LeArning and robuSt decision SupporT systems for agile mANufacTuring environments

Project Acronym:

ASSISTANT

Grant agreement no: 101000165

Deliverable no. and the title	**D2.3 - Management plan and ethics in/by design methodology**	
Work package	WP2	Ethic and human-centric toolbox
Task	T2.4	Value in and value by design methodology
Subtasks involved	T2.3, T2.4, and T2.5	
Lead contractor	Institut Mines-Telecom (IMT) Alexandre Dolgui, mailto: alexandre.dolgui@imt-atlantique.fr	
Deliverable responsible	University College Cork Eduardo Vyhmeister and Barry O' Sullivan	
Version number	V1.2	
Date	20/04/2022	
Status	Final	
Dissemination level	Public (PU)	

Authors

Participant no.	Part. short name	Author name	Chapter(s)
1	UCC	Eduardo Vyhmeister, Gabriel G. Castañé, Barry O'Sullivan	All

Document History

Version	Date	Author name	Reason
V1.0	31/01/2022	Eduardo Vyhmeister, Gabriel G. Castane	Preparation of D2.4
-	15/03/2022	Simon Thevenin	First Review of the Document
V1.1	31/03/2022	Eduardo Vyhmeister	Edited Version after First Review
-	06/04/2022	Christos Gkornelos	Second Review of the Document
V1.2	07/04/2022	Barry O'Sullivan	Comments after Second Review

Publishable Executive Summary

Incorporating ethics and values within the life cycle of an Artificial Intelligence (AI) component means securing the development, deployment, use, and decommissioning (i.e. life cycle) of it safely with considerations of the social perspectives and concerns that focus on a greater good over the agents and environment involved.

Even when ethics is essential to be incorporated in any human-based development, the AI assets possess specific characteristics that make them require further attention during their life cycle. Furthermore, ethics do not encompass all considerations involved in developing technically sound components that will secure the trust of their users and the market (i.e. values of the domain involved). Therefore, approaches that extend from the domain's ethical imperatives should be seen as a base to construct and develop any management or framework for AI.

As a differentiated approach, **Trustworthy AI** does not only imply the consideration of ethical use and development of AI Assets. Instead, it works as a general approach in which topics such as policies and regulations of AI are considered, including specifics related to regional considerations. Foremost among these are: agents' roles, societal issues - inclusion, diversity, universal access -, edition, reflection, analyses - positively and negatively -, and, given their discrepancy, the analysis and consideration of legal concerns.

Different organisations have generated various methods based on ethical principles to facilitate practitioners' develop AI components worldwide. These organisations include academia, trade union, business, government, and NGOs. Current trends also include setting out policy options on how to achieve promoting or the uptake and addressing the risks associated with particular uses of AI, the definition of a horizontal regulatory approach to AI that sets minimum requirements to address risks linked to AI, and different standards that cover topics related to AI, including big data, reference architecture, and artificial intelligence. However, independent of all these development, the implementation of all these approaches is cumbersome, and there is no straightforward actionability approach that shows how to perform such implementation.

In order to facilitate the process of developing, deploying, tracking performance, using, and continually improving the AI component on different systems, we propose the development of a vertical-domain framework for the manufacturing sector based on Trustworthy AI. This implies that the framework considers ethical perspectives, values, requirements and regulations (as established by the EC and users), and well-known risk management and decision-making approaches. Furthermore, specifically designed KPIs are embedded within the framework to facilitate the implementation, management, and track of the developed AI components, securing a continual evaluation of the system, sub-systems, or components state (concerning risk considerations).

Our framework is developed under regional considerations. As such, we consider the European context as the base to set the regulatory framework (and its objectives) and the scope (and requirements) on which AI systems should be based on. This context involves fundamental rights, Union values, investment, innovation, legality, governance, law enforcement, safety, and support in developing a single market for lawful, safe and trustworthy AI applications with a general perspective of preventing market fragmentation.

This framework will be tested in ASSISTANT by evaluating the frameworks and KPIs on the case studies and, more specifically, the AI components embedded within ASSISTANT work packages that produce AI assets tested by the case studies

Contents

List of figures

List of tables

1. About This Document

This document will provide a Management plan methodology for the AI components and a framework for developing and designing AI components within the Manufacturing sector (i.e. framework for developing ethics in/by design). We are proposing a well-structured approach based on risk management that would allow implementing ethical concerns in any life cycle stages of AI components (named development, deployment, use, and decommission).

The fundamental base of this approach is given by the idea that ethical considerations can and should be handled through a risk-based approach; more specifically, hazards and also referred to in this document as e-risks. The nature of the e-risk consideration is given by the technical, non-technical, and legal requirements and constraints that an AI component can face (and, therefore, its management requirements). If these requirements or constraints are not adequately fulfilled and managed, it is expected to have a severe negative impact on different sustainable pillars (e.g. economic, social, and brand) and legal implications. Even when this approach was initially focused on the manufacturing sector, its extension to other domains is expected to be easily performed.

The developed framework can be combined with other risk management processes to handle general processes or design within the manufacturing sector, allowing a global risk handling process. Further specification of extensions and evaluation of this approach will be given in the upcoming deliverables (i.e. D2.5), in which a final general framework for AI component development, deployment, use, and decommissioning will be generated.

The document is organized into two main components. The first component, which involves several sections, presents the framework by giving a sound introduction and state of the art review concerning Trustworthy AI, risk management, and AI standards (Section 2 and Section 3). Then, the methodology used in the framework is presented (Section 4). The methodology explains technical and non-technical components embedded within the proposed framework which includes a thorough explanation of (1) ethical risk (e-risks) and how to identify them, (2) a specific framework used for risk management (i.e. ISO 31000), (3) Fuzzy logic as a supporting tool for incorporation ethics-by-design, (4) Analytical Neural Process (ANP) and Analytic hierarchy process (AHP) as decision making components for the implementation of values within risk management process, and (5) Failure Mode and Effects Analyses (FMEA) and Failure Moded, Effects, and Criticality Analysis (FMECA) as a bottom-up, inductive analytical tools for performing risk analysis. After the methodological section, the framework (here and after also referred to as the e-framework) is presented in Section 5. The name e-framework is used to make a difference with external frameworks or general ones such as the ISO 31000.

The framework is presented in combination with suitable KPIs for Trustworthy AI management, which facilitates the tracking of e-risks and, at the same time, evaluates the continual improvement of the managed situation. Section 6 describes the implementation approach of the developed framework specifically for ASSISTANT (i.e. describes how ISO 31000, trustworthy guidelines and other referencing documentation are used in conjunction with the scope of the current project).

The first component ends with a conclusion section (Section 7) summarising the work performed, expected results, and future work. The second component corresponds to the documental policy in which e-risks are managed within ASSISTANT. Finally, a formal risk management document based on the RASP framework is presented. This document could be considered a documental example used to settle the architecture, structures and protocols required to manage AI components under the umbrella of Trustworthy AI under different

scenarios. The document is based on the general framework defined in the first component but has been created specifically for the ASSISTANT project and, therefore, contains specifications that would require to be adapted in other scenarios. Given the formality of this final document, it is presented as a stand-alone document in the annexe.

2. Introduction

The industry is becoming more automated in the Digital Era, with sensors and captors, advanced planning systems, process controls, and supervisory control systems. The main focus has been acquiring, collecting, and managing all data produced intelligently and efficiently during the last decade. However, the production activity of the manufacturing sector suffers from a lack of concrete and well-integrated solutions that empower the full potential of digitalisation.

The increased demand for new products with customized requirements and the digitalisation of the production processes drives Industry 4.0. Current factories blend the need for massive production with extensive customisation, increasing their product assortments [1]. Many of these advances have been supported by incorporating AI tools and techniques in manufacturing, reducing the number of lost sales, improving maintenance processes, and improving product and process quality (30\%, 29\%, and 27\%, respectively [2]).

Even though a direct adaptation of AI components could be made in the industrial sector, several technical, ethical, legal, and security challenges need to be overcome. One common denominator of these challenges is that they secure trust. The concept of trust is fundamental for a consumer of technology components to be confident in their use and adaptation, independent of the domain of implementation. Trust, as a general concept, can be achieved by the combination of specific definitions of the users (which can be translated into values, ethics, and technical requirements), robustness over time (under technical and social perspectives), and compliance with local and general regulations that establish, among others, accountabilities of users and developers.

As stated by Hegstler et al. [3], trust is the willingness to be vulnerable to the actions of another person. In the case of AI, this implies that an agent would be willing to interact and accept the outcomes of an AI. Therefore, trust is both a relevant and necessary quality that the users of AI will need if they are comfortable accepting the outcomes of an AI or using the products that have embedded AI in them. Therefore, trustworthiness is a vital necessity for incorporating the industrialization process of Industry 4.0, securing the marketability of the products in society.

The European ethical principles for AI presented by the AI4People group in 2018 defined five principles /ethical imperatives on which AI components should be relly on [4]. These imperatives include (1) non-maleficence, that state that AI should not harm people, (2) Beneficence, that state a worthwhile end goal for peoples, (3) Autonomy, which state the respect for people's goals and wishes, (4) Justice, that state that AI should act in a just and unbiased way, and (4) Explicability, that sates explanation on how an AI system arrives at a conclusion or result.

The European Commission's High-Level Expert Group on AI developed a set of ethics guidelines for Trustworthy AI, which has become a formal policy within the European Union. Trustworthy AI is defined as AI that is legal, ethical and robust. Its operationalisation is articulated through seven key requirements for AI systems that should be met to deem them trustworthy [5]. The recently draft AI Act goes further, building upon the work of the HLEG-AI,

and proposes a risk-based approach to the regulation of AI: the greater the risk, the greater the level of conformity assessment, testing, labelling, etc. that is required.

Trustworthy AI is a pillar that should be considered incorporated at every stage in the production of AI assets. Trustworthy AI does not only implies the consideration of ethical use of AI Assets. It works as a general approach in which topics such as policies and regulations of AI are considered, including specifics related to regional considerations [6]. Foremost among these are: agents' roles, societal issues (e.g. inclusion), diversity, universal access, prediction, reflection, analyses (positively and negatively), and given their discrepancy, the analysis and consideration of legal concerns.

The development of trusted AI assets requires the follow-through of the stated legal compliance. It can be seen that some aspects on which AI fundamentally rely are starting to be regulated by governments (e.g. the General Data Protection Regulation – GDPR – EU 2016/679) or can be complemented without prejudice (e.g. Law Enforcement Directive (Directive (EU) 2016/680)). In the lack of legal requirements and system control, ethical imperatives and requirements derived from them define a dimension space on which AI assets could and should be fundamentally based.

As specified by the European Commission (EC), AI entails potential Risks. These risks must be addressed by regulatory frameworks that should concentrate on minimising the various potential harms, particularly the most significant ones [7] [8]. These risks can be linked to three primary sources related to (1) fundamental rights (including those related to personal data, privacy protection, and non-discrimination), (2) safety and consistency, and (3) liability-related issues (including, among others, accountability and transparency).

To achieve this goal, the EC has settled in its Artificial Intelligence Act [8] the introduction of the regulatory framework in the EU that introduce binding rules for AI systems, a list of prohibited AI systems, extensive compliance obligations for high-risk AI systems, and definitions of fines (of up to €30 million or 6% annual turnover) [9].

The list of prohibited AI systems is defined based on a risk-based approach. As shown in Figure 1, Artificial Intelligence (AI) components can be categorized as unacceptable, high, limited, and minimal risk. This categorization is based on the AI functionality (i.e. what the AI component does) and on the AI domain and application area (i.e. the impact that an unexpected behaviour could produce on persons, critical infrastructures, social structure and its disruption and environment).

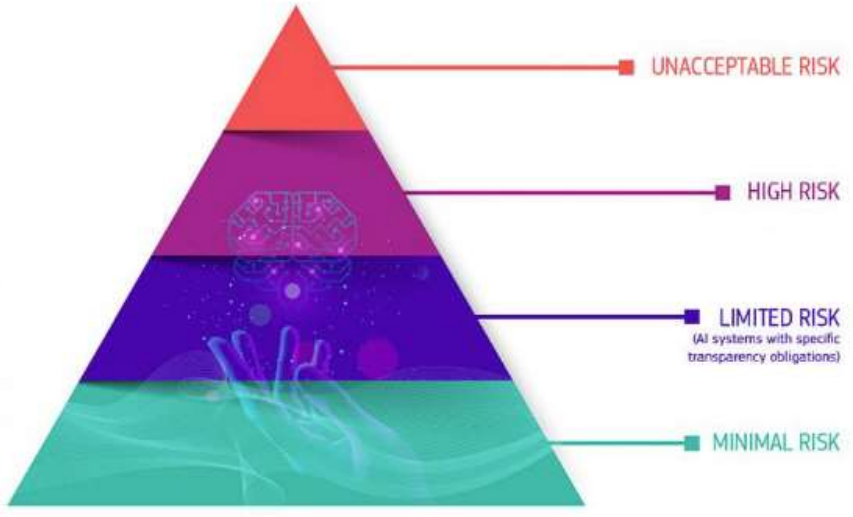


Figure 1 Diagram of the AI Risk categories

Even though the EC is arguable the most significant regulatory body in Europe, member state governments have also settled on the importance of regulations of AI components based on risk considerations. For example, the German Federal Governments Data Ethics Commission has settled recommendations and opinions regarding actions and suggestions for possible legislation and implementation [10]. As described in the documents, there are recommendations for a Risk-adapted regulatory approach and descriptions of the risk involved in algorithmic systems. It is clear up to this point that the consideration of Risk, and risk management, is fundamental during an AI life cycle.

The EC has presented a well-structured framework to foster an AI asset. The framework is shown in Figure 2, and it depicts the three main components (ethical, lawful, and robust) in which the framework is sustained. As pointed out, these three components should be operationalized, ideally, in harmony to secure a **Trustworthy AI**.

