tool**
Stage	Stage 3: EXPLORE		Stage 4: ENHANCE
Action:	a. The user reviews the THEMIS 5.0 solutions to mitigate vulnerabilities along with the estimated business/ legal/ethical benefits and risks for the Port Authority that the implementation of each solution entails [UR_46]. b. The user asks for explanations (e.g., input criteria used in the decision support process) in this process [SR_07, SR_36, SR_41, SR_42]. c. The user can experiment with alternative solutions to see their impact [SR_34]. d. The user identifies the optimal solution for the Port Authority and the involved actors based on the provided metrics and with the support of the conversational interface.		a. The user selects a solution [UR_62]. b. Upon request, the user reviews the new trustworthiness & risk assessment results, after the implementation of the solution [UR_67, UR_68] c. Upon request, the user is presented with analytical details on what has changed to the ETA system (e.g., fine-tuned in dataset X) [UR_70]

User Journey 7	**User type:** **A: Business Development Manager** **Use case: Media**	**Goal:** **The user wants to perform trustworthiness enhancement of an existing fake news/hate speech detection system.**	**Expectations:** **-Deep understanding of the risk assessment results.** **-Enhancement is tailored to the KPI objectives of all actors involved (media organization, AI tech company)**	
Stage	Stage 1: IDENTIFY		Stage 2: ASSESS	
			Trustworthiness Assessment (Optional)	Risk Assessment
Action:	1. The user logs in. 2. The user provides legal, ethical, and regulatory frameworks that the Media organization should comply with (if not provided already) [SR_43]. 3. The user is engaged in personalized dialogues through a chatbot and if they have not already used THEMIS 5.0, they provide their role, technical awareness [UR_1], attitude towards AI [UR_4a], Business development role requirements [UR_4b], objectives, capabilities, motives, decision support needs, targets, objectives and KPIs [SR_14, SR_21, SR_22]. Otherwise, they can proceed with one of the following: i) review & update the existing persona [UR_12], ii) add a new persona (e.g., if they have multiple roles – Business Development and Editor in Chief) [UR_14], iii) insert the user preferences of other users from the same organization [UR_19], iv) to the action. 4. The user reviews the trustworthiness preference prediction results as calculated by THEMIS 5.0, asks for explanations if necessary, and if needed, fine-tunes the results [UR_14, UR_17, UR_18].		a. The user selects the trustworthiness characteristics they want to be assessed (they can select also all trustworthiness characteristics) [UR_36]. b. Reviews a high-level report of the trustworthiness assessment results [UR_25]. c. Can obtain further information on the trustworthiness assessment results (e.g., timeline of trustworthiness assessments of the fake news/hate speech detection system over the various trustworthiness enhancement cycles) [UR_26]. d. Can obtain high-level explanations on the trustworthiness assessment results (e.g., why the accuracy is low, how it was calculated, sources used) [UR_32]. e. Can observe the performance of the system in various scenarios (e.g., what is the performance of the AI system in "difficult" texts where enhanced obfuscation	a. The user selects to perform risk assessment. b. Reviews the business-related risks displayed on their screen [UR_40]. c. The user is notified about potential ethical, legal risks that need to be checked [UR_7, UR_8]. d. Requests from the chatbot for more detailed information on the risk assessment results (e.g., for which political agendas the fake news detection system tends to overlook fake news) [UR_44]. e. Asks the chatbot for explanations on the risk assessment results (e.g., why is the fake news detection system in favour/against certain agendas? How certain risks are related to that?) [UR_42]. f. Asks the chatbot for further clarifications on terms that they do not understand [UR_2]. g. The user may add through the chatbot missing risks related to the media organization from certain trustworthiness vulnerabilities [UR_60].

	5. The user indicates the type of process (hate speech/fake news detection) that the AI system is used for and through a high-level discussion with the chatbot, provides their specific trustworthiness preferences for such processes. 6. The user reviews and fine-tunes the process-specific trustworthiness preferences [UR_18]. 7. The user indicates the fake news/hate speech detection system that wants to enhance its trustworthiness. 8. If the fake news/hate speech detection system has not been previously assessed by someone in the Media organization, the user creates a qualitative model for the media socio-technical environment with regards to the hate speech/fake news detection tool via the chatbot and with the help of a GUI, by inserting all the ethical/business/legal risks and implications on the organizational KPIs stemming from all potential fake news/hate speech detection system trustworthiness vulnerabilities [SR_18, SR_19, SR 24, SR_47]. For instance, the reputational damage caused by the low accuracy of a fake news detection tool (see Table 14). 9. The user reviews the qualitative socio-technical model generated by THEMIS 5.0 asks the chatbot for explanations if necessary, and, if needed, fine-tunes the results through the chatbot [SR_47].	techniques have been applied) [UR_39] f. Stores locally the trustworthiness assessment report [UR_30].	h. The user can store locally the final risk assessment report [UR_30].

User Journey 7 (cntd)	**User type:** **A: Business Development Manager** **Use case: Media**	**Goal:** **The user wants to perform trustworthiness enhancement of an existing fake news/hate speech detection system.**	**Expectations:** **-Deep understanding of the risk assessment results.** **- Enhancement is tailored to the KPI objectives of all actors involved (media organization, AI tech company)** **-Short interaction with the tool**
Stage	Stage 3: EXPLORE		Stage 4: ENHANCE
Action:	a. The user reviews the THEMIS 5.0 solutions to mitigate vulnerabilities along with the estimated business/ legal/ethical benefits and risks for the media organization that the implementation of each solution entails [UR_46]. b. The user asks for explanations (e.g., input criteria used in the decision support process) in this process [SR_07, SR_36, SR_41, SR_42]. c. If the user is familiar with AI, they can use the sandbox to design and deploy experiment variants and evaluate them [UR_39]. d. The user identifies the optimal solution with the support of the conversational interface.		a. The user selects a solution [UR_62]. d. The user reviews the new trustworthiness & risk assessment results, after the implementation of the solution [UR_67, UR_68] b. Upon request, the user is presented with analytical details on what has changed to the hate speech/fake news detection system (e.g., fine-tuned in dataset X) [UR_70]

User Journey 8	**User type:** **B: Journalist** **Use case: Media**	**Goal:** **The user wants to perform risk assessment of an existing fake news/hate speech detection system.** ****As the user is of type B, they cannot perform Stages 3, 4, which affect the AI system.***	**Expectations:** **-Deep understanding of the risk assessment results.** **-Short interaction with the tool**
Stage	Stage 1: IDENTIFY		Stage 2: ASSESS

		Trustworthiness Assessment (Optional)	Risk Assessment
Action:	a. The user logs in. b. The user is engaged in personalized dialogues through a chatbot and: 1. If the user has not already used THEMIS 5.0, they provide their role, technical awareness [UR_1], attitude towards AI [UR_4a], requirements stemming from the specific role (journalist)[UR_4b], personal and role capabilities, motives, objectives and KPIs [SR_14, SR_21, SR_22]. Otherwise, they can proceed with one of the following: i) review & update the existing persona [UR_12], ii) add a new persona (e.g., if they have multiple roles -journalist & business development manager) [UR_14], iii) insert the user preferences of other users from the same organization [UR_19], iv) to the next actions. 2. The user reviews the trustworthiness preference prediction results as calculated by THEMIS 5.0, asks for explanations if necessary, and if needed, fine-tunes the results [UR_14, UR_17, UR_18]. c. The user indicates the type of process (hate speech/fake news detection) that the AI system is used for and through a high-level discussion with the chatbot, provides their trustworthiness preferences for such processes. d. The user reviews and fine-tunes the process-specific trustworthiness preferences [UR_18]. e. The user indicates the hate speech/fake news detection system that wants to assess its trustworthiness.	a. The user selects the trustworthiness characteristics they want to be assessed (they can select also all trustworthiness characteristics) [UR_36] b. Reviews a high-level report of the trustworthiness assessment results [UR_25]. c. Stores locally the trustworthiness assessment report [UR_30].	a. The user selects to perform risk assessment. b. Reviews the risks related to the provided KPIs displayed on their screen [UR_40]. c. The user is notified about potential business, ethical, legal risks that need to be checked [UR_7, UR_8]. d. Requests from the chatbot for more detailed information on the risk assessment results (e.g., for which political agendas the fake news detection system tends to overlook fake news) [UR_44]. e. Asks the chatbot for further clarifications on terms that they do not understand [UR_2]. f. The user can store locally the final risk assessment report [UR_30].

User Journey 9	**User type: C: AI tech support for media organization who may work in a tech company Use case: Media**	**Goal: The user wants to perform trustworthiness enhancement of the fake news/hate speech detection system that the media organization is using.**	**Expectations: -Deep understanding of the trustworthiness & risk assessment results. - Enhancement is tailored to the KPI objectives of all actors involved (media organization, AI tech company) -Short interaction with the tool**	
Stage	Stage 1: IDENTIFY		Stage 2: ASSESS	
			Trustworthiness Assessment	Risk Assessment
Action:	a. The user logs in. b. The user provides legal, ethical, and regulatory frameworks their organization should comply with (if not provided already) [SR_43]. c. The user is engaged in personalized dialogues through a chatbot and if they have not already used THEMIS 5.0, they provide [UR_10c] their role, *trustworthiness preferences*, role requirements, objectives, capabilities, motives, decision support needs, targets, objectives and KPIs [SR_14, SR_21, SR_22]. Otherwise, they can proceed with one of the following actions: i) review & update the existing persona [UR_12], ii) add a new persona (e.g., if they have multiple roles)		a. The user selects the trustworthiness characteristics they want to be assessed (they can select also all trustworthiness characteristics) [UR_36] b. The user reviews the *technical (quantitative)* trustworthiness assessment results presented by THEMIS 5.0 [UR_28]. c. The user can obtain further information on the trustworthiness assessment results (e.g., timeline of trustworthiness assessments of the hate speech/fake news detection systems over the various	a. The user selects to perform risk assessment. b. Reviews the business-related risks displayed on their screen [UR_40]. c. The user is notified about potential ethical and/or legal risks that need to be checked [UR_7, UR_8]. d. The user requests, from the chatbot, more detailed information on the risk assessment results (e.g., timeline of risk assessments over the various trustworthiness enhancement cycles) [UR_44]. e. Asks the chatbot for explanations on the risk assessment results (e.g., which rules triggered a specific risk) [UR_60].

	[UR_14], iii) insert the user preferences of other AI-developers from the same organization [UR_19], iv) to the next action. d. The user indicates the system and, if the fake news/hate speech detection system has not been assessed before, *they upload the respective AI model & data cards*. e. The user provides their technical trustworthiness preferences for the hate speech/fake news detection system. f. If someone in the organization has already assessed the fake news/hate speech detection system, then the user reviews and refines if necessary the qualitative/quantitative model of the sociotechnical environment. Otherwise, the user creates a qualitative model for the media socio-technical environment via the chatbot and with the help of a GUI, by inserting all the ethical/business/legal risks and implications on the tech company KPIs stemming from all potential fake news/hate speech detection system trustworthiness vulnerabilities [SR_18, SR_19, SR 24, SR_47]. For instance, the reputational damage caused in the company by the low accuracy of the detection tools. g. The user reviews the qualitative socio-technical model generated by THEMIS 5.0 asks the chatbot for explanations if necessary, and, if needed, fine-tunes the results through the chatbot [SR_47].	trustworthiness enhancement cycles) [UR_26] d. The user can obtain detailed *technical* explanations on the trustworthiness assessment results (e.g., why the accuracy is low, how it was calculated, sources used) [UR_31, UR_32, UR_33]. e. The user obtains deeper understanding of the trustworthiness of the fake news/hate speech detection system by performing focused experiments, e.g., testing its performance in "hard" fake news datasets [UR_39, SR_33, SR_34, SR_49, SR_51]. f. Stores locally the trustworthiness assessment report [UR_30].	f. The user stores locally the final risk assessment report [UR_30].

User Journey 9 (cntd)	**User type: AI tech support for media organization who may work in a tech company** **Use Case:Media**	**Goal: The user wants to perform trustworthiness enhancement of the fake news/hate speech detection system that the media organization is using.**	**Expectations:** **-Deep understanding of the trustworthiness & risk assessment results.** **- Enhancement is tailored to the KPI objectives of all actors involved (media organization, AI tech company)** **-Short interaction with the tool**
Stage	Stage 3: EXPLORE		Stage 4: ENHANCE
Action:	a. The user reviews the THEMIS 5.0 solutions to mitigate vulnerabilities along with the estimated business/ legal/ethical benefits and risks for the tech company that the implementation of each solution entails [UR_46]. b. The user asks for explanations (e.g., input criteria used in the decision support process) in this process [SR_07, SR_36, SR_41, SR_42]. c. The user can experiment with alternative solutions to see their impact [SR_34]. d. The user identifies the optimal solution for the tech company and the media outlet based on the provided metrics.		a. The user selects a solution [UR_62]. b. The user reviews the new trustworthiness & risk assessment results, after the implementation of the solution [UR_67, UR_68] c. Upon request, the user is presented with analytical details on what has changed to the fake news/hate speech detection system (e.g., fine-tuned in dataset X) [UR_70]

User Journey 10	**User type:** Domain Expert	**Scenario:** The user wants to run the THEMIS 5.0 platform with an AI System built from scratch.	**Expectations:** The user can create its own AI model that can be used to further explore THEMIS 5.0 platform capabilities freely.

Stage	Stage 1: INITIALISE	Stage 2: EXPLORE	Stage 3: CREATE	Stage 4: RUN
Action	By launching the application, the user initialises the creation of a model that can be trained statically or dynamically with a real time data source.	At this point, there is a blank canvas in the centre of the screen, while on the left-hand side it is possible to choose algorithms, datasets or ready-to-use models from a special catalogue, the details of which will be customised in the menu on the right-hand side of the screen that appears after selecting them.	The elements available on the left can be transported with a simple drag-and-drop onto the white canvas available in the centre, then connected in the manner you deem most appropriate to create a functioning data flow.	Finally, click on 'Create application' to save the work performed. At this point, it will be possible to run the service, which will provide results based on what was previously created, e.g. processed datasets or trained models.

3.7. Mock-ups that visualise the co-created 'user journeys

3.7.1. UI Requirements

THEMIS 5.0 user interfaces were designed based on (i) the co-created User Journeys described in detail in the previous section and (ii) the user requirements collected by the Use Case partners and the project's co-creation activities. In addition, the following UI guidelines have been established based stemming from the partners' long-standing experience and strong expertise in the development of user-oriented software applications. These guidelines are based on the assumption that the applications and components will be either web-based or locally deployed, utilizing web technologies, and were taken into consideration for the design of the THEMIS interfaces and the overall User Experience alongside the requirements from certain EU legal and technical standards, such as the General Data Protection Regulation -GDPR[31], the ePrivacy Directive[32], the NIS directive[33], Web Content Accessibility Guidelines-WCAG[34].

- **Uniform behaviour and look-and-feel across all applications:** The UI must ensure consistent behaviour and a common look-and-feel of elements across all applications within the project. This uniformity is crucial for reducing bugs, improving maintainability, and providing a cohesive and seamless user experience. By enforcing both consistent behaviour and visual design, we can streamline development, minimize discrepancies, and enhance the overall reliability and cohesion of the entire project.
- **Reusability**. UI elements should be designed for easy reuse across different applications and to be easily shared among project teams. Reusability is challenging but crucial for the development of applications in heterogeneous environments and for ensuring UI components can be used in different frameworks. By designing UI components with reusability in mind, developers can ensure consistency and efficiency across various applications, reducing redundancy and minimizing the time and effort required for development.
- **Decoupling from frameworks:** The developed UI components should not be dependent on any specific JavaScript framework. This is important to allow different development teams to work on various applications and frameworks without the risk of vendor lock-in. It also facilitates maintenance in the event of a framework update or change, eliminating the need to rewrite or heavily modify existing components.
- **Interoperability with other web APIs**: The UI should work seamlessly with other web APIs, such as HTML5, CSS3, and JavaScript, without requiring additional wrappers or libraries.
- **Clear separation of responsibilities:** The UI design must ensure a clear separation between presentation (visual elements and layout) and functionality/logic (underlying code and behaviour). This separation is essential for making the codebase easier to manage, maintain, and scale as the project evolves.
- **Sharing resources**. UI elements developed should be shared and made accessible both technically and legally when possible. This will foster collaboration and innovation within the broader community, supporting FAIR practices, a goal to which this project is committed.
- **Style and behaviour isolation**: Components of the UI should encapsulate their styles and behaviour, preventing them from leaking out and affecting the rest of the application. This style encapsulation will prevent the risk of conflicting with global CSS that may be used in the applications.
- **Offline functionality**: Themis 5.0 applications may need to be deployed offline in a local, isolated environment. The UI and its components must be fully functional in this isolated setting, ensuring that all required resources, libraries, and dependencies are available locally.
- **Responsive design:** The UI must be designed to adapt seamlessly to a wide range of screen sizes and devices, ensuring a consistent and user-friendly experience across a wide range of devices such as desktops and tablets.

[31] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[32] https://eur-lex.europa.eu/eli/dir/2002/58/oj
[33] https://eur-lex.europa.eu/eli/dir/2022/2555
[34] https://www.w3.org/TR/WCAG21/

- **Cross-browser support**: The UI must be fully supported by all modern browsers and exhibit consistent behaviour across each of them.
- **Forward Compatibility**: UI resources should be based on open web standards, ensuring that the application's UI remains compatible with and supported by all modern browsers in the future. This approach is crucial for the longevity, stability, and maintainability of the components.
- **Performance**. The UI must not use unnecessary heavy frameworks or libraries that would affect the performance of the browser.
- **Accessibility Standards**: The UI should comply with accessibility standards, such as WCAG, to accommodate users with disabilities, ensuring that alternative text for images is provided, keyboard navigation is supported, and high-contrast colour schemes are available.
- **Internationalization and localization**: The UI should support multiple languages, enabling users to switch seamlessly between languages. It should also accommodate localization needs, including cultural factors like date formats and currency.
- **Real-time Feedback and responsiveness**: The UI should provide visual feedback, such as loading indicators, during data processing or loading to keep users informed. It should ensure that all interactions are responsive, minimizing delays to enhance the user experience.
- **User-friendly error messages**: The UI should display user-friendly error messages that are clear, informative, and provide actionable steps for resolution.
- **State preservation:** In case of a failure or an unexpected issue, the UI should preserve the user's state as much as possible, allowing them to continue from where they left off without losing data or context.
- **Guided tours and help sections**: Wherever necessary, the UI should include guided tours, tooltips, and help sections to assist users in navigating the application, especially for complex features.
- **Reduced initial load times**: The UI design and development should focus on techniques like lazy loading, resource optimization, and prioritizing critical content to reduce load times.
- **EU funding acknowledgment**: The EU funding acknowledgment statement should be visible in the UI, following the official guidelines.

3.7.2. THEMIS 5.0 Conversational Interface

THEMIS 5.0 framework is largely based on a conversational interface. Apart from being a means to facilitate and guide users while they are navigating through the THEMIS 5.0 framework and applications, certain essential THEMIS 5.0 features are foreseen to be implemented exclusively through a conversation-based Human-Machine interaction including:

- Identification of the user's persona and moral orientation to derive the users' trustworthiness preferences.

  "The THEMIS 5.0 personalised dialogues will capture and provide human values, preferences, requirements, human-defined objectives, capabilities, motives and behavioural patterns to all phases of the [THEMIS] methodology." (THEMIS 5.0, Grand Agreement, Description of Action, section 1.2.3.5, p.14)

- Creation (also with the use of a Graphical User Interface) of the socio-technical model by associating trustworthiness vulnerabilities to business-related risks and target KPIs, and support users to understand the impact of the alternative trustworthiness optimisation measures and to decide on which measure is the most suitable to be implemented.

  "… the Human-AI conversational agent will elicit knowledge related to human particular decision support needs and ethical values, as well as to the key success factors of the wider socio-technical system" (THEMIS 5.0, Grand Agreement, Description of Action, section 1.2.3.6, p.14)

  "… leverage trustworthiness optimization and, hence, decision improvement through AI-driven dialogues, …" (THEMIS 5.0, Grand Agreement, Description of Action, section 1.2.3.6, p.14)

"… the human user using the developed DI methodology and supporting graphics will create a qualitative model of the socio-technical environment. The model will contain all the possible actions that the AI system can recommend, all the KPIs that might be affected by any of the possible actions, any external factors that might influence the actions or the actions' effect on the KPIs, and how the actions and the external factors affect the KPIs through pairwise relations. (…) a DI AI simulation will be trained that generates a quantitative simulated model of the socio-technical system." (THEMIS 5.0, Grand Agreement, Description of Action, section 1.2.3.2, p.12)

"The provided human-interpretable explanations will inform human recipients about: the input criteria used in the decision support process (training data sets, AI models); the output of that process; and the perceived causal relationship between input and output, taking into consideration the characteristics of the human recipients of decision support, the context and circumstances triggered the decision support, and the objectives pursued thereby." (1.2.3.6, p.14)

Moreover, considering that THEMIS 5.0 is an ecosystem of interconnected components and applications giving rise to a complex set of functionalities and alternative paths that can be overwhelming, the conversational interface in the form of a conversational agent will also provide personalised user guidance and explanations as needed to ensure that all THEMIS 5.0 users will be able to take advantage of the personalised trustworthiness assessment and optimisation potential of the project.

"The THEMIS 5.0 ecosystem is composed of cloud-based, AI-based services that seamlessly engage with humans by means of AI-driven interactive Dialogues. Specifically, an AI-driven conversational agent will transmit sufficient but not excessive human-interpretable explanations on how the AI system takes a particular set of inputs and reaches a conclusion" (1.1.1, p.3)

3.7.3. THEMIS 5.0 Graphical Interface

In addition to the conversational aspect of the THEMIS 5.0 interface, a Graphical User Interface (GUI) will be implemented to (i) communicate the visual aspects of the THEMIS 5.0 platform's applications, (ii) visualise the overall progress throughout the THEMIS 5.0 platform, and (iii) allow users to provide the necessary human input for the socio-technical modelling and the risk-assessment based decision support tool.

Based on the User Journeys that were presented in section 3.6 above and the user requirements for the THEMIS 5.0 platform, a tentative mock-up design for its Graphical User Interface was produced, whose purpose is to be used alongside the co-created User Journeys as input material for co-creation phase B. Thus, participants of the co-creation living labs will have a concrete understanding of the THEMIS 5.0 purpose, structure, and functionalities, ensuring that the further definition of the THEMIS 5.0 user requirements will be based on input from sufficiently informed participants who are aware of the THEMIS 5.0 potential and limitations. This mock-up GUI design is presented in this and in the following subsections and while not binding, it is expected to drive the discussion and be used as a reference point for the development of the final THEMIS 5.0 GUI. The Graphical User Interface should feature a common structure and common visual elements across all components and applications of the THEMIS 5.0 ecosystem, fostering a unified and seamless User Experience. Therefore, the GUI layout comprises a set of static sections, whose configuration can be seen in the following Figure 17, Figure 18 and Figure 19, following the users throughout their journey across THEMIS 5.0. ecosystem. These sections are the (i) Header section, (ii) Footer section, (iii) Application control panel, (iv) Conversational interface panel, (v) Main Application window and they are presented in detail in the following subsections.

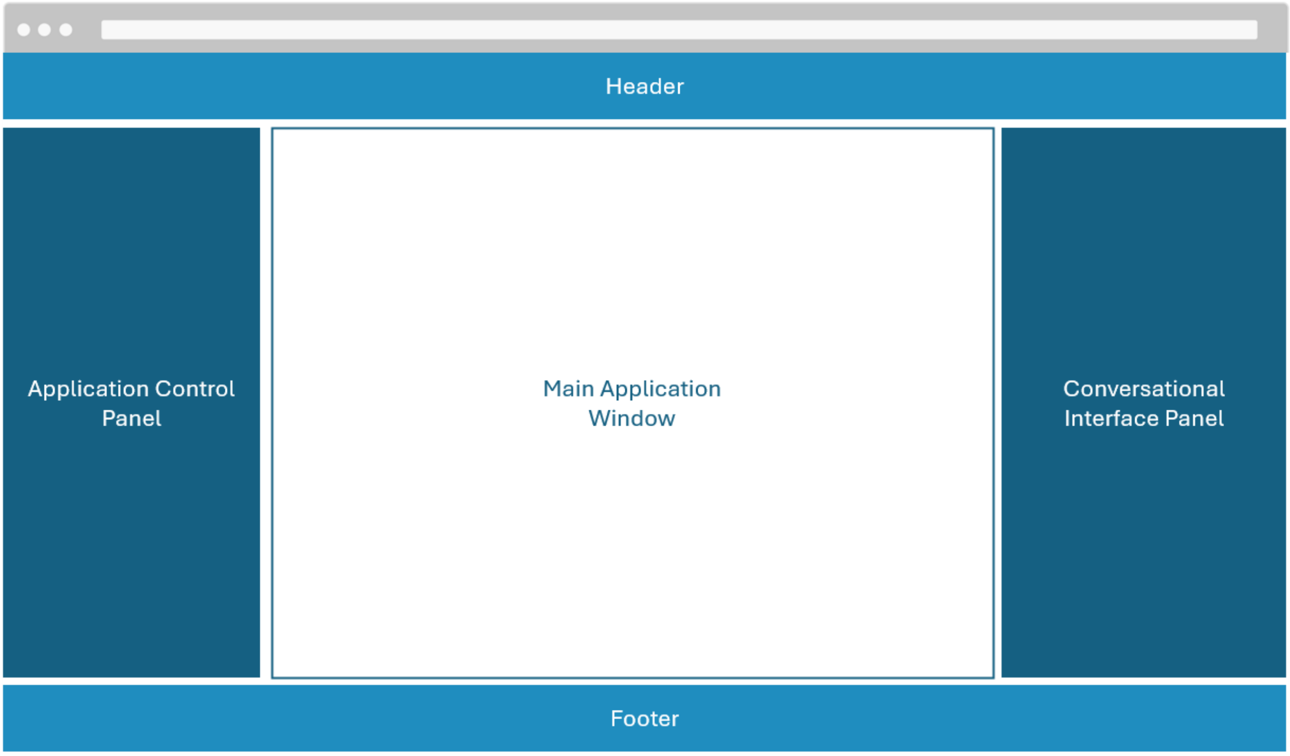

Figure 17: THEMIS 5.0 Graphical User Interface: Spatial Configuration of Basic UI Elements

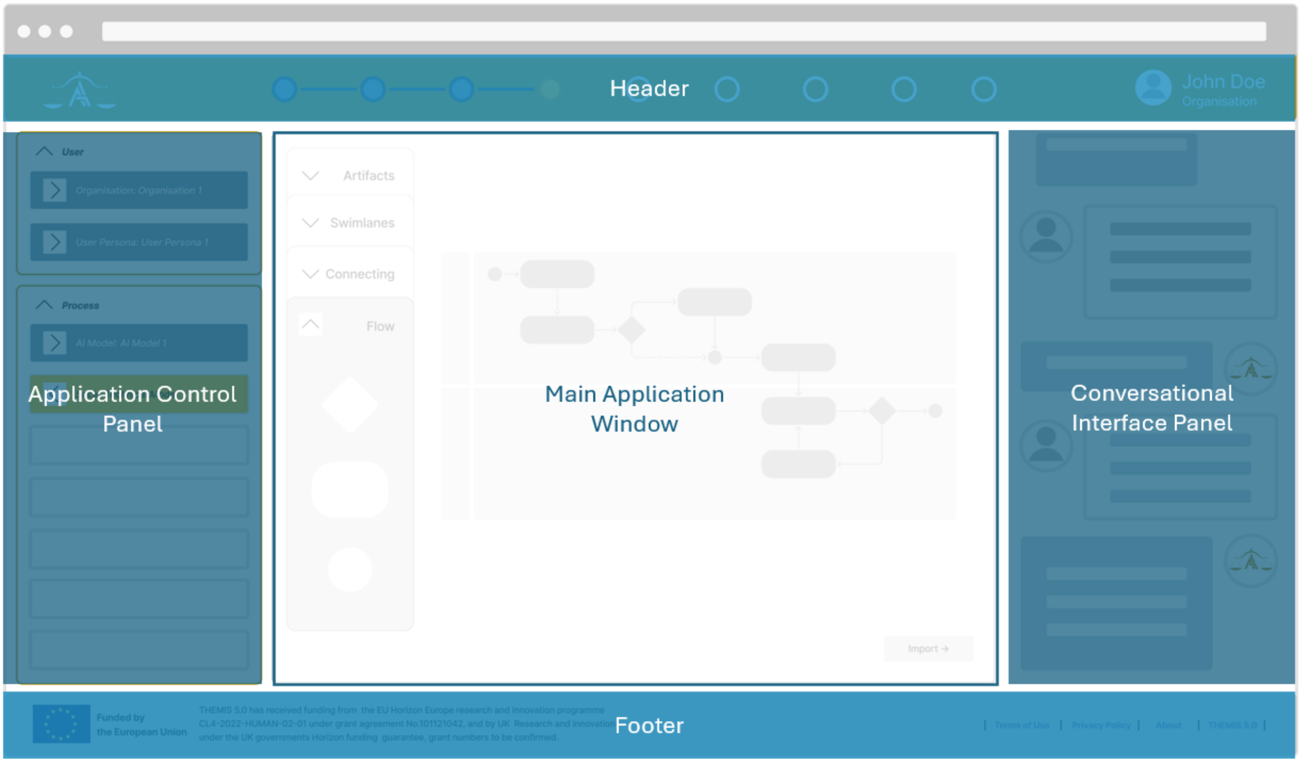

Figure 18: THEMIS 5.0 Graphical User Interface: Spatial Configuration of Basic UI Elements

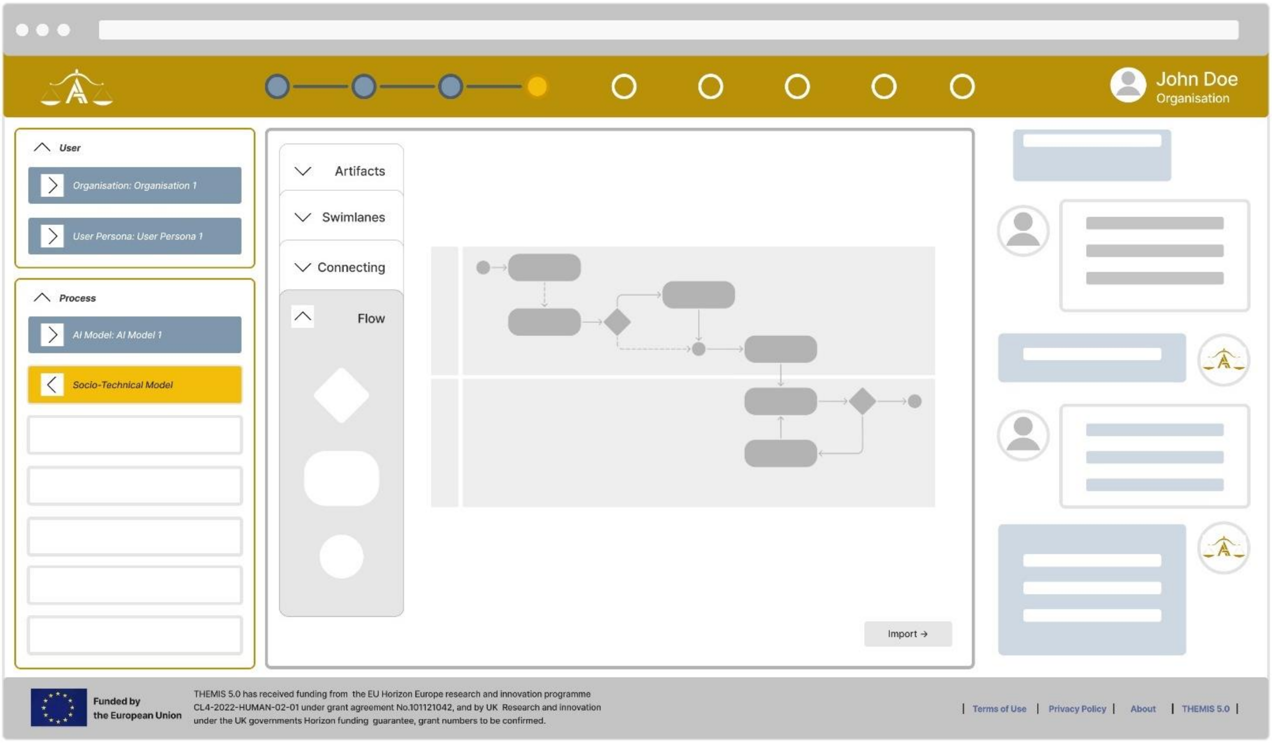

Figure 19: THEMIS 5.0 Graphical User Interface: Screen Example

3.7.3.1 Header

At the top of the page, the application *header* (see **Figure 20** below) is a static horizontal UI element, on top of all the other UI elements underneath, binding them with the THEMIS 5.0 context and brand. The header is separated in three main parts, namely (i) the left part hosting the THEMIS 5.0 logo, (ii) the central part that contains a dynamic visual element that informs the user about the progress and the remaining steps across the THEMIS 5.0 platform, and (iii) the right part hosting the logged-in user's basic information (Icon/image, name/alias and organisation).