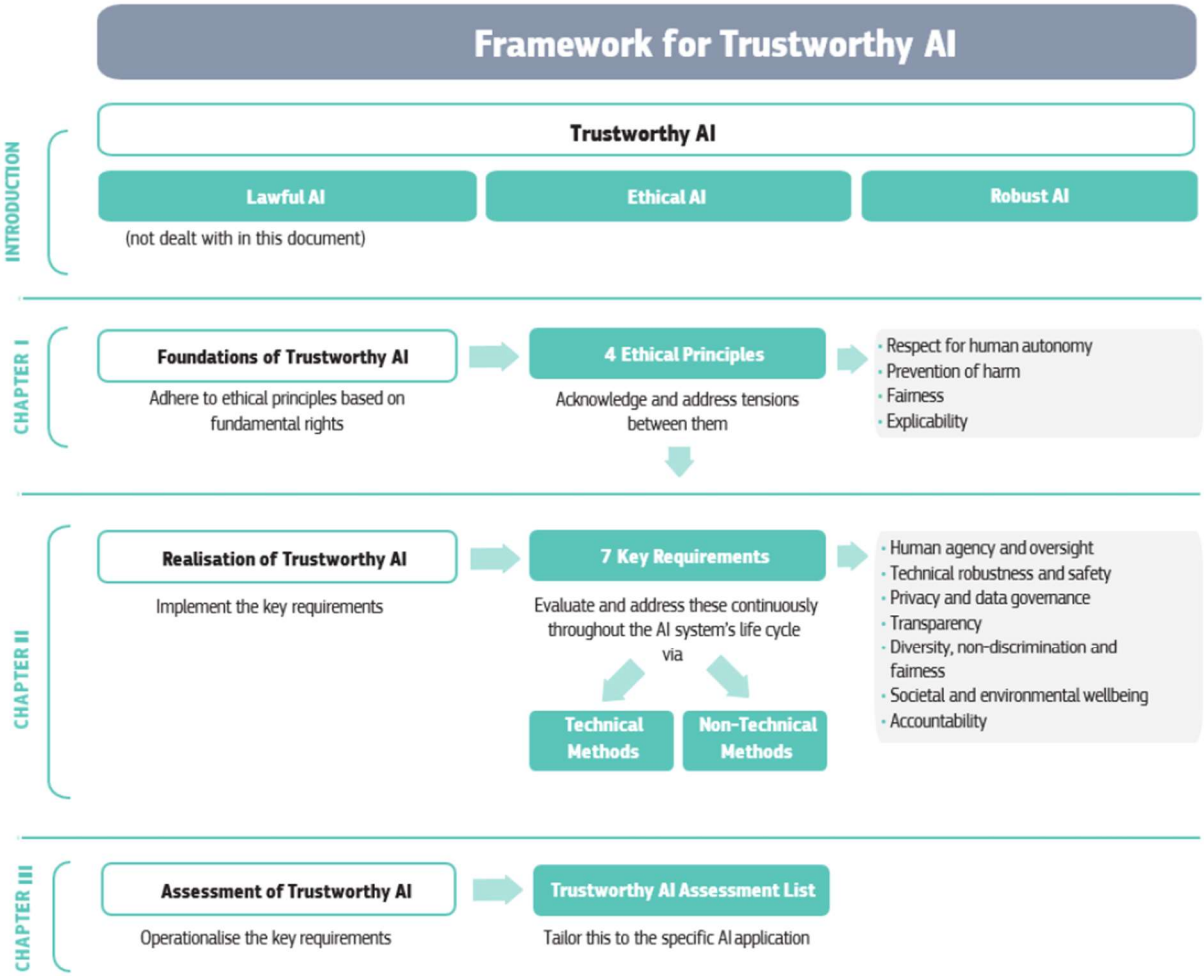


Figure 2 Framework for Trustworthy AI

As a stand-alone definition, trust involves uncertainty, vulnerability, and consistency over time. In other words, the trustor knows that they are making themselves vulnerable to the actions of the other agent and, in order to secure trust, the actions of the other agent (or AI in our case) should be reliable [11] or, in other words, with the lowest risk to produce adverse outcomes (i.e. a direct link between trust and risk management).

Therefore trust can be fostered by securing that the AI component fulfils lawful, ethical, and robust components throughout its entire life cycle. It should be ethical, ensuring adherence to the ethical imperatives and values that govern local and regional social behaviours and regulations. It should be lawful, securing compliance with local and global laws and regulations. Finally, it should be robust from a technical perspective (embedded within the trustworthy requirements) and a social perspective [11].

Even though the current state of the EC framework operationalization does not include the lawful component, the evolution and trends observed from its publication and the expected contributions from the High-Level Expert Group could lead to a set of binding rules relevant throughout the AI life-cycle.

Currently, AI assets development can be driven by the blend of some general frameworks, developed standards and ethical regulations (which are also considered as a backbone structure for legal considerations). However, given the lack of a broad specificity on the regulatory landscape, the gap between principles and actionable requirements is still considerable. Currently, pioneer companies are working hard to construct applied ethical frameworks in different sectors for using AI components that generate trust in their clients and workforce [12] [13]. Nevertheless, most companies/industries tend to fall into the same gap between principles and actionable requirements given by a lack of understanding of the impact of using (or not) Trustworthy AI components, the lack of trained human resources with the capabilities to handle each technical and non-technical aspects of Trustworthy AI, the level of technification within the company, a misunderstanding on how to apply AI within the companies, and other factors that could delay the development of AI components within the industrial sector.

The lack of standards, definitions, and regulations creates gaps that could lead to self-claimed implementations based on Trustworthy AI, but its concepts could not be adequately addressed. These AI assets could further misunderstand the industrial domains since they mislead how ethical concepts and moral values from the field's regulatory frameworks are integrated into the AI software components.

The rapid growth of AI components within the global market (as being incorporated within the different goods produced) and the manufacturing processes describe a current need of the sector to evaluate their AI development approaches, deployment and use under the umbrella of Trustworthy AI. AI components, specifically robots, are already in various manufacturing tasks.

On the other hand, even though AI is already used to make low-level decisions in the manufacturing sector, such as automated machine tuning or predictive quality, there is still significant room for improvement and extension into higher-level manufacturing decisions. Incorporating data-driven and AI approaches would facilitate the design, planning, control, testing, management, and integration of the product and processes. These products and processes involve the use of hybrid augmented intelligence by human-AI components/devices, cross-domain/cross-media reasoning, development of new models (including internet-based, customizable, flexible, collaborative, and service-oriented), means (including human-machine systems, IoT, virtualization, flexibility, service, customization, intelligence), and forms (including intelligent manufacturing, integrative, data-driven) [14]. Combining all these possibilities allows the creation of a new paradoxical manufacturing ecosystem that will require architectures and supporting technologies (e.g., data fabric) that should consider incorporating Trustworthy AI approaches during their development and implementation.

It is clear at this stage that the manufacturing sector would require the support of a suitable framework based on Trustworthy AI that will help integrate AI components within their product and processes. Therefore, this framework would focus on the considerations encapsulated as a management component. At the same time, such a framework should be adaptable to regulatory modifications, standards, the domain of implementation, and the different life cycles of the AI component.

Given the previous needs and considerations, the present document focuses into:

- Support developers to incorporate ethical principles and values within the AI in product life-cycle processes. It is key that AI developers are familiarized with the ethical principles at every stage of implementing/operating AI assets. Furthermore, it is key to distinguish between requirements -- could be needed by law to acquire commercial certifications -- and values -- that are societally imposed and can vary depending on the region and culture. Therefore the framework should be flexible enough to blend these.
- Modifications on the regulatory environment for weak AI assets, securing its use independent of legal and technical requirement changes, must be easily incorporated. Flexibility is required as there is heterogeneity in legislation to be applied by different countries to use AI.
- Facilitate the combination of the developed framework with other approaches used to handle risks by industrial stakeholders. The combination approach will enhance the adoption by companies that already have their own Risk Management Process (RMP). Therefore, the framework needs to be designed as a complementary asset and not a replacement.
- Facilitate a continual improvement in handling risk components within the AI assets. Many processes in software do not follow a sequential development but a spiral/ iterative development process - e.g. agile techniques. The framework should incorporate the benefits of these development cycles to ease developers' risk management and foster more secure and better products.
- Ensure that metrics and Key Performance Indicators (KPIs) can be tracked to register the evolution of ethical based risk management. For many companies, specifically, the business units, tracking KPIs is essential for their daily operations. In addition, managerial levels must use this tool to better understand the incorporation of ethical aspects into development in parallel to the existing process.
- Construct an architecture to support a better understanding of responsibilities and channels of communication between technical and non-technical stakeholders. For example, the legal departments of many companies do not have the technical knowledge to satisfy the legislation on some aspects of the AI life cycle. Similarly, technical users - developers, architects - do not know AI's ethical aspects that could be imposed by current or future regulations.
- Foster the replicability of outcomes for other use cases and domains with analogous ethical risk and AI functionalities. Replicability is key for advancing research and for companies to save revenues in future developments and incorporate new processes into the existing ones. In addition, a well-structured risk identification avoids repeating failing conditions to similar AI components.
- Facilitate the ethical-based risk evaluation using a pipeline-based approach. Having flowcharts to model the framework eases its understanding and implementation.

The framework developed here is explicitly constructed to identify the questions of responsibility to be asked during the development and deployment stages. This enables the teams producing AI components to reflect on potential outcomes and implement features based on pre-specified values influenced by what can be called ethical standards. This implies that abstract ethics must be projected to concrete use cases and applications. Therefore Ethics in AI can be envisioned as a funnel that gets more narrow; the more profound the user gets into it, the higher the impact on ethical considerations in AI components. This funnel starts quite

broad with a general overview of Ethics in AI and general frameworks. Then, narrowing down the scope, a domain-specific context that involves specific features to that domain.

There is no specification yet on how stringent the funnel should go, especially in the manufacturing and production planning domains that have not yet envisioned the impact of AI use for the multiple requirements of current (and future) scenarios. For example, in the manufacturing of healthcare devices, the data from customers' devices require a different sensibility than the social media posts of the communications department of the same company. This is also due to the overarching values and norms that can be assumed for a domain even when, as stated, they are based on the same principle - the principle of autonomy in bioethics and AI components should process the data differently.

In the following subsections, a general description of the different approaches and methodologies used for the development and deployment under the umbrella of Trustworthy AI is done. The objective is to give a broader perspective before defining specifications for the manufacturing sector, specifically digital twins.

3. State of the art - literature review

Even though a direct adaptation of AI components could be made in the industrial sector, several technical, ethical, legal, and security challenges need to be overcome.

The first challenge is associated with the need to process data at speed. The vast amount of information accumulated by the sensors adds complexity to the systems for taking decisions in near real-time. This challenge makes the system performance a critical feature required for an operational Manufacture 4.0. Most techniques to increase performance are hardware-based, such as utilising heterogeneous hardware accelerators - GPUs, FPGAs, and MICs architectures. However, several alternatives can be found in the literature around speeding up the performance by exploiting the capabilities of edge and cloud computing resources together. The increased performance is obtained from relieving communications pressures by moving computation closer to the sensors and control loops. However, some of the limitations of edge devices can be found in the amount of memory, CPU, and storage they can offer, which can hamper the execution of computational intense (e.g. digital twin within the required time boundaries for large amounts of data.

The second challenge is the need for skilled human resources to operate and understand AI techniques to support the acceleration of executions on heterogeneous systems. Ensuring expert operators and developers skilled in several specific topics related to distributed systems, high performance, and AI require heterogeneous teams with diverse skills to support the maintainability and sustainability of the software life-cycle.

The third one, and the main focus of this work, is related to the ethical dimension of using AI in manufacturing environments. Ensuring the adoption of the AI components - involving or not the workforce - requires an honest assessment of these components. The burden of lacking specific regulations and standards for the development and deployment of AI components and, at the same time, the relatively low level of understanding of both the algorithms and processes from key stakeholders impose further complications. Thence, there is a big difference between creating services business-to-client (B2C) or business-to-business (B2B), where applications and algorithms created later can lack the domain context, and therefore the same service can be used in a non-trustworthy manner.

Finally, the last challenge involves legal requirements that define a framework on which AI components can operate and construct a dimensional landscape in which the manufacturing

sector can provision or use techniques to fulfil such requirements. The legal components also provide liability and responsibilities of users and developers and therefore allow to establish the scopes in which the manufacturing sector should be focused to reduce operational failing conditions of their AI assets. Significantly, this challenge is not disconnected from the ethical dimension, as broadly discussed in this document, and therefore several factors could be considered ethical factors at present but could derive from legal requirements in the future.

3.1 Frameworks, Guidelines, and other approaches for Trustworthy AI implementation

The applicability of ethical use of AI in software development arises from its conception. Like every other component or system constructed by humans, technological artefacts always demand decisions and incorporations of flaws embedded by their creators [15]. This antecedent means that humans have a concrete view and responsibility of the outcome from designing a software component to the development, testing, production, and use until its decommissioning. In a scope in which artefacts, tools, and software are meant to automatise tasks, predictions and decision-making flaws are permanently embedded in the context. These decisions and predictions are guided and influenced by personal biases and values developers inscribe into the applications or come from the information source used for development.

Therefore, AI management regulations, frameworks and guidelines should contain technical and non-technical factors. These factors will come from the AI by itself, the data is used for their development and use (i.e. training information and supplied data for its per), and social, ethical and values factors that will come from stakeholders involved in each stage of the AI life cycle. Next, some critical regulations frameworks and guidelines currently set for the Trustworthy AI field are revised.

3.1.1 Ethical Frameworks and Guidelines

A framework can be defined as an open structure that gives shape and support to something. Therefore, Ethical frameworks and guidelines can be seen as an ethical-based generalization approach for the AI development, deployment, use, and decommissioning stages. Different organizations have developed various methods based on multiple ethical principles to facilitate practitioners in developing AI components worldwide. These organizations include academia, trade union, business, government, and NGOs. Examples include The Institute for Ethical AI and Machine learning, Microsoft's Responsible AI guidelines, UNI Global Unions, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, and the Ethics guidelines for trustworthy AI of the High-Level Expert Group on Artificial Intelligence.

An extensive list of guidelines and strategies based on critical AI issues can be seen in [16]. No approach covers all the 22 extracted issues as specified in this work. The most mentioned issues include privacy protection, fairness, non-discrimination, justice, accountability, transparency, openness, safety, and cybersecurity. Even though organizations show different interests in what principles should focus on, the most relevant concepts seen by organizations and companies includes privacy, fairness, accountability, transparency, explainability, and safety [17] [16]. These key components would have different importance and relevancy depending on how the developed components would be deployed. Therefore, strategies for using ethical guidelines and general frameworks should be seen as a supporting alternative to generate suitable (and more specific) frameworks.

Controversy has been generated related to specificity and industry influence in frameworks and guidelines [18] [16]. Even though some could agree with these considerations,

the specification of tools and domain-specific impact (and requirements) will be developed in due time as better examples and understanding are done during AI components' development and deployment. It should be reminded that even though some ethical principles overlap with other domains (e.g., transparency, justice, non-maleficence, and fairness), the interpretation is dependent on the scenario in which they will be implemented. Furthermore, AI ethical considerations would require enough time to reach the maturity level of ethical perspectives similar to those domains that have dealt with ethical considerations for a long time (e.g., medicine and Business). These considerations do not imply that principles, frameworks, and guidelines should be adopted and followed to checkboxes and security requirements or status (such as certification). Contrarily, the systematic inclusion of ethical principles should be done in a structured way by the industry, following techniques and implementation methods under development.

It is essential to highlight that ethical AI frameworks and AI implementation guidelines should consider the entire environment [18] in which these components are developed and deployed (including all the agents involved). Conditions could change over time as tools are integrated into dynamic environments and, therefore, challenges and concerns would not always be foreseen at the initial stages. Better risk identification can be performed by clearly identifying the environment where AI will be developed and deployed. Furthermore, given the dynamic nature of the systems in which AI could be deployed, monitoring the application throughout its lifecycle is necessary. Finally, as new legislative endeavours emerge, it might be essential to update the application (frameworks and tools used in the implementation), especially as some ethical concerns, values hierarchy, and decisions change over time.

The establishment of ethical AI frameworks in the industry sector is a considerable important step, and they are required to consider the complexity of the domain; otherwise, they will fail in the continuously changing environment [18]. This consideration implies that the industry requires frameworks that have to be generalizable but with enough specificity and aided tools to allow the sector to not struggle during its implementation.

Independent on how well-developed guidelines and frameworks are constructed and used, failures will arise by unintentionally negative consequences (i.e., when AI are developed and deployed without sufficiently robust governance and compliance [16]) and by the incorporation of these tools in systems controlled by agents that are not ready for them. The risks of failure could come from different sources, including company management, regulatory incentives, manufacturing practices, employee training, and quality assurance, in addition to the risks involved within the development of AI components (e.g., lack of understanding, biased information, improper combination and managing of data, misuse of algorithms).

3.1.2 **ART principle**

As specified by Dignum et al. [19], they proposed the principles of Accountability, Responsibility and Transparency (ART) as a design for values approach to ensure that values and ethical principles are included in the AI design process.

Even when chronologically the ART principle was presented before the European Union Trustworthy requirements and the European AI act [5] [8], its validity and point of view of the most relevant factors for developing social robots can be considered relevant at the moment of considering the focus of the more relevant trustworthy component.