Figure 20: THEMIS 5.0 Graphical User Interface: Header

3.7.3.2 Footer

At the bottom of the page, the application *footer* (see Figure 21 below) is another static horizontal UI element, that together with the header, frames the other UI elements between them. The footer is separated in two main parts. The left-side part hosts the EU emblem and a funding statement acknowledging and informing users that the project and its results are funded by the European Union, while the right-side part features hyperlinks allowing users to easily access the THEMIS 5.0 Framework's *Terms of Use* and *Privacy Policy*, as well as the *about* page providing more information about the framework and the *THEMIS 5.0 project official website*.

Figure 21: THEMIS 5.0 Graphical User Interface: Footer

3.7.3.3 Application control panel (Left)

At the left side of the page, the *application control panel* is a dynamic vertical UI element that allows the user to navigate between the steps of the THEMIS 5.0 pipeline and ecosystem's applications, while also providing a visual anchor to users about the current step, the overall progress across the pipeline and the remaining steps until reaching the end of the pipeline. The *application control panel* (see Figure 22 below) always occupies a predefined area of the screen as well as its sub elements whose visual state changes from (i) unavailable (empty placeholder frame), to (ii) available (light grey frame), to (iii) active (yellow frame) and eventually (iv) completed (dark grey blue frame) according to the transition from the respective application's state.

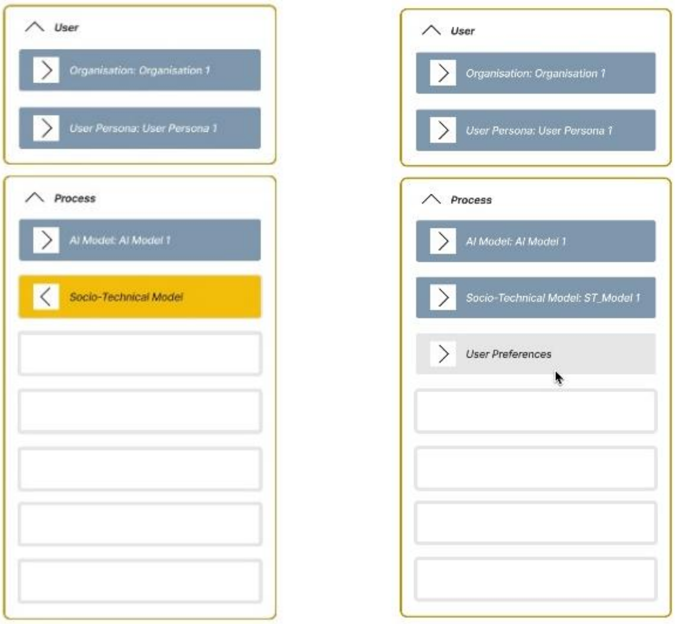

Figure 22: THEMIS 5.0 Graphical User Interface: Application Control Panel - different states

3.7.3.4 Conversational agent panel (Right)

Similarly, at the right side of the page, the *conversational agent panel* (see **Figure 23** below) is the dynamic vertical UI element that displays the conversations between the user and the conversational agent. This panel remains continuously active to guide users while they are navigating through the THEMIS 5.0 framework and applications and to support the THEMIS 5.0 features that are foreseen to be implemented exclusively through a conversation-based Human-Machine interaction.

Figure 23: THEMIS 5.0 Graphical User Interface: Conversational agent panel

3.7.4. Mock-ups

In this sub-section, the mock-ups of that were developed to help explain the THEMIS 5.0 ecosystem, flow and functionalities are presented, including visual mock-ups to describe the foreseen interaction that is facilitated via a Graphical User Interface (GUI) as well as mock-up (example) dialogues to describe the foreseen interaction with the Conversational Interface (CI). Together, they aim to provide the participants of the co-creation phase B with a clear understanding of the THEMIS 5.0 platform's purpose, potential and limitations.

The configuration of the THEMIS 5.0 ecosystem and the functional pipeline that supports the anticipated functionalities was defined based on the *Description of Action* as well as the user requirements and the co-created user journeys that are presented in the previous sections. As presented in section 3.6 above, this pipeline includes 12 key functionalities that are distributed in the 4 phases of the *Trustworthiness Optimization Process (TOP)* as follows:

1. IDENTIFY (I)
    - I1. User indicates AI system
    - I2. User models AI system
    - I3. User models the socio-technical environment
    - I4. Formulation of the user persona
    - I5. User trustworthiness preferences prediction
    - I6. User provides/adjusts trustworthiness preferences

2. ASSESS (A)
    - E1. User reviews trustworthiness assessment
    - E2. User reviews risk assessment

3. EXPLORE (X)
    - X1. User reviews suggested solutions
    - X2. User experiments to design and deploy experiment variants

4. ENHANCE (N)
    - N1. User selects solution to implement
    - N2. User reviews updated trustworthiness and risk assessment

3.7.4.1 GUI Mock-ups

These key functionalities appended with other necessary functionalities, and they were broken down into detailed steps and they were grouped into 12 key flow steps, to produce a comprehensive description of the THEMIS 5.0 user experience. In **Figure 24** below, the complete map of the **50 keyframes** is presented, organising the 12 major steps in the THEMIS 5.0 pipeline flow from the top to the bottom and the internal progress for each pipeline step from the left to the right. In this section, a sub-set composed of the most important amongst the developed mock-ups will be presented corresponding to the aforementioned key functionalities and key flow steps, while the complete set of the developed mock-ups is presented in Annex 4.

Figure 24: THEMIS 5.0 mock-ups map (Vertical: pipeline steps, Horizontal: internal progress for each pipeline step)

The numbering convention of the mock-ups presented hereinafter, follows the *Step X.Y.* format, with *X* indicating the pipeline step and *Y* indicating the internal step for each pipeline step that each mock-up belongs to.

As a starting point, users are requested to insert their credentials to sign in and access the THEMIS 5.0 services or to sign up (Figure 25 below).

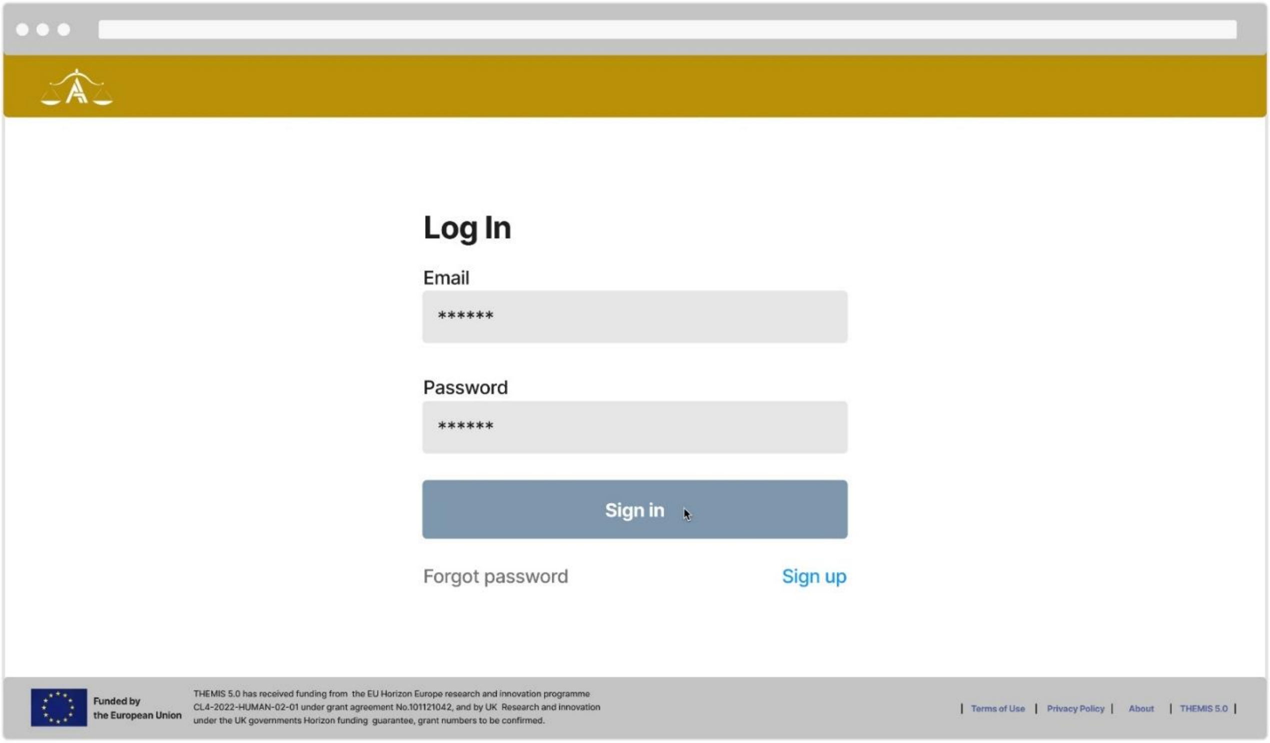

Figure 25: Step 0.0 User Log-in

Upon signing in, users are redirected to the welcome screen (Figure 26 below) where they can start an interaction cycle by pressing the *Start* button, or by asking the conversational agent to do so.

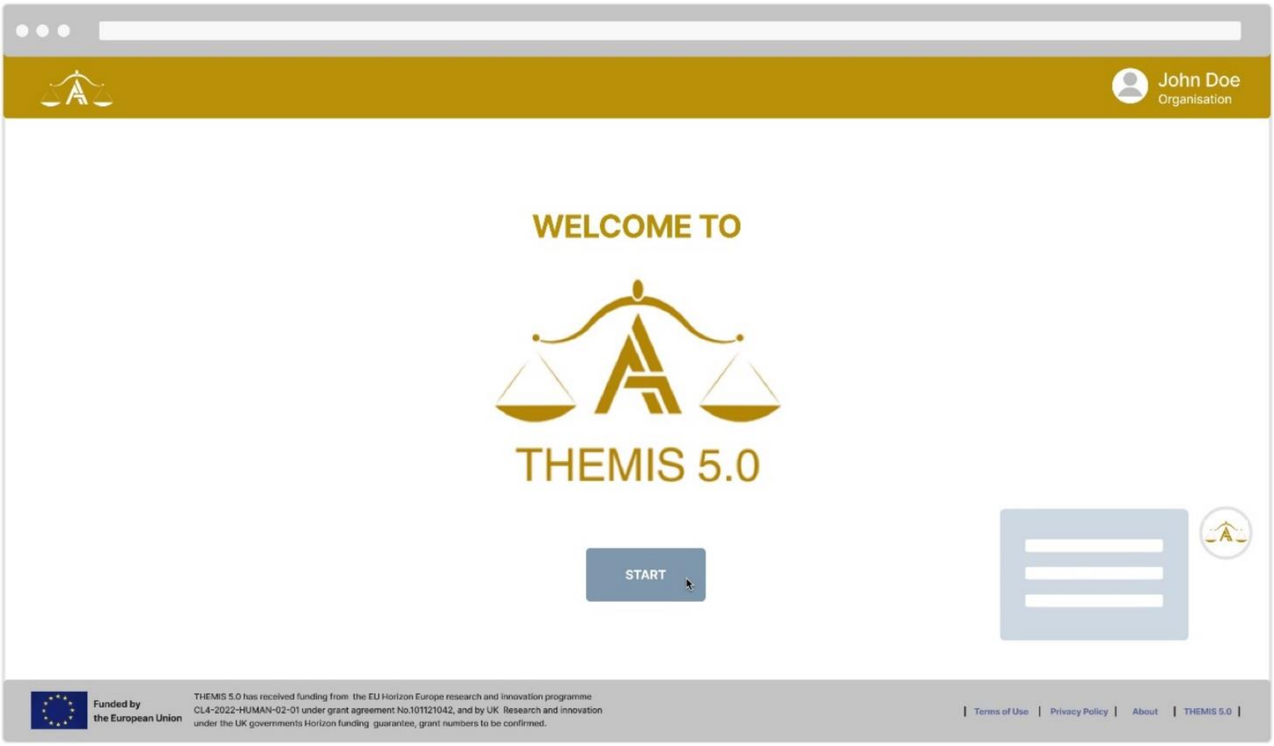

Figure 26: Step 1.0 Start THEMIS 5.0 Pipeline

The second step in the actual THEMIS 5.0 pipeline manages the organisation requirements. Users with elevated organisational rights, can indicate the legal, ethical, and regulatory frameworks that their organisation should comply with (Figure 27 below) while other users can review this information without being able to make changes.

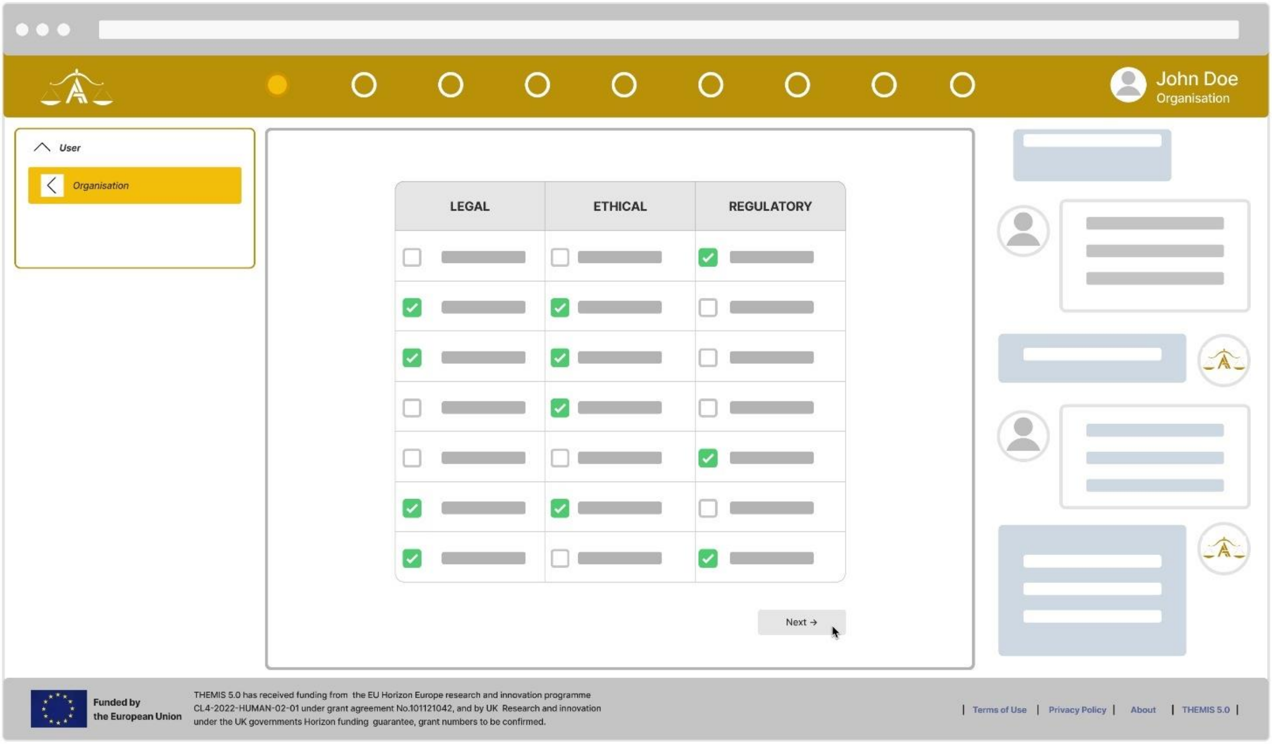

Figure 27: Step 2.2 Organisation requirements

The next step manages the user's persona. In this step, the conversational agent (Figure 28 below, for mock-up dialogues see chapters *3.7.4.2.1 User Persona* and *3.7.4.2.2 Moral orientation*) interviews the user to derive the users' persona, moral orientation and users' group that they belong to.

Figure 28: Step 3.2 Set user persona (Key Functionality I4)

As users may be using and managing a number of different AI systems, at the next step, users can add or create a new process that all subsequent pipeline steps will refer to (Figure 29 below). Thus, trustworthiness assessment and optimisation will be process-specific while user organisation, personas and moral orientation persist across all processes.

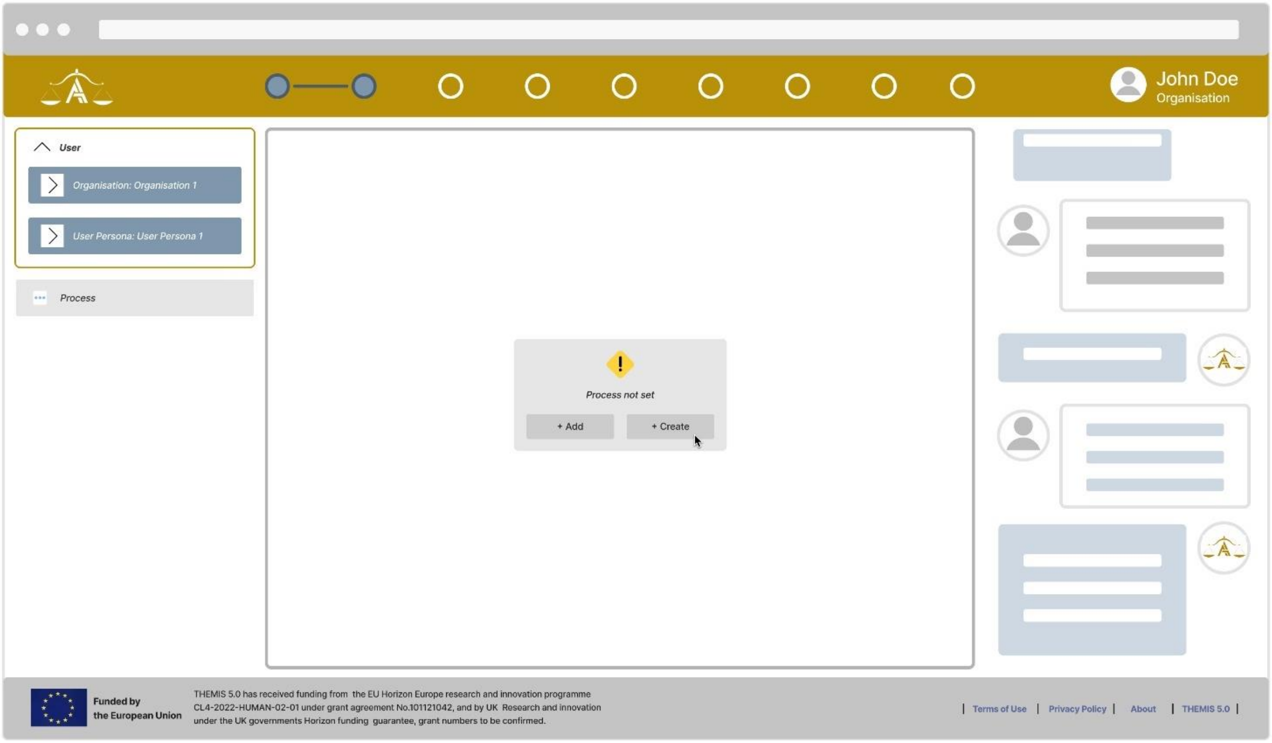

Figure 29: Step 4.0 Create a new process

Upon creating or adding an existing process, in the next step users indicate the AI system to be assessed. Users can scroll through the already available AI-models and there's also the option to create a new one. Upon selection of a certain model, users can overview its details and can proceed with its assessment (Figure 30 below).

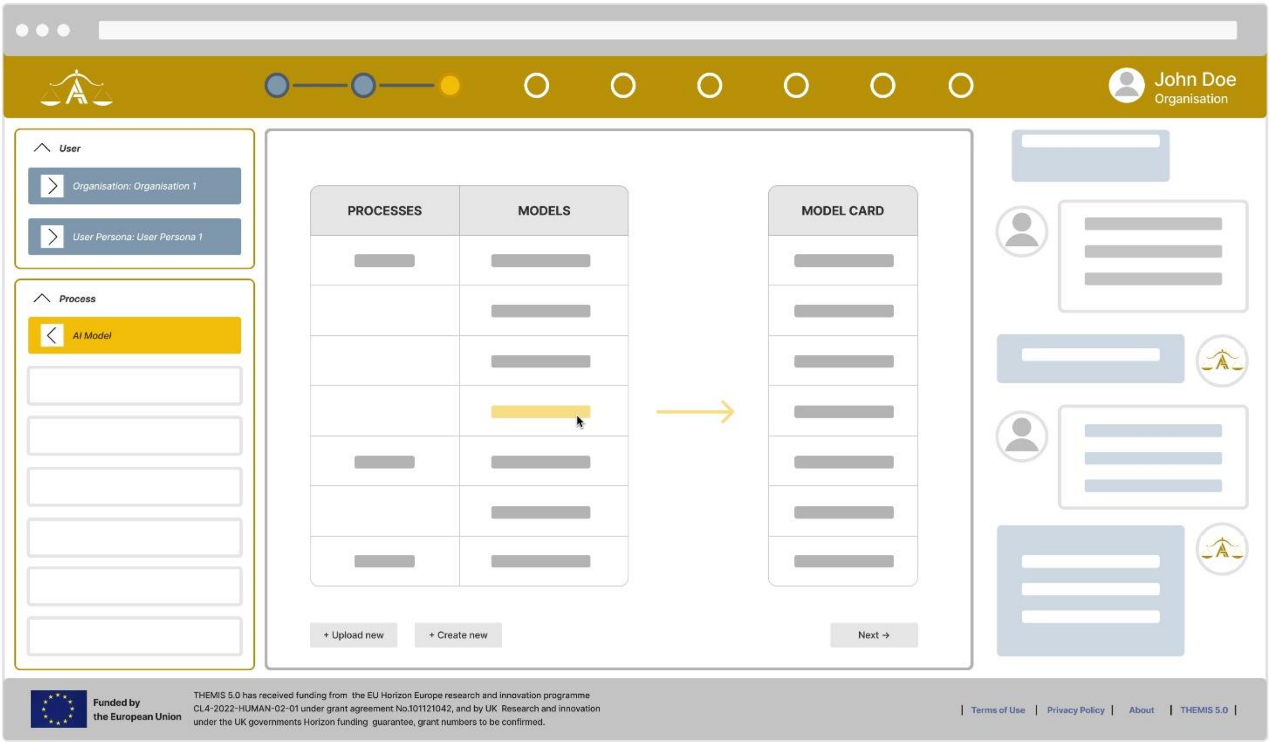

Figure 30: Step 5.5 User indicates AI model to be assessed (Key Functionality I1)

Within the same step, if the *create new model* feature is selected, users can use a toolbox with Models, Datasets and services, within a drawing environment, that allows them to create a new AI-model from scratch (Figure 31 below). Once finished, the new model can be imported into the AI Models' library.

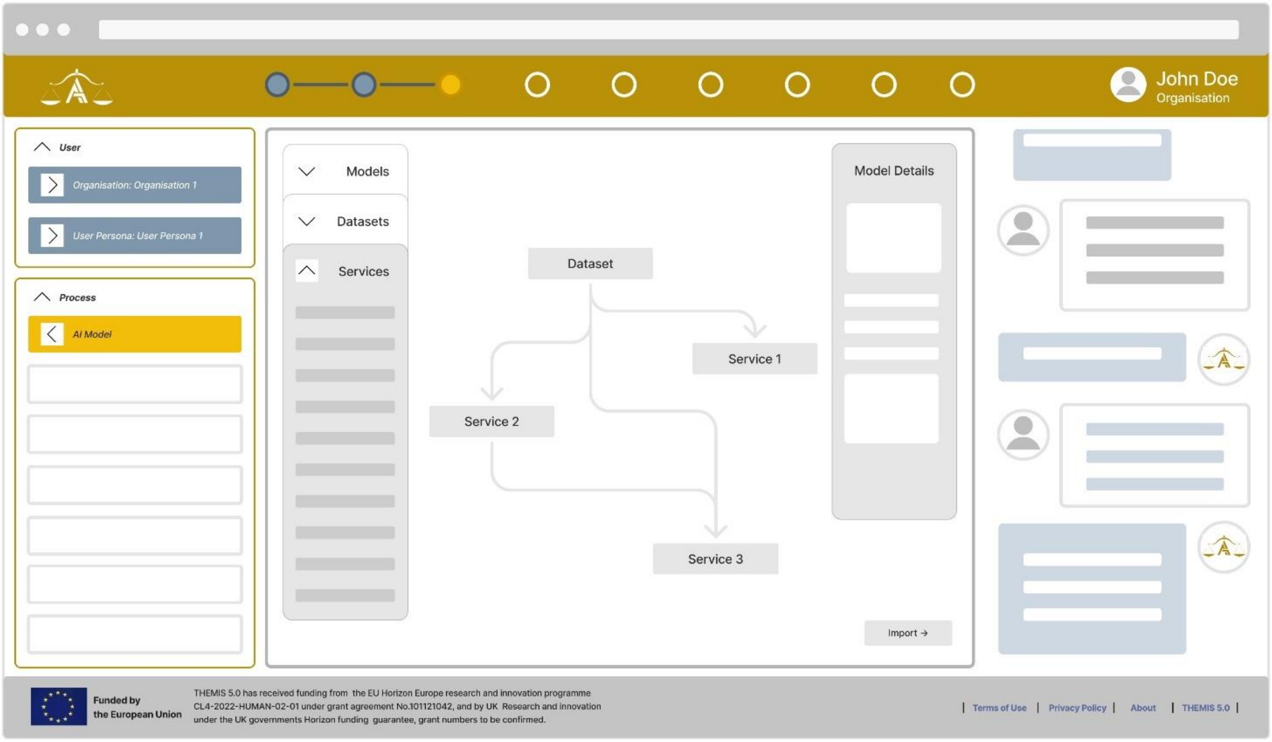

Figure 31: Step 5.2 User models AI model (Key Functionality I2)

The next step manages the socio-technical context for the System Under Test. In this step, users use a toolbox that allows them to model the socio-technical environment, following the Business Process Modeling and Notation (BPMN), to define how the trustworthiness characteristics influence the processes and the business objectives (Figure 32 below).

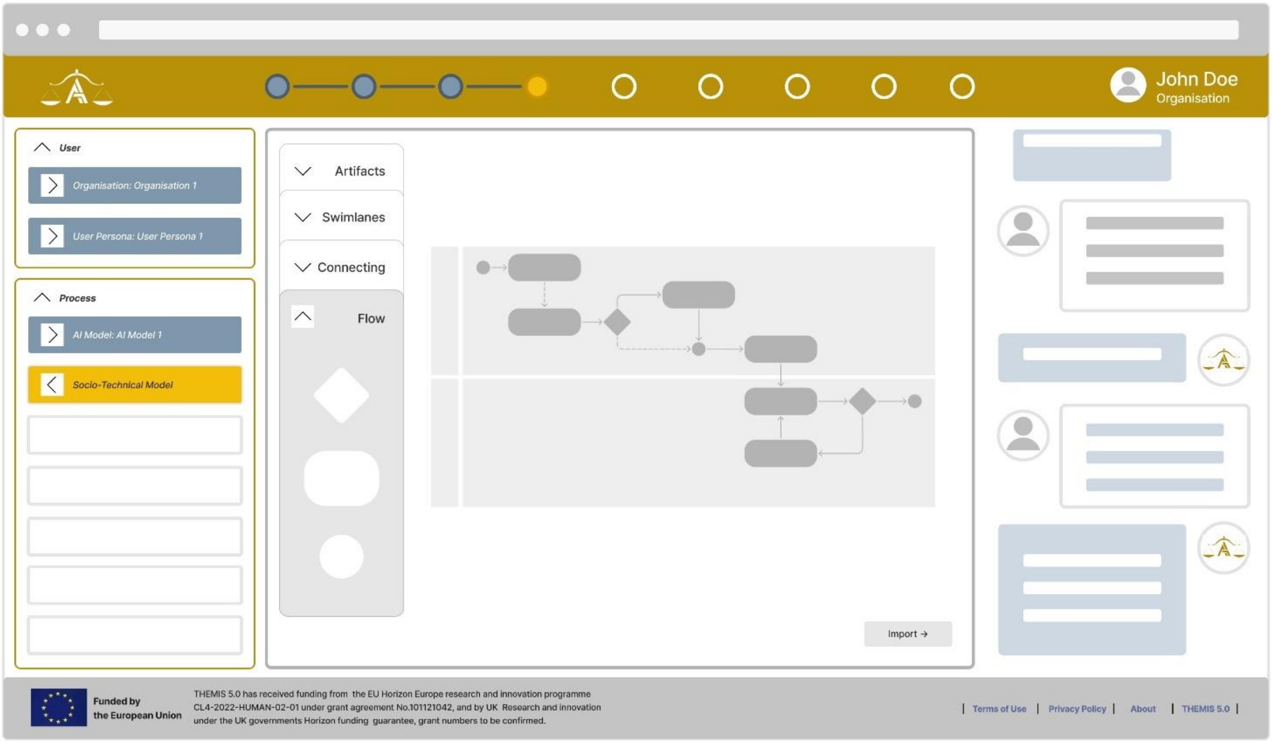

Figure 32: Step 6.2 User models the socio-technical environment (Key Functionality I3)

In the next step, the users' trustworthiness preferences for this particular process are derived based on their persona and moral orientation and users have the opportunity to further fine-tune the settings (Figure 33 below).

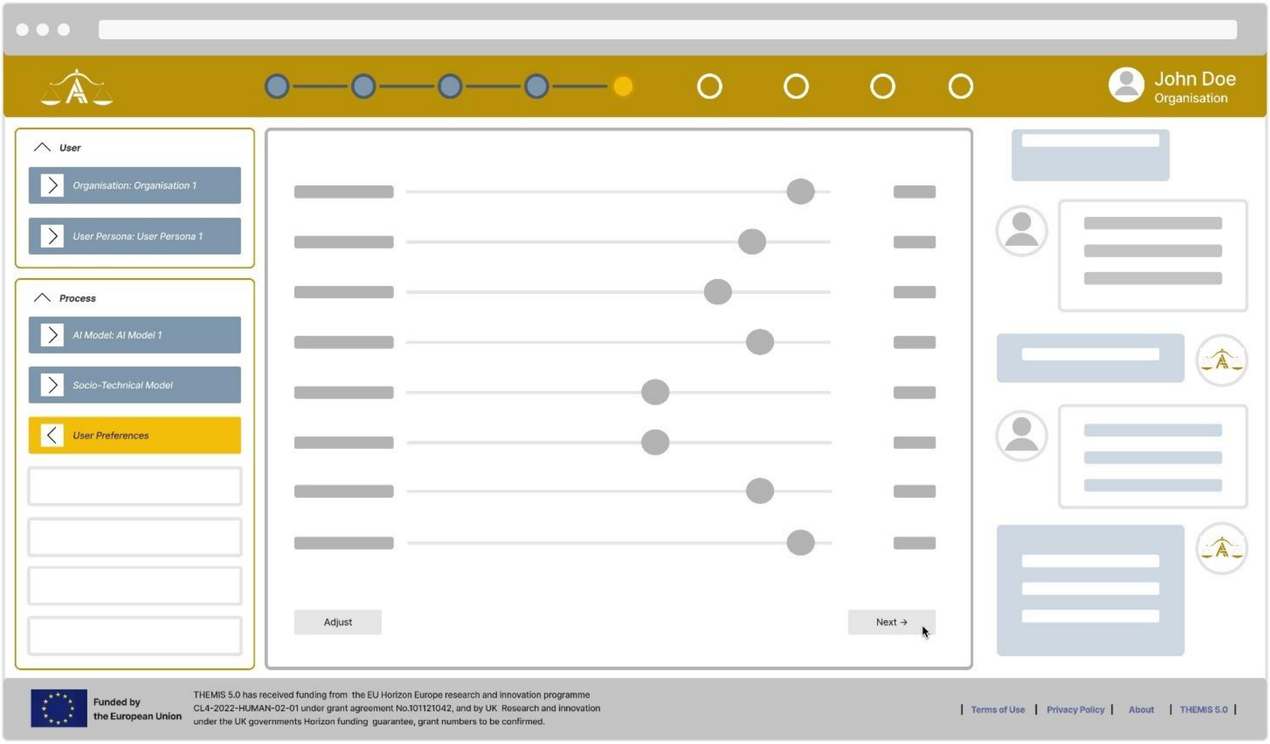

Figure 33: Step 7.2 User reviews, provides or adjust trustworthiness preferences (Key Functionalities I5 & I6)

Within the next step, following the trustworthiness assessment of the AI System Under Test, the assessment results are presented to the user (Figure 34 below).

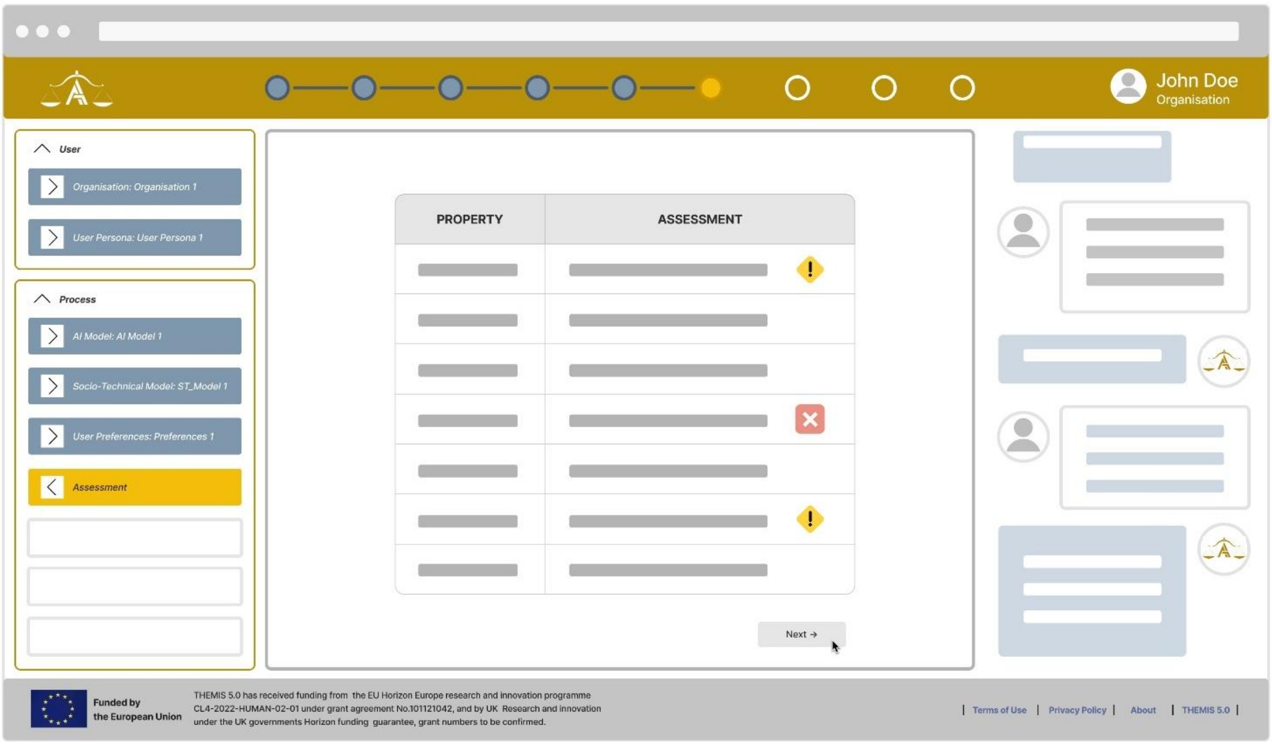

Figure 34: Step 8.2 User reviews the trustworthiness assessment results (Key Functionality E1)

The next step is implemented through the conversational agent (Figure 35 below; for mock-up dialogues see chapter *3.7.4.2.3 Trustworthiness optimisation solution exploration*). The user is informed about the impact of the trustworthiness assessment results and is assisted in the exploration of potential solutions also considering the estimated impact.

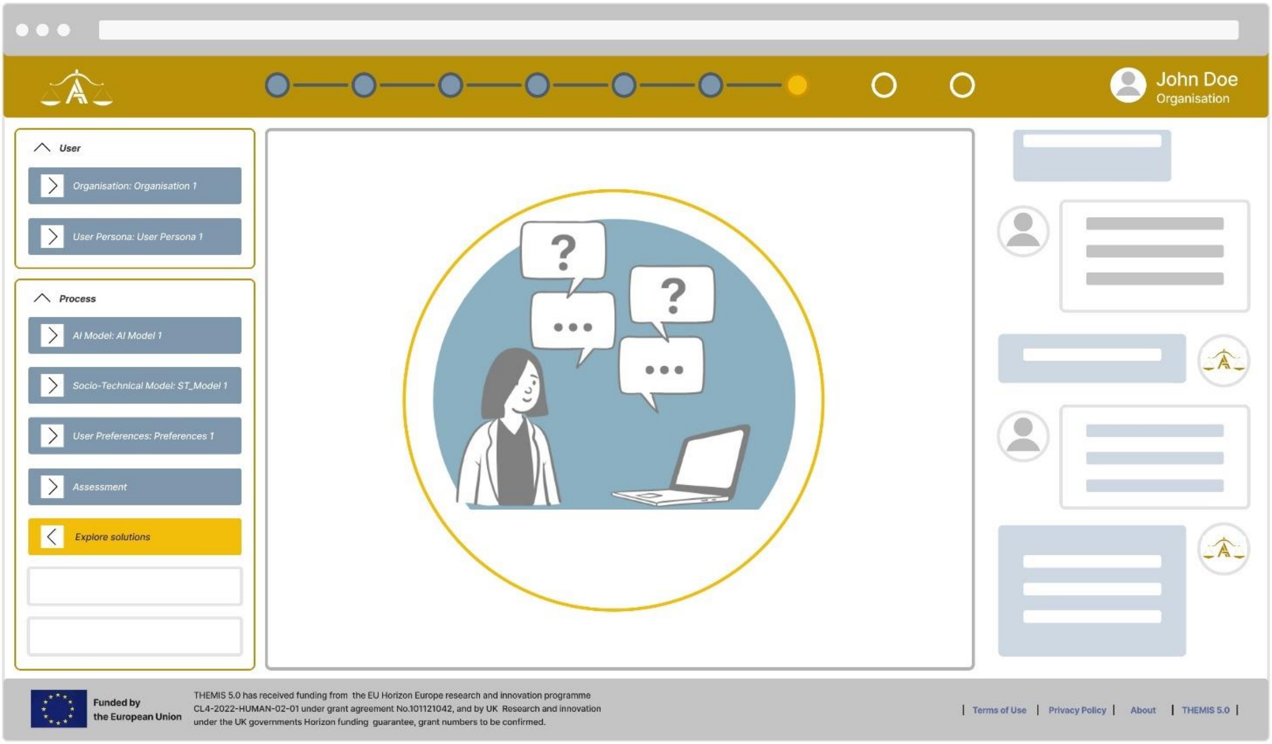

Figure 35: Step 9.2 User reviews the risk assessment results and explores solutions (Key Functionality E2)

Thereupon and within the same step, the various solutions are presented to the user by order of suitability to the user's needs and preferences (Figure 36 below) where the user can select the solution(s) to be promoted.

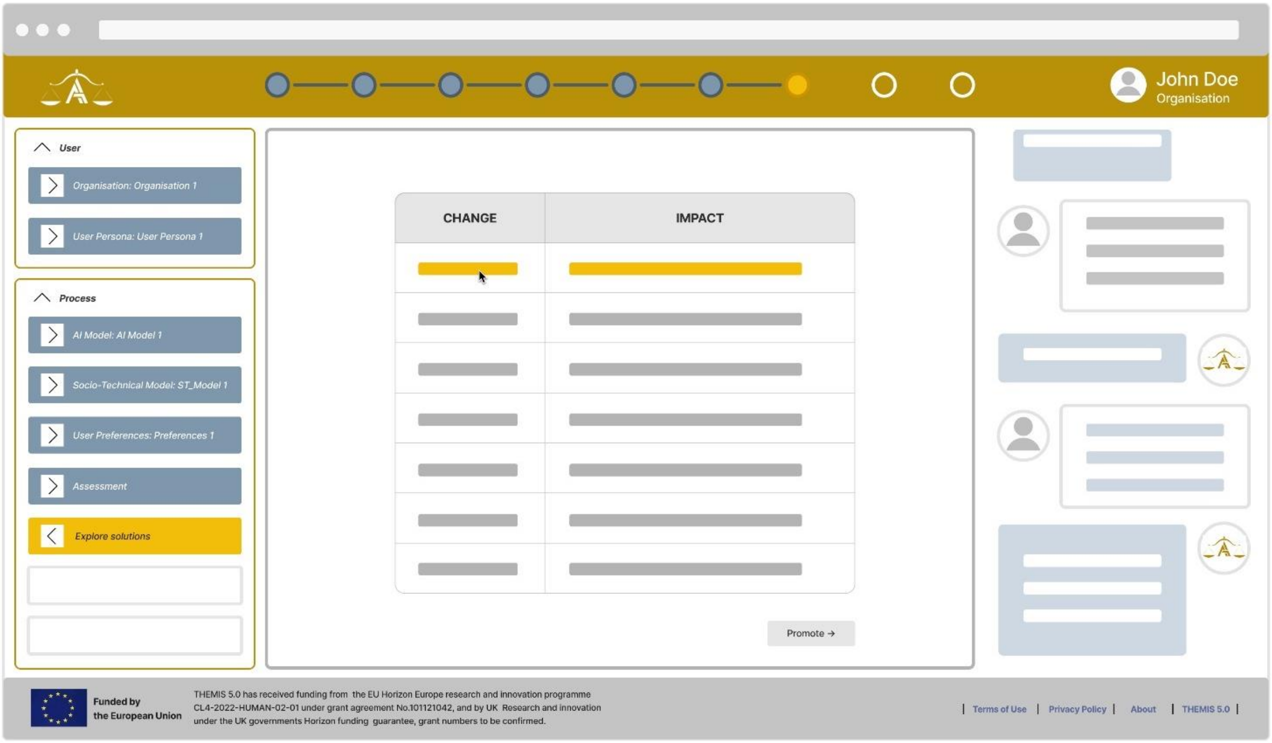

Figure 36: Step 9.4 User explores solutions (Key Functionality X1)

In the next step, experienced users have the option to further experiment with the selected solution. Users can see the available models and can also upload a new model to the model library. They can also select a model to inspect its details (Figure 37 **Error! Reference source not found.**).

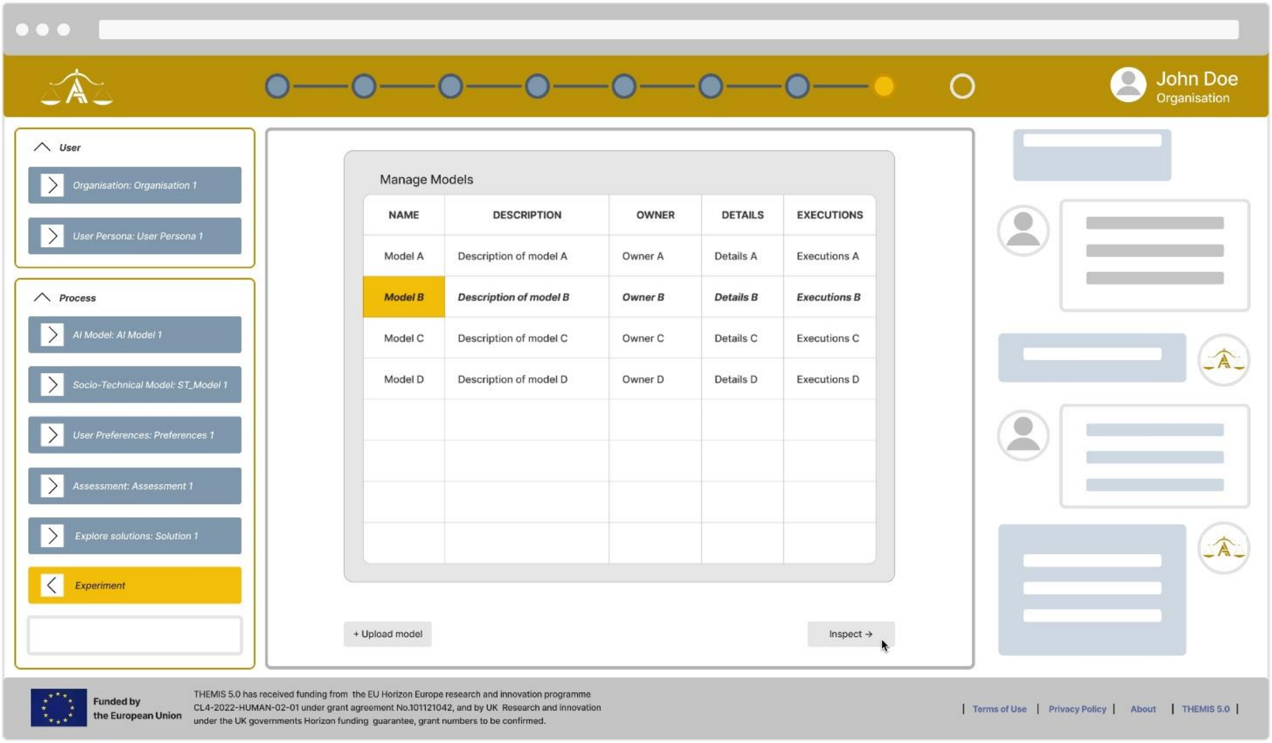

Figure 37: Step 10.2 User explores alternative experiment versions (1) (Key Functionality X2)

In the next screen of this step, users can inspect the selected models' details and can optionally select to configure the model (Figure 38).

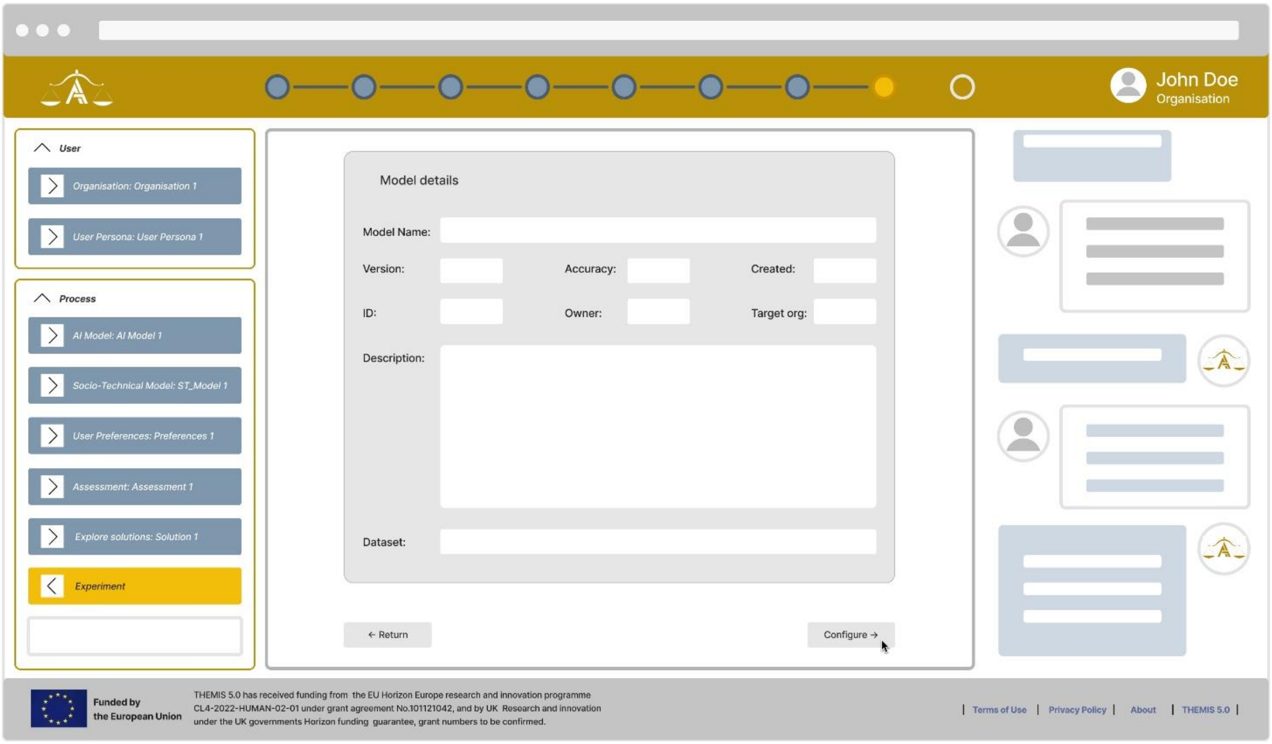

Figure 38 Step 10.3 User explores alternative experiment versions (2) (Key Functionality X2)

Furthermore, users can add, delete and select algorithms to and from the algorithm library and upon selection they can proceed to the configuration of the model's dataset (Figure 39 below).

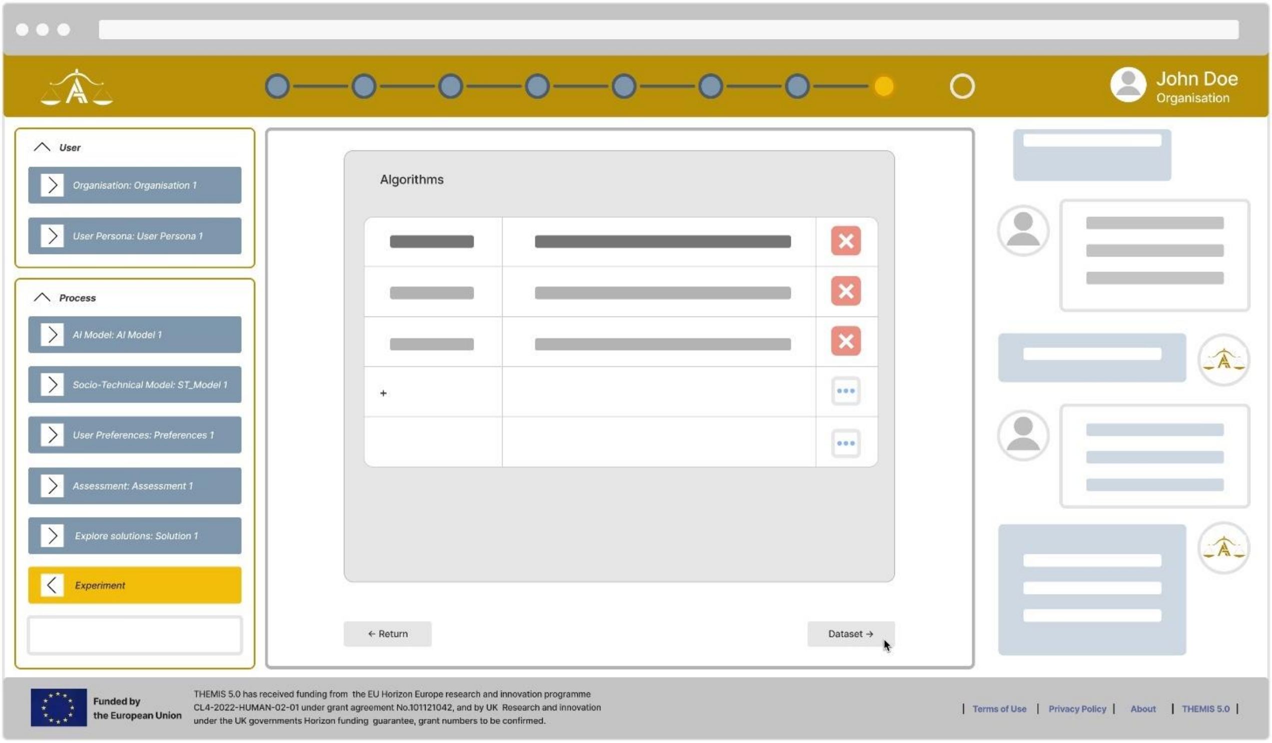

Figure 39: Step 10.4 User explores alternative experiment versions (3) (Key Functionality X2)

Then, users can review the selected dataset in a tabular format, and they can select the dataset fields to be included in the model while another dataset can also be included (Figure 40 below).

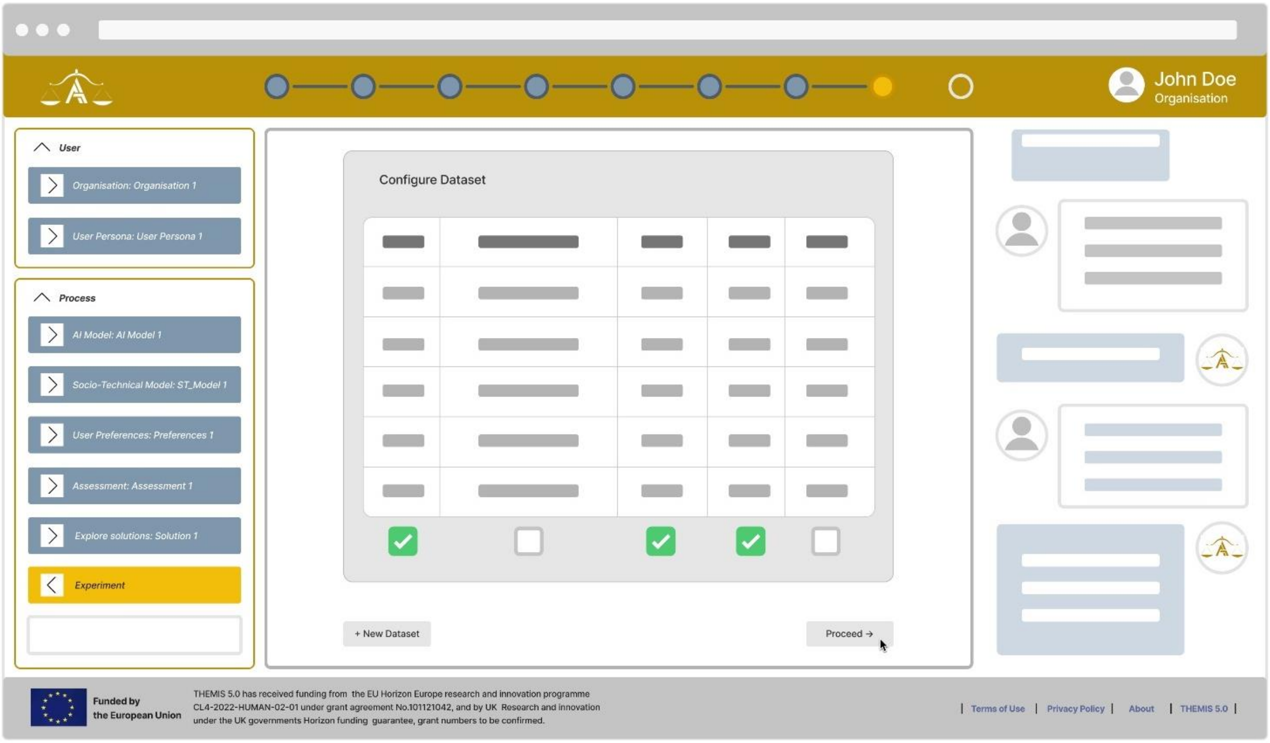

Figure 40: Step 10.5 User explores alternative experiment versions (4) (Key Functionality X2)

Finally, users go through the available solutions (pre and post experimentation) and select the one to be implemented (Figure 41 **Error! Reference source not found.**).

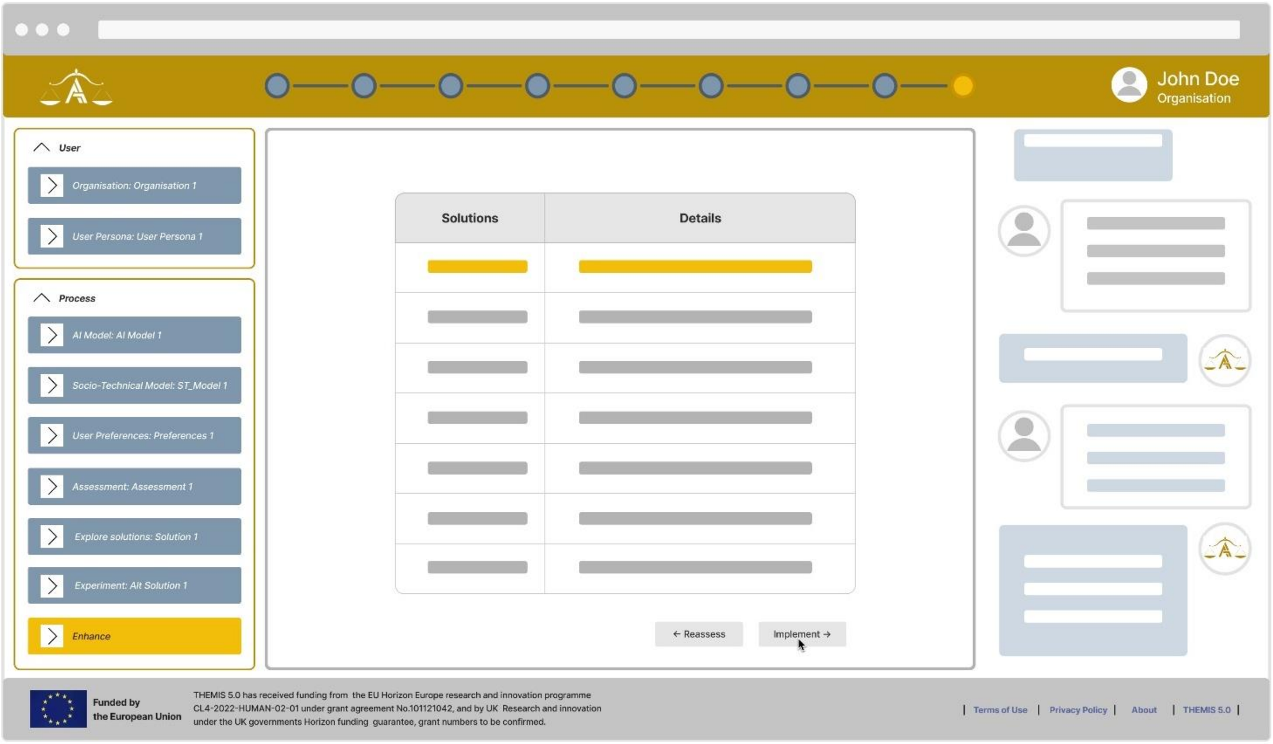

Figure 41 :Step 11.2 User selects solution to implement (Key Functionality N1)

The users are then informed that the selected solution will be implemented, and that they will be informed once the trustworthiness optimisation is completed (Figure 42 below) where they can either quit without proceeding or confirm the implementation of the selected improvement before quitting the application.