3.1.3 **Human-Centric**

Human-centric can be considered a sub-class or a particular AI framework that focuses on the interaction and collaboration with human agents. The algorithms (and learning processes) can continually be updated and consider the human agents' state, needs, experiences, and human-AI physical component interactions. For the algorithms to perform this way, a combination of sensed and historical information can be entwined to extract behavioural data such as patterns and choices, among other trends.

Since the AI component is deployed in a physical structure (Human-Centric consideration), the system could require, depending on functionalities, an understanding of the environment in which the interactions are performed. Under this umbrella, for an AI component to be considered human-centred, it requires to be: explainable, verifiable (that can be linked to six generic properties: reliability, safety, availability, confidentiality, integrity, and maintainability [20]), physical, collaborative, and integrative [6]. The perspectives on which Human-Centric considerations are based can be linked to other ethical frameworks, but given the nature of a specific human-AI physical component, it can be classified as a particular case. Furthermore, there are some challenges, given by the possibility of the direct physical and dynamic interaction with humans, that make it relatively harder to be applied (depending on the goal of the developed AI element). These considerations/challenges could include, among others:

- Understanding of the human uncertainties by the AI. Understanding humans from the AI point of view would allow an understating of the whole system environment from the AI perspective. This approach could be made, for example, by methods that predict the user's trajectory [21] or estimation of zone occupancy [22], among others (e.g. human emotion state reading).
- Understanding of AI uncertainties by humans. This consideration could lead to an increase in unsafe practices given the lack of understanding or misinterpretation of how AIs work [23]. Solutions to these problems are directly linked to transparency and trustworthiness considerations.
- Intrinsic/cognitive human biases (such as confirmation bias, in-group bias, availability bias, and anchoring biases) can modify the perception and behaviour of human agents in a multi-agent environment-specific systems
- Processing of multi-sensorial systems to combine information from agents and the environment. The information should be captured with a dynamic granularity homogenized, so the sensed information captures behaviours and significant trends
- Explainability of black box AI elements that, for example, are intrinsic in the case of image processing
- Make the system reliable, especially for critical applications (e.g., human surgery, automatic driving). Therefore, the system would not produce erroneous estimates and be safe to a broad extent (including noisy information and cyberattacks) - in other words, verifiable to the extent to which performance surpass the current state.
- Defining standards and protocols for general/specific applications and domains for the AI elements that will interact with human agents, independent of the method of communication (e.g., verbal)
- Models or techniques to improve understanding of human behaviours (individually and aggregated) and under AI interactions. These models could be used to forecast human reactions and actions and, at the same time, improve verifiable and collaborative perspectives
- A suitable link between non-interpretable formalism with interpretable formalism. In other words, data and machine learning components with symbolic models and specifications are interpretable by human agents deployed by encoding processes.

Again, chronologically, the Human-Centered approach was defined before the trustworthy requirements [5]. Thus, under that contextualisation, the latest one can be considered a framework that takes a broader perspective over the considerations of human and AI interactions.

3.1.4 Human-in-the-loop (HITL), human-on-the-loop (HOTL), and human-in-command (HIC)

HITL, HOTL, and HIC can be considered another sub-class of AI consideration that extends to the autonomy and collaboration of AI regarding hun agents. HITL considers the interaction of humans within the decision-making process, allowing to take advantage of intelligent automation efficiency while remaining amenable to the interactions of human oracle feedback. The benefits of HITL include relative incorporation of transparency within the systems, incorporation of human judgment (i.e. accountability) and, among others, removing pressure from perfect algorithms [24]. Dignum [6] mentioned that HITL is often the most appropriate since they allow for more clear responsibility attribution. Nevertheless, the decision made by the agent is affected by societal, legal, and physical infrastructures.

Some important considerations within the HITL approach are that (1) it is dependent on the system granularity and functionalities. Strongly dynamic systems, for example, would not allow human participation in the system, therefore restricting human participation in real-time processes and (2) it is a perspective similar to human-centric approaches.

HITL can easily be seen that a human still has complete control over starting or stopping any system action. If human agents are pushed outside the process cycle but with system oversight capabilities, the system achieves higher participation in the decision process, achieving actions at the required processing speed. This approach is called HOTL. HOTL has more substantial benefits for highly dynamic systems (e.g. manufacturing system control). Furthermore, it can easily be foreseen that implementing such systems would not be possible for a system with high intrinsic risk unless some approaches are developed to reduce the likelihood of events to materialize lower than those in which AI systems are not participating. Traditionally, safety analyses do not focus on user-related or user induced hazards [25].

Finally, HIC can be seen as the human agent's approach to making all interventions and decisions. The early specification of human control was built on the perception that humans and machines have different capabilities [26]. This process involves incorporating several human-related weaknesses that could derive from technical and non-technical failing conditions. For example, boredom at routine monitoring, bias incorporation, and alert fatigue, among other considerations, establish that humans perform poorly as supervisors of automated technical systems [26] (i.e. a restriction to be used on complex manufacturing systems).

A sound risk management framework contributes to any AI system that involves decision making and requires human oversight. First, by incorporating it within the AI management process, HOTL considerations are secured since, independent of the system dynamic considerations, a risk management process places oversight on the overall system and secures intervention stages when needed.

Second, by including trustworthy and values considerations, it can foster the inclusion of the entire system environment in the analyses, thus securing the inclusion of users on the focus of system security.

3.1.5 In-, By-, and For-Design (ethics) and Design for Values

AI elements can be developed and deployed under different criteria that possess diverse impacts on the functionalities and legal concerns involved. These criteria lead to diverse opportunities to incorporate responsible considerations on AI elements under different scopes. One scope involves that components can be built, deployed, and integrated under some pre-specified concepts or approaches (e.g., ethics) and embed them with specified concepts or methods (i.e., in design). Another alternative implies that the component will be built with intrinsic capabilities that are part of the concepts or approaches of interest (i.e., by-design). Finally, the concepts can be specified as part of codes, standards, and regulations that ensure the integrity of the different components and stakeholders under the selected considerations' umbrella (i.e., for-design). For a thorough understanding of these concepts, readers are encouraged to check [6] [27] [28]. A critical benefit of the -in -by and -for design approaches is that they can be implemented transversally in the process of developing, deployment and use of AI elements but always under the umbrella of a specific scope (in our case, ethical and social requirement – i.e., Ethics-by-Design, Ethics-in-Design, and Ethics-for-Design).

Ethics can be seen as a human-related discipline concerned with behaviours that classify them in labels recognized as "morally good" and "morally bad". Independent of what could be considered good, wrong, correct, or incorrect, the final decisions on the actions to perform are usually driven by an agent's values, principles, and purpose in a system that could consider multiple agents in a complex environment.

Therefore the theory and disciplines of ethics are strongly involved in understanding agents' actions and values. However, one difference highlighted when considering ethical-driven actions and values is that the first involves a generalization of concepts that will derive systematizing behaviours under "right" and "wrong". Contrarily, values influence agents' behaviours and attitudes and reflect their sense of "right" and "wrong". This implies that even though approaches of -in -by and -for design could be implemented based on ethics, it should consider the domain and environment in which these approaches will be implemented (e.g., cultural differences could possess similar values, but the hierarchy in which these values are pondered could be different).

In terms of ethics, the normative, virtue, and applied subdomains can be considered for implementation within the approaches of -in, -by, and -for design. This consideration is given by the scope on which these subdomains fundamentally focus. Even though some classical theories are extensively known in normative ethics (e.g., consequentialism or utilitarianism and deontology or Kantianism), specific applications tend to favour some theories over others. Different ethical frameworks have already been settled in their establishment as a solution for AI development and deployment [29] [30] [31]. It is also worth noticing that among the different theories, those based on the concepts of consequential approaches tend to be favoured, probably given the more accessible methodologies involved in using metrics that can be optimised to determinate behaviours (i.e., based on the premise that "an action depends on the consequences it has").

Values and principles are dependent on the context of the application. Additionally, several values could be incorporated within the implementation context that could be contradictory. For example, personal values tend to behave in such a way (e.g., benevolence and universalism over personal power or achievement enhancement). Therefore development and deployment stages could follow a structured methodology guided by hierarchically organized values [27]. The design-for-values approach allows incorporating values rationally guided by a process that involves identifying relevant values, generating a normative practice for incorporating such values and linking such normative systems with concrete functionalities [6].

As specified by Lason [32], AI elements can be aligned in different behaviours (these alignments are: The agent does what is instructed to do, The agent does what it is intended to do, the agent does what the behaviours reveal its preferred, the agent does what it is in the interest or best to do, objectively, and the agents does what its morally ought to do, as defined by values). As specified in this work, the alignment based on morals and values is one of the most suitable alternatives to impress with ethical considerations different components.

Two approaches, parallel or series, can be used to secure the integration of values during the use or development of AI elements. A parallel structure implies merging or fusing values with the technical components (e.g., merging with current approaches such as architectures – see [6]). On the other hand, a series structure can be done control active as in a series structure (e.g., as part of a pipeline in which functionalities and normative sets are applied outside the AI element, but it works as an ethical screening device).

The main benefits of the Design-for-Values approach its three-fold. From one part, it allows the integration with technical components – this facilitates the incorporation of values (by specifying derived norms) into legacy approaches (that could be modified) or in components under development. Second, even though generic, considerations such as wealth, health, safety, and other values can be linked to metrics representing the system's state. The last benefit is significantly important since it allows monitoring, given a pre-specification of suitable indicators, of the state of a given condition. Finally, they transform abstract concepts into norms, which leads to specific requirements that different stakeholders can understand.

A clear example of implementing the Design-for-Values approach can be seen in [27] [28]. In that work, the Design-for-Values approach works as a filtering component around the developed AI element to map moral values into explicit, verifiable norms that constrain the system inputs and outputs.

3.1.6 Bottom-Up, Top-Down, and hybrid systems

Top-Down and Bottom-Up approaches are methods used to analyze, extract, and implement specific "concepts", such as human goals and values, into and from the systems. These "concepts" can be broadly different, depending on whether they are analyzed or extracted. In addition, the domains of applicability of these approaches are broad and can include, for example, security, business, and ethics.

The Top-Down approach is linked to using a general understanding (Top) of the system and its components. In the Top-Down definitions and analyses, the system is evaluated as a whole in which specific components (Down) can interact. A general example of a Top-Down approach is macroeconomics.

On the other side, the Bottom-Up approach focuses on understanding specific characteristics and attributes (Bottom) that could be used for a better understanding and specification of the whole system (Up) (e.g., microeconomics).

The hybrid systems combine the previous approach to develop the best decisions and actions possible based on an approach fed by different stakeholders and information that can contribute to a thorough understanding of the whole system. In addition, these Top-Down and Bottom-Up approaches have also been beneficial in designing indicators that help evaluate the systems' state [21].

In terms of AI ethics, the Top-Down approaches have been linked to the availability of the system to use and deploy pre-structured ethical approaches within the system and

frameworks (implying an overlapping and mixing opportunities of strategies with several previously specified approaches - e.g., ethics-in design by Top-Down approaches). On the other hand, Bottom-Up approaches correlate to using existing system information to extract values and behaviours from agents. This consideration implies deriving the intrinsic rules that describe agents' intentions, but that does not imply agreement with the domain's ethics and values. Furthermore, data could contain biased trends that must be removed or thoroughly analyzed before defining and constructing system models.

The hybrid approach considers the mixing of Top-Down and Bottom-up approaches, given the capabilities to regularize the system with the systems' and agents' goals and behaviours. Independent of the approach to be used, there are still definitions that will have a broader impact on their outcomes and systems models - who will define rules for the case of the Top-Down approach? Moreover, based on what values? Or what data to use to extract such information in the Bottom-Up approach? What variables will be selected for such a task? [32].

The Top-Down approach could be considered relatively easier to implement. This consideration is based on the idea that experts and theories are used comprehensively during the development of the system or similar domains (e.g., a utilitarian AI would perform over the overall system (Top) by creating the greatest happiness for the most significant number of agents - a rule broadly known that can be implemented). In addition, different studies currently focus on extracting and identifying agents' objectives, values, biases, and rewards [31] [33]. For example, Inverse Reinforcement Learning is a relatively recently developed machine-learning framework that focuses on solving the reverse problem of Reinforcement Learning. This approach can learn human values and biases from data (Bottom), based on the idea that agents' actions "seek" to obtain available rewards.

3.2 Trustworthy AI considerations in the manufacturing sector

Ethical considerations within the manufacturing sector should be focused on two different approaches depending on the nature of the goods produced. The goods could or could not have included AI elements that will interact with secondary stakeholders. In the latter (not embedded AI element), the reach of AI is limited to those stakeholders within the manufacturing sector, and, therefore, a difference between the reach of guidelines and frameworks should be clearly stated. In the case of ASSISTANT, the approach is limited within the manufacturing sector domain (i.e., most of the following discussions would follow with these considerations).

This consideration implies that the interactions between AI elements and the stakeholders previously mentioned should be differentiated. Furthermore, since the AI would be mainly implemented in a higher-level technified domain (i.e. industry instead of any regular social domain), transparency, accountability, and safety considerations could be easily defined and implemented.

Manufacturing can be considered the production of goods using different transformation techniques over raw or intermediate materials, including machines, tools, labour, chemical, and biological processes. Incorporating AI in manufacturing expects to value U$ 16.7 billion by 2026 [34]. This trend is driven by an increasing number of large and complex datasets, the revolution of interconnectivity and sensing provided in Industry 4.0 and IoT, and the improvement of computational power and automation performance and capabilities.

As can fastly be reviewed in the literature, by specifying the combinations of "AI ethics" and "manufacturing" or "Trustworthy AI" and "manufacturing" in search engines, there is a considerable discrepancy in the number of works that focus only on "AI" and "manufacturing" or

"machine learning" and "manufacturing" keywords. For example, as of 05-08-2021 (google scholar), only 777 and 304 publications were obtained for the first two combinations, while the last two produced 1,260,000 and 2,070,000 results, respectively.

A significant requirement to incorporate AI in the factories is to keep processes robust and low risk (optimally free). Thus, real-time AI techniques enable the exploration of what-if scenarios without interrupting production. In addition, processes such as future actions An example of new enablers for exploration and exploitation in the manufacturing sector are the digital twins.

A digital twin is a higher-level model corresponding to a digital representation of physical objects or processes. A digital twin is encompassed by three main components that include the physical component (e.g., shop floor, robots, and operational units), the digital representation (encompassed by domain models, metamodels, and constraints), and its communication (e.g., the data fabric). This specification means that digital twins are tied back to their physical build through the sensors' information: the state, the working condition, or the position integrated with a physical item. These can support complex processes such as autonomic resource reconfiguration, replacement, or movement, in factories, increasing the flexibility of the resources and the. However, an inadequate application can lead to a production break that is economically catastrophic for factory plants. For example, in 2018, a 30-minute unplanned break (power outrage) at Samsung's Pyeongtaek fab resulted in some 60,000 NAND flash wafers being scrapped, with an estimated $43.3 million in damages.

However, what could go wrong? Catastrophic implementations of AI elements are seen in the short time when AI has experienced a relatively colossal boom. To name a few: An AI-powered tool designed to identify a person's gender by analyzing its email address was shut down after its lunch - several controversies were generated based on their failed estimations [35]. A facial recognition mass surveillance AI approach was implemented and considered illegal based on scraping images from social media and other public sites [36]. Criminals used AI-based software to impersonate a chief executive's voice and demand fraud [37]. In Tempe, Arizona, an autonomous Uber car struck and killed a victim related to the first pedestrian death associated with self-driving technologies [38]. A chatbot was corrupted in less than 24 hours, fed by controversial concepts such as racist and misogynistic talks [39].