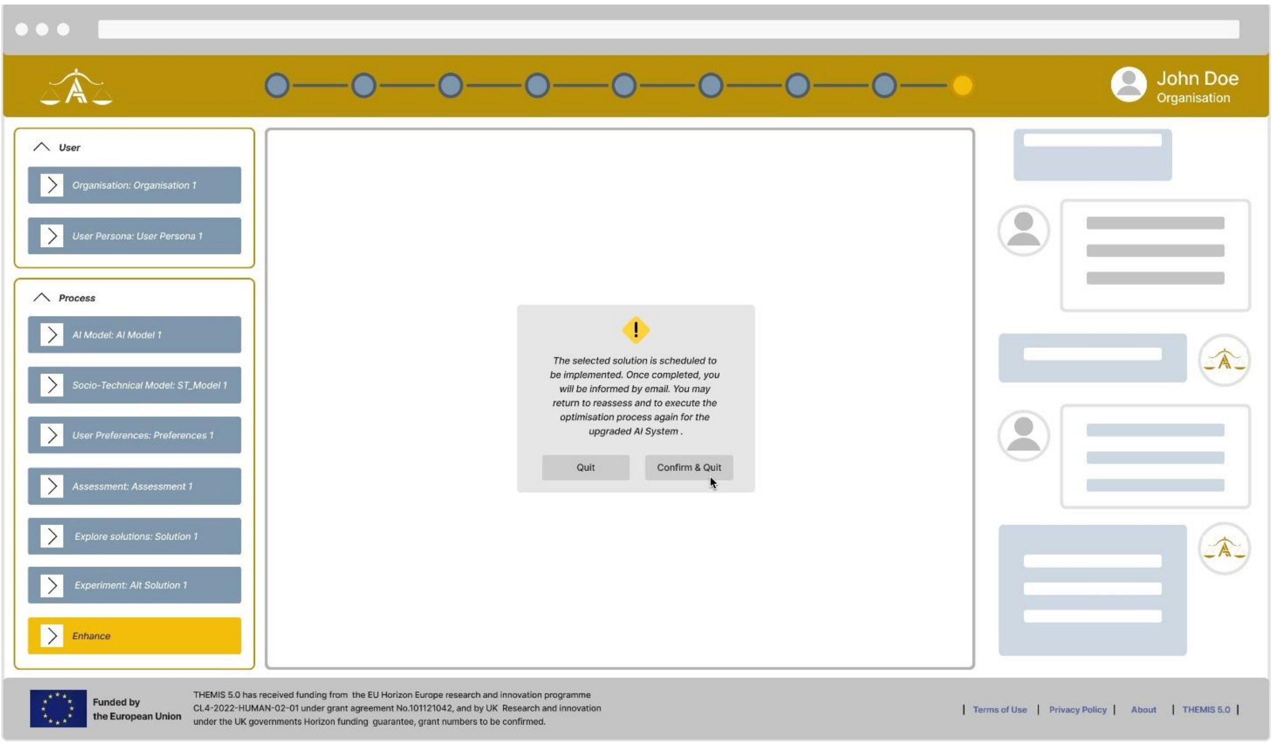

Figure 42: Step 11.3 User selects solution to implement and to reassess (Key Functionality N2)

3.7.4.2 Mock-up dialogues

In this section, mock-up example dialogues are presented aiming to showcase how the THEMIS 5.0 functionalities that depend exclusively on a conversational interface are foreseen to interact with the users of the THEMIS 5.0 platform also taking into consideration the differences between the three *use cases* of the project.

3.7.4.2.1 User Persona

The example dialogues hereinafter demonstrate how the user's persona is derived based on dialogue between the users and the THEMIS 5.0 conversational agent. (**T5**: THEMIS 5.0 Conversational Agent, **U**: User)

T5 1.1: "Can you tell me about your professional experience in your current field?"

- **U 1.1.1:** "Less than 5 years of experience."
  [Likely to be clustered into "Young Tech-Savvy Professionals" in the healthcare sector or "Operational Staff with Basic Technical Skills" in the port management sector.]
- **U 1.1.2:** "5-15 years of experience."
  [May indicate mid-career professionals who are more experienced but still open to new technologies. Could fall into "Technically Proficient Managers" or "Sceptical Journalists" based on their attitudes towards AI.]
- **U 1.1.3:** "More than 15 years of experience."
  [Likely to be clustered as "Experienced Sceptical Professionals" in healthcare or "Experienced Managers" who require high transparency and reliability in AI tools.]

T5 1.2: "How comfortable are you with using new technologies, particularly AI tools, in your work?"

- **U 1.2.1:** "Very comfortable."
  [Could indicate "Young Tech-Savvy Professionals" in healthcare, "Technically Proficient Managers" in port management, or "Fact-Checkers and Analysts" in disinformation who are open to using new technologies.]
- **U 1.2.2:** "Somewhat comfortable."
  [May suggest users who are adaptable but cautious, potentially falling into clusters like "Operational Staff with Basic Technical Skills" or "Sceptical Journalists."]
- **U 1.2.3:** "Not comfortable at all."
  [Likely to be grouped into clusters requiring more support and training, such as "Operational Staff with Basic Technical Skills" or "Experienced Sceptical Professionals" who are resistant to AI.]

T5 1.3: "What are your thoughts on the adoption of AI in your industry? Are you generally in favour, sceptical, or neutral?"

- **U 1.3.1:** "In favour."
  [Users who are enthusiastic about AI adoption are likely to be "Young Tech-Savvy Professionals" or "Technically Proficient Managers."]
- **U 1.3.2:** "Sceptical."
  [Indicates a cautious approach, likely clustering into "Experienced Sceptical Professionals" in healthcare or "Sceptical Journalists" in disinformation.]
- **U 1.3.3:** "Neutral."
  [May suggest a balanced view, potentially placing them in clusters like "Operational Staff with Basic Technical Skills" who are open to AI but not particularly enthusiastic.]

T5 1.4: "How important is transparency in AI decision-making processes to you?"

- **U 1.4.1:** "Very important."
  [Likely to fall into "Experienced Sceptical Professionals" or "Sceptical Journalists" who value transparency and ethical standards in AI.]

- **U 1.4.2:** "Somewhat important."
  [Could indicate users who value transparency but are not solely driven by it, such as "Young Tech-Savvy Professionals" or "Technically Proficient Managers."]

- **U 1.4.3:** "Not important."
  [Users who do not prioritize transparency might be more results-oriented, potentially fitting into "Young Tech-Savvy Professionals" or "Operational Staff with Basic Technical Skills."]

T5 1.5: "What motivates you most in your professional role? Is it innovation, efficiency, ethical considerations, or something else?"

- **U 1.5.1:** "Innovation."
  [Suggests a cluster like "Young Tech-Savvy Professionals" or "Technically Proficient Managers" who are eager to integrate new technologies.]

- **U 1.5.2:** "Efficiency."
  [Could fit into "Technically Proficient Managers" or "Operational Staff with Basic Technical Skills," who focus on improving processes.]

- **U 1.5.3:** "Ethical Considerations."
  [Likely to be "Experienced Sceptical Professionals" or "Sceptical Journalists" who prioritize ethical standards and transparency.]

- **U 1.5.4:** Other (e.g., "patient care", "safety")
  [Could fall into more specialized clusters based on sector-specific motivations, like "Fact-Checkers and Analysts" focusing on accuracy and integrity.]

T5 1.6: "How do you typically handle uncertainty in your work environment?"

- **U 1.6.1:** "Adapt well to uncertainty."
  [Likely to be in clusters that value flexibility and quick decision-making, such as "Young Tech-Savvy Professionals" or "Technically Proficient Managers."]
- **U 1.6.2:** "Prefer predictable environments."
  [Indicates a preference for stability and may suggest clusters like "Experienced Sceptical Professionals" or "Sceptical Journalists" who require clarity and reliability.]

- **U 1.6.3:** "Avoid uncertainty when possible."
  [Could align with clusters that need more structured support and predictable outcomes, such as "Operational Staff with Basic Technical Skills.]

T5 1.7: "How important are ethical standards and integrity in your work with AI and technology?"

- **U 1.3.1:** "Extremely important."
  [Suggests a high alignment with clusters like "Sceptical Journalists" and "Fact-Checkers and Analysts," who prioritize ethics and accuracy.]

- **U 1.3.2:** “Moderately important.”
  [Could indicate a balanced approach, potentially clustering into groups that value ethics but are also pragmatic, such as "Technically Proficient Managers.]

- **U 1.3.3:** “Not important.”
  [Unlikely to be a significant portion in sectors like disinformation; if present, might suggest a need for ethical training and awareness.]

Moreover, the example dialogues hereinafter demonstrate how the user persona is created in cases where human preferences deviate from the straightforward, linear approach that is based on simple trait identification. For instance, a participant may be both experienced and sceptical yet very comfortable with technology. To determine more accurately the persona of an experienced and sceptical user who is very comfortable with technology, further probing questions could be asked to determine which attribute (scepticism or comfort with technology) has a more significant influence on their preferences. (**T5**: THEMIS 5.0 Conversational Agent, **U**: User)

T5 2.1: “Even though you are comfortable with technology, how important is it for you to understand the inner workings and decision-making process of AI systems you use?”

- **U 2.1:** *[If the user emphasizes the importance of understanding and transparency, this indicates a preference for Fairness.]*

T5 2.2: “Given your comfort with technology, would you be willing to use an AI tool that is highly efficient and reliable but offers limited insight into how it reaches its conclusions?”

- **U 2.2:** *[If the user is willing to use such tools despite limited transparency, this suggests a preference for Robustness.]*

T5 2.3: “How do you weigh the need for ethical guidelines and transparency against the efficiency and reliability of AI tools?”

- **U 2.3:** *[A focus on ethical guidelines and transparency leans towards Fairness, while prioritizing efficiency and reliability points towards Robustness.]*

T5 2.4: “Imagine an AI system that improves your workflow but occasionally produces outcomes that are hard to explain. Would you still prefer to use it if it consistently delivers high performance?”

- **U 2.4:** *[A willingness to use the tool despite unexplained outcomes indicates a preference for Robustness. Conversely, reluctance or concern about the lack of explanation suggests Fairness.]*

T5 2.5: “If an AI system made an unexpected recommendation that you didn't fully understand, would you trust it based on its past reliability, or would you require a detailed explanation before using its recommendation?”

- **U 2.5:** *[Trusting the system based on reliability indicates a preference for Robustness. Requiring a detailed explanation indicates a preference for Fairness.]*

T5 2.6: “Are you more inclined to adopt new technologies quickly to stay ahead, even if they sometimes lack full transparency, or do you prefer to wait until there is complete clarity and understanding of the technology?”

- **U 2.6:** *[A tendency to adopt quickly aligns with Robustness, while a preference for waiting for full clarity aligns with Fairness.]*

3.7.4.2.2 Moral orientation

The example dialogues hereinafter demonstrate how the user's moral orientation is derived based on dialogue between the users and the THEMIS 5.0 conversational agent, differentiated according to the use case. (**T5**: THEMIS 5.0 Conversational Agent, **U**: User)

T5 3.1: "Hello! Let me briefly introduce the three ethical theories of TECC.

1. Virtue ethics prioritizes moral agency, focusing on the ethical virtues at a personal level of the moral agent, empowering and cultivating traits of character, ethical qualities, and values such as fairness, transparency, and honesty, typically regardless of moral agent's duties and obligations or the ethical consequences of the moral action.

2. Deontological Ethics focuses on moral actions, emphasizing the duties and obligations to comply with ethical rules, ensuring that moral actions align with established ethical guidelines and policies, typically regardless of the ethical consequences of the moral action or the ethical virtues of the moral agent.

3. Utilitarian Ethics guides moral actions based on maximizing overall outcomes and benefits, while minimizing risks and harms, aiming to achieve the greatest good for the greatest number of people, typically regardless of the moral agent's ethical duties and obligations or the ethical virtues of the moral agent."

<u>Use Case 1: Health Care</u>

T5 3.2: "Is your ethical motivation as a healthcare professional to provide rapid and accurate diagnoses driven by your personal values, your duty to comply with regulatory standards or the overall consequences for most people?"

- **U 3.2.1:** *Personal Values (=Moral Agent)*
- **U 3.2.2:** *Regulatory Standards (=Moral Action)*
- **U 3.2.3:** *Consequences (=Moral Action)*

T5 3.3 [for U 3.2.1]: "Do you believe that AI users in healthcare should be personally empowered with traits like integrity and fairness?"

- **U 3.3.1:** *Yes*
- **U 3.3.2:** *No*

T5 3.4 [for U 3.3.1]: "Do you agree that fostering ethical virtues, such as prudence and responsibility, is of primary importance for the use of AI software in healthcare?"

- **U 3.4.1:** *Yes*
- **U 3.4.2:** *No*

T5 3.4 [for U 3.2.2]: "Would you primarily advocate for specific ethical rules and procedures in the use of AI software in healthcare to protect human rights, such as privacy and informed consent?"

- **U 3.4.1:** *Yes*
- **U 3.4.2:** *No*

T5 3.5 [for U 3.4.1]: "In the healthcare sector, should the use of AI software adhere to ethical guidelines and standards to ensure that fair treatment of all stakeholders?"

- **U 3.5.1:** *Yes*
- **U 3.5.2:** *No*

T5 3.5 [for U 3.2.3]: "Would you primarily advocate for specific ethical rules and procedures in the use of AI software in healthcare to protect human rights, such as privacy and informed consent?"

- **U 3.5.1:** *Yes*
- **U 3.5.2:** *No*

T5 3.6 [for U 3.5.1]: "Should you minimize the risks of AI software by focusing on methods to mitigate problems and harms for as many healthcare professionals and stakeholders as possible?"

- **U 3.6.1:** *Yes*
- **U 3.6.2:** *No*

Use Case 2: Port Management

T5 3.7: "Is your ethical motivation as a port manager to optimize operations and improve efficiency driven by your personal values, your duty to comply with regulatory standards or the overall consequences for most people?"

- **U 3.7.1:** *Personal Values (=Moral Agent)*
- **U 3.7.2:** *Regulatory Standards (=Moral Action)*
- **U 3.7.3:** *Consequences (=Moral Action)*

T5 3.8 [for U 3.7.1]: "Do you believe that AI users in port management should be personally empowered with traits like fairness and prudence?"

- **U 3.8.1:** *Yes*
- **U 3.8.2:** *No*

T5 3.9 [for U 3.8.1]: "Do you agree that fostering ethical virtues, such as integrity and responsibility, is of primary importance for the use of AI software in port management?"

- **U 3.9.1:** *Yes*
- **U 3.9.2:** *No*

T5 3.10 [for U 3.7.2]: "Would you primarily advocate for specific ethical rules and procedures in the use of AI software in port management to protect human rights, such as transparency and equality?"

- **U 3.10.1:** *Yes*
- **U 3.10.2:** *No*

T5 3.11 [for U 3.10.1]: "In the port management sector, should the use of AI software adhere to ethical guidelines and standards to ensure that fair treatment of all stakeholders?"

- **U 3.11.1:** *Yes*
- **U 3.11.2:** *No*

T5 3.12 [for U 3.7.3]: "Would you primarily assess whether the benefits of AI software are maximized for as many people involved in the port as possible?"

- **U 3.12.1:** *Yes*
- **U 3.12.2:** *No*

T5 3.13 [for U 3.12.1]: "Should you minimize the risks of AI software by focusing on methods to mitigate problems and harms for as many port professionals and stakeholders as possible?"

- **U 3.13.1:** *Yes*
- **U 3.13.2:** *No*

Use Case 3: Media

T5 3.14: "Is your ethical motivation as a journalist to provide credible information to the public driven by your personal values, your duty to comply with regulatory standards, or the overall consequences for most people?"

- **U 3.14.1:** *Personal Values (=Moral Agent)*
- **U 3.14.2:** *Regulatory Standards (=Moral Action)*
- **U 3.14.3:** *Consequences (=Moral Action)*

T5 3.15 [for U 3.14.1]: "Do you believe that AI users and media experts should be personally empowered with traits like fairness and accuracy?"

- **U 3.15.1:** *Yes*
- **U 3.15.2:** *No*

T5 3.16 [for U 3.15.1]: "Do you agree that fostering ethical virtues, such as transparency and honesty, is of primary importance for the use of AI software in media and journalism?"

- **U 3.16.1:** *Yes*
- **U 3.16.2:** *No*

T5 3.17 [for U 3.14.2]: "Would you primarily advocate for specific ethical rules and procedures in the use of AI software in journalism and media to protect human rights, such as privacy and safety?"

- **U 3.17.1:** *Yes*
- **U 3.17.2:** *No*

T5 3.18 [for U 3.17.1]: "In the media sector, should the use of AI adhere to ethical guidelines and standards to ensure that published information is fair and accurate?"

- **U 3.18.1:** *Yes*
- **U 3.18.2:** *No*

T5 3.19 [for U 3.14.3]: "Would you primarily calculate whether the benefits of AI software are maximized for as many people in journalism and media as possible?"

- **U 3.19.1:** *Yes*
- **U 3.19.2:** *No*

T5 3.20 [for U 3.19.1]: "Should you minimize the risks of AI software by focusing on methods to mitigate problems and harms for as many media professionals and stakeholders as possible?"

- **U 3.20.1:** *Yes*
- **U 3.20.2:** *No*

3.7.4.2.3 Trustworthiness optimisation solution exploration

The example dialogue snippets hereinafter demonstrate how THEMIS 5.0 supports users to understand the impact of the alternative trustworthiness optimisation measures and to decide on which measure is the most suitable to be implemented, differentiated according to each Use Case. (**T5**: THEMIS 5.0 Conversational Agent, **U**: User)

Use Case 1: Health Care

U 4.1: I want to optimize the fairness of the diagnosis-support AI, by balancing both medical and ethical considerations. What factors does the AI consider when assisting in diagnosis?

T5 4.1: "The AI takes into account a range of medical factors, including the patient's symptoms, medical history, radiology results, lab data, and current clinical guidelines."

U 4.2: "Help me identify potential ethical risks in the AI's diagnosis recommendations, especially in terms of fairness."

T5 4.2: "One ethical risk is the possibility of bias in how the AI processes patient data. For instance, if the AI places a higher weight on certain medical conditions based on typical demographic patterns, it might overlook critical factors specific to the individual patient. This could lead to unequal consideration of patients from diverse backgrounds or lifestyle."

[….]

U 4.3: "I want to optimize the Diagnosis Engine so that it ensures that lifestyle aspect is not overlooked or neglected during the route-cause analysis."

T5 4.3: "Lifestyle evaluation is marked in the "DD" tool as a dimension to evaluate in context by a Human being therefore medical doctors should be aware of that."

[….]

Use Case 2: Port Management

U 4.4: "I want to optimize the fairness of the ETA calculation, based on a better balance between business and ethical aspects. What are the factors of ETA calculation?"

T5 4.4: "Major factors that weigh in the ETA are terrestrial congestion, shipping agency, the nature of the cargo, service level agreements and the weather. Scheduling choices are made through a calculation based on those variables, and many others."

U 4.5: "Help me identify the ethical risks in terms of fairness."

T5 4.5: "Due to the business logic in place, discrimination based on nationality could appear if some companies are systematically privileged due to the nature of their cargo."

U 4.6: "Tell me more on this"

[….]

U 4.7: "I want to optimize the ethical aspect of decision-making prior to the ETA calculation, to ensure that no discrimination based on nationality."

T5 4.7: "The best option should be to take nationality as an explicit KPI in reports, for further configuration of the ETA calculation."

U 4.8: "What would be the business impacts of this?"

T5 4.8: "That could violate some service level agreements and therefore lead to financial penalties."

[….]

U 4.9: "I want to increase the contract value with transportation companies by ___%. How can an improvement of the ETA calculation AI-system help me to achieve this?"

T5 4.9: "Well, based on the trustworthiness assessment of the AI-tool, the fairness of the system is __%. This is a vulnerability that could reduce the system's trustworthiness. If transportation companies don't trust the fairness of the ETA result against all the involved stakeholders, they will be reluctant to rely on it for the scheduling of their routes. This could increase traffic and congestion in the port area, which would result in longer vessel servicing times while docked. Therefore, fewer vessels will be serviced and reduced cargo volume will be transferred by the transportation companies.

U 4.10: "Give me some options that would improve the trustworthiness of the system."

T5 4.10: "Ok, based on the risks and the target that you provided, the AI tool's trustworthiness could be improved significantly either by increasing its fairness or by making the decision-making process more transparent to the users. For the former, the AI model needs to be improved to account for a number of parameters that affect the fairness of the results, while for the second, a sufficient explanation of the results should be offered. Note, however, that the major shortcoming of the AI system in terms of trustworthiness is its fairness. Making the tool more transparent could improve the overall trust of the users, but in the long term, by increasing its fairness the users' trust would improve more and would help in the creation of a more stable image for the port ETA prediction tool. Based on your profile, the second option might be more suitable."

U 4.11: "Indeed, it is more suitable. Can both options be implemented? What would be the impact of such an approach?"

T5 4.11: "Yes, definitely. This would improve the users' trust both in the short and in the long term. It will be more expensive to implement and more complex to manage instead of gradual improvements."

U 4.12: "Ok. Let's stick with improving the system's fairness and once this is done, we can reassess the situation and consider further improvements."

Use Case 3: Media

U 4.13: "I want to optimize the fairness of the AI's detection of hate speech and fake news, balancing technical accuracy with ethical considerations, such as journalistic integrity. What factors does the AI take into account when flagging content?"

T5 4.13: "The AI analyses linguistic patterns, the context of speech, user behaviour, and cross-references with verified sources. It also considers the historical accuracy of the source, engagement metrics, and the presence of inflammatory or offensive language."

U 4.14: "Help me identify the potential ethical risks in terms of fairness when it comes to journalistic work."

T5 4.14: “A major ethical risk lies in the potential for over-censorship or bias. For example, the AI could flag legitimate journalistic content if it contains sensitive topics, controversial opinions, or nuanced discussions that involve strong language but do not qualify as hate speech or fake news.”

[….]

U 4.15: “Help me refine the input datasets and fine-tune the model so that the hate speech detection rate increase does not weaken the journalist's ability to inform the public of legitimate and nuanced issues.”

T5 4.15: “Based on the contextual knowledge brought within the DD framework, the AI detection tool necessarily fosters an automation bias which contradicts some tenets of journalist's deontology.”

[….]

4. IMPLICATIONS AND RECOMMENDATIONS

4.1. Recommendations based on groundwork and user requirements

The groundwork presented in Section 2 and the user requirements work presented in Section 3 have provided the basis for a number of recommendations important for the further development of THEMIS 5.0.

In the following, we summarize key recommendations from the two strains of work. First, in Table 17 below, we summarize recommendations from the groundwork in Section 2 and then, in Table 18 below, we summarize recommendations from the user requirements work in Section 3.

Table 17: Recommendations for THEMIS 5.0 based on groundwork presented in Section 2

#	Recommendation	Description	Source section
1.01	THEMIS 5.0 should support optimization for **a broad range of systems under test**	The THEMIS 5.0 approach should be applicable for trustworthiness optimization for a broad range of AI systems, referred to as systems under test. In line with the definitions of OECD and the AI Act, AI systems are defined through a functional approach rather than as a list of technologies. Hence, the system under test in THEMIS 5.0 should not be limited to specific AI technologies but should cover a range of current and future technologies. The breadth of relevant AI technologies is reflected in the three THEMIS 5.0 use cases	2.3.1
1.02	Optimization through THEMIS 5.0 should take into account the **socio-technical environment**	In line with existing guidelines and frameworks (AI HLEG, 2019; ENISA, 2023; NIST, 2023), THEMIS 5.0 trustworthiness optimization will be conducted with regard to the socio-technical environment of the system under test. Specifically, THEMIS 5.0 will address the AI system in the socio-technical environment from a business perspective with concern for *decision impact*, in addition to addressing the technical and ethical / legal perspectives on Trustworthy AI.	2.3.2
1.03	Optimization in THEMIS 5.0 depends on **detectable trustworthiness characteristics**	In order to assess and optimize the trustworthiness of an AI system under test, THEMIS 5.0 identifies and analyses trustworthiness characteristics corresponding to key ethical requirements in AI systems. To achieve this, trustworthiness characteristics need to be detectable, automatically or semi-automatically.	2.3.3
1.04	The THEMIS 5.0 approach should consider **trustworthiness characteristics from three perspectives**	Trustworthiness characteristics concern a range of aspects of an AI system. In line with ENISA (2023), THEMIS 5.0 will address trustworthiness characteristics from (a) a technical perspective concerning the technical performance of the system under test, (b) a business perspective on decision impact, and (c) an ethical/legal perspective concerning ethical/ legal requirements. These perspectives are not mutually exclusive. Rather trustworthiness characteristics may be addressed from some or all of these perspectives.	2.3.4-2.3.6
1.05	The THEMIS 5.0 approach should provide **flexible support of selected trustworthiness characteristics**	A wide range of relevant trustworthiness characteristics exist (e.g. ENISA, 2023; NIST, 2023). However, to allow for in-depth research, the THEMIS 5.0 project will address a limited number of characteristics. These are chosen so as to reflect relevant perspectives: accuracy and robustness reflecting a technical perspective, decision impact reflecting a socio-technical perspective, and fairness reflecting an ethical/legal perspective. However, the THEMIS 5.0 approach should be developed so as to flexibly enable coverage also of other trustworthiness characteristics.	2.3.4-2.3.6
1.05	THEMIS 5.0 should support **human-centred explanations**	THEMIS 5.0 should be developed in line with principles for human-centred AI. This allows for the involvement of domain-experts and users in trustworthiness assessment and optimization. Following from this, THEMIS 5.0 should facilitate human-centred explanations of key aspects of the system under test. That is, trustworthiness characteristics should be explainable to domain-experts and users of the domains for which THEMIS 5.0 is applied.	2.4.1
1.06	THEMIS 5.0 should **adapt to fit user personas**	Trustworthiness optimization in THEMIS 5.0 should take into account the specific personas of the domain-experts and users of the domains for which THEMIS 5.0 is applied. Such adaptation should be semi-automatic, following the capture of domain-expert and user characteristics and preferences in the form of personas. Persona development should draw on data captured from relevant user and domain-expert groups.	2.4.2
1.07	THEMIS 5.0 should encompass a **risk management approach to trustworthiness optimization**	Risk management is a key approach to assessing and optimizing the trustworthiness of AI systems (AI HLEG, 2020; ENISA, 2023; NIST, 2023). An innovative approach to risk management, based on the Spyderisk platform, has been shown relevant to support risk management for AI trustworthiness in line with the guidelines proposed by ENISA (2023) and NIST (2023), and an approach towards such tool-based risk management support has been proposed. THEMIS 5.0 should further develop a risk management approach to AI trustworthiness optimization based on the Spyderisk platform.	2.5
1.08	THEMIS 5.0 should support **assessment of accuracy through machine learning**	A broad range of assessment approaches for the accuracy of a system under test exists. However, much-used approaches may have limitations for complex, real-world contexts and, as part of the groundwork, a machine learning approach to the analysis of accuracy has been proposed. THEMIS 5.0 should aim to include assessment of accuracy through a machine learning approach, as a supplement to established approaches for assessing accuracy.	2.6.1

1.09	THEMIS 5.0 should support **assessment of robustness through out of distribution detection and by use of counterfactuals**	Robustness is a relatively complex concept, concerning the ability of an AI system's ability to perform as expected under varying conditions. Promising approaches for assessing robustness in the context of the THEMIS 5.0 use cases concern out-of-distribution detection and stability under perturbation of input by the use of counterfactuals. THEMIS 5.0 should aim to further refine and include these approaches for assessing the robustness of a system under test.	2.6.2
1.10	THEMIS 5.0 should enable **tailoring of fairness assessments**	Fairness is a highly complex concept concerning multiple aspects of equal treatment and avoidance of undesirable bias and discrimination. As part of the groundwork, a range of approaches to the assessment of fairness has been identified, several of which may have relevance to the THEMIS 5.0 use cases as well as future users of the THEMIS 5.0 approach. At the same time, as noted also in the groundwork on ethical and legal requirements, different contexts may pose highly different requirements for fairness assessment. THEMIS 5.0 should aim to provide an approach to tailor fairness assessments to a given system under test and socio-technical environment.	2.6.3
1.11	THEMIS 5.0 should provide a **lifecycle-approach to trustworthiness optimization**	The trustworthiness of an AI system concerns the combined outcome of relevant trustworthiness characteristics for the system within its socio-technical environment. Trustworthiness optimization is a continuous process throughout the AI lifecycle. Hence, trustworthiness optimization concerns the continuous balancing of relevant trustworthiness characteristics. As noted in the groundwork, trustworthiness optimization is a concept only treated to a limited degree in the literature. THEMIS 5.0 should provide a conceptual framework and technology support for human-centred trustworthiness optimization of AI systems.	2.8
1.12	THEMIS 5.0 should make a **unique contribution of human-centred trustworthiness optimization**	While a number of technologies and tools exist to support ethical and trustworthy AI development and deployment, as detailed in Section 2.6.2, there is a lack of approaches supporting human-centred involvement of domain-experts and users as part of a comprehensive assessment and optimization approach. THEMIS 5.0 should address this gap in the current state of the art to make a unique contribution to the field.	2.6.4
1.13	THEMIS 5.0 should **comply with ethical and legal requirements**	Ethical and legal requirements are relevant for THEMIS 5.0 at two levels: (a) As trustworthiness characteristics for the system under test and (b) as requirements for the THEMIS 5.0 approach itself. This is because both the system under test and the THEMIS 5.0 approach are AI-based systems. Ethical and legal requirements are provided through guidelines, soft law, legislation, and case law, based on the EU AI Act, GDPR and others. THEMIS 5.0 should comply with all relevant ethical and legal requirements. An overview is provided in Section 2.7.	2.7

Table 18: Recommendations for THEMIS 5.0 based on the human-centric requirements analysis presented in Section 3

#	Recommendation	Description	Source section
2.01	THEMIS 5.0 should **comply with the system requirements stemming from the DoA.**	The design and development of the THEMIS 5.0 framework and ecosystem should address the system requirements stemming directly from the Description of Action. The complete set of 80 system requirements can be found in Appendix A.2, alongside the relevant passage from the DoA, and have been classified in 5 categories according to their scope: (i) Overall, (ii) Identify, (iii) Assess, (iv) Explore, and (v) Enhance. Using this as a starting point and as there is much overlap between them, a consolidated list of 27 System Requirements is included in Section 3.4 which also references the initial set of system requirements for a more detailed view of the system requirements translated from THEMIS 5.0 DoA.	3.4
2.02	THEMIS 5.0 should **address the user requirements.**	The design and development of the THEMIS 5.0 framework and ecosystem should address the user requirements as they resulted from the analysis of the end users' input collected from (i) the co-creation workshops and (ii) the use case partners. The complete set of user requirements can be found in Appendix 0, in Table 24, Table 25, Table 26, Table 27, classified in 5 categories according to their scope: (i) Overall, (ii) Identify, (iii) Assess, (iv) Explore, and (v) Enhance. Using this as a starting point and as there is much overlap between them, a consolidated list of 81 unique user requirements is included in Section 3.5 which also references the complete set of user requirements for a more detailed view if needed.	3.5
2.03	Trustworthiness evaluation and optimisation of the SUTs should **address the end-users' expectations from AI Tools for each sector.**	Stemming from the 193 participants of the 1st co-creation workshop of Phase A, the co-created attention points that formulate the AI users' anticipation for the implementation of AI tools in their respective domains and determine the users' preferences affecting the perceived trustworthiness of the AI Systems Under Test, are included in Section 3.3.13.3.1 per sector, alongside the related users' expectations that have been extracted. THEMIS 5.0 should consider these expectations while designing and implementing the THEMIS 5.0 framework, whose ultimate goal is to help AI-users enhance an AI-system's trustworthiness.	3.3.1
2.04	Trustworthiness evaluation and optimisation of the SUTs should **be based on the users' expectations regarding fairness, accuracy, and robustness**	The end-users' expectations and concerns regarding fairness, accuracy, and robustness in AI systems used in their professional context are presented in Section 3.3.2.3.3.2 For each entry, the respective source, the categories that they fall within and their relevance to THEMIS 5.0 is listed. The assigned categories can be used as key words to swiftly sort and filter the expectations for a specific category (e.g. Training datasets, Adaptation to the needs of end-users, Explainability, End-users' engagement, Bias, etc.). THEMIS 5.0 should consider these expectations for each use case separately for the design and implementation of the THEMIS framework whose ultimate goal is to	3.3.2

		help AI-users enhance an AI-system's trustworthiness in terms of their actual as well as perceived fairness, accuracy, and robustness.	
2.05	Trustworthiness evaluation and optimisation of the SUTs should be **personalised according to the users' trustworthiness preferences.**	In Section 3.3.3.1, an initial analysis is presented of the end-users' preferences against the three major AI trustworthiness categories that THEMIS 5.0 addresses, namely Fairness, Accuracy and Robustness, collected from a small sample of domain experts/end-users (four from each use case). Each participant (i) ranked the trustworthiness parameters according to the needs of their sectors, (ii) assigned a number from 1-10 indicating their importance, and (iii) provided the minimum level of each characteristic that they could tolerate using. These preferences against trustworthiness aspects should be used as pointers for the initial steps of the development of the relevant personalization components and should also be used as a basis for the collection of preferences from a substantial sample of end-users/experts during the co-creation phases B (Living Labs) and C (Pilots).	3.3.3.1
2.06	Trustworthiness evaluation of the SUTs should also be based on the **users' considerations for AI systems' Trustworthiness Assessment.**	For the collection of informed end users' input, they were provided with working definitions for fairness, accuracy and robustness as well as examples for what could be the meaning of these characteristics in AI tools used in respective sectors. Participants and interviewees provided additional criteria for assessing an AI system's trustworthiness in terms of fairness, accuracy and robustness that are also presented in Section 3.3.3.2.These criteria should also be considered for the assessment of the trustworthiness of an AI system in the respective sectors.	3.3.3.2
2.07	THEMIS 5.0 Trustworthiness optimisation decision support should **calculate the trustworthiness related Business, Ethical and Legal Risks**.	In addition to 2.06 above, use case partners and co-creation participants were asked to provide trustworthiness-related business, ethical and legal risks. In Section 3.3.4, the main business, ethical and legal risk that, according to the end-users, are related to the AI system's fairness, accuracy and robustness are presented. THEMIS should consider the domain-specific risks that are associated with each trustworthiness characteristic as a pointer for the design and development of the solutions that (i) will help users to build an AI system's socio-technical model and (ii) will support users to take an informed decision also considering the impact of each potential trustworthiness optimization measure on the business targets and objectives.	3.3.4
2.08	THEMIS 5.0 ecosystem should **allow users to adapt their experience according to their needs.**	As THEMIS 5.0 addresses different use cases-sectors, as well as various types of users with different capabilities and authorization levels, the THEMIS 5.0 ecosystem should be able to accommodate an adaptive user experience and flow, allowing users to achieve their objectives in a flexible and user-friendly manner. Based on the analysis of the system and the human-centric requirements, user journeys were compiled to document and visualize the various step-by-step user experiences from beginning to end, listing the different actions users take to accomplish a particular goal. In Section 3.6, these user journeys that the design of the THEMIS 5.0 ecosystem should be able to accommodate are presented.	3.6
2.09	THEMIS 5.0 should **follow the UI guidelines and the legal and technical standards.**	Having as a starting point the mock-ups presented in this deliverable, THEMIS 5.0 user interfaces should also follow the UI guidelines presented in Section 3.7.1 that have been established stemming from the partners' long-standing experience and strong expertise in the development of user-oriented software applications alongside the requirements from certain EU legal and technical standards.	3.7.1
2.10	THEMIS 5.0 Conversational Agent should **guide the user to navigate the interconnected applications and to achieve their objectives.**	THEMIS 5.0 conversational agent will be the interface for (i) the identification of the user's persona and moral orientation to derive the users' trustworthiness preferences, (ii) the creation of the socio-technical model, and (iii) users' support to understand the impact of the alternative trustworthiness optimisation measures and to decide on which measure is the most suitable to be implemented. Moreover, the conversational interface should guide users while they are navigating through the THEMIS 5.0 framework, also providing personalised user guidance and explanations as needed to ensure that all THEMIS 5.0 users will be able to take advantage of the personalised trustworthiness assessment and optimisation potential of the project.	3.7.2

4.2. Implications for further work

This deliverable plays a key role in the definition of the conceptual basis for the innovative approach of THEMIS 5.0 as it documents the theoretical groundwork (Section 2 above) and the human-centric requirements (Section 3 above) for a human-centred AI trustworthiness optimisation environment. The groundwork details the concept of trustworthiness for AI systems, considers specific trustworthiness characteristics, expands on the involvement of human users and AI system adaptation to human values and business targets and discusses the use of risk management approaches for assessing AI trustworthiness. The user requirements will be identified and summarized on the basis of extensive user involvement. The requirements will further be instantiated in mock-ups illustrating user journeys.

The work documented in this deliverable is already and may be used as input in the following tasks and work packages:

- **WP1:** The theoretical groundwork performed in Task 2.1, as well as the collection of user and system requirements and the compilation of the user journeys performed in Task 2.2, run in parallel with Task 1.4 that used these outputs as they were being produced as a starting point for the work on the Legal and Ethical Compliance to assess the impact of THEMIS 5.0 activities on the legal and ethical requirements and especially for *D1.3 Template and guidance for legal and ethical impact assessment* being a collaborative questionnaire and a guidance manual for assessing the legal and ethical impacts of THEMIS 5.0 activities.

- **WP2 (T2.3):** This deliverable is the first step towards the definition of the conceptual basis for the innovative approach of THEMIS 5.0 as it documents the theoretical groundwork (Task 2.1) and the human-centric requirements (Task 2.2) for a human-centred AI trustworthiness optimisation environment. Based on this work, Task 2.3 will seek to establish a comprehensive conceptual model. Having as input the user journeys and the mock-ups, Task 2.3 will prepare the material for the execution of Co-creation Phase B to define the architectural considerations and decision-making processes needed to support the foreseen functionalities and user experience. Task 2.3 has, for a big part, run in parallel with Tasks 2.1 and 2.2 and has been using their outputs as they were being produced, while input from T2.3 has been fed back to Tasks 2.1 and 2.2 in the form of a proposed Trustworthiness Optimization Process that will be the subject of D2.2.

- **WP3:** In co-creation Phase B, citizens (AI users) related to the key domains selected to be piloted are invited into targeted co-creation living labs in four countries across Europe, (Greece, Bulgaria, Spain, Denmark) and are presented with the mock-ups of the THEMIS 5.0 'user journeys', produced in Phase A. The aim of this co-creation Phase B is to transform the user journeys captured in the mock-ups into decision-making specifications for THEMIS 5.0 AI ecosystem services, also using material from Task 2.3 as described above.

- **WP4:** The work documented in this deliverable impacts directly and indirectly the work performed in the framework of WP4. T4.1, in particular, runs in parallel with Tasks 2.1 and 2.2 and has been using their outputs as they were being produced to gauge early on the users' needs and the functionalities that the technical solutions developed in WP4 will aim to address. This deliverable, as well as the upcoming D2.2 that will document the THEMIS 5.0 conceptual modelling, will be the solid foundation for the update and the finalisation, respectively, of the preliminary technical architecture and system software design specifications that are being under definition since the beginning of Task 4.1.

- **WP5:** Although WP2 and WP5 are not directly linked, the work documented in this deliverable can be useful for WP5 particularly for the definition of the pilot scenarios (Task 5.1) and the definition of tasks per use case that will be used to observe how the users interact with the AI system (Task 5.2) that can use as a starting point the user journeys documented here.

5. CONCLUSION

5.1. Key insights

The presented work in this deliverable constitutes the outcomes of Tasks 2.1 and 2.2, concerning the groundwork and user requirements for THEMIS 5.0. The insights gained in this work, in part, serve to ground the coming research in THEMIS 5.0 in the state of the art, as well as in the needs and requirements of the user groups which are intended to benefit from the THEMIS 5.0 approach.

From the groundwork in Task 2.1, presented in Section 2, key insights are made in the following areas:

- **AI system and socio-technical environment:** THEMIS 5.0 concerns the assessment and optimization of trustworthiness for an AI system under test. In the groundwork, we have aligned our understanding of an AI system with the current EU and OECD definitions and detailed the implications of considering the AI system as embedded in a specific socio-technical environment.
- **Detectable trustworthiness characteristics:** In the groundwork we addressed AI trustworthiness as possible to assess and optimize by way of detectable trustworthiness characteristics. With basis in the state of the art, we have detailed three perspectives on such characteristics – technical, socio-technical, and ethical/legal – and identified example characteristics to address specifically in the THEMIS 5.0 project.
- **Human-centred AI and user preferences:** The groundwork has also summarized the state of the art on human-centred AI and how to align trustworthiness assessments and optimization with the preferences of specific groups. The human-centred approach and the consideration of user preferences are key to enabling sufficient consideration of the socio-technical environment of the AI system under test.
- **Risk assessment for Trustworthy AI:** In line with the state of the art, risk assessment is considered key to assess and optimize AI trustworthiness. The groundwork has further shown the relevance and feasibility of Spyderisk as a tool for conducting such risk assessment as it aligns with major frameworks of relevance to trustworthy AI, such as those by NIST (2023) and ENISA (2023).
- **Technologies for Trustworthy AI:** A number of relevant approaches and technologies to leverage the assessment and optimization of Trustworthy AS has been identified through the groundwork, specifically for accuracy, robustness, and fairness.
- **Ethical and legal considerations:** Ethical and legal aspects are key to trustworthy AI, particularly in the context of the AI Act (EU, 2024). The groundwork allows for thorough consideration of ethical and legal aspects of the THEMIS 5.0 approach, with key ethical and legal requirements explicated.

From the human-centric requirements analysis in Task 2.2, presented in Section 3, key insights are made in the following areas:

- **THEMIS 5.0 use cases' definition:** In the framework of Task 2.2, the THEMIS 5.0 use cases have been further defined to include the user types that are involved as AI users in each use case, the scenarios of use, the technical implementation of the AI systems under test, the AI trustworthiness considerations and risks. This work was used throughout this Task and in this deliverable, an updated description of the user types, the scenarios of use and the KPIs for the validation of THEMIS 5.0 in WP5 are included in Section 3.2.
- **THEMIS 5.0 system requirements:** The complete set of 80 system requirements translated from THEMIS 5.0 DoA can be found in Appendix A.2. A consolidated list of 27 System Requirements is included in Section 3.4.
- **THEMIS 5.0 user requirements:** The complete set of 178 user requirements can be found in Appendix 0 in Table 24, Table 25, Table 26, Table 27. A consolidated list of 81 unique user requirements is included in Section 3.5.
- **End-users' expectations from AI Tools for each sector:** The users' expectations for the implementation of AI tools in their respective domains that determine the users' preferences affecting the perceived trustworthiness of the AI Systems Under Test are presented per sector in Section 3.3.1.
- **End users' expectations and concerns regarding trustworthiness characteristics in AI systems used in their professional context:** The end-users' expectations and concerns regarding fairness, accuracy, and robustness in AI systems used in their professional context are presented in Section 3.3.2.

- **End users' trustworthiness preferences:** An initial analysis is presented of the end-users' preferences (rank, importance, minimum acceptable level) against the three major AI trustworthiness categories that THEMIS 5.0 addresses, namely Fairness, Accuracy and Robustness, collected from a small sample of domain experts/end-users, is presented in Section 3.3.3.1.
- **Users' considerations for AI systems' Trustworthiness Assessment:** End-users' criteria for assessing an AI system's trustworthiness in terms of fairness, accuracy and robustness are presented in Section 3.3.3.2.
- **Trustworthiness related Business, Ethical and Legal Risks:** The main business, ethical and legal risk that according to the end-users are related to the AI system's fairness, accuracy and robustness are presented Section 3.3.4.
- **THEMIS 5.0 user journeys:** Based on the analysis of the system and the human-centric requirements, user journeys were compiled to document and visualize the various step-by-step user experiences from beginning to end, listing the different actions users take to accomplish a particular goal. The user journeys are presented in Section 3.6.
- **THEMIS 5.0 UI guidelines and mock-ups:** The THEMIS 5.0 user interfaces and requirements are presented in Section 3.7.1 followed by mock-ups and mock-up dialogues that aim to visualise the user journeys and showcase the user experience while interacting with the THEMIS 5.0 ecosystem.

5.2. Significance of outcomes

THEMIS 5.0 is guided by a number of key challenges. Among these, the THEMIS 5.0 Description of Action, outlined some key challenges for the work presented in this deliverable – as well as for the overall project. When completing this deliverable, we summarize our perspectives on each of these.

a) **What constitutes a trustworthy AI decision support system within the dynamic reality of the socio-technical system that is operated in?** For THEMIS 5.0, a trustworthy AI decision support system is one where detectable trustworthiness characteristics have been assessed and optimized in line with the preferences of users within the systems' intended socio-technical environment. As detailed in this deliverable, trustworthiness characteristics may be considered through different methods and technologies, and parts of the challenge will be to identify the set of methods and technologies that provide the needed input to the assessment and optimization process. This assessment and optimization itself will be conducted through a human-centred risk assessment process involving relevant users, in line with ethical and legal requirements. The work on this challenge will be continued in Task 2.3 on conceptual modelling of THEMIS 5.0.

b) **What are the human characteristics that contribute to the evaluation of the trustworthiness of an AI decision support system?** The human characteristics concern a broad spectrum of information, including user needs, preferences, behaviours, and demographic detail. In THEMIS 5.0, we will establish insight into user preferences for the specific socio-technical environments in which the AI systems under test are deployed. Specifically, we will address factors associated with technology acceptance and behavioural intentions to use technology, such as perceived usefulness and ease of use. Furthermore, measures of trust will be key to understanding user preferences for this purpose. The work on this challenge will be continued in the WP3 work on co-creation.

c) **How can we evaluate levels of trustworthiness and what are the effective approaches to trustworthiness optimisation in view of human preferences and values?** In the groundwork, it is clarified that estimates of trustworthiness are conducted through balancing of detectable trustworthiness characteristics. Also, based on input from users and domain experts, the human-centric requirements provide insights for trustworthiness assessment and optimisation with consideration for user preferences and values. This will further require estimates of key trustworthiness characteristics as input to risk assessment, and a method for matching trustworthiness characteristics with established user preferences.

d) **How to understand and incorporate the socio-technical system's decision-making risks in the evaluation of the trustworthiness of a hybrid decision-support AI system?** The socio-technical risks in the evaluation of AI trustworthiness will be understood and incorporated in part through a human-centred risk assessment approach where users and domain experts are involved in the assessment and optimization of AI trustworthiness.

e) **How to intelligently engage and converse with humans in the optimisation of the trustworthiness of decision-supporting AI systems.** Engagement and conversation with users and domain experts will be conducted by making available the THEMIS 5.0 approach of human-centred risk assessment through a conversational user interface.

The work on challenges c-e will be continued in Task 2.3 on the conceptual modelling of THEMIS 5.0, Task 4.1 on the specification of THEMIS 5.0 services, platform and architecture and Task 4.4 on the development of the THEMIS 5.0 platform.

6. REFERENCES

[1]	AI HLEG. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[2]	AI HLEG. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment.* European Commission. Directorate General for Communications Networks, Content and Technology. https://data.europa.eu/doi/10.2759/002360
[3]	AI Seoul Summit. (2024, May 20). [Press Release] Only global AI standards can stop a race to the bottom. *AI Seoul Summit*. https://aiseoulsummit.kr/press/?mod=document&uid=39
[4]	Akbar, M. A., Khan, A. A., Mahmood, S., Rafi, S., & Demi, S. (2024). Trustworthy artificial intelligence: A decision-making taxonomy of potential challenges. *Software: Practice and Experience*, *54*(9), 1621–1650. https://doi.org/10.1002/spe.3216
[5]	Allen, G., C., & Adamson, G. (2024, May 23). *The AI Seoul Summit*. CSIS | Center for Strategic and International Studies. https://www.csis.org/analysis/ai-seoul-summit
[6]	Artificial Intelligence Act, No. 2024/1689, European Union (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[7]	Achimugu P., Selamat A., Ibrahim R., and Mahrin M. N., (2014) A Systematic Literature Review of Software, Information and Software Technology, 584, 568-585.
[8]	Axel Springer v. Germany, No. Application no. 39954/08 (European Court of Human Rights 7 February 2012). https://hudoc.echr.coe.int/fre#{%22itemid%22:[%22001-109034%22]}
[9]	Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. https://fairmlbook.org/
[10]	Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*, 671–732. https://doi.org/10.15779/Z38BG31
[11]	Baudel, T., Colombet, G., & Hartmann, R. (2023). *AI Decision Coordination: Faciliter l'appropriation de la décision automatisée par les utilisateurs métier*. IHM'23 - 34e Conférence Internationale Francophone sur l'Interaction Humain-Machine, AFIHM. https://hal.science/hal-04046408/document
[12]	Baudel, T., Verbockhaven, M., Cousergue, V., Roy, G., & Laarach, R. (2021). ObjectivAIze: Measuring Performance and Biases in Augmented Business Decision Systems. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen (Eds.), *Human-Computer Interaction – INTERACT 2021* (Vol. 12934, pp. 300–320). Springer International Publishing. https://doi.org/10.1007/978-3-030-85613-7_22
[13]	Bertuzzi, L. (2023, November 9). *OECD updates definition of Artificial Intelligence 'to inform EU's AI Act'*. Www.Euractiv.Com. https://www.euractiv.com/Section/artificial-intelligence/news/oecd-updates-definition-of-artificial-intelligence-to-inform-eus-ai-act/
[14]	Bevilacqua, M., Berente, N., Domin, H., Goehring, B., & Rossi, F. (2023). *The Return on Investment in AI Ethics: A Holistic Framework* (No. arXiv:2309.13057). arXiv. http://arxiv.org/abs/2309.13057
[15]	Biasin, E. (2023, June 23). About Accuracy (and its Meaning in Healthcare). *KU Leuven CiTiP*. https://www.law.kuleuven.be/citip/blog/about-accuracy-and-its-meaning-in-healthcare/
[16]	Bietti, E. (2020). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency—FAT* '20* (pp. 210–219). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3351095.3372860
[17]	Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, article no. 377. https://doi.org/10.1145/3173574.3173951
[18]	Braiek, H. B., & Khomh, F. (2024). *Machine Learning Robustness: A Primer* (No. arXiv:2404.00897). arXiv. http://arxiv.org/abs/2404.00897

[19]	Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law*, *25*(3), 273–291. https://doi.org/10.1007/s10506-017-9214-9
[20]	Buruk, B., Ekmekci, P. E., & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, *23*(3), 387–399. https://doi.org/10.1007/s11019-020-09948-1
[21]	Calegari, R., Castañé, G. G., Milano, M., & O'Sullivan, B. (2023). Assessing and Enforcing Fairness in the AI Lifecycle. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence Survey Track—IJCAI 2023* (Vol. 6, pp. 6554–6562). https://www.ijcai.org/proceedings/2023/735
[22]	California Consumer Privacy Act (CCPA), State of California Department of Justice (2018). https://oag.ca.gov/privacy/ccpa
[23]	Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-07939-1
[24]	Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., ... & Lagendijk, R. L. (2023). Meaningful human control: actionable properties for AI system development. AI and Ethics, 3(1), 241-255.
[25]	Charter of Fundamental Rights of the European Union, No. C 326/391, European Union (2012). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT
[26]	Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., & Yeung, K. (2021). Trustworthy AI. In B. Braunschweig & M. Ghallab (Eds.), *Reflections on Artificial Intelligence for Humanity* (pp. 13–39). Springer International Publishing. https://doi.org/10.1007/978-3-030-69128-8_2
[27]	Chen, P., Wu, L., & Wang, L. (2023). AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications. *Applied Sciences*, *13*(18), Article 18. https://doi.org/10.3390/app131810258
[28]	Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), article no. 6. https://doi.org/10.1186/s12864-019-6413-7
[29]	Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047
[30]	Coeckelbergh, M. (2020). *AI Ethics*. MIT Press. https://mitpress.mit.edu/9780262538190/ai-ethics/
[31]	Consolidated Version of the Treaty of the European Union, No. C 326/13, European Union. https://eur-lex.europa.eu/resource.html?uri=cellar:2bf140bf-a3f8-4ab2-b506-fd71826e6da6.0023.02/DOC_1&format=PDF
[32]	Consolidated Version of the Treaty on the Functioning of European Union, No. C 326/47, European Union (2012). eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012E/TXT
[33]	Council Of Europe. (2023). *Guidelines on the responsible implementation of artificial intelligence systems in journalism*. Council of Europe. https://rm.coe.int/cdmsi-2023-014-guidelines-on-the-responsible-implementation-of-artific/1680adb4c6
[34]	Cybersecurity Act, No. 2019/881, European Union (2019). https://eur-lex.europa.eu/eli/reg/2019/881/oj
[35]	Data Protection Commissioner v Facebook Ireland Limited and Maximillian Schrems, No. Case C-311/18 (European Court of Justice 16 July 2020). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62018CJ0311
[36]	Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/15192
[37]	Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008
[38]	De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. *Patterns*, *3*(6), 100489. https://doi.org/10.1016/j.patter.2022.100489
[39]	Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62. https://doi.org/10.1145/2844110

[40]	Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, *99*, 101896. https://doi.org/10.1016/j.inffus.2023.101896
[41]	Directive of the European Parliament and of the Council on on Liability for Defective Products, European Parliament (2024). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0132_EN.pdf
[42]	Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning* (No. arXiv:1702.08608). arXiv. http://arxiv.org/abs/1702.08608
[43]	Duenser, A., & Douglas, D. M. (2023). Whom to Trust, How and Why: Untangling Artificial Intelligence Ethics Principles, Trustworthiness, and Trust. *IEEE Intelligent Systems*, *38*, 19–26. https://doi.org/10.1109/MIS.2023.3322586
[44]	EC. (2018, April 25). *Artificial Intelligence for Europe*. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237
[45]	EC. (2020). *White Paper on Artificial Intelligence. A European approach to excellence and trust* (No. COM(2020) 65 final). European Commission. https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf
[46]	EC. (2021a). *Proposal for a Regulation, Artificial Intelligence Act*. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
[47]	EC. (2021b). *Proposal for a Regulation, Artificial Intelligence Act (COM(2021) 206 final)*. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
[48]	EC. (2022a). *Proposal for a Directive on liability for defective products*. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0495
[49]	EC. (2022b). *Proposal for a Directive on liability for defective products (COM/2022/495 final)*. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0495
[50]	EC. (2022c). *Proposal for an AI Liability Directive*. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0496
[51]	EC. (2022d). *Proposal for an AI Liability Directive (COM/2022/496 final)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496
[52]	EC. (2022e, December 2). *TTC Joint Roadmap for Trustworthy AI and Risk Management | Shaping Europe's digital future*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management
[53]	EC. (2023a, May 31). *EU-U.S. Terminology and Taxonomy for Artificial Intelligence*. European Commission. https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence
[54]	EC. (2023b, October 30). *Hiroshima Process International Guiding Principles for Advanced AI system*. European Commission. https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system
[55]	Eitel-Porter, R., Corcoran, M., & Connolly, P. (2018). *Responsible AI Principles to Practice | Accenture*. Accenture. https://www.accenture.com/gr-en/insights/artificial-intelligence/responsible-ai-principles-practice
[56]	Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned From Human–Automation Research. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350
[57]	ENISA. (2020). *AI cybersecurity challenges: Threat landscape for artificial intelligence.* European Union Agency for Cybersecurity - ENISA. https://data.europa.eu/doi/10.2824/238222
[58]	ENISA. (2021). *Securing Machine Learning Algorithms* [Report/Study]. https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms
[59]	ENISA. (2023). *A multilayer framework for good cybersecurity practices for AI: Security and resilience for smart health services and infrastructures.* European Union Agency for Cybersecurity - ENISA. https://data.europa.eu/doi/10.2824/588830

[60]	ETSI. (2022). *Securing Artificial Intelligence (SAI)—AI Threat Ontology*. ETSI. https://www.etsi.org/deliver/etsi_gr/SAI/001_099/001/01.01.01_60/gr_SAI001v010101p.pdf
[61]	European Convention on Human Rights, Council Of Europe (1950). https://www.echr.coe.int/documents/d/echr/Convention_ENG
[62]	European Equality Law Network. (2023). *A comparative analysis of non-discrimination law in Europe 2023*. https://op.europa.eu/en/publication-detail/-/publication/0624900d-e73b-11ee-9ea8-01aa75ed71a1/language-en
[63]	European Equality Law Network. (2024). *A comparative analysis of gender equality law in Europe 2023*. https://www.equalitylaw.eu/publications/comparative-analyses
[64]	European Parliament. (2024a, August 20). *AI liability directive. In "A Europe Fit for the Digital Age"*. Legislative Train Schedule. hhttps://www.europarl.europa.eu/legislative-train/theme-legal-affairs-juri/file-ai-liability-directive
[65]	European Parliament. (2024b, August 20). *New Product Liability Directive. In "A Europe Fit for the Digital Age"*. Legislative Train Schedule. https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-new-product-liability-directive
[66]	Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
[67]	Farzadmehr, M., Carlan, V., & Vanelslander, T. (2023). Contemporary challenges and AI solutions in port operations: Applying Gale–Shapley algorithm to find best matches. *Journal of Shipping and Trade*, *8*(1), article no. 27. https://doi.org/10.1186/s41072-023-00155-8
[68]	Feldman, T., & Peake, A. (2021). *End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning* (No. arXiv:2104.02532). arXiv. https://doi.org/10.48550/arXiv.2104.02532
[69]	Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Addison-Wesley.
[70]	Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, *1*(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y
[71]	Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford University Press.
[72]	Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1
[73]	Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Council Of Europe (2024). https://rm.coe.int/1680afae3c
[74]	Freiesleben, T., & Grote, T. (2023). Beyond generalization: A theory of robustness in machine learning. *Synthese*, *202*(4), article no. 109. https://doi.org/10.1007/s11229-023-04334-9
[75]	Freiman, O. (2023). Making sense of the conceptual nonsense 'trustworthy AI'. *AI and Ethics*, *3*(4), 1351–1360. https://doi.org/10.1007/s43681-022-00241-w
[76]	Frey, C. B., & Osborne, M. (n.d.). Generative AI and the Future of Work: A Reappraisal. *Brown Journal of World Affairs*, *30*(1), 1–17.
[77]	G7. (2018). *Charlevoix Common Vision for the Future of Artificial Intelligence*. Global Affairs Canada. https://open.canada.ca/data/en/info/3220321f-aa77-4a4b-aacd-c070a6b058d1/resource/1876fb3d-faf9-41fc-a57d-6f3f37190b68
[78]	G7. (2024). *Apulia G7 Leaders' Communiqué*. Council of Europe. https://www.consilium.europa.eu/media/fttjqncg/apulia-g7-leaders-communique.pdf
[79]	G20. (2019, June 9). *Ministerial Statement on Trade and Digital Economy*. https://www.g20.utoronto.ca/2019/2019-g20-trade.html
[80]	G20. (2023). *G20 New Delhi Leaders' Declaration*. https://www.mea.gov.in/Images/CPV/G20-New-Delhi-Leaders-Declaration.pdf

[81]	Gagnon, G. P., Henri, V., & Gupta, A. (2020). *Trust me!: How to use trust-by-design to build resilient tech in times of crisis*. https://montrealethics.ai/wp-content/uploads/2020/07/PI_Trust-by-design.pdf
[82]	General Data Protection Regulation (GDPR), No. Regulation (EU) 2016/679, European Union, EUR-Lex (2018). https://eur-lex.europa.eu/EN/legal-content/summary/general-data-protection-regulation-gdpr.html
[83]	Giordano, C., Brennan, M., Mohamed, B., Rashidi, P., Modave, F., & Tighe, P. (2021). Accessing Artificial Intelligence for Clinical Decision-Making. *Frontiers in Digital Health*, *3*, article no. 645232. https://doi.org/10.3389/fdgth.2021.645232
[84]	Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
[85]	GOV.UK. (2023a, August 3). *AI regulation: A pro-innovation approach*. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach
[86]	GOV.UK. (2023b, November 1). *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023
[87]	Granese, F., Romanelli, M., Gorla, D., Palamidessi, C., & Piantanida, P. (2021). Doctor: A simple method for detecting misclassification errors. Advances in Neural Information Processing Systems, 34. https://proceedings.neurips.cc/paper_files/paper/2021/file/2cb6b10338a7fc4117a80da24b582060-Paper.pdf
[88]	Grobelnik, M., Perset, K., & Russell, S. (2024, March 6). *What is AI? Can you make a clear distinction between AI and non-AI systems?* OECD.AI Policy Observatory. https://oecd.ai/en/wonk/definition
[89]	Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, *10*, 178–206. https://doi.org/10.1162/tacl_a_00454
[90]	Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8
[91]	Hallinan, D., & Zuiderveen Borgesius, F. (2020). Opinions can be incorrect (in our opinion)! On data protection law's accuracy principle. *International Data Privacy Law*, *10*(1), 1–10. https://doi.org/10.1093/idpl/ipz025
[92]	Hamon, R., Junklewitz, H., & Sanchez Martin, J. I. (2020). *Robustness and explainability of Artificial Intelligence: From technical to policy solutions.* Publications Office of the European Union. https://data.europa.eu/doi/10.2760/57493
[93]	Heilinger, J.-C. (2022). The Ethics of AI Ethics. A Constructive Critique. *Philosophy & Technology*, *35*(3), 61. https://doi.org/10.1007/s13347-022-00557-9
[94]	Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136.
[95]	Herrmann, T., & Pfeiffer, S. (2023). Keeping the organization in the loop: A socio-technical extension of human-centered artificial intelligence. *AI & SOCIETY*, *38*(4), 1523–1542. https://doi.org/10.1007/s00146-022-01391-5
[96]	Hillier, M. (2023, February 20). Why does ChatGPT generate fake references? *TECHE*. https://teche.mq.edu.au/2023/02/why-does-chatgpt-generate-fake-references/
[97]	Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912
[98]	Hohma, E., & Lütge, C. (2023). From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems. *AI*, *4*(4), Article 4. https://doi.org/10.3390/ai4040046
[99]	ICO UK. (2017). *Big data, artificial intelligence, machine learning and data protection*. Information Commissioner's Office, UK.
[100]	Irish SA v. TikTok Technology Limited, No. Binding Decision 2/2023 (European Data Protection Board 7 February 2012). https://www.edpb.europa.eu/our-work-tools/our-documents/binding-decision-board-art-65/binding-decision-22023-dispute-submitted_en
[101]	Irwin, T. (1990). *Aristotle's First Principles*. Oxford University Press.
[102]	ISO. (n.d.). *ISO 2700x standards*. https://www.iso.org/search.html?q=27000