Even though some of these implementations involve misuse of AI technologies or a lack of understanding of social trends, values, and moral concerns, the implications that could be produced in the case of failing AI elements in the manufacturing sector can be equally dangerous for the users, workers, and companies. The trends show that around 85% of AI projects could deliver erroneous outcomes due to bias in data, algorithms, or poor management by the teams involved in their development and implementation [17]. Interestingly, Lauer [18] stated that most critical considerations and the failure to implement AI elements (as exemplified before) are not given only by technological considerations. Instead, there is a fundamental lack of ethics within companies (that includes poor requirements, governance, and processes) that would inevitably lead to the development, deployment, and use of AI elements with several responsible issues that will, in the end, lead to adverse outcomes.

Even though these considerations could be alarming, the industry sector would not stop implementing AI elements within their processes or products. Some commonalities could be followed by industrial companies' other technologies or operations. Nevertheless, these considerations focused on traditional roles where the industry sector only has broad experience (e.g., safety regulations and business ethics). Contrarily, incorporating AI technologies comes in hand with broader responsibilities for each participant involved in developing, deploying, and using such components.

To foster a correct implementation of AI elements within the industrial sector under the well-named Trustworthy-AI umbrella, stakeholders must understand the approaches and frameworks proposed for incorporating ethical considerations, values, and other requirements. A better understanding would facilitate, among other benefits, reducing biased information, improving monitoring, improving performance, and reducing failing systems. As Accenture analyzes, 63% of stakeholders consider it relevant to monitor AI systems but are unaware of achieving such tasks [17].

In general, the incorporation of AI in manufacturing can be characterized by autonomous intelligent sensing, interconnection, collaboration, learning, analysis, decision-making, and execution of human, machine, material, environment, and information processes in the whole system and its life cycle [40]. Implementing these assets will require Trustworthy AI integration, mainly to keep human agency and security, among other concerns.

Even though some applications could be transferred to other domains, they require considerable adaptation or rework. Adaptation of any new technology does come with different challenges that do not only involve technological ones. A common denominator of these challenges is the development of trust. The concept of trust is critical for technology providers allowing consumers to be confident in their use, independently of the market segment.

Therefore, incorporating AI in the manufacturing sector should not be treated differently or separately from the challenges involved in incorporating AI elements' Trustworthy AI considerations in any domain. This consideration is especially true since the manufacturing sector algorithms are not different from those used in other domains. Nevertheless, it is necessary to establish why it is essential to incorporate Trustworthy AI frameworks and methods within the manufacturing sector to grasp the challenges. In other words, the approaches used in general for AI components should be merged and possibly improved by the values that are intrinsic to the manufacturing sector.

It can be foreseen that the relevance of incorporating Trustworthy AI would be seen during failing conditions. The system failing conditions can produce diverse outcomes (including production halt, safety, loss of revenue, and brand name damage) with different levels of severity. Independent of the results, acting reactively over failing systems prevents future problems since they tend to be systemic and involve cascading or multi-system failures [18]. Therefore, proactive methodological approaches to reducing failing conditions should be the primary drive for developing, deploying, and using AI elements.

The failing considerations put a general framework in which analysis and implementation of ethical considerations could be driven. In addition, the combination of methodologies and frameworks previously established (in section 2) with well-known system failure analysis (e.g., what-if analyses, event tree analyses, HAZOP, fault tree analysis) could facilitate the incorporation of ethical considerations within the manufacturing sector.

The challenges involved in estimating system failures do require (1) an accurate definition of the problem, (2) identification of potential failure causes, (3) objectively evaluate the likelihood of each failure cause (including its impact analysis), and (4) implementation steps that define the approaches to prevent this failure causes from occurring, prioritizing those with higher impact and likelihood.

3.3 Risk Management as a source of trust

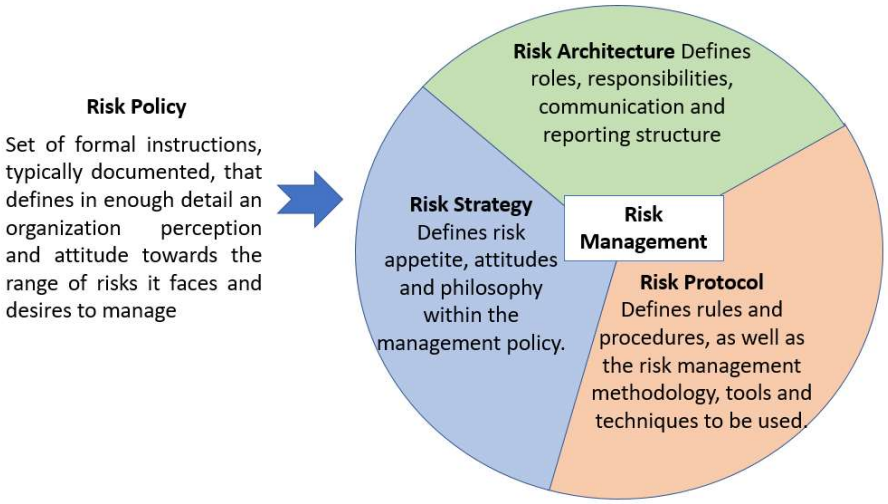


Figure 3 Risk Management Structure

In the present document, even though the general framework for the risk management process is presented with a scope in general applicability, a Risk Policy Definition, specific for ASSISTANT, is presented as annexes. This risk policy definition can set a precedent in managing ethical risk and, therefore, could be used to define some definitions incorporated into current risk management strategies or as initial risk policy definitions for companies.

As described in **Erreur ! Source du renvoi introuvable.**, risks could be correlate to ethical, legal, and robustness considerations. This definition implies that a suitable risk management process that contemplates these components will improve the AI asset's perception of their users and improve trust.

Even though legal components could be handled based on a risk management framework, legal constraints should not allow uncertainties in their range of applications. Therefore, their considerations should not be perceived with soft boundaries that would allow relative violations of the settled regulations.

Based on this perspective, the constructed framework would not consider legal requirements within the risk management process. Instead, legal requirements and definitions are recommended to be handled in the early stages of AI development. In the present framework, as described in section 5, current and future regulatory considerations of AI elements are tested initially following a pipeline evaluation structure that secures that the minimal constraints are fulfilled before starting the development stage of any AI element.

Robustness considerations follow two categories, one category related to technical robustness and another category to social robustness. The first, composed of the requirements for trustworthiness, will be covered together with the other ethical-based requirements. Social robustness corresponds to the considerations of present and future conditions and points of view from a social perspective that can drive modifications behaviours (e.g. acceptance or rejection), requirements (e.g. legal), policies, trends, and outcomes of the object under consideration. Social robustness can be achieved if “a strategy and its consequences on the fulfilment of needs are considered acceptable from different present and future points of view (perspectives)” [41]. However, the definition of present and future points of view consideration

place a heavier burden on social robustness considerations since, as explained by Beumer et al. [41], change in strategies tends to be costly and less effective.

It is mentioned that policies and technological solutions that do not enjoy widespread acceptance can lead to damaging rather than positive impacts [42]. It is essential to consider that social trends and conceptions are cyclical concerning the implementation of strategies. Therefore, whatever framework or strategy is implemented for managing ethical risks, its nature should be adaptable for modifications of social perspectives. For example, as described by Beumer, the globalisation process is shaped by and in return with cultural values, assumptions and policy discourse which have specific outcomes and impact on sustainability and quality o life which, at the same time, are shaped by the trends and processes involved in globalisation.

In terms of social robustness, the present framework possesses two characteristics that allow the “fulfilment of needs from present and future points of view”.

First, the framework developed focuses on incorporating ethical requirements and values within the risk management process. This consideration implies that future perspectives or requirements derived from social perspectives, ethical concerns, or definitions and regulatory frameworks can easily be incorporated within a comprehensive modification of the proposed framework. This characteristic is crucial since it allows the incorporation of ethical considerations derived from the domain in which the AI will be implemented (e.g. medical ethics). Furthermore, values defined by the different stakeholders (including companies or those derived by regional definitions) can also be incorporated as long as they do not contradict the legal requirements and ethical concerns established for the AI elements.

Ethical and values definitions often involve managing tradeoffs between principles that cannot be satisfied simultaneously. For example, some fatality rates are acceptable within most risk works environments with the beneficence of higher-paying loads; tradeoffs explain this between beneficence and non-maleficence considerations (common in different domains). Therefore, the framework constructed includes a process to define an optimal set of values derived from the perspective of several stakeholders, and that can be contradictory. Well-known decision-making tools are integrated for defining these sets (i.e. ANP and AHP processes).

Second, the framework developed is based on well-defined risk management strategies. Therefore incorporation of them in the manufacturing sector can be “considered acceptable” as a strategy since it would not impose a considerable change of strategies that the stakeholders can currently use.

Finally, from the risk management perspective, trust involves handling at least the ethical requirements established in the trustworthy guidelines.

Accountability corresponds to the fact or condition of being responsible. Ethics and governance are equated with answerability, blameworthiness, liability, and the expectation of account giving. In terms of AI, accountability places and distributes responsibilities within its life cycle. Therefore, it is reasonable to deduce that if a clear establishment of uncertainty and responsibilities are provided (or clarified) to the AI users, the AI developing stakeholders could be considered to act in an accountable manner. Furthermore, the users could better understand the AI assets outcomes (e.g. predictions or forecasts), improving trustworthiness. Thus, accountability and user trust are linked.

Based on Human agency and oversight, “ AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same

time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches. “. Oversight over the AI elements does, by itself, help in the accountability process and, at the same time, improve the trust of the system. Nevertheless, these approaches of enforced oversight do impose a necessity to make humans able to “ catch “ mistakes made by the AI elements. Human overseeing does not solve the problem [43]. One of the significant concerns about human oversight is that people's growing dependence on algorithms could erode their ability to think for themselves. This consideration is critical in fields in which the outcomes of potential risk are considerable (e.g. healthcare). A risk management process could help secure a constant oversight from users since it could enforce metrics and procedures within the AI use and implementation to keep track of the evaluation process, securing human agency and oversight and, thus, trust.

AI systems need to be resilient and secure based on technical robustness and safety. Thus, technical robustness and safety considerations are crucial for ensuring that fallback plan cases exist if something goes wrong and AI assets are accurate, reliable, and reproducible. The trustworthy guidelines state that a way to minimise harm is through technical robustness and safety. As previously mentioned, risk management does involve the managing of hazards. Thus a consistent implementation of risk assessment, which includes the definitions of treating and terminating risk conditions, would secure a constant improvement of the AI assets and their robustness and safety.

Based on Privacy and data governance, the EC sets regulatory requirements regarding data management and handling. The data protection regulations set rules for businesses and organisations and, at the same time, set rights for citizens regarding their data rights and redress. The risk management process can secure better security over data access and, at the same time, incorporate fallback planes in case a violation takes place. A constant evaluation process over the implemented security approaches would also allow updates over the existing trends that would affect data and, at the same time, the AI by itself.

Transparency indicates the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions. Therefore, it is necessary to determine responsibilities and hold the responsible people accountable. Transparency increases trust, as people have to trust and ground their faith on a sophisticated understanding of how algorithms work. Making the algorithms transparent allows stakeholders to criticize what is going on. A risk management process that secure that the algorithms involved run transparently and keep transparent to their users would allow, thus, an increase and secure an increase of trust.

Based on Diversity, non-Discrimination and Fairness (DnDF), bias is defined as the risk of a systematic error or deviation from the truth [44]. One important consideration is that bias is natural to humans, and therefore most of the social information and analyses performed could describe some form of bias. Bias could have multiple negative implications [45], but it is essential to recognize those linked explicitly to the marginalization of vulnerable groups and exacerbate prejudice and discrimination. A framework developed with such biases considerations should, in the first instance, evaluate the possibility that such biases are part of the information handled by the AI assets. Second, the AI's developers do not impose over the system cognitive biases that can drive the AI behaviour in biased directions. Third, estimate the risks that the outcome of the AI, in the form of forecasting or recommendation, could be used negatively or perpetuate biases.

Since it is well documented that biases can exist in complex historical data, AI-based risk scoring systems could perpetrate such biases. Some applications show significant disparities in accuracy – e.g. examination of facial analysis shows errors of 0.8 % for light-skinned men, while for dark-skinned women, the error rate is 34.7% [46]. The EC has foreseen this

consideration and has defined some High-Risk or prohibitive AI applications that include scoring systems.

Even though a risk management process would not eliminate the biases, its nature would allow setting mechanisms to prevent its prevalence. The risk management process will be considerably helpful for securing standards for High-Risk AI elements and could, maybe in the future, allow those considered prohibitive to change their condition to High-Risk.

Finally, based on Societal and environmental well-being, AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Regarding risk related to societal well being, AI has the potential to tackle some of the most challenging social problems. As mentioned by [46], there are at least 18 capabilities from AI that could be used to benefit society. They are linked to computer vision, natural-language processing, speech and audio processing, reinforcement learning, content generation, and structured deep learning with domains of implications that include equality and inclusion, education, health and hunger, security and justice, info verification and validation, crisis response, economic empowerment, public and social sector, environment, and infrastructure.

It is clear that the impact, and therefore the AI risks, would depend on the implementation scale. This consideration implies that evaluating and managing the risk involved in environmental, social and societal impact should be carefully considered. In the case of the manufacturing sector, the impact of AI assets embedded in products can be considerable. For example, autonomous vehicles will produce a modification to societal trends that should be considered at the moment of developing and deploying such products. Even though the importance of these considerations can be seen to affect society in the long term, the framework proposed in this work does propose to take into consideration these perspectives as a recommendation and within the general risk management process, nevertheless given the higher level effect, there is no clear identification of process recommended to manage and ensure sustainable protocols.

3.4 Standards and approaches for risk management

Organizations face external and internal factors that affect them, making achieving their objectives a process that includes uncertainties. However, managing risk in an iterative way helps in decision making in an informed way, allowing the organizations to settle strategies and achieve objectives more robustly.

Managing risk is part of all activities associated with an organization and, thus, should include stakeholders from the different domains. A critical consideration of risk management is external and internal context, including human behaviour and cultural factors [47]. These considerations are essential for objectives that incorporate ethics and values.

Different established risk management standards and frameworks recommend different structures to handle risk. Nevertheless, they are based on the same principles described later. The Australian Standards Body developed the first standard in 1995 [48]. Other countries, later on, followed this standard. However, even though the Australian standards were highly recognized, they were withdrawn using the ISO 31000 [49]. Other recognized frameworks include the ERM version of the Committee of Sponsoring Organization of the Treadway Commission (COSO) framework. It was published for the Internal Control-Integrated Framework (ICIF) and describes risk assessment processes, control activities, information and communication, control environment, and monitoring activities. COSO, given its roots, has widely been used in the USA. Additionally, the British Standards BS 31100:2011 “Risk

management – code of practice". The latest version of these standards explains developing, implementing, and maintaining risk management processes.

Finally, the ISO (International Organization for Standardization), a worldwide federation of national standard bodies, has continually released several risk management standards. The latest versions of these documents include ISO31000:2018, Risk Management - Guidelines, IEC31010: 2019 Risk Management - Risk Assessment Techniques and the ISO GUIDE 73:20009 Risk Management – Vocabulary. The first provides principles, a framework and a process for managing risk.