[103]	ISO. (2015). *Quality management systems -Fundamentals and vocabulary.* (No. ISO 9000:2015). ISO. https://committee.iso.org/sites/tc176sc1/home/news/content-left-area/iso-9000-benefits/iso-9000.html
[104]	ISO. (2018). *Risk management – Guidelines* (No. ISO 31000:2018). https://www.iso.org/standard/65694.html
[105]	ISO. (2020). *Information technology—Artificial intelligence—Overview of trustworthiness in artificial intelligence* (Technical Report No. TR 24028). ISO. https://www.iso.org/standard/77608.html
[106]	ISO. (2022a). *Information security, cybersecurity and privacy protection – Guidance on managing information security risks.* (No. ISO/IEC 27005:2022). ISO. https://www.iso.org/standard/80585.html
[107]	ISO. (2022b). *Information technology—Artificial intelligence—Artificial intelligence concepts and terminology* (No. ISO/IEC 22989:2022). ISO. https://www.iso.org/standard/74296.html
[108]	ISO. (2022c). *Trustworthiness—Vocabulary* (No. ISO/IEC TS 5723:2022). ISO. https://www.iso.org/standard/81608.html
[109]	JAPANGOV. (2024, February 9). *The Hiroshima AI Process: Leading the Global Challenge to Shape Inclusive Governance for Generative AI*. https://www.japan.go.jp/kizuna/_userdata/pdf/2024/spring2024/hiroshima_ai_process.pdf
[110]	Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
[111]	Johnsen, P. V., & Remonato, F. (2024). SafetyCage: A misclassification detector for feed-forward neural networks. In *Proceedings of the 5th Northern Lights Deep Learning Conference* (pp. 113–119). PMLR. https://proceedings.mlr.press/v233/johnsen24a.html
[112]	Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185
[113]	Kashive, N., Powale, L., & Kashive, K. (2020). Understanding user perception toward artificial intelligence (AI) enabled e-learning. *The International Journal of Information and Learning Technology*, *38*(1), 1–19. https://doi.org/10.1108/IJILT-05-2020-0090
[114]	Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys*, *55*(2), 1–38. https://doi.org/10.1145/3491209
[115]	Kirlidog, M., & Kaynak, A. (2011). Technology Acceptance Model and Determinants of Technology Rejection: *International Journal of Information Systems and Social Change*, *2*(4), 1–12. https://doi.org/10.4018/jissc.2011100101
[116]	Korzybski, Alfred (1988). Une carte n'est pas le territoire, Prolégomènes aux systèmes non-aristotéliciens et à la Sémantique générale, Editions de l'éclat, 1998.
[117]	Kőszegi, B., & Rabin, M. (2007). Reference-Dependent Risk Attitudes. *American Economic Review*, *97*(4), 1047–1073. https://doi.org/10.1257/aer.97.4.1047
[118]	Larsson, S., Bogusz, C. I., & ; Schwarz, J. A. (Eds.). (2020). *Human-centred AI in the EU: Trustworthiness as a strategic priority in the European member states*. European Liberal Forum (ELF) ; Fores.
[119]	Laugros, A., Caplier, A., & Ospici, M. (2019). *Are Adversarial Robustness and Common Perturbation Robustness Independant Attributes*. the IEEE/CVF International Conference on Computer Vision - ICCV 2019. https://openaccess.thecvf.com/content_ICCVW_2019/html/RLQ/Laugros_Are_Adversarial_Robustness_and_Common_Perturbation_Robustness_Independant_Attributes__ICCVW_2019_paper.html
[120]	Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, *18*(1), 3–32. https://doi.org/10.1111/rego.12512
[121]	Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html
[122]	Lemaire, P. (1999). *Psychologie cognitive*. Éditions De Boeck Université.

[123]	Leslie, D., Rincón, C., Briggs, M., Pereni, A., Jayadeva, S., Borda, A., Bennett, S., Burr, C., Aitken, M., Matell, M., Fischer, C., Wong, J., & Garcia, I. K. (2023). *AI Fairness in Practice*. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2023-12/aieg-ati-fairness_1.pdf
[124]	Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, *55*(9), 1–46. https://doi.org/10.1145/3555803
[125]	Liang, S., Li, Y., & Srikant, R. (2020). Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690.
[126]	Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3313831.3376590
[127]	Liao, Q. V., & Varshney, K. R. (2022). *Human-Centered Explainable AI (XAI): From Algorithms to User Experiences* (No. arXiv:2110.10790). arXiv. http://arxiv.org/abs/2110.10790
[128]	Lin, T. (2021, September 16). *A Human-centric Approach to Fairness in AI*. Quasilinear Musings. https://www.timlrx.com/blog/a-human-centric-approach-to-fairness-in-ai
[129]	Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., & Tang, J. (2022). Trustworthy AI: A Computational Perspective. *ACM Trans. Intell. Syst. Technol.*, *14*(1), 4:1-4:59. https://doi.org/10.1145/3546872
[130]	Lunney, A., Cunningham, N. R., & Eastin, M. S. (2016). Wearable fitness technology: A structural investigation into acceptance and perceived fitness outcomes. *Computers in Human Behavior*, *65*, 114–120. https://doi.org/10.1016/j.chb.2016.08.007
[131]	Maham, S., Tariq, A., Khan, M. U. G., Alamri, F. S., Rehman, A., & Saba, T. (2024). ANN: adversarial news net for robust fake news classification. Scientific Reports, 14(1), 7897.
[132]	Mannion, P., Karimpanal, T. G., Heintz, F., & Vamplew, P. (2021). Multi-Objective Decision Making for Trustworthy AI. In *Proceedings of the 1st Multi-Objective Decision Making Workshop (MODeM 2021)*. http://modem2021.cs.nuigalway.ie/
[133]	Mattioli, J., Sohier, H., Delaborde, A., Pedroza, G., Amokrane-Ferka, K., Awadid, A., Chihani, Z., & Khalfaoui, S. (2023). Towards a holistic approach for AI trustworthiness assessment based upon aids for multi-criteria aggregation⋆. In *SafeAI 2023-The AAAI's Workshop on Artificial Intelligence Safety (Vol. 3381)*. CEUR-WS. https://ceur-ws.org/Vol-3381/15.pdf
[134]	Maximillian Schrems v. Data Protection Commissioner, No. Case C-362/14 (European Court of Justice 6 October 2015). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62014CJ0362
[135]	Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792
[136]	McNamara, D., Ong, C. S., & Williamson, R. C. (2019). Costs and Benefits of Fair Representation Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society—AIES '19* (pp. 263–270). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3306618.3317964
[137]	Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
[138]	Moravec, V., Hynek, N., Skare, M., Gavurova, B., & Kubak, M. (2024). Human or machine? The perception of artificial intelligence in journalism, its socio-economic conditions, and technological developments toward the digital future. *Technological Forecasting and Social Change*, *200*, 123162. https://doi.org/10.1016/j.techfore.2023.123162
[139]	Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5
[140]	Mosley v. the United Kingdom, No. Application no. 48009/08 (European Court of Human Rights May 2011). https://hudoc.echr.coe.int/fre#{%22itemid%22:[%22002-533%22]}
[141]	Neumann, T., & Wolczynski, N. (2023). Does AI-Assisted Fact-Checking Disproportionately Benefit Majority Groups Online? *2023 ACM Conference on Fairness, Accountability, and Transparency*, 480–490. https://doi.org/10.1145/3593013.3594013

[142]	Newman, J. (2023). *A Taxonomy of Trustworthiness for Artificial Intelligence* (CLTC White Paper Series). UC Berkeley - Center for Long-Term Cybersecurity. https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/
[143]	NIS 2 Directive, No. 2022/2555, European Union. https://eur-lex.europa.eu/eli/dir/2022/2555
[144]	NIST. (n.d.). *Privacy Framework*. Retrieved 30 August 2024, from https://www.nist.gov/privacy-framework
[145]	NIST. (2021). *Guide for Conducting Risk Assessments.* (No. NIST SP 800-30). National Institute of Standards and Technology (U.S.). https://www.nist.gov/privacy-framework/nist-sp-800-30
[146]	NIST. (2022, March 17). *AI Risk Management Framework: Initial Draft*. National Institute of Standards and Technology (U.S.). https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf
[147]	NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (No. NIST AI 100-1). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.AI.100-1
[148]	Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342
[149]	OECD. (n.d.). *Metrics for Trustworthy AI*. Retrieved 30 August 2024, from https://oecd.ai/en/catalogue/metrics
[150]	OECD. (2022). *OECD Framework for the Classification of AI systems* (OECD Digital Economy Papers No. 323; OECD Digital Economy Papers, Vol. 323). OECD. https://doi.org/10.1787/cb6d9eca-en
[151]	OECD. (2023, October 30). *G7 Hiroshima Process on Generative Artificial Intelligence (AI): Towards a G7 Common Understanding on Generative AI*. OECD. https://doi.org/10.1787/bf3c0c60-en
[152]	OECD. (2024a, March 4). *Explanatory memorandum on the updated OECD definition of an AI system*. OECD. https://www.oecd.org/en/publications/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51.html
[153]	OECD. (2024b, May 3). *OECD updates AI Principles to stay abreast of rapid technological developments*. OECD, Newsroom. https://www.oecd.org/en/about/news/press-releases/2024/05/oecd-updates-ai-principles-to-stay-abreast-of-rapid-technological-developments.html
[154]	OECD. (2024c, May 3). *Recommendation of the Council on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449
[155]	Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A., & Rinzivillo, S. (2023). Co-design of Human-centered, Explainable AI for Clinical Decision Support. *ACM Trans. Interact. Intell. Syst.*, *13*(4), 21:1-21:35. https://doi.org/10.1145/3587271
[156]	Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., & Gomez, E. (2023). The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency—FAccT '23* (pp. 1139–1150). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3593013.3594069
[157]	Pasquale, F. (2018, August 20). Odd Numbers: Algorithms alone can't meaningfully hold other algorithms accountable. *Real Life*. https://reallifemag.com/odd-numbers/
[158]	PDPC Singapore. (2018). *PDPC Discussion Paper—Artificial Intelligence and Personal Data*. https://www.pdpc.gov.sg/help-and-resources/2020/03/discussion-paper-artificial-intelligence-and-personal-data
[159]	Phillips, S. C., Taylor, S., Boniface, M., Modafferi, S., & Surridge, M. (2024). Automated Knowledge-Based Cybersecurity Risk Assessment of Cyber-Physical Systems. *IEEE Access*, *12*, 82482–82505. https://doi.org/10.1109/ACCESS.2024.3404264
[160]	Pomerleau, W. P. (2023). Western Theories of Justice. In *Internet Encyclopedia of Philosophy*. https://iep.utm.edu/justwest/
[161]	Potter, N. N. (2002). *How Can I Be Trusted?: A Virtue Theory of Trustworthiness*. https://rowman.com/ISBN/9780742511507/How-Can-I-Be-Trusted?-A-Virtue-Theory-of-Trustworthiness

[162]	Powles, J. & Nissenbaum, H. (2018, December 7). The Seductive Diversion of 'Solving' Bias in Artificial Intelligence. *OneZero*. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53
[163]	Pratt, L. (2019). *Link: How decision intelligence connects data, actions, and outcomes for a better world* (First edition). Emerald Publishing.
[164]	Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. AI and Ethics, 3(3), 699-716.
[165]	Quelle, D., & Bovet, A. (2024). The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, *7*, article no. 1341697. https://doi.org/10.3389/frai.2024.1341697
[166]	Rafique, H., Almagrabi, A. O., Shamim, A., Anwar, F., & Bashir, A. K. (2020). Investigating the Acceptance of Mobile Library Applications with an Extended Technology Acceptance Model (TAM). *Computers & Education*, *145*, article no. 103732. https://doi.org/10.1016/j.compedu.2019.103732
[167]	Regulation of the European Parliament and of the Council on Horizontal Cybersecurity Requirements for Products with Digital Elements, European Parliament (2024). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0130_EN.pdf
[168]	Regulation of the Europen Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence, European Parliament (2024). https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf
[169]	Regulation on General Product Safety, No. 2023/988, European Union (2023). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2023.135.01.0001.01.ENG&toc=OJ%3AL%3A2023%3A135%3ATOC
[170]	Regulation on in Vitro Diagnostic Medical Devices, No. 2017/746, European Union (2017). https://eur-lex.europa.eu/eli/reg/2017/746/oj
[171]	Regulation on Machinery, No. 2023/1230, European Union (2023). https://eur-lex.europa.eu/eli/reg/2023/1230/oj
[172]	Regulation on Medical Devices, No. 2017/745, European Union (2017). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745
[173]	Regulation the European Parliament and the Council of the European Union on European Health Data Space, European Parliament (2024). https://www.europarl.europa.eu/news/en/press-room/20240419IPR20573/eu-health-data-space-more-efficient-treatments-and-life-saving-research
[174]	Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. *AI and Ethics*, *3*(3), 735–744. https://doi.org/10.1007/s43681-022-00200-5
[175]	Richey, R. G., Chowdhury, S., Davis-Sramek, B., Giannakis, M., & Dwivedi, Y. K. (2023). Artificial intelligence in logistics and supply chain management: A primer and roadmap for research. *Journal of Business Logistics*, *44*(4), 532–549. https://doi.org/10.1111/jbl.12364
[176]	Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, *1*(1), 33–36. https://doi.org/10.1002/hbe2.117
[177]	Ronanki, K., Cabrero-Daniel, B., Horkoff, J., & Berger, C. (2023). RE-centric Recommendations for the Development of Trustworthy(er) Autonomous Systems. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems—TAS '23* (pp. 1–8). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3597512.3599697
[178]	Rosenthal, D. (2024, May 21). *The most important compliance requirements for AI.* Vischer. https://www.vischer.com/en/knowledge/blog/part-18-the-most-important-compliance-requirements-for-ai/
[179]	Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, *26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y
[180]	Sadek, M., Calvo, R. A., & Mougenot, C. (2023). Designing value-sensitive AI: a critical review and recommendations for socio-technical design processes. AI and Ethics, 1-19.
[181]	Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot is more informative than the ROC Plot when evaluating Binary Classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), article no. e0118432. https://doi.org/10.1371/journal.pone.0118432

[182]	Salminen, J., Wenyun Guan, K., Jung, S.-G., & Jansen, B. (2022). Use Cases for Design Personas: A Systematic Review and New Frontiers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3491102.3517589
[183]	Schlegel, M., & Sattler, K.-U. (2023). Management of Machine Learning Lifecycle Artifacts: A Survey. *SIGMOD Rec.*, *51*(4), 18–35. https://doi.org/10.1145/3582302.3582306
[184]	Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (No. NIST SP 1270; p. NIST SP 1270). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.SP.1270
[185]	Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, *27*(12), 2176–2182. https://doi.org/10.1038/s41591-021-01595-0
[186]	Shneiderman, B. (2020a). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Trans. Interact. Intell. Syst.*, *10*(4), 26:1-26:31. https://doi.org/10.1145/3419764
[187]	Shneiderman, B. (2020b). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118
[188]	Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press.
[189]	Simion, M., & Kelp, C. (2023). Trustworthy artificial intelligence. *Asian Journal of Philosophy*, *2*(1), 8. https://doi.org/10.1007/s44204-023-00063-5
[190]	Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–115. https://doi.org/10.2307/1884852
[191]	Skournik, P. (2023). *Applicability of Artificial Intelligence and Machine Learning in the Maritime Industry and Port Management*. The American Association of Port Authorities. https://www.aapa-ports.org/unifying/landing.aspx?ItemNumber=21705&navItemNumber=20808
[192]	Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act* (SSRN Scholarly Paper No. 3899991). https://doi.org/10.2139/ssrn.3899991
[193]	Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305
[194]	The Employment Equality Directive, No. 2000/78, European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0078
[195]	The Gender Equality Directive, No. 2006/54, European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32006L0054
[196]	The Gender Goods and Services Directive, No. 2004/113, European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32004L0113
[197]	The Race Equality Directive, No. 2000/43, European Union. Directive - 2000/43 - EN - EUR-Lex (europa.eu)
[198]	The White House. (2020, November 17). *Guidance for Regulation of Artificial Intelligence Applications*. The White House, USA. https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf
[199]	The White House. (2022, October 4). *Blueprint for an AI Bill of Rights | OSTP*. The White House. https://www.whitehouse.gov/ostp/ai-bill-of-rights/
[200]	Thiebes, S., Lins, S. & Sunyaev, A. Trustworthy Artificial Intelligence. *Electron Markets* 31, 447–464 (2021). https://doi.org/10.1007/s12525-020-00441-4
[201]	Toreini, E., Aitken, M., Coopamootoo, K. P. L., Elliott, K., Zelaya, V. G., Missier, P., Ng, M., & van Moorsel, A. (2022). *Technologies for Trustworthy Machine Learning: A Survey in a Socio-Technical Context* (No. arXiv:2007.08911). arXiv. https://doi.org/10.48550/arXiv.2007.08911
[202]	Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2024).

	Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology*, *42*(1), 3–15. https://doi.org/10.1007/s11604-023-01474-3
[203]	UK Parliament. (2023). *Artificial Intelligence (Regulation) Bill*. https://bills.parliament.uk/bills/3519
[204]	UN. (2024a, March 11). *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*. UN. https://documents.un.org/doc/undoc/ltd/n24/065/92/pdf/n2406592.pdf
[205]	UN. (2024b, March 21). *General Assembly adopts landmark resolution on artificial intelligence*. UN News. https://news.un.org/en/story/2024/03/1147831
[206]	UNESCO. (2022). *Recommendation on the Ethics of Artificial Intelligence* (No. SHS/BIO/PI/2021/1). https://unesdoc.unesco.org/ark:/48223/pf0000381137
[207]	UNESCO. (2023). *Ethical impact assessment: A tool of the Recommendation on the Ethics of Artificial Intelligence*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386276
[208]	Unity. (2018, November 28). *Introducing Unity's Guiding Principles for Ethical AI*. https://unity.com/de/archive/blog/engine-platform/introducing-unitys-guiding-principles-for-ethical-ai
[209]	Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190498511.001.0001
[210]	Vallor, S. (2024). *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford University Press.
[211]	Varona, D., & Suárez, J. L. (2022). Discrimination, Bias, Fairness, and Trustworthy AI. *Applied Sciences*, *12*(12), Article 12. https://doi.org/10.3390/app12125826
[212]	Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412
[213]	Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness* (pp. 1–7). ACM. https://dl.acm.org/doi/10.1145/3194770.3194776
[214]	Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009
[215]	Von Hannover v. Germany, No. Application no. 59320/00 (European Court of Human Rights 24 September 2004). https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-61853%22]}
[216]	Von Neumann, J., & Morgenstern, O. (1953). *Theory of Games and Economic Behavior.* (3rd edition). Princeton University Press.
[217]	Walter, S. (2022). User Journey Mapping. Publisher(s): SitePoint, ISBN: 9781925836493
[218]	Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 842–886. https://doi.org/10.2139/ssrn.3063289
[219]	Wenar, L. (2021). John Rawls. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/rawls/
[220]	WHO. (2023). *Regulatory considerations on artificial intelligence for health*. World Health Organization - WHO. https://iris.who.int/bitstream/handle/10665/373421/9789240078871-eng.pdf
[221]	Wing, J. M. (2021). Trustworthy AI. *Communications of the ACM*, *64*(10), 64–71. https://doi.org/10.1145/3448248
[222]	Wu, C., Li, Y.-F., & Bouvry, P. (2023). *Survey of Trustworthy AI: A Meta Decision of AI* (No. arxiv.2306.00380). https://arxiv.org/pdf/2306.00380
[223]	Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *ACM Interactions*, *26*(4), 42–46. https://doi.org/10.1145/3328485
[224]	Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., & Yu, H. (2020). *Federated learning*. Springer.

[225]	Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems—CHI '19* (p. article no. 279). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3290605.3300509
[226]	Zanotti, G., Petrolo, M., Chiffi, D., & Schiaffonati, V. (2023). Keep trusting! A plea for the notion of Trustworthy AI. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01789-9
[227]	Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 325–333). PMLR. https://proceedings.mlr.press/v28/zemel13.html
[228]	Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, *3*(4), article no. 100455. https://doi.org/10.1016/j.patter.2022.100455
[229]	Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society—AIES '18* (pp. 335–340). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3278721.3278779

APPENDIX A

In this Appendix we present a) all user stories collected from the use case partners and the 2nd round of co-creation workshops, and b) the user requirements collected from the use case partners and their technical counterparts along with the user requirements translated from the user stories.

A.1. User Stories

Table 19. User stories from Use case 1 partners and technical counterparts.

User Story ID#	UC Partner	As person/role X...	... I want functionality Y...	... so I get business benefit Z.	Testable (Completed by WP2 partners)	Scope
US1_ T01	MUP	General Practitioner/ Pre-Hospital Healthcare Provider	I want to have a clear explanation on the health risk level value - what is the basis of the assessment made	so that the assessed health risk level will be accepted	Y	SUT
US1_ T02	MUP	General Practitioner/ Pre-Hospital Healthcare Provider	I want to have a clear explanation on what benefit is expected by the proposed preventive program	so that the proposed prevention program will be accepted and implemented.	Y	SUT
US1_ T03	MUP	General Practitioner/ Pre-Hospital Healthcare Provider	I want to have a justification for the required parameters to be monitored, their frequency and values	so that secondary data parameters for monitoring will be accepted and set as a program.	Y	SUT
US1_ T04	MUP	Emergency Medicine Specialists and Healthcare Providers	I want to have a clear explanation on the health risk level value - what/why	so that the assessed health risk level will be accepted.	Y	SUT
US1_ T05	MUP	Emergency Medicine Specialists and Healthcare Providers	I want to have a justification on the proposed consultations and tests	so that the proposed prevention program will be accepted and implemented.	Y	SUT
US1_ T06	MUP	Emergency Medicine Specialists and Healthcare Providers	I want to have a justification on the need for monitoring	so that secondary data parameters for monitoring will be accepted and set as a program.	Y	SUT

Table 20. User stories from Use case 2 partners and technical counterparts (US[X]_T[Y] collected via the User Stories template and US[X]_Q[Y] collected via the User Requirements collection questionnaires)

User Story ID#	UC Partner	As person/role X...	... I want functionality Y...	... so I get business benefit Z.	Testable (Completed by WP2 partners)	Scope Relevant Relevance to THEMIS 5.0 (Completed by WP2 partners)
US2_ T01	VPF	Port Traffic Control	I want access to the assessment of the AI SUT against trustworthiness characteristics.	so I can make informed decisions with regards to the adoption of the AI system's output.	Y	THEMIS Y

US2_ T02	VPF	Pilot	I want to be informed about the risks that the current trustworthiness assessment entails	so I can make informed decisions with regards to the adoption of the AI system's output. Risks may involve wrong planning of the following berths to terminals, as Pilots do collaborate with Port Traffic Control in scheduling the pilotage service	Y	THEMIS Y
US2_ T03	VPF	Port Terminal	I want to be informed about the risks that the current trustworthiness assessment entails	so I can make informed decisions with regards to the adoption of the AI system's output. Risks may involve wrong resource management in the terminal, incurring additional costs or congestion at the terminal entrance	Y	THEMIS Y
US2_ T04	VPF	AI System developer	I want to be informed about the current level of trust considering a predefined time frame backward thanks to the conversations of users with the chatbot	so I can adjust and improve the models of the AI System	Y	THEMIS Y
US2_ T05	VPF	Port Traffic Control / Port Terminal / Pilot / Towing Services /Mooring Services / Transport Company	I want to be explained in a user-friendly manner the main factors that influenced the prediction results	so I can trust better the output of the AI System	Y	SUT N (refers to SUT)
US2_ T06	VPF	AI System developer	I want the system to explain in a user-friendly manner the main factors that influenced the prediction results	so that the AI System will be better adopted by the users	Y	THEMIS + SUT N (refers to SUT)
US2_ T07	VPF	Port Traffic Control / Port Terminal / Pilot / Towing Services /Mooring Services / Transport Company	I want to be informed about the current accuracy of predictions made by the AI System in comparison to different scenarios (vessels of similar type, weekday, shipping agency, etc.)	so I will not place excessive trust in predictions or be swayed by false negative predictions due to inaccurately estimated prediction accuracy	Y	THEMIS + SUT Y
US2_ T08	VPF	Port Traffic Control	I want to know about possible biases in the AI System against other agents	so I can take this into account and mitigate the associated risks	Y	THEMIS + SUT Y
US2_ T09	VPF	Port Traffic Control / Port Terminal / Pilot / Towing Services /Mooring Services / Transport Company	I want to read predictions as accurate as possible especially when the ETA predicted differs considerably with respect to the one given by the vessel's captain	so I can feel confident in taking desicions derived from this prediction	-	SUT N (refers to SUT)
US2_Q01	VPF	Port Authority Manager-Port Community System	I would like to have a high-level report with most important results in a nutshell based on my preferences. It can be through the chat but in one single response.	So that I don't spend too much time in collecting results.	Y	Y
US2_Q02	VPF	Port Authority Manager-Port	I would like to dig into some details of a specific result from the above report by asking through the chat or by	So that I can go to what it really matters to us.	Y	Y

		Community System	navigation within the report (hyperlink).			
US2_Q03	VPF	Port Authority Manager-Port Community System	I would like to have a high-level report with most important vulnerabilities in a nutshell based on my preferences. It can be through the chat but in one single response.	So that I don't spend too much time in collecting results.	Y	Y
US2_Q04	VPF	Port Authority Manager-Port Community System	I would like to dig into some details of a specific vulnerability from above's report by asking through the chat or by navigation within the report (hyperlink).	So that I can go to what it really matters to us.	Y	Y
US2_Q05	VPF	Vessel Traffic Service Supervisor	I would like the system to display a comparison between shipping agents' ETA, and AI ETA vs ATA.	To show the shipping agents the benefits of providing accurate data.	Y	N (refers to SUT)
US2_Q06	VPF	Transportation Traffic Manager	I would like to receive the current level of accuracy with respect to the comparison of all previous predictions vs observations during last days	As it contributes to the transparency of ETA predictions in terms of accuracy.	Y	Y
US2_Q07	VPF	Transportation Traffic Manager	I would like to receive the current level of accuracy with respect to the comparison of previous predictions vs observations during x last days for the port calls that fall into a specific group (port call terminal, shipping agency, etc.)	As it contributes to the transparency of ETA predictions in terms of accuracy.	Y	Y
US2_Q08	VPF	Transportation Traffic Manager	I would like to receive the current level of accuracy with respect to the comparison of previous predictions vs observations during x last days for the port calls that fall into a specific group of external characteristics (meteorological conditions, holidays, etc.)	As it contributes to the transparency of ETA predictions in terms of accuracy.	Y	Y
US2_Q09	VPF	Transportation Traffic Manager	I would like to initiate a chatbot conversation with the system to gather information about the system's trustworthiness and Risk Assessment	As it makes the system more user friendly and in line with today's similar tools: ChatGPT, Gemini, etc.	Y	Y
US2_Q10	VPF	Transportation Traffic Manager	The chatbot may offer predefined questions once started the conversation. "Do you want to know how factor X has influenced with the that risk?"	To facilitate and speed up the process.	Y	Y
US2_Q11	VPF	Transportation Traffic Manager	I would like to provide feedback about my level of trust regularly in terms of accuracy and if I have really used the ETA predictions or not	As It will inform AI developers about the usability of the ETA prediction service.	Y	N (refers to SUT)

Table 21. User stories from Use Case 3 partners and technical counterparts.

User Story ID#	UC Partner	As person/role X...	... I want functionality Y...	... so I get business benefit Z.	Testable (Completed by WP2 partners)	Relevant Relevance to THEMIS 5.0 (Completed by WP2 partners)
US3_01	ANA, ATC	Journalist, Commerical Director, AI System developer	I want access to the assessment of the AI SUT against trustworthiness characteristics in a user friendly manner.	so I can make informed decisions with regards to the adoption of the AI system's output.	Y	Y
US3_02	ANA, ATC	Journalist, Commerical Director, AI System developer	I want to know which parts of the SUT suffer	so to get an idea about why the percentages that I was presented with are such	Y	Y
US3_03	ANA	Journalist, Commerical Director	I want to be able to explore all solutions and choose based on explanations	so I can make informed decisions with regards to the adoption of the enhancement solution	Y	Y
US3_04	ANA	Journalist, Commerical Manager I	I want the system to perform the optimal solution and then explain to me why this solution is the best, based on my persona.	so to assess if my preferences have been identified correctly	Y	Y
US3_05	ANA, ATC	Journalist, Commerical Director, AI System developer	I want to know the new trustworthiness results after the enhancement of the system	So I can make informed decisions about the THEMIS 5.0 platform and the AI System	Y	Y
US3_06	ATC	AI System Developer	I want to be informed about the impact (positive or negative) that each optimisation measure will have in other aspects of the AI SUT.	So I can take informed decisions with regards to the implementation of optimisation measures	Y	Y
US3_07	ATC	AI System Developer	I want measurable indications about the effectiveness of the trustworthy methods applied.	So that I know the improvements per TW property.	Y	Y
US3_08	ATC	AI System Developer	I want a testbed environment to apply the trustworthiness methods in a pipeline.	So that I have a controlled system to evaluate the changes in trustworthiness properties.	Y	Y

Table 22. User stories from co-creation with AI-users and business stakeholders of UC1 (Health)

Use Case	As person/role X... (who)	... I want functionality Y... (what)	... so I get business benefit Z (why)	Testable	Relevant
Name the relevant Use Case	*Name the contributor's role in the organisation*	*Describe a functionality that you would like the THEMIS system to have.*	*Describe the benefit that you would like to achieve through this functionality*	*(Completed by WP2 partners)*	*Relevance to THEMIS 5.0 (Completed by WP2 partners)*
INFORMATION FOR PREFERENCE PROFILES					
US1_C1	Healthcare professionals (practitioners and administrative)	I want the chatbot to ask about my attitude towards AI	So the chatbot can adjust to my preference profile and support me by giving information in a way that can make me approach AI in a more neutral way or	Y	Y

			provide counterarguments that challenges my attitude		
US1_C2	Healthcare professionals (practitioners and administrative)	I want the chatbot to ask how many years of experience I have in my field of work	So the THEMIS 5.0 tool can adjust the optimization suggestions to my level of experience within the field	Y	Y
US1_C3	Healthcare professionals (practitioners and administrative)	I want the chatbot to ask into details about the professional field I work in	So the THEMIS 5.0 tool can adjust its optimization suggestions to the area I am working with (working with diagnostics, prevention, or research, etc.)	N	N
US1_C4	Healthcare professionals (primarily and practitioners)	I want the chatbot to ask details about my specialty in the field of medicine	So the THEMIS 5.0 tool can take into consideration that there might be specific circumstances in my specialty that needs to be considered when generating the optimization suggestions (e.g. if I work with pregnancy, the lack of male data in the data set might not be a problem)	Y	Y
US1_C5	Healthcare professionals (practitioners and administrative)	I want the chatbot to ask details about my professional interests	So the THEMIS 5.0 tool can take into consideration if my professional interest might impact how I weigh decisions	Y	Y
US1_C6	Healthcare professionals (practitioners and administrative)	I want the chatbot to ask for details about my qualifications/competencies	So the THEMIS 5.0 tool can adjust its preference profiling based on my specific skillset	UNCLEAR	UNCLEAR
CHATBOT INTERACTION					
US1_C7	Healthcare professionals (primarily administrative)	I want the interaction with the chatbot to take place once when I start using the THEMIS 5.0 tool	So I make a comprehensive user preference profile the first time I use the THEMIS 5.0 tool and not every time I use it	Y	Y
US1_C8	Healthcare professionals (primarily administrative)	I want the interaction with the chatbot to take around 30-60 min.	So I only have to make the preference profile once but know that it will be thorough and useful for future uses of the THEMIS 5.0 tool to assess and optimize other AI tools	Y	Y
US1_C9	Healthcare professionals (practitioners and administrative)	I want the interaction with the chatbot to ask the same questions in different ways (with variation in wording)	So it my answers can be processed by the chatbot to avoid and counter any potential misunderstandings and my own potential bias	Y	Y
US1_C10	Healthcare professionals (primarily practitioners)	I want the interaction with the chatbot to focus on 20 questions to begin with and have the ability to answer more later	So I can get started with using the THEMIS 5.0 tool quickly, and decide to expand my user preference profile analysis later on if I want more tailored assessment and optimization suggestions	Y	Y
US1_C11	Healthcare professionals (practitioners and administrative)	I want the chatbot/THEMIS 5.0 tool to work on a computer	So I can easily use it – but it should also work on a smartphone	Y	Y
US1_C12	Healthcare professionals (practitioners and administrative)	I want the interaction to be text-based	So I can quickly go through the interaction with the chatbot on a computer	Y	Y
US1_C13	Healthcare professionals (practitioners and administrative)	I want the interaction to have a multiple-choice option to choose my profession from	So I can make sure the profession fit the knowledge of the system and no mistakes happens if there are spelling errors or specifications of profiles that the system does not recognize	Y	Y
US1_C14	Healthcare professionals (practitioners and administrative)	I want to be able to update the user preference profile	So it is always updated on changes in my professional field, changes to my workplace or new competencies/new fields of work/change of role	Y	Y
US1_C15	Healthcare professionals (practitioners and administrative)	I want to be able to create several user preference profiles	So I can choose which profile the THEMIS 5.0 tool should consider in its assessment and optimization, depending on what field of work I am working in (in case of having several fields of work)	Y	Y
OPTIMISATION SUGGESTION					

US1_C16	Healthcare professionals (practitioners and administrative)	I want the optimization suggestions to show in a rating system (e.g. five stars or percentages)	So I can easily see a visual disclaimer on various optimization parameters (accuracy, robustness, etc.)	Y	Y
US1_C17	Healthcare professionals (practitioners and administrative)	I want to have 4-5 optimization suggestions to choose from	So I can pick the one I find most relevant and read more details about it	UNCLEAR	UNCLEAR
US1_C18	Healthcare professionals (practitioners and administrative)	I want to be able to read details about each of the optimization suggestions	So I can consider all of them and choose what optimization to prioritize	Y	Y
US1_C19	Healthcare professionals (practitioners and administrative)	I want a user-friendly interface	So I can easily access the THEMIS 5.0 tool while working in a busy working environment	UNCLEAR	UNCLEAR

Table 23. User stories from co-creation with AI-users and business stakeholders of UC2 (Port)

Use Case	As person/role X... (who)	... I want functionality Y... (what)	... so I get business benefit Z (why)	Testable	Relevant
Name the relevant Use Case	*Name the contributor's role in the organisation*	*Describe a functionality that you would like the THEMIS system to have.*	*Describe the benefit that you would like to achieve through this functionality*	*(Completed by WP2 partners)*	*Relevance to THEMIS 5.0 (Completed by WP2 partners)*
INFORMATION FOR PREFERENCE PROFILES					
US2_C1	General opinion by employees the port sector	I want the user preference profile to be generic and not personalized to each profession	So I can rely on the same optimization suggestions for anyone working in the sector	Y	Y
US2_C2	On the ground employees in the port sector	The preference profile should not be based on personal or role specific characteristics but rather based on the port authorities' decisions	So the THEMIS 5.0 tool does not suggest optimizations that are contradictory between different roles	Y	N
US2_C3	On the ground employees in the port sector	The chatbot should know of the different roles that will use the AI tool	So the THEMIS 5.0 tool can take them all into consideration and not make contradictory optimization suggestions	Y	Y
US2_C4	AI developers in the port sector	The chatbot should ask about my attitude towards the use of AI	So the THEMIS 5.0 tool can take into consideration that I might need arguments to counter my positive or negative attitude towards AI and the data set in the AI tool	Y	Y
US2_C5	AI developers in the port sector	The chatbot should ask about my preferences on accuracy	So I know that the optimization suggestions include suggestions on how to reach high level of accuracy because accuracy is crucial for the ETA calculation	Y	Y
US2_C6	AI developers in the port sector	I want the chatbot to ask about the internal port rules	So I am sure that the optimization suggestions consider existing rules, both local, national and international	Y	Y
US2_C7	AI developers in the port sector	I want the chatbot to ask about important matters related to port security	So I know that the optimization suggestions take into consideration the laws and procedures related to security at the port	Y	Y
OTHER					
US2_C8	Employees in the port sector	I want the optimization suggestions to consider "just-in-time" (JIT) arrival	So I can make sure that we work towards the international new standard concept/aim for more efficient vessel ETA in the port sector to help decrease cost, reduce waiting time and other factors that might benefit from using "Just-in-time" arrival	Y	Y

Table 22. User stories from co-creation with AI-users and business stakeholders of UC3 (Media).

Use Case	As person/role X... (who)	... I want functionality Y... (what)	... so I get business benefit Z (why)	Test-able	Relevant
Name the relevant Use Case	*Name the contributor's role in the organisation*	*Describe a functionality that you would like the THEMIS system to have.*	*Describe the benefit that you would like to achieve through this functionality*	*(Completed by WP2 partners)*	*Relevance to THEMIS 5.0 (Completed by WP2 partners)*
INFORMATION FOR PREFERENCE PROFILE					
US3_C1	Media professionals	I want the chatbot to ask me about what profession/field of work I work within	So the chatbot can consider the differences in responsibility between various professions in the sector (e.g, news vs entertainment)	Y	Y
US3_C2	Media professionals	I want the chatbot to ask me about the ethical guidelines in the company I work within	So I know that the ethical guidelines I need to follow as an employee is being met as there are differences between the ethical guidelines of companies	Y	Y
US3_C3	Media professionals	I want the chatbot to ask me nuanced questions on how I understand 'Trustworthy AI'	So I am sure that the final optimization suggestions fit my own understanding of 'Trustworthy AI', and also fit the understanding within the sector	Y	Y
US3_C4	Administrative/decision makers in the media sector	I want the chatbot to ask me about my concerns about the climate impact of AI tools	So THEMIS 5.0 can consider my standpoint related to climate when making optimization suggestions	Y	Y
US3_C5	Media professionals	I want the chatbot to ask me where in the work process, I am using the AI-tool	So the chatbot understands what is important in this step of the work process	Y	Y
US3_C6	Media professionals	I want the chatbot to ask me about my previous experience with using AI tools	So the optimization suggestions take into consideration my knowledge about functionality, bias, data use, and transparency, and explains the benefits of the suggested optimization in a manner that is suitable	Y	Y
US3_C7	Media professionals	I want the chatbot to ask me about copyright challenges in my field of work	So I avoid any infringement on copyright laws which is an important topic in the sector	*Y*	*N*
US3_C8	Media professionals	I want the chatbot to ask about security measures in my company	So I can make sure security will not become an issue in the optimization suggestions	*Y*	*Y*
US3_C9	Media professionals	I want the chatbot to ask about who the different users of the AI tool are	So THEMIS 5.0 can understand how different users may need to interact with the AI tool and take it into consideration when it gives me optimization suggestions so it does not generate recommendations that might become a disadvantage for other types of users	*Y*	*Y*
CHATBOT INTERACTION					
US3_C10	Media professionals	I want the interaction with the chatbot to be short (e.g., to last maximum 15/30 min.)	So I can find time for it in a busy work life	*Y*	*Y*
US3_C11	Media professionals	I want the interaction with the chatbot in the THEMIS 5.0 tool to take place on a laptop	So it fits with my primary working tool (laptop) that I am always using for my daily work	*Y*	*Y*

US3_C12	Media administrative/decision makers	I want the chatbot to provide me with parameters/categories that I need to fill in	So I can easily fill the professional user profile needed for the optimization suggestions without having to write lengthy answers to the chatbot	*Y*	*Y*
OPTIMISATION SUGGESTIONS					
US3_C13	Media professionals	I want the optimization suggestions to consider the data source and how data is used by the AI tool	So I can get information about whether the AI tools data sources is compliant with the (ethical/climate/etc.) standards/guidelines of my company	*Y*	*Y*
US3_C14	Media professionals	I want the THEMIS 5.0 tool to be able to suggest if the use of an AI tool is not needed in a given situation	So I can make sure to only use an AI tool when it is providing enough positive impact for the task it is used for	*Y*	*N*
US3_C15	Media professionals	I want THEMIS 5.0 to make suggestions about how I can communicate the way AI tools are used in my company	So I can clearly indicate that an AI tool has been involved in the production of the outcome in a way that does not impact the trustworthiness of me or the company	*Y*	*Y*

A.2. System Requirements

Table 23. System requirements from THEMIS 5.0 project objectives.

THEMIS 5.0 Pipeline	ID	System Requirement	Passage form the DoA
OVERALL	SR_01	THEMIS cloud-based services should interact with the user via an AI-driven conversational interface that will be able to explain in the necessary level of detail the outputs of AI systems.	*"The THEMIS 5.0 ecosystem is composed of cloud-based, AI-based services that seamlessly engage with humans by means of AI-driven interactive Dialogues. Specifically, an AI-driven conversational agent will transmit sufficient but not excessive human-interpretable explanations on how the AI system takes a particular set of inputs and reaches a conclusion*" (1.1.1, p.3)
	SR_02	The trustworthiness optimisation should be a continuous process implemented in improvement cycles, where users select improvements suitably selected to address their personalised trustworthiness assessment.	*"This interaction will enable the execution of continuous trustworthiness improvement processes where, at each trustworthiness improvement cycle, human-centered assessment of the trustworthiness of the AI system will take place and corrective actions will be determined ahead of the next improvement cycle.*" (1.1.1, p.3)
	SR_03	Human-centric trustworthiness assessment and optimisation should (i) follow a risk-assessment approach and (ii) be transparent to the users.	*"Human-centered evaluation and optimisation of trustworthiness (…) based on a dynamic, inherently transparent risk-assessment approach …"* (1.1.1, p.3)
	SR_04	THEMIS 5.0 AI approaches and models comply with the European legal and ethical framework.	*"…THEMIS 5.0 AI approaches and models comply with the European legal and ethical framework" (1.1.2, p.4)*
	SR_07	THEMIS conversational interface must (i) adjust to the users' preferences and traits, and (ii) support enhanced transparency on the AI outcomes.	*"…enhanced transparency of the hybrid decision supporting AI system based on an intelligent conversational environment which adjusts to the human profiles and traits (…)"* (1.1.2, p.5)
	SR_11	THEMIS should be implemented as a flexible tool that users can use to enhance the fairness, transparency, and accountability of the AI-systems they are using.	*"… a flexible tool that users can use to enhance trust in their AI systems and achieve fairness, transparency, and accountability."* (1.2.4, p.14)
	SR_12	THEMIS should be implemented as a micro-service-based platform supporting cycles of design, deployment, evaluation, tuning of experiment variants of AI systems.	*"… a micro-service-based platform supporting cycles of design, deployment, evaluation, tuning of experiment variants of AI systems, …"* (1.2.4, p.14)
	SR_13	Personalised AI Trustworthiness assessment should consider (i) the legal, moral, and ethical principles of its human users, (ii) their position and role in their organisation, (iii) the business objectives.	*"… for an AI system to be trustworthy, the legal, moral, and ethical principles of its human users as well as the organisational responsibility and liability in relation to the business objective of the consideration embedding socio-technical system needs to be taken into consideration"* (1.1.1, p.3)
	SR_14	The AI-driven conversational interface will capture the users' decision-support needs, moral values and business goals.	*"… intelligently elicit the knowledge related to human particular decision support needs and moral values, as well as to the key business goals of the embedding socio-technical system."* (1.1.1, p.3)

IDENTIFY	SR_15	THEMIS should identify (i) the users' needs, (ii) their moral values, and (iii) success factors.	"*Effective elicitation of human users' decision support needs and moral values and key success factors …*" (1.1.1, p.3)
	SR_16	Training datasets will be created through co-creation sessions and purpose -specific web-based tools, respecting the participants' privacy.	"*benchmark data sets for training the THEMIS 5.0 AI models (…) through appropriately set-up, privacy-preserving co-creation sessions and web-based tools, taking care to minimise human curation during co-creation.*" (1.1.2, p.5)
	SR_17	THEMIS should incorporate methods and approaches to capture and model users' utility, needs and business priorities.	"*… design methods and approaches to help elicit and model user utility in such decision settings;*" (1.1.3, p.7)
	SR_18	THEMIS DI methodology, should enable the user to create a model of the socio-technical environment using a combination of high-level knowledge specification and machine learning.	"*… a DI methodology will enable the user to create a model of the socio-technical environment in which the AI system operates, using a combination of high-level knowledge specification and machine learning.*" (1.1.3, p.8)
	SR_19	THEMIS should allow the users to model an AI system in the context of a socio-technical environment in accordance with decision support needs.	"*… enabling the user to model an AI system in the context of a socio-technical environment in accordance with decision support needs and moral values.*" (1.1.3, p.8)
	SR_20	THEMIS personalised trustworthiness assessment will be based (i) on the users' behavioural traits based on a system of user preference categories with specific attributes and measurement scales: Personality, Social-behavioural, Technical awareness, Motivation, and Trigger, as well as on (ii) the users' moral values based on moral judgement taking into account applied ethical theories.	"*… incorporate the human perspective to the evaluation of trustworthiness by assessing both human behavioural and moral value aspects. (…) human profiles can be proposed using five (5) different categories of traits with specific attributes and measurement scales: Personality, Social-behavioural, Technical awareness, Motivation, and Trigger. (…) Moral value evaluation of the human factor in THEMIS 5.0 will take place via a critical exploration of moral judgement of the human and in terms of a self-reflective dialogical awareness. In particular the development of the moral values evaluation will take into account applied ethical theories.*" (1.2.3.4, p.12-p.13)
	SR_21	THEMIS should feature a conversational interface that will engage in personalised dialogues that will capture the users' values, preferences, requirements, objectives, capabilities, motives and behavioural patterns.	"*The THEMIS 5.0 personalised dialogues will capture and provide human values, preferences, requirements, human-defined objectives, capabilities, motives and behavioral patterns to all phases of the [THEMIS] methodology.*" (1.2.3.5, p.14)
	SR_22	THEMIS conversational interface must be able to capture (i) the users' decision support needs, (ii) the users' ethical values, and the targets, objectives and KPIs that the optimised AI system will need to support.	"*… the Human-AI conversational agent will elicit knowledge related to human particular decision support needs and ethical values, as well as to the key success factors of the wider socio-technical system*" (1.2.3.6, p.14)
	SR_23	THEMIS should make available key information on users' preferences and socio-technical contexts.	"*• Make available key information on user (anonymous user profiling) and socio-technical contexts (circumstances and objectives of the decision support);*"
	SR_24	THEMIS should (i) feature a GUI that will allow the users to create a qualitative model of the socio-technical environment by inserting (i.a) all the possible actions that the AI system can recommend, (i.b) all the KPIs that might be affected by any of the possible actions, (i.c) any external factors that might influence the actions or the actions' effect on the KPIs, and (i.d) how the actions and the external factors affect the KPIs through pairwise relations; THEMIS should be able to (ii) generate a qualitative model.	"*… the human user using the developed DI methodology and supporting graphics will create a qualitative model of the socio-technical environment. The model will contain all the possible actions that the AI system can recommend, all the KPIs that might be affected by any of the possible actions, any external factors that might influence the actions or the actions' effect on the KPIs, and how the actions and the external factors affect the KPIs through pairwise relations. (…) a DI AI simulation will be trained that generates a quantitative simulated model of the socio-technical system.*" (1.2.3.2, p.12)
	SR_56	THEMIS will enable the user to model an AI system in the context of a socio-technical environment in accordance with decision-support needs and moral values.	"*Themis_AI_TMF will innovate by enabling the user to model an AI system in the context of a socio-technical environment in accordance with decision support needs and moral values" (1.1.2, p. 8)*
	SR_25	THEMIS must provide explanations for the trustworthiness assessment which will be based on anomaly detection indicators related to (i) fairness, (ii) technical accuracy and robustness, and (iii) the embedding socio-technical environment.	"*THEMIS 5.0 empower human recipients of AI-based decision support with: Enhanced explainability over the trustworthiness of AI-based decision support based on anomaly detection indicators related to fairness, technical accuracy and robustness as well as to the embedding socio-technical environment*" (1.1.1, p.3)
ASSESS	SR_26	AI trustworthiness assessment should be based on at least 50 indicators and metrics.	"*KPI2.2: 50+ indicators and metrics for trustworthiness evaluation*" (1.1.2, p.4)

	SR_27	THEMIS should incorporate alternative methods and algorithmic approaches for the trustworthiness assessment.	*"… analyze the alternative algorithmic approaches and metrics for evaluating the solutions …"* (1.1.3, p.7)
	SR_28	THEMIS trustworthiness assessment will be dynamic and will be updated based on the implementation (or the projection) of trustworthiness optimisation measures.	*"THEMIS 5.0 trustworthiness evaluation methodology, (…), will adopt a dynamic risk management approach following the lifecycle of managing socio-technical threats …"* (1.2.2, p.10)
	SR_29	THEMIS fairness assessment will be (i) implemented as an AI-driven anomaly detection system, whose (ii) fairness bias indicators will be defined according to NIST, including (ii.a) systemic bias, (ii.b) computational bias and (ii.c) human bias.	*"THEMIS 5.0 will base its assessment of fairness bias indicators as defined by NIST including systemic bias (in the use of AI), computational bias (in algorithms, training datasets), and human bias (in system use). The THEMIS 5.0 'AI based Assessment of Fairness' module will compute metrics for these 3 major categories of AI bias and it will be implemented on the principles of an AI-driven anomaly detection system."* (1.2.3.1, p.11)
	SR_30	THEMIS technical accuracy & robustness assessment will be (i) implemented as an AI-driven anomaly detection system, whose (ii) metrics will be able to adapt to the severity of risks on the business environment as well as on the users' needs and preferences.	*"The assessment of technical accuracy & robustness in the 'AI-based Assessment of Technical Accuracy & Robustness' module will be implemented on the principles of an AI-driven anomaly detection system. (…) needs to take into account the degree of potential severity or harm of potential failures. (…) the required levels of robustness and accuracy needs to be matched to the actual characteristic of the hybrid decision systems."* (1.2.3, p.11-p.12)
	SR_31	THEMIS will implement trustworthiness assessment based on a risk assessment approach.	*"Our approach to trustworthiness assessment is based around risk assessment."* (1.2.3.5, p.13)
	SR_32	THEMIS should feature AI / ML components to monitor the operational system under test, that based on an anomaly detection approach will be able to assess the AI system's accuracy and fairness their impact on the business-related risks and KPIs.	*"AI / ML components monitor the operational system under test and are trained to recognize situations of raised unfairness/ reduced accuracy and the impact of decisions, based on an anomaly detection approach."* (1.2.3.5, p.13)
	SR_33	THEMIS will provide access to datasets for experimentation, Open APIs for experiment development, ML/DL algorithms models, advanced analytics, visualization configurations, human/machine interactions schemas, as well as the legal framework for trusted AI.	*"This environment will provide access to datasets for experimentation, Open APIs for experiment development, ML/DL algorithms models, advanced analytics, visualization configurations, human/machine interactions schemas, as well as the legal framework for trusted AI."* (1.2.4, p.15)
	SR_34	THEMIS will feature a fully equipped sandbox to design and deploy experiment variants and evaluate them against defined sets of criteria.	*"… a fully equipped sandbox (…) to design and deploy experiment variants and evaluate them against defined sets of criteria."* (1.2.4, p.15)
	SR_35	THEMIS should ensure human involvement in the AI-based decision support.	*"… enabling decision makers to benefit from AI decision support while maintaining human involvement needed for trustworthy decision-making."* (1.1.3, p.7)
	SR_36	THEMIS should support enhanced transparency on the AI decision support outcomes.	*"… enabling transparent communication of the decision process (…) to stakeholders of the hybrid decision system, …"* (1.1.3, p.7)
EXPLORE	SR_37	THEMIS should implement negotiation strategies for the selection between multiple potential solutions, resulting from the different actors and objectives there may be.	*"… real-world problems involve both multiple actors and multiple objectives that should be considered when making a decision (…) how to develop negotiation strategies for selecting between multiple potential solutions, to learning about the behavior or objective preferences of other agents."* (1.1.3, p.7)
	SR_38	THEMIS will be able to inform the user beforehand for the impact of the optimisation measures on the established business related KPIs.	*"… when a user gets a recommendation, they will be able to see the projected effects of the recommended decision on the different KPIs. (…) simulate the effect of a decision on KPIs can be used before a new system, or a change in an existing system, …"* (1.1.3, p.8)
	SR_39	Taking into account the users' profile and preferences, THEMIS will (i) forecast risks and vulnerabilities and will (ii) propose mitigation measures.	*"… proposes a Trust Management Framework for AI (…) that will forecast, assess and mitigate socio-technical AI- threats considering the profiles and preferences (culture, behavior, psychology, societal) of the AI- users (individuals, communities, enterprises)."* (1.1.3, p.8)
	SR_40	THEMIS AI-driven conversational interface should support decision improvement for the users and trustworthiness optimisation.	*"… leverage trustworthiness optimization and, hence, decision improvement through AI-driven dialogues, …"* (1.2.3.6, p.14)
	SR_41	THEMIS conversational interface should provide understandable and human-interpretable explanations on AI system functionality.	*"The agent will provide sufficient but not excessive human-interpretable explanations on how the AI system took a particular set of inputs and supported a particular decision."* (1.2.3.6, p.14)
	SR_42	THEMIS conversational interface should provide users with (i) input criteria used in the decision	*"The provided human-interpretable explanations will inform human recipients about: the input criteria used in the decision support process*

		support process (training data sets, AI models); (ii) the output of that process; and (iii) the perceived causal relationship between input and output, taking into consideration (iii.a) the characteristics of the human recipients of decision support, (iii.b) the context and circumstances that triggered the decision support, and (iii.c) the user defined objectives.	*(training data sets, AI models); the output of that process; and the perceived causal relationship between input and output, taking into consideration the characteristics of the human recipients of decision support, the context and circumstances triggered the decision support, and the objectives pursued thereby.*" (1.2.3.6, p.14)
	SR_43	Trustworthiness optimisation measures should be presented to the user through the THEMIS conversational agent based on multi-actor and multi-objective interaction methods.	"*… implement the multi-actor, multi-objective interaction methods supporting 'Personalized Interactive Dialogues' acting upon "Generation of Improvements" …*" (1.2.3.6, p.14)
	SR_44	THEMIS-enabled cyclic improvement of AI systems will enrich the THEMIS knowledge base with trustworthiness-enabling interactions with human users.	"*… cyclic deployments of the AI system in the form of experiments build up the knowledge base through intelligent elicitation of trustworthiness-enabling interactions with human users, reducing the complexity associated with manual tuning of AI models and algorithms.*" (1.2.4, p.14)
	SR_46	THEMIS should be able to use feedback from users on the produced results to fine-tune the implemented AI models.	"*In interactive reinforcement learning, an agent learns e.g. from human evaluative feedback, i.e., evaluations of the quality of the agent's behaviour provided by a human user, or advice/instruction*." (1.1.3, p.7)
ENHANCE	SR_47	THEMIS should (i) receive feedback from the users with respect to their trust in the DI simulation results and (ii) based on this input THEMIS should update either the qualitative or the quantitative model of the socio-technical environment.	"*During the iterative trustworthiness assessment by the user, they may also need to evaluate their trust in the DI simulation results. Feedback from the user could highlight problems in the DI model, or in the DI simulation. In turn, this feedback will be acted on by either correcting the socio-technical environment model, or the generated AI based DI simulation.*" (1.2.3.3, p.12)
	SR_48	THEMIS users should be able to choose when and if a trustworthiness optimisation measure will be applied to the AI system under test.	"*The human operator may optionally apply those controls to the real system, which should reduce the occurrence of raised unfairness or reduced accuracy.*" (1.2.3.5, p.13)
	SR_49	A ModelOps approach should be followed for the operationalisation of the produced experimental variables.	"*• Support the operationalizing of all experiment variants according to a ModelOps approach.*" (1.2.4, p.15)
	SR_50	THEMIS will monitor the AI systems' trustworthiness optimisation at each cycle by calculating quantitative and qualitative KPIs, benchmarking and measures on progress monitoring based on a defined set of criteria on the various process assets (AI models, algorithms, datasets, as well as specific configurations).	"*At each cycle of experiment instance deployment, the solution will compute a set of criteria on the various process assets (AI models, algorithms, datasets, as well as specific configurations) and will produce quantitative and qualitative KPIs, benchmarking and measures on progress monitoring.*" (1.2.4, p.15)
	SR_51	THEMIS will allow the design of improved experiments (versions of AI systems) through a no-code Big Data Analytics as a Service cloud-based designer (Interface).	"*… this environment will allow the design of improved experiments through a no-code Big Data Analytics as a Service cloud-based designer.*" (1.2.4, p.15)
	SR_52	THEMIS should improve the trustworthiness and explainability of results of an AI system for (cancer) risk identification and assessment to be integrated in healthcare systems.	"*… how best an AI system used to analysis "omics" measures can be more explainable and trustworthy so its outputs can be trusted and integrated in already established healthcare systems? … the AI system for (cancer) risk identification and assessment needs to be optimised in terms of becoming more explainable and trustworthy, …*" (1.2.6, p.17)
	SR_53	THEMIS should enhance the results' explainability of the AI system for the ETA calculation.	"*… to enhance the current system by providing explanations for predictions at a comprehensible level for all users. This involves elucidating key input variables that significantly impact prediction outcomes.*" (1.2.6, p.17)
	SR_54	THEMIS should ensure that the AI system for the ETA calculation complies to all ethical considerations (e.g. fair and non-discriminatory) on decision-making.	"*THEMIS 5.0 will consider ethical considerations in decision-making, such as potential discrimination against specific shipping companies.*" (1.2.6, p.18)
	SR_55	THEMIS should improve the robustness and fairness of the AI the AI system for the ETA calculation	"*THEMIS 5.0 will ensure the robustness and fairness of the AI model.*" (1.2.6, p.18)

A.3. User Requirements

Table 24. User requirements translated from user stories collected from THEMIS 5.0 use case partners and technical counterparts.