Even though there is a wide gamut of standards, an organization needs to select those that are more relevant to their particular circumstances. In this regard, The ISO is currently developing different standards with a specific focus on AI. These include the ISO/IEC JTC 1 family that focuses on creating 31 standards that are currently eight already being published. These includes:

- ISO/IEC 20546:2019 Information technology — Big data — Overview and vocabulary;
- ISO/IEC TR 20547-1:2020 Information technology — Big data reference architecture — Part 1: Framework and application process
- ISO/IEC TR 20547-2:2018 Information technology — Big data reference architecture — Part 2: Use cases and derived requirements
- ISO/IEC 20547-3:2020 Information technology — Big data reference architecture — Part 3: Reference architecture
- ISO/IEC TR 20547-5:2018 Information technology — Big data reference architecture — Part 5: Standards roadmap
- ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
- ISO/IEC TR 24029-1:2021Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
- ISO/IEC TR 24030:2021 Information technology — Artificial intelligence (AI) — Use cases

The backbone framework used for AI ethical considerations in Europe is the Ethic Guidelines for Trustworthy AI [5]. These guidelines put forward seven essential requirements that AI systems should be met to be considered trustworthy.

Even though there are no current legal regulations for the development, deployment, use, and decommissioning of AI elements (under the considerations of Trustworthy AI), it is clear that such regulations will be implemented soon. These regulations will be driven by recommendations made by groups that have actively discussed and developed general frameworks (e.g., the High-Level Expert Group on Artificial Intelligence also has the responsibility to recommend legal frameworks for implementation within the EU).

Independent of the specification of the Ethics Guidelines for Trustworthy AI as a global implementation framework, other components, methods, and frameworks will be used to consider the main framework in this work. To be more specific, an ethical risk management framework is proposed that considers the ISO31000 family, the trustworthy guidelines as ethical requirements, the white paper on artificial intelligence, the classification of AI elements based on the Artificial Intelligence Act, and different techniques that are supporting for the use of the framework. Given the importance of the ISO standard as a base for risk management, a description is made next. Finally, the full description of the framework is done in Section 5.

3.4.1 **ISO 31000**

The ISO 31000 is a standard that provides principles and guidelines for risk management. It can be seen as the framework of frameworks since it provides minimal considerations to develop risk management approaches applied to different types of risks (i.e. hazards, control risks, and opportunity / speculative risks).

As shown in Figure **4**, the general Risk Management framework is based on the considerations of principles, a framework, and a process. The principles guide effective and efficient risk management characteristics, communicating its values and explaining its intention and purpose. As described later, this documentary is expressed the risk policy, which is included in the Annex section for ASSISTANT. As described, the principles should be ***integrative*** to other activities and, therefore, in the case of AI, should be integrative to developing, deploying, using, and decommissioning in the manufacturing sector.

For the case of ***structured and comprehensive*****,** it implies that it should contribute to consistent and comparable results and, therefore, should include metrics that will allow measuring its integration into the processes or systems. Furthermore, it should be ***customized*** to the proportional level of risk, securing that the costs involved in the risk management process are level to the possible consequences. Thankfully, the EC has settled the risk levels first layer [8]. Therefore, the framework's risk management process efforts and considerations are based on this classification.

It should be ***inclusive*** of appropriating and timely involve stakeholders enabling the incorporation of the different knowledge from their area. However, this imposes, at least in the case of AI, a challenge since it will imply a combination of experts, at least for those AI with High-Risk, that have the domain the AI, risk management, the domain involved by itself, and understanding additional values that want to be incorporated on the AI assets.

It should be ***dynamic*** to emerging changing risks. This consideration allows the implementation of the current framework in a gamut of domains with the capabilities to be dynamic to values and ethical considerations (considerations given the domain of implementation - e.g. medical ethics) integration.

It should use the ***best available information*** (documental with historical and current data). The current stage of AI in the different domains is currently in an early stage of integration, and, therefore, there is still considerable space for generating historical information that could allow improvement in the risk management process. Furthermore, it should consider human and cultural factors and that, as described before, this is a positive trait to incorporate values and ethical considerations within the frameworks. Finally, it should be continual and therefore allow a ***continual improvement***, making the probabilities of risk materialize lower over time.

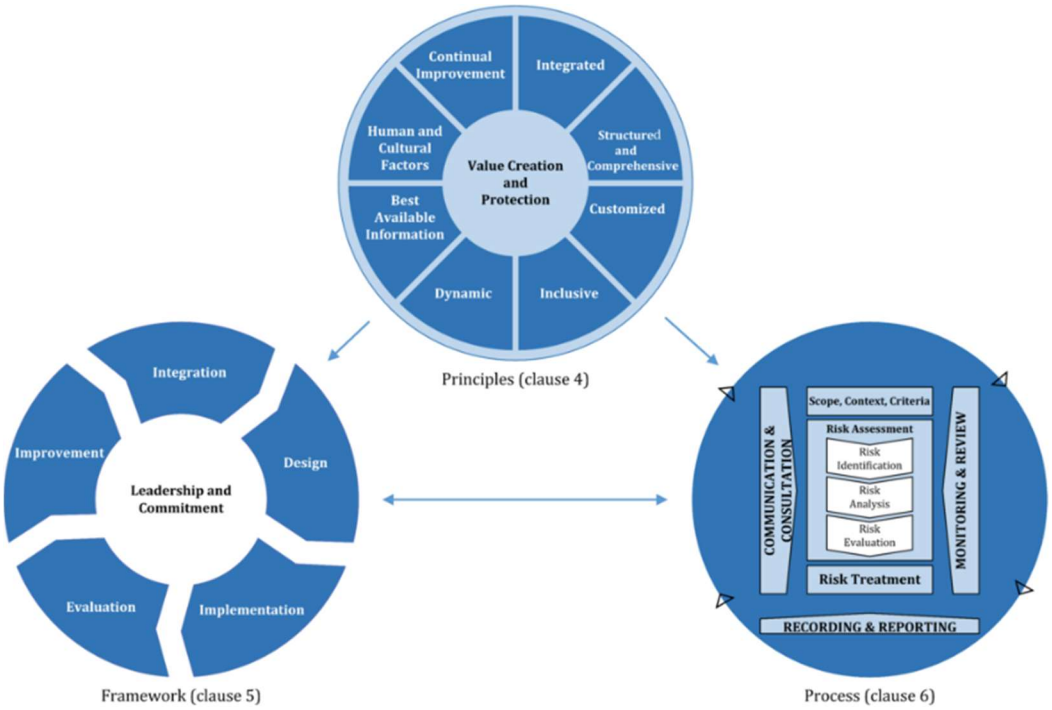


Figure 4 ISO 31000 General Framework

The ISO 31000 framework helps integrate risk management into organizations' activities and functionalities. The ***leadership and commitment*** ensure that risk management is integrated into organization activities. An exemplification of leadership and commitment is given in the ASSISTANT policy specification. In it, the responsibilities are defined for each contributor. In addition, the ***integration*** framework specifies that risk management relies on the organizational structure and, thus, should be dynamic.

For the case of ASSISTANT, the structure is organized based on the internal assistant structure, and therefore, it uses the current organisation/hierarchy and extends responsibilities based on it. The ***design*** consideration in the framework defines that risk management should consider the internal and external context that includes external considerations: **social, cultural, political, legal**, financial, technological, national, regional, and environmental factors. Internal considerations include vision, mission, **values**, strategy, policies, organization culture, **standards**, capabilities, **data**, relationships with stakeholders, contractual relationships, and interdependencies, including communication, commitment, roles, resources, and accountabilities. Bold words marked some of the factors incorporated in the framework specified for ASSISTANT.

Given how the ethical risk management process is formulated, the other factor previously described can easily be integrated into parallel, as described in section 5. The organization should define methods for implementing the risk management framework—an exemplification of these specifications is given in the ASSISTANT policy in the annexe section.

The ***Evaluation*** implies that both risk and the framework should be measured to check the suitability of the approach used. This consideration implies the necessity of KPIs to measure the state of the management of risk conditions and, at the same time, a feedback process to check the risk management framework performance against its purpose, implementation plans, and expected behaviour. These KPIs also improve the organization by continually monitoring and adapting risk management.

Finally, the ISO 31000 Process involves a systematic application of procedures and practices that establish the application of assessing, treating (that implies the application of

the 4T's of risk management – treat, tolerate, transfer, and terminate), monitoring, reviewing, recording and reporting. As specified in the standard, there can be many applications of the risk management process within an organization and, therefore, suitable for different processes involved in the life cycle of an AI. Notably, the risk treatment options are not necessarily mutually exclusive or appropriate in all circumstances and depend on the institution's risk appetite that should, for AI elements, consider regulatory considerations.

The main components of the ISO 31000 process, as shown in Figure **5**, are: (1) risk identification, which focuses on recognizing and describing risks, (2) risk analysis which focuses on comprehending the nature of the risk and its characteristics, including the sources, consequences, likelihood, identification, scenarios and control of risk, and (3) the risk evaluation, that consider the comparing the results of the risk analysis with the established risk criteria to determinate actions.

Notably, different techniques can be used to perform (1) and (2) that depend on the interest of the company/user/stakeholders, the standards used for the processes, and the ability to use the information that exist to perform such task.

For example, the ISO 14971- Medical devices – application of risk management to medical devices specifies that the risk management process involves nine steps, while the IEC 60812 is based on an Effect and Criticality Analysis (FMEA) approach, which defines only six steps. In fact, some standards are contradictory and can lead to violations of other approaches in comparison [50].

Some of these techniques include the Root Cause Analysis (RCA), the Strenght, Weakness, Opportunities and Threats analysis (SWOT analysis), the Delphi technique, the Failure Mode and Effects Analysis (FMEA), the Failure Mode, Effects and Criticality Analysis (FMECA), among others. A list of some relevant risk analysis processes (except for FMEA, which is used fundamentally in the current framework) is shown in the following table.

The actions are intrinsically dependent on the component evaluated under the risk assessment; nevertheless, these actions seek for hazards:

Table 1 Risk Analyses Process Approaches

Type	Description
HAZOP	A Hazard and Operability (HAZOP) study is a qualitative, structured, and systematic examination of processes. HAZOP follows a top-down examination between events and their causes.
Root Cause Analyses	RCA is a process designed to investigate and categorise the fundamental cause of an event with safety, health, environmental, quality, reliability, and production impacts. This analysis helps to identify how and why something happened and develop recommendations
Event Tree Analysis	Event Tree Analysis Is a technique to evaluate the sequence of events. The analysis is performed by creating event trees that follow a logical sequence. The objective is to determine if there is sufficient control of the system and procedures.
Bow Tie Analysis	This approach graphically displays the relationship between hazardous events, their causes and consequences and the risk control barriers to stopping the accident sequence.

- To avoid the risk by deciding not to start or continue with the activity (Terminate)
- To Remove the risk source
- To Change the likelihood of occurrence
- To Change the consequences
- To Transfer the risk (e.g. through contracts or buying insurance)

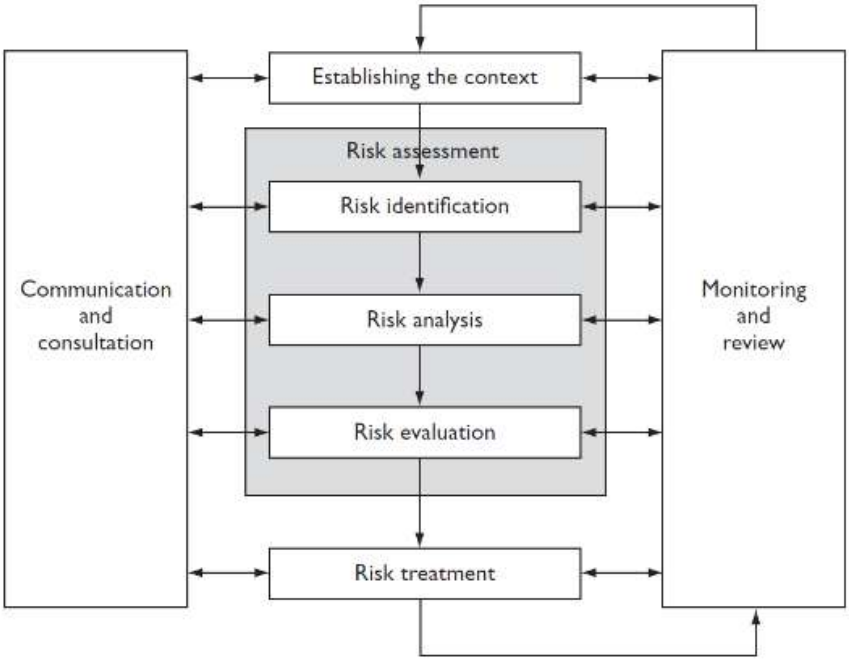


Figure 5 ISO 31000 process

3.4.2 **Communication and Decision-Making**

Communication within the risk assessment process can be seen as dependent on the hierarchical level involved of the stakeholders. For example, at the corporate level, this might involve the identification of risk issues through dialogue with key stakeholders and partners; the communication of priorities set by the organization as to which risk issues will be analyzed; dialogue on the allocation of scarce resources, among programs, analysis and capacity building in the organization, expected benefits of activities, among others.

At the policy and program planning level, risk communication will likely be focused on the validity of the data used, the applicability of the models used, the identification of stakeholder benefits and concerns, evaluation of treatment options, and likely biases in the analysis.

At the operations level, risk communication will be focused on implementation details for new programs regulations. For example, the ongoing effort to modify behaviour (e.g., promote healthy lifestyles), explain how to use products to the best advantage, and comply with regulations.

Many risk issues are identified based on Monitoring and Quality Control performed periodically. However, this does not imply that all the risks will be treated since several factors will strategically define the most critical risks to be managed first. Moreover, in terms of ethics and values, several of them are contradictory or, in some cases, perform synergically under companies' goals. Therefore a decision making process is required to be incorporated into the framework (The incorporation process will be defined later on, but we will recommend the implementation of ANP/AHP approaches, seen in the methodological section, to select the most relevant e-risk components based on regulatory considerations and values).

The DecisionMaker will select a limited number of e-risk issues for Preliminary Analysis (Identification) (the remainder will be set aside for several reasons, including, among other considerations, low risk, no feasible treatment, few benefits and lack of jurisdiction).

After obtaining further information on risks in the Preliminary Analysis (Identification), the Decision-Maker will establish the Context for the risk issues that have not been set aside. If necessary, the Decision-Maker will return to Preliminary Analysis (Identification) for additional information, and this iterative cycle may be repeated until there is sufficient information to decide to commit resources.

Once each risk issue that has been selected for further consideration has a proper context, the issue then proceeds through the entire risk management process. The Decision-Maker may set the risk issue aside, go back for more analysis, or select a treatment option for Implementation. Following implementation, the whole system returns to the ongoing Monitoring and Quality Control, which will generate new risk issues and new opportunities. Stakeholder Relations (e.g. risk communication and public involvement) are ongoing throughout the process, at a level determined by the decision-maker depending on the urgency of the issue and the resources available.

4. Methodology for Trustworthy AI management in industrial environments

This section describes a specification and definition of techniques used in the proposed framework. Additionally, a definition of ethical risk is formally given to clarify the scope of the techniques.

4.1 Ethical-risks (e-risks)

Risk is a general concept that represents a combination of probabilities or likelihood of an event to happen and the outcome (positive or negative) that this event has over the systems. There are several definitions, and some of the most remarkable are described below.

PRINCE2 glossary of terms – A risk is an uncertain event or set of events that, should it occur, will affect the achievement of objectives. Risk is measured by combining the probability of a perceived threat (i.e. opportunity of occurring) and the magnitude of the objective‘s impact.

ISO31000 – Risk is an effect of uncertainty on objectives. Note that an effect may be positive, negative, or a deviation from the expected. Also, the risk is often described by an event, a change in circumstances or consequences

Institute of Risk Management – Risk is a combination of the probability of an event and its consequence. These can range from positive to negative.

One important consideration is that an event's outcomes can be positive or negative if materialised, as stated by the Institute of Risk Management. A classification given by the nature of the outcome and its occurrence is normally used. Risk is classified as Opportunity or Speculative if the outcome can be positive or negative, implying an intrinsic nature related to investment, marketplace, and commerce.