THEMIS 5.0 Pipeline stage	ID	User Requirement	Role	Use case	User Story ID
Assess	UR_S1	The trustworthiness assessment results should be presented in a user-friendly manner.	Port Traffic Control, Journalist, Commercial Director	Media Port	US3_01, US2_ T01
	UR_S2	The AI system's trustworthiness assessment needs to be presented to the users via a chatbot based interface, that is initiated by them.	Transportation Traffic Manager	Port	US2_Q09
	UR_S3	The discussion with the chatbot could be based on a set of predefined questions to understand the level of detail that the user wants to have about the Trustworthiness Assessment.	Transportation Traffic Manager	Port	US2_Q10
	UR_S4	The trustworthiness assessment results should be presented at once (chatbot or document) in the form of a high-level report with most important results.	Port Authority Manager-Port Community System	Port	US2_Q01
	UR_S5	More details on specific results from the Trustworthiness Assessment overview should be provided on demand (chatbot or hyperlink in the document)	Port Authority Manager-Port Community System	Port	US2_Q02
	UR_S6	The accuracy measurement should be based on the AI system's previous predictions vs observations. The comparison period and conditions should be customisable by the user.	Transportation Traffic Manager	Port	US2_Q06 US2_Q07 US2_Q08
	UR_S7	The user should be presented with which parts of the SUT suffer (e.g., model, deployment, datasets)	Journalist, Commerical Director, AI System developer	Media	US3_02
	UR_S8	The user should be presented with analytical results about bias: which are the favoured and the discriminated groups.	Port Traffic Control	Port	US2_ T08
	UR_S9	The user should be presented with the related risks stemming from trustworthiness vulnerabilities.	Pilot, Port Terminal	Port	US2_ T02, US2_ T03
	UR_S10	The risk assessment results should be presented at once (chatbot or document) in the form of a high-level report with most important results.	Port Authority Manager-Port Community System	Port	US2_Q03
	UR_S11	More details on specific results from the Risk Assessment overview should be provided on demand (chatbot or hyperlink in the document)	Port Authority Manager-Port Community System	Port	US2_Q04
	UR_S12	The user should be able to provide scenarios and check the accuracy of the predictions of SUT for these scenarios.	Port Traffic Control / Port Terminal / Pilot / Towing Services /Mooring Services / Transport Company	Port	US2_ T07
	UR_S13	The user should be able to access previous trustworthiness assessments	AI System developer	Port	US2_ T08
Explore	UR_S14	The user should be informed about positive and negative business impact impacts for each enhancement measure	Journalist, Commercial Director	Media	US3_03
	UR_S15	The user should be able to choose between the provided solutions.	Journalist, Commercial Director	Media	US3_004
	UR_S16	The user should be informed about positive and negative impacts in SUT's trustworthiness for each enhancement measure.	AI developer	Media	US3_07
	UR_S17	Testbed environment should be provided to allow the application of the provided solutions in a pipeline.	AI developer	Media	US3_08
Enhance	UR_S18	The accuracy of the enhanced SUT should be as high as possible.	Port Traffic Control / Port Terminal / Pilot / Towing Services /Mooring Services / Transport Company	Port	US2_ T09
	UR_S19	The new trustworthiness results should be communicated to the user	Journalist, Commerical Director, AI System developer	Media	US3_ T05

Table 25. User Requirements from the co-creation workshop with AI-users (1st co-creation workshop).

Sector	ID	Attention Points	User Requirement	Scope	THEMIS 5.0 Pipeline stage
Healthcare	UR_W1	***Responsibility:*** *AI developers of healthcare related AI tools should consider the way their tools deliver output to the healthcare professional to ensure that it is functioning as a support tool for human decision rather than the tool taking decision for the user.*	AI systems for the Healthcare sector should not function independently and should work as decision support tools assisting users to reach a decision.	THEMIS + SUT	EXPLORE
	UR_W2	***Responsibility:*** *AI-generated recommendations should be transparent, ensuring that healthcare professionals understand the basis and limitations of AI suggestions*	The user should be aware of how an AI system has produced a recommendation.	THEMIS + SUT	EXPLORE
	UR_W3	***Responsibility:*** *AI developers of healthcare related AI tools should develop their tools with a patient-centric approach*	AI systems for the Healthcare sector should be developed based on a patient-centric approach.	THEMIS + SUT	OVERALL
	UR_W4	***Transparency & Accuracy:*** *AI developers of healthcare related AI tools should communicate clearly about the data used to train their AI tools and use evidence-based data*	The users of AI systems for the Healthcare sector need to be informed about the quality of the training datasets that have been used.	SUT	-
	UR_W5	***Transparency & Accuracy:*** *AI developers of healthcare related AI tools should investigate potential biases in their training data*	Potential bias in the training datasets used for healthcare related AI tools should be investigated.	SUT	-
	UR_W6	***Transparency & Robustness:*** *AI developers of healthcare related AI tools should clearly define the limitations of their tools, and enable ongoing improvements of AI tools to address contextual and evolving healthcare challenges*	Healthcare related AI tools' limitations should be clearly communicated to the user.	THEMIS + SUT	EXPLORE
	UR_W7		AI tools need to be adaptable to address contextual and evolving healthcare challenges.	THEMIS + SUT	OVERALL
	UR_W8	***Transparency & Accuracy:*** *AI developers of healthcare related AI tools should find a succinct way to present transparency due to limited time of healthcare professionals*	The AI system should ensure explainability and transparency of results while respecting the users' time constraints. Explanations and details should be brief and to the point.	THEMIS + SUT	EXPLORE
	UR_W9	***Cost vs. Efficiency:*** *AI developers of healthcare related AI tools should consider the cost vs. efficiency of their tools*	The development of AI systems for the Healthcare sector should consider cost vs. efficiency.	THEMIS	EXPLORE
	UR_W10	***Cost vs. Efficiency:*** *AI developers of healthcare related AI tools should consider and minimise the climate impact of their tools*	The development of AI systems for the Healthcare sector should minimise climate impact.	THEMIS + SUT	OVERALL
	UR_W11	***Cost vs. Efficiency:*** *AI developers of healthcare related AI tools should align the development of AI tools with what the healthcare sectors and professionals' need*	The development of AI systems for the Healthcare sector should be based on the needs of healthcare sector and professionals.	THEMIS + SUT	OVERALL
	UR_W12	***Attitudes towards AI:*** *AI developers of healthcare related AI tools should ensure that AI tools are beneficial for healthcare professionals and do not become more time consuming than not using AI*	AI systems for the Healthcare sector should facilitate and optimise the work of healthcare professionals.	THEMIS + SUT	OVERALL
	UR_W13	***Attitudes towards AI:*** *AI developers of healthcare related AI tools should ensure that the use of AI tools does not negatively impact healthcare professionals' capabilities and the possibility to develop competencies*	AI systems for the Healthcare sector should not hamper healthcare professionals from acquiring and developing competences.	SUT	-
Port	UR_W14	***Responsibility:*** *AI developers of port management related AI tools should consider the way their tools deliver output to the port management professional to ensure it is functioning as a support tool for human decision making rather than the tool taking decisions for the user.*	AI systems for the port management sector should not function independently and should work as decision support tools assisting users to reach a decision.	THEMIS + SUT	EXPLORE
	UR_W15	***Responsibility:*** *AI developers of port management related AI tools should foster a collaborative relationship between their tools and its users.*	AI systems for the port management sector should foster a collaborative way of working, so that users are involved in the AI-supported decision-making process.	THEMIS + SUT	OVERALL

	UR_W16	***Responsibility:*** *AI developers of port management related AI tools should be transparent about who has the responsibility of AI tools contextual output, considering existing laws and limitations of the law.*	AI systems for the port management sector should be able to communicate the roles that are accountable for the AI tool's output according to the legal and regulatory framework.	SUT	-
	UR_W17	***Transparency & Accuracy:*** *AI developers of port management related AI tools should consider how the AI tools will provide explanations for their outputs to foster trust*	The AI system should ensure explainability and transparency of results to foster trust amongst users.	THEMIS + SUT	EXPLORE
	UR_W18	***Transparency & Accuracy:*** *AI developers of port management related AI tools should consider how to strike a balance between the need for objectivity and subjectivity in port management*	AI systems for the port management sector should be developed balancing the need for objectivity and subjectivity in the decision-making process.	SUT	-
	UR_W19	***Contextual Nuances:*** *AI developers of port management related AI tools should ensure constant availability of their tools*	AI systems for the port management sector should be constantly available.	SUT	-
	UR_W20	***Contextual Nuances:*** *AI developers of port management related AI tools should be able to take into account the complexity and different protocols and politics of individual ports*	AI systems for the port management sector should be able to model the structure, protocols and policies of each port.	THEMIS	IDENTIFY
	UR_W21	***Contextual Nuances:*** *AI developers of port management related AI tools should ensure that their tools benefit the many, not the few*	AI systems for the port management sector should ensure that all stakeholders are equally (or proportionally) benefited.	THEMIS + SUT	EXPLORE
	UR_W22	***Contextual Nuances:*** *AI developers of port management related AI tools should consider how commercial interests can affect how ports function*	AI systems for the port management sector should model how commercial interests interact with the port's operation.	THEMIS	IDENTIFY
	UR_W23	***Attitudes towards AI:*** *AI developers of port management related AI tools should consider how the work between the tool and the users becomes a collaborative effort to ensure users do not become too dependent on AI*	AI systems for the port management sector should foster a collaborative way of working, so that users are involved in the AI-supported decision-making process, as well as ensuring that users can still operate without the use of AI.	THEMIS + SUT	OVERALL
	UR_W24	***Attitudes towards AI:*** *AI developers of port management related AI tools should take into consideration how their tools can interfere with the current cultural norms and established hierarchy in ports.*	AI systems for the port management sector should be able to adapt to the cultural norms and established hierarchy of each port.	THEMIS	IDENTIFY
	UR_W25	***Attitudes towards AI:*** *AI developers of port management related AI tools should consider ways to encourage the AI uptake from end-users not familiar with AI technologies.*	AI systems for the port management sector should facilitate users with limited or no experience of AI tools to use them in their line of work.	THEMIS	OVERALL
Media	UR_W26	***Responsibility:*** *AI developers of media related AI tools should consider the way their tools deliver output to the journalists and fact-checkers to ensure that it is functioning as a support tool for human decision making rather than the tool taking its own decisions.*	AI systems for the media sector should not function independently and should work as decision support tools assisting users to reach a decision.	THEMIS + SUT	EXPLORE
	UR_W27	***Responsibility:*** *AI developers of media related AI tools should be transparent about who is involved in the development of AI tools and what data is being used.*	The users of AI systems for the media sector need to be informed about the parties that have been involved in their development and the training datasets that have been used.	THEMIS + SUT	OVERALL
	UR_W28	***Responsibility:*** *AI developers of media related AI tools should be clear about what the intended purpose and use of their AI tool is.*	The user should be aware about the purpose and use of the AI system.	THEMIS + SUT	OVERALL
	UR_W29	***Transparency & Accuracy:*** *AI developers of media related AI tools should make it possible for users to examine the 'line of thought' going on in the AI tool to ensure trust.*	The AI system should ensure explainability and transparency of results, informing the user about how a recommendation has been produced, to foster trust amongst users.	THEMIS + SUT	EXPLORE
	UR_W30	***Transparency & Accuracy:*** *AI developers of media related AI tools should understand how transparency, accuracy and trust are closely connected in the work of journalists and fact-checkers.*	AI systems for the media sector should emphasise accuracy and transparency of results to foster trust.	SUT	-

	UR_W31	***Transparency & Accuracy:*** *AI developers of media related AI tools should inform the end-user about the data used for training their AI tools.*	The users of AI systems for the media sector need to be informed about the training datasets that have been used.	SUT	-
	UR_W32	***Contextual Nuances:*** *AI developers of media related AI tools should consider that most of work in media in heavily dependent on the context and real-time state of the world.*	AI systems for the media sector should be able to adapt their results to the everchanging context and state of the world.	THEMIS + SUT	IDENTIFY
	UR_W33	***Contextual Nuances:*** *AI developers of media related AI tools should make it possible to still adhere to and support working with industry ethical standards and norms.*	AI systems for the media sector should adhere and promote the industry's ethical standards and norms.	THEMIS + SUT	OVERALL
	UR_W34	***Contextual Nuances:*** *AI developers of media related AI tools should not make AI tools that are subjective and make moral judgements but rather focus on tasks were objectivity is the goal.*	AI systems for the media sector (i) should focus on tasks where objectivity is the goal and (ii) should refrain from making moral judgements and taking subjective decisions.	SUT	-
	UR_W35	***Attitudes towards AI:*** *AI developers of media related AI tools should make their AI tools supportive in their functionality for journalists and fact-checkers.*	AI systems for the media sector should facilitate and optimise the work of journalists and fact-checkers.	THEMIS + SUT	OVERALL
	UR_W36	***Attitudes towards AI:*** *AI developers of media related AI tools should consider how to balance time optimisation by enabling use of AI tools without impacting the quality of work.*	AI systems for the media sector should support time-optimisation while not sacrificing the quality of work.	SUT	-

Table 26. User Requirements from the co-creation workshop with AI-users and business stakeholders.

THEMIS 5.0 Pipeline stage	ID	User Requirement	Role	Use case	User Story ID
Identify	UR_C1	The interaction with the chatbot should be text-based	Healthcare professionals (practitioners and administrative)	Health	US1_C12
	UR_C2	The chatbot should ask the same questions in different ways (with paraphrases)	Healthcare professionals (practitioners and administrative)	Health	US1_C9
	UR_C3	The chatbot must ask about end-user's role in their organization	Media professional	Media	US3_C1, US3_C9
	UR_C4	The chatbot should provide multiple-choice option for the various roles of the end-user (to eliminate vagueness)	Healthcare professionals (practitioners and administrative)	Health	US1_C13
	UR_C5	The chatbot must ask about end-user's years of experience	Healthcare professionals (practitioners and administrative)	Health	US2_C3
	UR_C6	The chatbot must ask about end-user's professional interests	Healthcare professionals (practitioners and administrative)	Health	US2_C5
	UR_C7	The chatbot must ask about organizational ethical guidelines that the end-user must comply with.	Media professional	Media	US3_C2
	UR_C8	The chatbot must ask about organizational security measures.	Media professional	Media	US3_C8
	UR_C9	The chatbot must ask about end-user's requirements with regards to trustworthiness characteristics	Media AI developer/ Port Sector AI developers	Media Port	US3_C3 US2_C5
	UR_C10	The chatbot must ask about end-user's climate concerns	Administrative/ decision makers in the media sector	Media	US3_C4
	UR_C11	The chatbot must ask where in the work process is the end-user using the SUT	Media professional	Media	US3_C5
	UR_C12	The chatbot must ask the purpose of use of the SUT	Healthcare professionals (practitioners and administrative)	Health	US1_C4
	UR_C13	The chatbot must ask about the AI- familiarity level of the end-user	Media professional	Media	US3_C6

	UR_C14	The chatbot should first make the 20 most crucial questions for a quick start, and later refine if necessary	Healthcare professionals (practitioners and administrative)	Health	US1_C10
	UR_C15	The interaction with the chatbot should last max 15’-30'	Media professional	Media	US3_C10
	UR_C16	The interaction with the chatbot should last max 30’-60’	Healthcare professionals (practitioners and administrative)	Health	US1_C8
	UR_C17	The questions about end-user's persona should run only once, in the initialization of the platform	Healthcare professionals (practitioners and administrative)	Health	US1_C7
	UR_C18	Add functionality for update of the persona stored in THEMIS 5.0 platform (e.g., change of role)	Healthcare professionals (practitioners and administrative)	Health	US1_C14
	UR_C19	Add functionality for creating multiple personas (e.g., different roles)	Healthcare professionals (practitioners and administrative)	Health	US1_C15
	UR_C20	Trustworthiness Preferences personas should be the same for all user types	Port sector employees	Port	US2_C1
	UR_C21	The chatbot should ask about the different roles that will use the AI tool	Port Ground Employees	Port	US2_C3
	UR_C22	The chatbot must ask about the user’s attitude towards AI	Port Sector AI developers	Port	US2_C4
	UR_C23	The chatbot must ask about organisation’s rules and established procedures	Port Sector AI developers	Port	US2_C6
	UR_C24	The chatbot must ask about organisation’s security related information	Port Sector AI developers	Port	US2_C7
	UR_C25	Trustworthiness assessment must take into account end-user's trustworthiness requirements.	Media professional	Media	US3_C3
	UR_C26	Fairness assessment should consider the purpose of use of the SUT	Healthcare professionals (practitioners and administrative)	Health	US1_C4
	UR_C27	THEMIS 5.0 should infer the prioritisation of trustworthiness characteristics from where in the work process is the SUT used	Media professional	Media	US3_C5
	UR_C28	THEMIS 5.0 should provide a printable trustworthiness assessment report in a user-friendly way.	Media professional	Media	US3_C15
	UR_C29	Enhancement suggestions must be accompanied by details to help the end-user to choose.	Healthcare professionals (practitioners and administrative)	Health	US1_C18
	UR_C30	Enhancement suggestions must be accompanied by the respective climate impact rating	Administrative/ decision makers in the media sector	Media	US3_C4
	UR_C31	Enhancement suggestions must be accompanied by explanations confirming that ethical/legal/security/climate organization’s restrictions are respected.	Media professional	Media	US3_C13
Explore	UR_C32	Enhancement suggestions should be based on end-user's years of experience	Healthcare professionals (practitioners and administrative)	Health	US1_C3
	UR_C33	Decision-making on enhancement should take into consideration end-user's professional interests	Healthcare professionals (practitioners and administrative)	Health	US1_C5
	UR_C34	Enhancement suggestions should be based on all different types of roles that are using the SUT within the organization	Port Ground Employees	Port	US2_C3
	UR_C35	Enhancement suggestions should be presented in a rating system	Healthcare professionals (practitioners and administrative)	Health	US1_C16

	UR_C36	Suggestions to include "Just In Time" concept compliant optimisations	Port sector employees	Port	US2_C8
	UR_C37	Accessibility to, and usability of, the various THEMIS 5.0 functionalities should be based on end-user's role within their organization and on organizational security/ethical guidelines	Media professional	Media	US3_C1, US3_C9
Enhance		-	-	-	-
Overall	UR_C38	Different user profiles should be created based on end-user's familiarity level with AI	Media professional	Media	US3_C6
	UR_C39	THEMIS 5.0 should facilitate different types of interaction based on the various AI-familiarity-levels	Media professional	Media	US3_C6
	UR_C40	Interaction with end-user should be made in a structured way and not on free dialogues.	Media administrative/decision makers		US3_C12
	UR_C41	THEMIS 5.0 functionalities must be in line with the organizational ethical guidelines provided by the end-user.	Media professional	Media	US3_C2
	UR_C42	THEMIS 5.0 functionalities must be in line with the organizational security guidelines provided by the end-user.	Media professional	Media	US3_C8
	UR_C43	THEMIS 5.0 platform installation requirements must comply with standard laptops, computers and smartphones	Healthcare professionals (practitioners and administrative), Media professional	Health, Media	US3_C11, US1_C11
	UR_C44	THEMIS 5.0 interactions should be based on user's attitude towards AI	Healthcare professionals (practitioners and administrative)	Health	US1_C1

Table 27. User Requirements collected from Questionnaires distributed to use case partners and technical counterparts.

THEMIS 5.0 Pipeline stage	ID	User Requirement	Role	Use case	Question ID
Overall	UR_Q1	The user must be provided with a user-manual including explanation of the various functionalities and the terminology used by THEMIS 5.0	Business Development	Media	3.2.2
Identify	UR_Q2	Display the prediction of trustworthiness preferences to the user	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), Port Authority Manager, Vessel Traffic Service Supervisor, Transportation Traffic Manager Fact-checkers	Health, Port, Media	3.1.1
	UR_Q3	Users should be able to give feedback on the displayed predictions of preferences	Technical Lead, AI R&D (UC1), Business Development (UC3), Fact checkers	Health, Media	3.1.1
	UR_Q4	User preferences should also consider the users' need for sufficient explanation and transparency around the AI system's results.	Transportation Traffic Manager, Business Development (UC3)	Port	3.1.1
	UR_Q5	Calculate and display explanation about the prediction of trustworthiness preferences.	Technical Lead, AI R&D (UC1), Vessel Traffic Service Supervisor, Port Authority AI developer (UC2), Technical Lead (UC3)	Health, Port, Media	3.1.2

	UR_Q6	The user should be able to adjust the estimated trustworthiness preferences.	Port Authority Manager, Vessel Traffic Service Supervisor, Port Authority AI developer, Transportation Traffic Manager	Port, Media	3.1.3
	UR_Q7	User should be able to copy the preferences of colleagues.	Technical Lead (UC1)	Health	3.1.3
	UR_Q8	User-specific information should be stored and be accessible by end-users so they can review their historical trustworthiness preferences.	Health care professional, general practitioner, internal medicine specialist, Port Authority Manager, Vessel Traffic Service Supervisor, Port Authority AI developer, Transportation Traffic Manager Fact-checkers	Port, Media	3.1.4
	UR_Q9	Accessibility to historical results must be based on role	Business Development (UC3)	Media	3.1.4
	UR_Q10	THEMIS should have different UIs for different AI-familiarity levels/roles.	Fact-checkers	Media	3.1_S1
	UR_Q11	User should provide familiarity level	Fact-checkers	Media	3.1_S2
	UR_Q12	User should be provided with definitions of terminology upon request	Fact-checkers	Media	3.1_S3
	UR_Q13	THEMIS 5.0 platform should ask me my tolerance on two different cases: a) false positive, b) false negative	Fact-checkers	Media	3.1_S4
	UR_Q14	User preferences within the organisation are not expected to differ significantly	Port Authority Manager, Vessel Traffic Service Supervisor, Port Authority AI developer, Transportation Traffic Manager	Port	3.1.5
	UR_Q16	The identify stage should not last more than 5'	Fact checkers	Media	3.1.6
	UR_Q17	The identify stage should not last more than 15'	Business Development	Media	3.1.6
	UR_Q18	The identify stage should not last more than a couple of weeks	Port Authority Manager	Port	3.1.6
	UR_Q19	The identify stage should not last more than **one day**	Technical Lead, AI R&D (UC1)	Health	3.1.6
	UR_Q20	The identify stage should be fast, **less than a minute**	Vessel Traffic Service Supervisor	Port	3.1.6
	UR_Q21	The identify stage should not last more than an hour	Transportation Traffic Manager	Port	3.1.6
Assess	UR_Q22	Trustworthiness assessment should be displayed to the end-user	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), Technical Lead, (UC2) Port Authority Manager, Business Development (UC3), Fact checkers	Health, Port, Media	3.2.1
	UR_Q23	The presentation of the trustworthiness assessment should be based on end-user's familiarity with AI	Fact checkers	Media	3.2.2
	UR_Q24	Ethical assessment results should be presented in a high-level but crystal-clear way	Port Authority Manager	Port	3.2.2
	UR_Q25	Trustworthiness assessment results could be displayed in a data-visualisation like dashboard (e.g. PowerBI) with descriptions	Technical Lead, AI R&D (UC1), Port Authority Manager	Health, Port	3.2.2
	UR_Q26	Trustworthiness assessment results should be presented as a percentage.	Technical Lead,	Health, Port	3.2.2

			AI R&D (UC1), Vessel Traffic Service Supervisor, Port Authority AI developer,		
	UR_Q27	Trustworthiness assessment should be displayed in context-specific way (General error track record, Correlation of error with conditions and events, Training datasets details, e.g., 20% of the patients may be wrongly identified as of high risk)	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), Transportation Traffic Manager, Business Development (UC3)	Health, Port, Media	3.2.2
	UR_Q28	Users should be informed about which parts of the AI System suffer the most (training data/model/test data/deployment etc.)	Technical Lead, AI R&D (UC1), Port Authority Manager, Port Authority AI developer, Fact-checker	Health, Port, Media	3.2.3
	UR_Q29	Only in a form of a high-level report should users be informed about which parts of the AI System suffer the most	Business Development (UC3)	Media	3.2.3
	UR_Q30	Trustworthiness assessment should be accompanied with high-level and understandable explanations.	Port Authority Manager	Port	3.2.4
	UR_Q31	Trustworthiness assessment should be accompanied with explanations.	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), Port Authority AI developer, Transportation Traffic Manager, Fact Checkers	Health, Port, Media	3.2.4
	UR_Q32	The system's current trustworthiness assessment could be displayed against previous trustworthiness assessments.	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), , Port Authority Manager, Port Authority AI developer	Health, Port	3.2.5
	UR_Q33	User should have access to previous trustworthiness assessments along with the corresponding state of the AI system	Fact-checkers	Media	3.2.5
	UR_Q34	The system could be able to comparatively analyse multiple AI systems and services for the same task (selected by the end-user).	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), Port Authority Manager, Port Authority AI developer, Business Development (UC2), Fact checkers	Health, Port, Media	3.2.6
	UR_Q35	Detailed explanation for the assessment of selected trustworthiness characteristics should be offered on demand	Port Authority Manager, Port Authority AI developer, Transportation Traffic Manager	Port	3.2.7
	UR_Q36	The user should be able to choose which trustworthiness characteristics would like to be assessed	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), Fact checkers, Business Development (UC3)	Health, Media	3.2.7
	UR_Q37	The factors that have mostly influenced the trustworthiness assessment should be presented	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), Port Authority Manager, Port Authority AI developer, Transportation Traffic Manager, Fact checkers, Business Development (UC3)	Health, Port, Media	3.2.8

	UR_Q38	The Trustworthiness Assessment should run automatically every time the SUT is updated	Fact checkers	Media	3.2.9
	UR_Q39	The Trustworthiness Assessment should not last more than a week	Technical Lead, AI R&D (UC1)	Health	3.2.9
	UR_Q39	The Trustworthiness Assessment should not last more than a couple of weeks	Port Authority Manager	Port	3.2.9
	UR_Q40	The Trustworthiness Assessment should not last more than an hour	Vessel Traffic Service Supervisor	Port	3.2.9
	UR_Q41	The Trustworthiness Assessment should not last more than a day	Port Authority AI developer	Port	3.2.9
	UR_Q42	The Trustworthiness Assessment should not last more than 10h	Fact checkers	Media	3.2.9
	UR_Q43	User should be able to perform tests for different scenarios (e.g., articles when trained only in posts)	Fact checkers	Media	3.2_ US1
	UR_Q44	User should be able to perform tests on own datasets (e.g., sets of "hard" articles)	Fact checkers	Media	3.2_ US2
	UR_Q45	The system should be able to display how the changes made to the system affect the risk assessment against previous assessments.	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), Port Authority Manager, Port Authority AI developer, Business Development (UC3), Fact-checker	Health Port, Media	3.3.1
	UR_Q46	The system could have a BI tool to analyse risk assessment results	Technical Lead, AI R&D (UC1)	Health	3.3.1
	UR_Q47	Explanations on how the presented risks have been calculated could be offered on demand.	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), Port Authority Manager, Port Authority AI developer, Transportation Traffic Manager, Business Development (UC3), Fact-checker	Health Port, Media	3.3.2
	UR_Q48	User must be informed about the factors have mostly influenced the risk assessment	Health care professional, general practitioner, internal medicine specialist, Technical Lead, AI R&D (UC1), Business Development (UC3), Fact-checker	Health, Media	3.3.3
	UR_Q49	The Trustworthiness Assessment should not last more than a month	Technical Lead, AI R&D (UC1)	Health	3.3.4
	UR_Q50	The Risk Assessment should not last more than a couple of weeks	Port Authority Manager, Vessel Traffic Service Supervisor	Port	3.3.4
	UR_Q51	The Risk Assessment should not last more than a day	Port Authority AI developer	Port	3.3.4
	UR_Q52	The Risk Assessment should not last more than a week	Port Authority AI developer	Port	3.3.4
	UR_Q53	The Risk Assessment should not last more than few seconds	Port Authority AI developer	Port	3.3.4
Explore	UR_Q54	Detailed information about the solutions to mitigate the vulnerabilities and the business impacts of the solutions.	Port Authority AI developer	Port	3.4.1
	UR_Q55	The solutions should be provided in tabular form with the pros and cons	Technical Lead,AI R&D (UC1)	Health	3.4.1
	UR_Q56	The user should be provided with visual representation of the trade-offs (when	Fact-checkers	Media	3.4.1

		increasing fairness-accuracy drops, etc) for each suggestion			
	UR_Q57	Exploration of solutions should be accessible only by specific roles	Business Development	Media	3.4.1
	UR_Q58	The user should be provided with the required time, and cost for the implementation of each solution.	Business Development	Media	3.4.1
	UR_Q59	The user should be provided with the estimated trustworthiness measurements for each solution	Business Development	Media	3.4.1
	UR_Q60	The user should be provided with the estimated changes in the usability of the tool (e.g., certain functionalities become slower by 10%).	Business Development	Media	3.4.1
	UR_Q61	The user should be able to add new potential risks of trustworthiness vulnerabilities.	Technical Lead,AI R&D (UC1), Port Authority Manager, Port Authority AI developer, Transportation Traffic Manager, Fact-checkers	Health, Port, Media	3.5.1
	UR_Q62	Only specific roles should be able to add new potential risks of trustworthiness vulnerabilities.	Business Development	Media	3.5.1
	UR_Q63	The several solutions (e.g., change of model/data/deployment) and their related benefits and risk should always be presented, even if they go against the estimated user's trustworthiness preferences.	Health care professional, general practitioner, internal medicine specialist, Technical Lead,AI R&D (UC1), Port Authority Manager, Port Authority AI developer, Fact Checkers, Business Development	Health, Port, Media	3.5.2 3.5.5
	UR_Q64	The optimal solution should be directly implemented without asking the user	Health care professional, general practitioner, internal medicine specialist	Health	3.5.3
	UR_Q65	The user should be able to choose between allowing THEMIS 5.0 to perform the optimal solution automatically and allowing the user the one that prefers	Business Development	Media	3.5.3
	UR_Q66	The optimal solution should be annotated alongside with the factors that have mostly influenced this suggestion.	Health care professional, general practitioner, internal medicine specialist, Port Authority Manager, Port Authority AI developer, Fact checker	Port, Media	3.5.4
	UR_Q67	The factors that have mostly influenced the enhancement suggestion should be presented with a user-friendly report, depending on AI and organizational role	Business Development	Media	3.5.4
	UR_Q68	The explore stage should not last more than one week	Technical Lead,AI R&D (UC1)	Health	3.5.6
	UR_Q69	The explore stage should not last more than a couple of weeks	Port Authority Manager, Port Authority AI developer	Port	3.5.6
	UR_Q70	The explore stage should not last more than a month	Vessel Traffic Service Supervisor	Port	3.5.6
	UR_Q71	The explore stage should not last more than a few days	Transportation Traffic Manager	Port	3.5.6
Enhance	UR_Q72	An overview of the updated trustworthiness assessment of the AI system against previous values should be presented.	Health care professional, general practitioner, internal medicine specialist, Technical Lead,AI R&D (UC1), Port Authority Manager, Port Authority AI developer, Transportation Traffic Manager,	Health, Port, Media	3.6.1

		Fact checkers		
UR_Q73	The user should be provided with visual representation of trustworthiness parameters changes through time, given a predefined timeframe.	Fact checkers	Media	3.6.1
UR_Q74	The presentation should be in tabular form along with timestamps	Technical Lead,AI R&D (UC1),	Health	3.6.1
UR_Q75	Upon request, the user should be provided with analytical details on what has changed to the SUT (e.g., fine-tuned in dataset X)	Fact checkers	Media	3.6.1
UR_Q76	The sources used for generating the enhanced results should be presented.	Vessel Traffic Service Supervisor	Port	3.6.1
UR_Q77	The enhancement stage should not last more than 24-hour	Health care professional, general practitioner, internal medicine specialist	Health	3.6.2
UR_Q78	The enhancement stage should not last more than one week	Technical Lead,AI R&D (UC1),	Health	3.6.2
UR_Q79	The enhancement stage should not last more than a couple of months	Port Authority Manager, Port Authority AI developer	Port	3.6.2