Risks are classified as Control if their nature is related to uncertainties of process and procedures. This consideration implies that they are related to budgets, timeframes, and management, where the outcomes can lead to different results. Independent of this, the objective related to this type of risk is to minimize the potential consequences of materializing risk conditions (e.g. project management). Finally, risks are classified as hazards if the only

outcome they can have is negative. They can be considered operational or insurable risks and always have a level of tolerance intrinsic to them. They are related to processes, dependencies, management and, in general, to any area in which the regular operation of the system is disrupted, the operational costs are increased, or there are adverse legal and social outcomes linked to the materialization of the risk condition.

In terms of Trustworthy AI, the ethical requirements imposed on AI, the values that would like to be branded on them, and the social, societal, legal, and environmental constraints should, among other considerations, should be considered as ethical objectives of AI assets functionalities. In addition, several conditions, processes, and statuses with different probabilities or likelihood of materializing can cause damper or restrain the expected AI behaviours. We call the combination of these events' probabilities to materialize and the impact on the AI objectives ethical risks or e-risks.

We foresee that this e-risk can only have negative consequences if materialised, and therefore, they should be considered hazards. Therefore techniques for risk management that are recognized for hazard management could be modified to handle these risks.

For the case of ASSISTANT, the e-risk considered would be those considered relevant for the manufacturing sector, emphasising the case of studies incorporated within the project. Therefore the intrinsic values of the industrial partners would be considered to be included in the ASSISTANT framework (Not in the general framework described in Section 5). Furthermore, the framework is constructed based on a European perspective; its considerations regarding the ethical concepts and the legal AI considerations (e.g. Trustworthy requirements [5] and constraints [8]) are implemented within the framework.

4.2 Fuzzy logic

Fuzzy logic can be seen as an approach to computing based on degrees of truth. Fuzzy logic scape the traditional boolean logic and focuses on approximate human reasoning by including different levels of reality between the classical yes and no representation of a given process's inputs and outputs. A fuzzy system behaves like a black box (not in the sense of an AI black box since it follows well defined and pre-established rules) for processing the mapping of input space into an output space. The benefits involved in fuzzy logic include that it is conceptually easy to understand, is flexible to extensions, is tolerant to inaccurate data, can model nonlinear functionalities with arbitrary complexity, is based on expert judgment, and can be blended with conventional techniques (e.g. fuzzy-controllers, fuzzy-ANP and fuzzy-AHP), and is based on natural language, making its implementation transparent to users.

The first component of understanding fuzzy logic is understanding the concept of membership functions. Figure 6 shows how to transform a variable to a level of significance in a linguistic shape under a classical boolean perspective (upper figure) and a fuzzy logic process (bottom figure). Since the classical perspective only allows two levels, a person can be classified as tall or not tall based on height. Under the fuzzy perspective, several member levels are classified based on their heights.

The fuzzification process implies converting a real variable into a level of belonging of a linguistic variable (e.g. tall, average, positive, or others). The fuzzification process is against a specific function " shape ", which describes the logic behind the degree of pertinence to this function. These functions are named membership functions (MF). The MF does include the boolean operators, but they also allow to integrate functions such as gaussian shapes, linear tendencies, logarithmic based shapes, or any functionality that, under an expert's perspective, represents the behaviour of the variables.

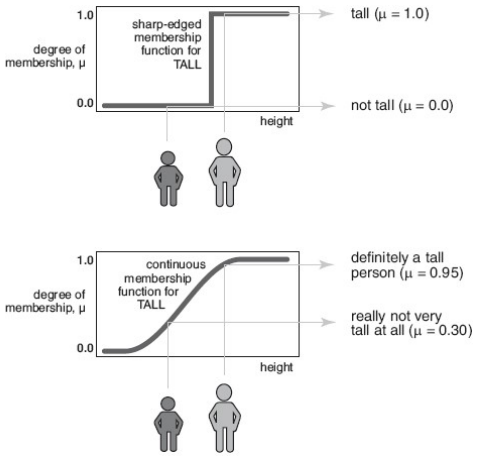


Figure 6 Membership representation in boolean and fuzzy processes

Figure 7 shows the overall fuzzy logic process when more than one variable is evaluated simultaneously (i.e. there are several fuzzification processes over different membership functions that can or not belong to the same linguistic variable. As observed in the figure, there are six fuzzification processes, one over the linguistic variable service (left column) and another over the linguistic variable food (following column). These transformations are over different MF or, in some cases, there is no MF to evaluate the variable (see middle row second column).

The following process involves mapping a transformation of the input variables into an output linguistic meaningful system. Therefore, the fuzzified system in ***if .... then*** evaluations using alternative operators such as *AND* and *OR* are performed. As observed in the figure, this evaluation maps into another linguistic variable (in this example tip) connected to each evaluation process (each row of the figure). Each of these evaluations produces a relative belonging or existence of the aggregated linguistic variables (fourth column of the figure). The aggregated result is then defuzzified to produce a practical numerical value for interpretation.

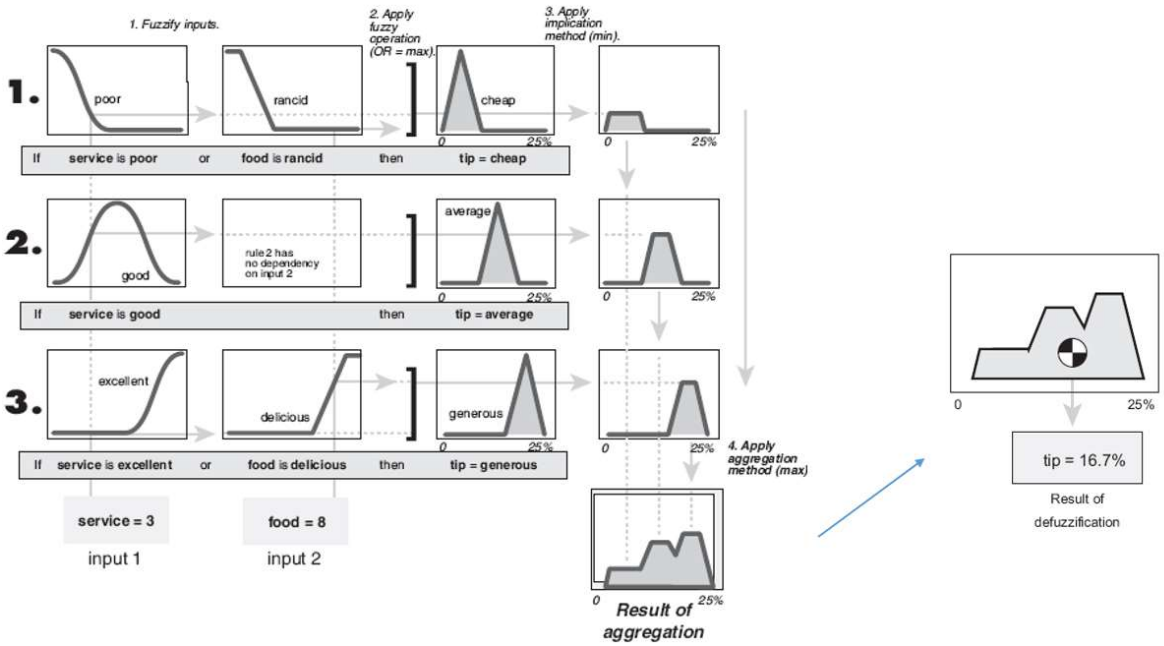


Figure 7 Fuzzyfication, Evaluation, Aggregation, and defuzzification.

As a framework tool related to Trustworthy AI, we have foreseen that fuzzy logic could have two main contributions to the system.

First is the possibility of being combined in decision-making processes (ANP and AHP, as will be explained next) in order to improve the representability of uncertainties and, at the same time, map systems that cannot be difficult to represent stakeholders' point of view at the moment of defining values systems to be incorporated on AI elements.

The second use of fuzzy logic is incorporating it as an ethical-by-design approach. Just recently, Sholla et al. [51] used a neuro-fuzzy system to define actions in function of input conditions. In this work, each of the processes previously defined (fuzzification, evaluation, aggregation, and defuzzification) is performed by an artificial neural network hidden layer, whit the advantage, or possibility, of giving some feedback information (i.e. training information) to construct a map of ethical representation.

The idea behind this approach is to allow the incorporation of ethical safeguards into the system that will be highly automatized and thus, construct a fuzzy-logic based approach with well defined constrain over the system (i.e. not trained from data) that will allow the AI to perform tasks in an automatized way. This consideration would be crucial in a system that Human – in –loop, human–in–control is not possible to be incorporated given the nature of the system (e.g. fast dynamics) but have high considerations over their human agency and oversight requirement.

A thorough definition of this approach is incorporated in the Annexe section and included within the framework pipeline constructed here.

4.3 ANP-AHP

In general, it is possible to confirm that the logistic management levels require a constant measurement of the objectives established by each unit that conforms to the Company. However, the definition of what objectives are considered strategic would be given by the critical analysis of experts. Therefore the uncertainties and biased strategies could be present in each definition.

The Analytic Network Process (ANP) proposed by Saaty [52] is a generalization of the Analytic Hierarchy Process (AHP). Its formulation is based on the priority generation or relative importance of the elements belonging to a complex network model by considering interdependencies between the elements.

One important characteristic is that ANP and AHP processes can be combined with fuzzy logic to deal with uncertain data and imprecision knowledge. Then, when the system process considerably inconsistencies, the fuzzy system is incorporated into a named Fuzzy-AHP or Fuzzy-ANP process.

The AHP process solves the decision-making process by separating the problem into three parts. The first part corresponds to an issue that is desired to be resolved, the second part is the alternative solutions, and finally, the criteria used for finding the solution. In so doing, there are fundamentally five steps in the AHP process.

The first step is to define the alternatives. These alternatives can possess different criteria (e.g. aroma, colour, velocity, or other features innate to the problem to be solved) that the solution should consider.

The second step is to define the problem and criteria based on AHP and ANP problem conceptualization. A decision making problem can be seen as a compilation of subproblems

(i.e. divide the problems into a hierarchy of subproblems). As the problem is divided ins smaller systems, consistency could be lost in subjective considerations.

Combining the first and second steps allows the creation of a hierarchical set problem represented as a networked system. To accelerate the understanding of this tool on ethical considerations, Figure 8 shows a diagram of the objective of setting the most suitable values and ethical considerations over a system. In this example, and at the first level, the objective is to set the most suitable values and ethical considerations within the risk management process.

At the second level of Figure 8, all the criteria that could impact the implementation, budgeting, legality, or impose a limitation or a preference of one value are defined. In addition, these criteria could have sub-criteria that will add additional hierarchy to the system, thus an additional layer that should be connected to the value, ethical considerations, and requirements, as seen in the figure.

In the third level of Figure 8, the alternatives are defined. Here a complete set of values, requirements, or ethical perspectives subject to the criteria are considered. Each of the values, requirements or ethical considerations (from the implementation domain – e.g. medical ethics) does need an associated numerical value within the criteria. For clarity, the values, ethical considerations, and requirements are not connected in the figure, but in reality, they are. A critical concept for AHP is that its process only allows hierarchical analysis, and therefore, it does not allow interdependencies between values (set in the red arrow in the Figure). If such dependencies exist, a networked-based analysis is required (i.e. An ANP evaluation).

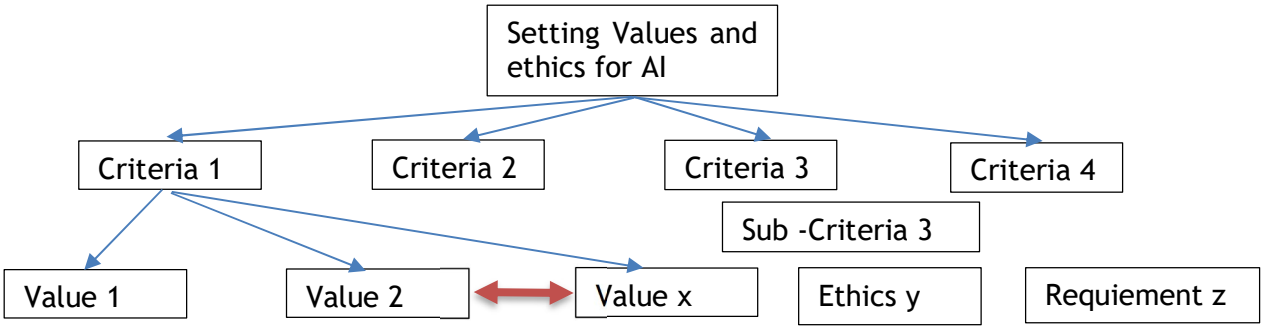


Figure 8 ANP/AHP exemplification

The third step is to establish priority amongst criteria using pair-wise comparison. The pair-wise comparison allows the creation of a valuable matrix for representing the system and performing the mathematical analysis. This matrix describes the relative importance of the different attributes concerning the system's goal (e.g. how critical is criteria one over criteria 2 for setting the value system).

The matrix has a predefined importance scale, as shown in the following table. An exemplification of the constructed matrix is shown in Figure 9. One characteristic of the ANP and AHP is that they allow the integration of perspectives of several stakeholders and, therefore, would be crucial for its implementation in Trustworthy AI evaluation with several perspectives that should be taken into account.

Once the table is constructed, it is normalized by dividing each value by the column sum (

Figure 10).

Table 2 Importance Level Scale

Scale	Importance Level
1	Equal Importance
3	Moderate Importance
5	Strong Importance
6	Very Strong Importance
9	Extreme Importance
2,4,6,8	Intermediate values
1/3, 1/5, 1/7, 1/9	Values for inverse comparison

	Tech	Econ	Social	Enviro
Technological	1	a	b	c
Economic	1/a	1	d	e
Social	1/b	1/d	1	f
Environmental	1/c	1/e	1/f	1

Figure 9 Criteria Evaluation Matrix.

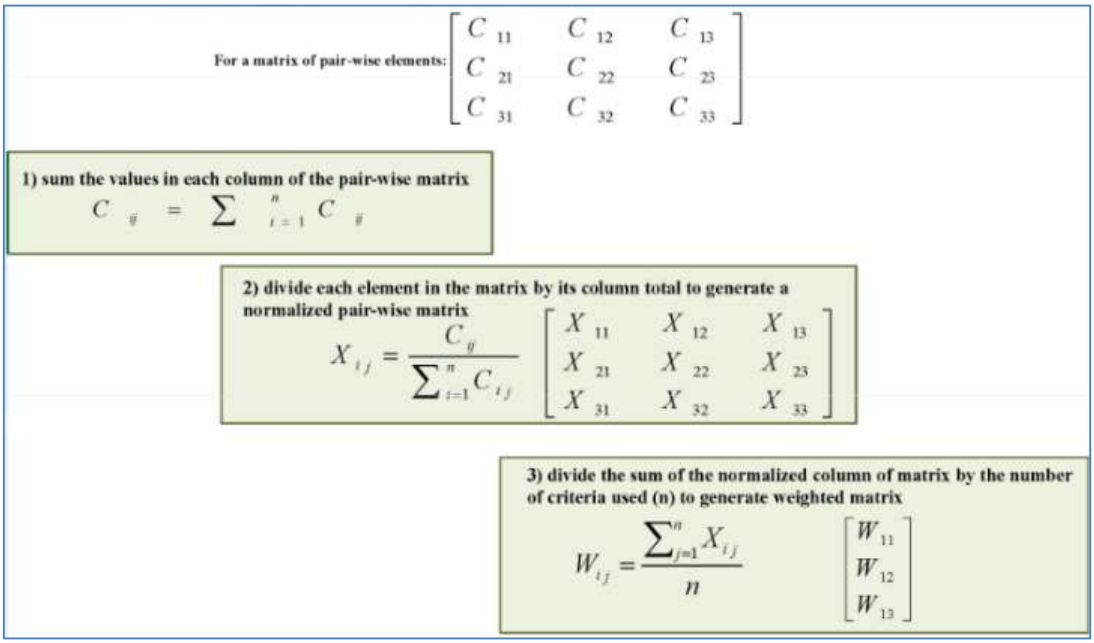


Figure 10 Normalization Process

The fourth step consists of getting the relative importance or weights of the elements constituted by the different criteria. Each row of the normalized matrix is averaged (for reference, we call this the averaged values the weighted factors). This construct gives us the relative criteria, or weight (Cw), that specify the more critical system criterion.

The fifth step is to evaluate the consistency of the information accumulated once the problem is subdivided into different criteria. First, the pairwise constructed matrix (not normalized) is elementwise multiplied by the criteria weight constructed (columns wise per weight). These values are next row-wise summed up to determine a consistency criteria value (λ_{max}) averaged between all the resulting ones. Finally, the overall consistency, consistency ratio, is estimated using the following equation, where *n* represents the number of criteria under consideration and *R* is a random index to correct for the number of criteria used in the analysis (R=0, 0, 0.58, 0.9, 1.12, 1.24, 1.32, 1.41, 1.45, 1.49 for 1,2,3,4,5,6,7,8,9, and 10 criteria, respectively).

$$\mathrm{CR} = \frac{l_{max} - n}{(n-1)R} \quad (1)$$

As a standard, if CR is lower than 0.1, it can be defined that there is consistency within the constructed matrix.

The previous process has only been used for defining the relative importance or weights of the criteria of the system but in order to define the most suitable values, requirements, or ethical considerations.

The process must be combined by performing a repetitive process of those previously described over each criterion with a pairwise comparison of the values, requirements or ethical considerations (e.g. define if ***Value 1*** or ***Value 2*** is more critical concerning criteria 1). This process will create different relative importance of weighted values ($C_{v,c}$) that can be used later on to define the essential values *v*, under the criteria *c*, to be incorporated into the system

Each of these analyses will create a set of weighted factors combined to create combined criteria–to–alternatives matrix. The previous matrix is element-wise weighted by the criteria weights (C_c) (i.e. C_c*$C_{v,c}$ $\forall$ v in each c) to construct a final matrix. In this matrix, the column sum gives us the relative importance of the system values. Therefore, this matrix gives us a decision-making approach for selecting the predominant values incorporated within the risk management process.

This result implies that no interdependences exist on values and that the system can be represented only as a hierarchical process. In case this is not feasible, the ANP process must be performed. The readers are encouraged to check the following references if an ANP process is required [53].

Finally, the same approach named here can be used to select, within the risk management approach, the most suitable treatment options (if more than one) for e-risks. Again, this case will allow considering criteria such as budgets, policies, and others that could impact incorporating risk management treatment options.

4.4 Failure Mode and Effects Analysis (FMEA) and Failure Mode, Effects, and Criticality Analysis (FMECA)

The Failure modes and effects analysis (FMEA) is a well known and documented engineering activity that supports fault-tolerant design, testability, safety, logistic support, and related functions. The FMEA tools analyse potential failure modes within a system to determine the impact of those failures.

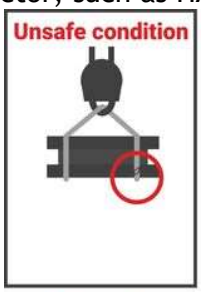
When evaluating alternatives to analyse risks, top-down (also known as a functional approach to risk) and bottom-up (also known as hardware approach) approaches can be used. Typically, the system complexity and data availability will define the approach used. The hardware approach is used when a system concept has been decided. Each component on the lowest level is studied one by one. The analysis is considered highly complete since all components are considered and evaluated. To facilitate this process, inductive questions, such as **What happens if?** are used for the analyses.

On the other hand, the functional approach recognises that each item has several functionalities classified as outputs. The output conditions to be produced are the core analysis.

Figure 11 shows the two perspectives that can be used for evaluating risk. The First alternative (option A) focuses on evaluating each component's action, condition, or status and, based on the collected information, defines the risks and, more specifically, the hazardous conditions. In this scenario, a bolt (a system component) is evaluated under all its failing "forms", or Failure Modes, that can produce loose and fall of the retained objects and thus lead to a hazardous situation. The Failure Modes, in this case, could include, for example, material fatigue, incorrect positioning and inappropriate positioning. This inductive approach from the elements to risk corresponds to a bottom-up approach directly linked to the FMEA.

The other alternative, the top-down approach, corresponds to the deductive approach of seeking the reason for a hazardous situation. In this case, the accident (Figure B). In this example, the accident has to be explained by its causes, related to the bolt breaking caused by several modes. This deductive approach is intrinsic to methods well known in the manufacturing sector, such as HAZOP and HAZAN.

(A)



(B)

Figure 11 Bottom-Up and Top-Down Risk Evaluation processes

Even though both methods could be used for performing risk assessment, the Bottom-Up approaches are recommended in the design process from concept throughout development (i.e. FMEA) [54]. This decision could be fostered by considering that the top-down approach requires extensive data (including previous risks to materialize) and the system's knowledge. Nevertheless, when complex systems are analysed (as we consider the case for ASSISTANT), a combination process (i.e. hybrid between top-down and bottom-up approach) is recommended. Based on the ASSISTANT architecture, a combination analysis can be performed by first analysing each component within the general architecture and then following the analysis based on increasing and decreasing system hierarchy and functionalities. In other words, a component within ASSISTANT should be evaluated as to their impact over hierarchical higher system elements and, at the same time, evaluate the functionalities involved in the components (e.g. tool) involved in the analyses.

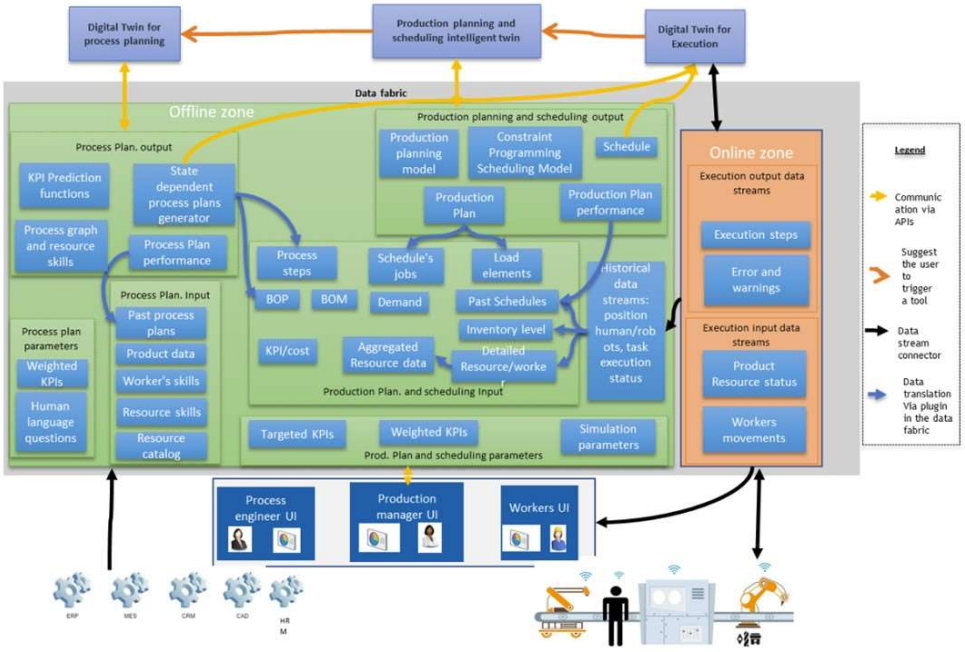


Figure 12 ASSISTANT Arthictecure Referencing Diagram

The US Department of Defense first developed the FMEA for systems design. This approach is defined as a standard, "Military Standard (MIL-STD-1629A), Procedures for Performing a Failure Mode, Effects and Criticality Analysis – Department of Defense, United States of America, and was adopted by commercial industries as an approach to reduce failing conditions, reducing their impacts (social, economical, and environmental).

An analytical process performs the FMEA to identify system weaknesses in all expected system functionalities and modes. A multidisciplinary expert team usually carries out the FMEA over a well-delimited scope and boundaries system. Information used to carry out the FMEA is diverse and SHOULD include schematics, procedures, manuals, systems configurations, and methods. In addition, the expert team evaluates and defines potential failure modes, their detection methods, their effects, and the corrective actions needed throughout a brainstorming process. Finally, based on the results, recommendations can be made using a ranking system that considers the risks considerations (i.e. severity and probabilities of occurrence). The general elements of the FMEA process are described next. Nevertheless, the specific definitions of FMEA for the framework usage and the Ethical-Based FMEA and FMECA processes are defined in other sections.

To correctly perform an FMEA process, the standard steps involved in it include:

4.4.1 Preparation:

A sound preparation of the FMEA, or any risk analysis, reduces the time and effort taken to perform FMEAs. Therefore the preparation of the FMEA considers:

a. **Methodological Understanding**: The team involved in the risk management process should understand the methodological approach involved in the FMEA. For that reason, different standards and guidelines can be used for preparation [54] [55] [56].

b. **Scope Definitions**: Generally speaking, the FMEA will depend on the type of system to be implemented and the goals set by the owners and the stakeholders. Since the current framework focuses on evaluating e-risks and minimising the event or their consequences, a homogenization of the technique is possible. Therefore, as the process of performing e-risk is

updated and applied in more manufacturing companies, it will be more straightforward and better-defined strategies to specify characteristics of the domain, which could include: failure modes, failure criteria and types, physical, functional, and operational boundaries, depth of analysis, operational philosophies and risk appetites, and criticality ranking (if FMECA is performed). Clearly, the Scope should be defined before performing any analytical process. These should be initially defined by the e-Risk management committee and provision with enough information to the e-Risk board. Notably is the understanding of the failure criteria and types that define (1) if failure modes are single or multiple (i.e. a failure mode can have many objects), (2) hidden failures (i.e. those failures that are not well defined but know their occurrence), (3) common-cause failures, treatment of unavailable systems, (4) failure of passive and active components, and (5) external factors (from the system and the parts under evaluation – interconnectivity with other software components and UIs, for example).

c. **Team Definition**: There are two ways to address FMEA. The first is using a workshop in which experts with updated information analyse the system. The alternative is to use external practitioners to develop the FMEA. Given the nature of the e-Risk management group, it would be required for a manufacturing application of AI components to have: (1) Experts with knowledge on FMEA techniques (2), Experts with knowledge on risk assessment and risk management (3), Experts on AI, (4) Experts on regulatory considerations and trustworthy AI, (5) Experts of the manufacturing process or system in which AI is deployed, and (6) experts on the data managed by the AI systems. As expected, one expert could cover more than one of these needs. Nevertheless, as the number of experts in the areas increases, the number of failure modes detected for the different components should increase.

d. **Ideal Timing and FMEAs conduit:** Given the nature of the FMEA approach (as an inductive/top-down approach), it is advantageously applicable in the early stages of design, allowing to catch design or system issues. An early implementation of FMEA allows safer design, indirectly tackling one of the AI trustworthy requirements, and economical by eliminating costly retrofits or systems upgrades post-build [57]. For ongoing systems, management needs to be aware that changes may be required arising from the FMEA recommendations and that enough funds and time must be available to meet the modifications. Once the modifications are performed, or regulatory processes are enforced, the current framework should be performed again. The extent of the recommendations will strongly depend on the maturity of the system developed and the changes involved.

As typically performed in industrial practices, FMEAs, are continual processes held in living documents. This information should be kept and maintained throughout the system's life. The reports should be updated to reflect the latest information and system status, including system upgrades, configuration modifications, or operational setup changes. The finding of the FMEAs should be incorporated within systems operations, manuals, emergency and training [57].

Even though the FMEA, or similar approaches, has been recognized as a design tool, effectiveness depends on proper communication for early design attention. As mentioned in MIL-STD-1629A [58], the most significant criticism of the FMEA is its limited use in improving designs, which is driven by poor inputs to the design process and time factors.

4.4.2 **Developing the FMEA:**

a. Data Management: Data management involves the processes of data collection, previous or other risk analysis performed so far, and initial data analyses of the previous processes.

b. FMEA study: The basic FMEA structure process is outlined in the following Figure. One important consideration that should be included, independent of the stage, is the technical,

physical and social concepts within the whole process. Instead of focusing only on each technical perspective (usually driven by mechanical or control systems in the manufacturing sector), each social aspect that could impact each hazardous situation (i.e. e-risks) should be considered. The first component, **Define The Analysis**, focuses on system physical and operational boundaries. These definitions set the system boundaries, and defining, in the current pipeline, some definitions based on ethical requirements and regulations are enforced to secure the considerations of Trustworthy AI within the system.

Further considerations should be driven by the scope and depth of the analysis, system functions, interfaces, expected system performance, system constraints, and failure definitions. Since some of the system boundaries will be regulated by external components (e.g. artificial intelligence act), an understanding of local regulatory conditions that are not foreseen in the current status of the framework should be considered in this stage if not implemented within the framework pipeline process. As covered in the framework presentation, several components are used before the FMEA process to Define The Analysis; therefore, this step will be enforced in the current framework.

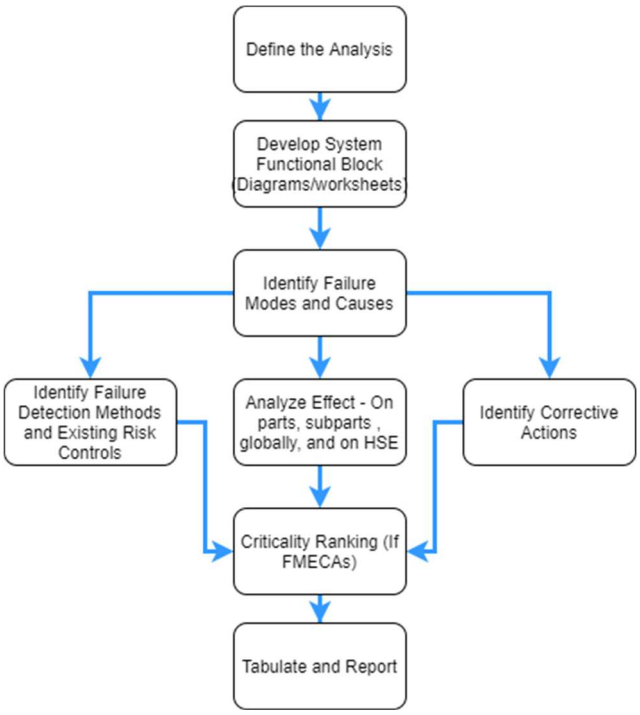


Figure 13 FMEA General pipeline (extracted from [55])

The second component, **Develop System Functional Block Diagrams**, focuses on constructing functional block diagrams, reliability block diagrams, and failure mode worksheets to help analyse the system. Failure Mode Worksheets are described in the following sections. At the same time, the developed system functional block diagram has been merged with the first step and other ones since these processes could be considered standard at the moment of performing an FMEA study. Ideally, the diagrams should provide a high-level hierarchical system structure to understand dependencies. If these are specified as functional dependencies, the diagram is known as Reliability Block Diagram (RBD).

For example, in the RBD information of interconnections in series or parallel structure, the series indicates that if any components fail, the whole system connected in series fails, while those connected in parallel can still be run.

The system functionality block is a requisite for performing the risk management process to facilitate this stage in the current framework. i.e. the hierarchical structure is not linked exclusively to the FMEA process. For example, an exemplification of an RBD is shown in the following figure. As seen on it, if element B1 fails, the overall process can still occur by running it throughout element B2; nevertheless, if Element A or C fails, the process cannot continue.

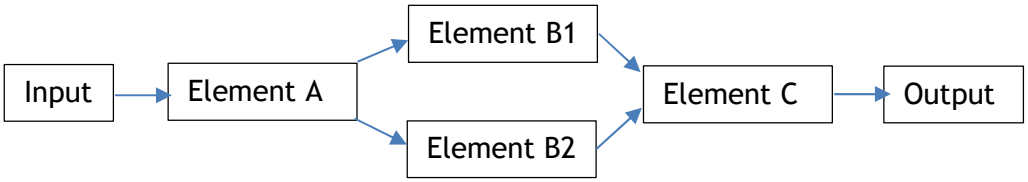


Figure 14 Reliability Block Diagram (RBD) exemplification

A list of supporting documents should be given to execute a correct FMEA analysis based on the level of detail needed for a correct implementation. This level of detail should be proportional to the intrinsic risk level of the components involved (i.e. high-risk components should have comprehensive coverage of information at hand). These documents could include:

i. FMEA worksheet and previous FMEA analyses
ii. System/AI boundary description
iii. System/AI design specification
iv. RBD
v. Safety, security, safeguards, and control system details
vi. Cause and effect matrix
vii. Operating procedures manuals
viii. Emergency procedures
ix. Maintenance, inspection, and testing procedures

Among these previously mentioned documents, the cause and effect matrix can be linked to the safety, security, and safeguard documentation since it helps identify possible causes (more than one). In a Cause and Effect Matrix, each vertical axis (rows) lists system deviations' possible causes. The system responses, and their effect, are listed in columns across the top. Although tools such as this one will help identify technical components (e.g. robustness and safety); They are not expected to drive solutions to non-technical ones, societal well-being. Independent of this, general structures can be used for ethical considerations in which the columns will be replaced by each of the 11th elements defined as **Ethical-Based General Failure Modes** later on. This list can be further expanded as a thorough understanding of the system's inherent ethical risks and the values to be incorporated into the system.

The third step is named **Identify Failure Modes**. Failure modes are determined based on relevant data and functional elements outputs (main component to analyse [55]). There are common failure modes considerations that include, among others

i. Premature or spurious operation
ii. Failure to operate when required
iii. Intermitent operation
iv. Failure to stop operation when required
v. Loss of output or failure during operation
vi. Degraded output or degraded operational capability

Even though these are generic failure modes for systems and equipment, specific failure modes are described in the domain of interest. Failure Modes analyses are standard in software

development, and consideration from linked topics can be translated almost directly to AI. For example, the following tables describe two well descriptive failure modes for IT security based on intentionally motivated and unintended failure modes [59].

Table 3 Safety Failure Modes – Intentionally Failures

Intentionally Failures		
1	Perturbation Attack	The attacker modifies the query to get an appropriate response
2	Poisoning Attack	Attacker contaminates the training phase of ML systems to get the intended result
3	Model Inversion	The attacker recovers the secret features used in the model
4	Membership Inference	Attacker infer if the given data record was part of the model's training data set
5	Model Stealing	The attacker can recover the model by constructing careful queries
6	Reprogramming ML system	Repurpose the ML system to perform a non-programmed activity
7	Adversarial Example in Physical Domain	Attacker brings adversarial examples into the physical domain to subvert ML system
8	Malicious ML Provider Recovering Traning Data	Malicious ML providers can query the model used by the customer and recover the customer's training data
9	Attacking the ML Supply Chain	Attacker compromises the ML model as it is being downloaded for use
10	Backdoor ML	Malicious ML provider backdoors algorithm that does not work unless triggered
11	Exploit Software Dependencies	The attacker uses traditional software exploits to confuse ML systems

Table 4 Safety Failure Modes – Unintended Failures

Unintended Failures		
1	Reward Hacking	Reinforcement learning systems act in unintended ways because of a mismatch between stated rewards and true rewards
2	Side Effects	System disrupts the environment as it tires of attaining its goal
3	Distributional shifts	The system is tested in one kind of environment but is unable to adapt to changes in other kinds of environment
4	Natural adversarial examples	Without attacker perturbations, the ML system fails to owe to hard harmful mining
5	Common corruption	The system is not able to handle common corruption and perturbations such as tilting, zooming, or noisy images
6	Incomplete testing	The Ml systems are not tasted in realistic conditions that it is meant to operate

These failure modes are readily available to be transferred for AI considerations. Furthermore, as described by Jason Millar [60], two types of social failures can be used to define social failure modes. These include (1) "**Absence of supportive norms**. An artefact can fail socially when its design requires a certain social norm to be held by its user(s) in order for it to work (i.e., be used) as intended, but that required norm is not held by its users(s)" and (2) "**Norm transgression**. An artefact can socially fail when a norm designed into it transgresses and accepts social norm held by its user(s)".

As an example for the first of these failure modes, Jason has described the Google Glass device that failed the norm that it is NOT acceptable to wear cameras that record surreptitiously (users norms). Google Glass developers defined this approach as acceptable, failing specific social norms. Furthermore, the same device also failed in the norm transgression failure mode since it transgressed privacy norms.

Even though FMEA processes are not well documented over non-technical components, few works have described the application of these approaches from social perspectives. As defined in [61], Social Responsibility could be improved by understanding the risk factors of social considerations within a company. They define several Social Responsibility Failure Modes (See the following Table adapted from [57] to define failure modes from social responsibility considerations). These failure modes could be included for any manufacturing company about their trustworthy considerations regarding societal well-being. Nevertheless, these failure modes should be directly translated for AI considerations.

Table 5 Social Responsibility Failure Modes – does require a link to AI trustworthiness

Social Responsibility Failure Modess		
1	Lack of pollution metrics	No systems for tracking and reporting on social and environmental results - Governance
2	Lack of protective policies	No organizational policy for the protection of property, which is to prevent the theft of technical resources
3	Human Rights communication	Lack of clear message about the importance of human rights in the organization
4	General communication problems	Lack of processes for resolving grievances
5	Lack of regulation compliance	Conditions of work do not comply with national law
6	Lack of environmental corrective actions	Lack of identification and action associated with protecting the natural environment
7	Lack of fair operating practices	Lack of identification of risk associated with corruption

Finally, typical failure modes for software components can be fed within the failure mods of system robustness. A list of these can be found in [57]. These includes:

Table 6 Robustness Failure modes – General for software

General Failure Modes		
1	Lack of Functionality	The software provides no output or control action not provided when expected
2	Improper Functionality	The programmed control system software performs an unexpected action as defined by the operator of the equipment
3	Timing	Software event happens too late, too early, or control action is mistimed
4	Sequence	Sofware event occurs in wrong order or control action with incomplete sequence concept error
5	False alarm/action	Sofware detects an error when there is no error or control action provided when not expected
6	Fault logic and Ranges	Concept error where the software or control actions contain incomplete or overlapping logic
7	Incorrect Algorithm	The software computes incorrectly based on some or all inputs or control action is based on wrong computation
8	Memory management	The software runs out of memory, or memory leakage or control actions stops due to lack of memory
9	Interface Failure	Software failure due to failure of hardware interfaces such as power supply
10	Software virus	The software did not function on demand due to a software virus

Significantly, from the previous list shown in Table 6, some of these failure modes should be extended or shifted to other considerations, given the trustworthy requirements. For example, software viruses should be considered a security consideration, while

The natural extension of the failure modes with ethical considerations involves the construction of failure modes in the function of trustworthy requirements. Given this natural link, the present framework defines 11 **Ethical-Based General Failure Modes Families** for AI.

i. Failure to Robustness

ii. Failure to Safety
iii. Failure to Transparency
iv. Failure to Accountability
v. Failure to Societal wellbeing
vi. Failure to environmental wellbeing
vii. Failure to Human agency and oversight
viii. Failure to Privacy
ix. Failure to Data Governance
x. Failure to bias (diversity, non-discrimination, and fairness)
xi. Failure to Users Values

Content specification of each of these failure modes domains will be performed during the life span of the ASSISTANT. However, some Examples of Robustness and Safety are already provided in Table 3 and Table 4. Furthermore, Failure modes of societal wellbeing could be classified based on Milllar's definitions [60], and therefore only a clear understanding of local and regional social norms would be required to evaluate the failure modes.

Extensions of these previous examples could include natural extensions or adaptation from software-based failure modes that include, for example, failure to data curation and failure to achieve desirable tolerance in training conditions, among others.

Notably, A common approach to performing these analyses in FMEA and defining failure modes is to analyse failures related to a particular system's functionality or its part by considering not performing or performing incorrectly. Such an approach will be the driver of estimating failure modes based on suitable defined software architecture. However, specified by [62], there are ground rules that have to be specified to settle the definitions of failure modes; These include:

- ***What type of failure (functional or component levels) should be discussed?***. In the case of AI, this will depend on the system architecture since, if an individual component is analysed, the functional failure modes will be enough. However, in ASSISTANT, functional and component levels should be analysed.
- **If multiple failures materialize simultaneously, shall and how that scenario is analysed?**. Given the algorithmic nature of the AI component, each failure mode will be rooted in specific system functionalities; Therefore, in ASSISTANT, the focus will be only on individual failure mode analysis.
- **How might a common initiating cause result in simultaneous failures (e.g. loss of power supply)?**. Common initiating causes should be considered in ASSISTANT (and the AI framework) since they could strongly impact considerations of Robustness and Safety (e.g. human robotics interactions).
- **What if a failure mode is not detected but is hidden hierarchical within another failure mode?**. It could be foreseen that technical failure modes will be foreseen given the algorithmic nature of the AI component. Nevertheless, those derived from non-technical considerations (e.g. biased information) will be rooted on a level that could not be foreseen beforehand. By using the current framework, considerations of these aspects are enforced if the risk level of the AI asset does deserve its attention.

The exercise of finding failure modes will be based on the assumption that if a component stops operating as expected, it can affect the component, sub-system or system performance, functionality or behaviours.

The performance, functionalities, and behaviours are translated to AI technical **Operations Modes** (OMs) and system or processes technical and non-technical OMs. For example, some technical functionalities could include training after 100 working hours or

creating neighbour solutions from a founded optimal for explainability systems. For a process, non-technical behaviour could imply that safeguards for an AI use or development are correctly implemented to secure human oversight as an end up solution. Therefore, a clear definition of all the operations modes is required to define the components or system failure conditions (i.e. failure modes).

In other words, there is a need to specify first the OMs that secure each of the eleven previously defined failure modes before defining the components' failure modes or, even worse, the nonexistence of their consideration.

An AI element should consider different failure drivers to facilitate this analysis. Under the current framework, we propose a clear subdivision of failures driver. These drives can be given physical, user and system interfaces, internal social, or data considerations.

The following figure schematises this approach in which each of the eleven previously defined failure modes domains (in this case, robustness) can be tested over the driver of the failure. The figure shows that these drivers can produce a flag that will “define” a failure mode throughout the OMs. The failure mode will be identified as the driver of the operation mode failure plus the specification over which Ethical-Based General Failure Modes Families are linked (e.g. failure of robustness by misuse of the interface system).

Several AI parts could be tasted, and once commonality is observed between AI components, a failure mode type/family can be defined (which will be driven by applying the current framework in ASSISTANT).

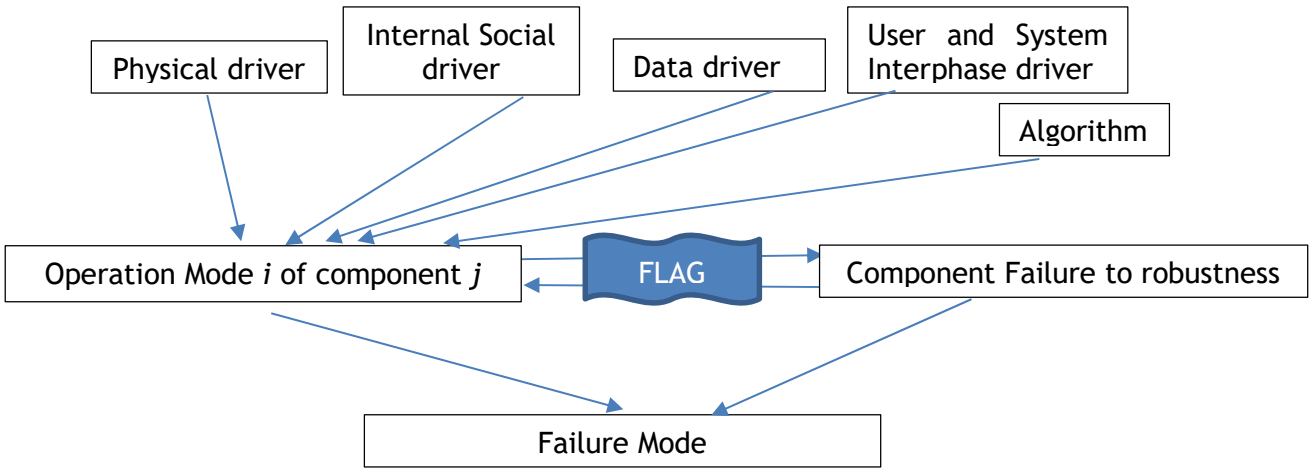


Figure 15 Schematic of Failure Mode Detection on AI components

The reference to physical driver includes, among others, power supply, communication/data link cables, robot parts, wearables, lenses, and sensors. The reference to the internal social driver is related to their developers' values and biases imposed over the AI component. At the same time, this implies all the values and approaches defined by developers, deployers and companies in general that could transcend, among others, social communication, social wellbeing, and social responsibility.

The reference to data drivers is related to data biases, quality, quantity, corrupted data, data security, or any data consideration that could drive an underperformance of the AI element. The user and system interphase driver is linked to poor use of the interphases by the users or hackers. Furthermore, it can include the lack of adequate information displayed in the interphases in order for the system to perform as expected. Furthermore, this also implies any interaction between AI and humans that are not physically based (i.e. including cognitive biases

– e.g. overtrust of the AI results). Further drivers could be considered. For example, as the framework is tested in ASSISTANT, further drivers will be included, if needed, for specification of the involved domain (i.e. manufacturing). Each FMEA requirement and process should be adequately documented. An exemplification of this document is given in the corresponding documentation section.

Other considerations within the process of Identify Failure modes are that the process usually considers single failures and their effects (i.e. two simultaneous independent failures are not considered). Nevertheless, an exception is the criteria of hidden failures that their presence is undetectable. In such cases, single failures should be combined with the hidden failure and their combined consequence for the analyses. Another consideration is redundant systems that both should be available during normal operations. This would impact if AI is embedded within redundant physical systems and, therefore, this consideration of redundancy should apply to them too (e.g. if two algorithms perform in parallel for security or robustness considerations, but if there is a problem with one of them, the process should be stopped until both are online – e.g. one of them is under maintenance). Among the other considerations is that of external events as failure modes. Traditionally FMESa focuses on the impact of the failures that originate within the equipment, however, giving requirements such as governance considerations or unexpected external physical factors.

Finally, concerning additional analyses regarding failure modes, the considerations should be extended to all physical and non-physical control, instrumentation, and safety devices and algorithms. A clear understanding is that no safeguards will function unless coupled with a demand for functionality (i.e. an activation need). Therefore, attention should also be placed on failing conditions (e.g. proof that it works as expected) regarding the activation of safeguards and control systems.

A list of failure modes with the specification of the source and families is given in the following table. This table has the first proposition of casualties that could be considered within the ASSISTANT evaluation and is derived by using previous table information and including different failure modes through a brainstorming process. Nevertheless, it will be extended as experts of each AI functionality component participate in the risk management process.

Table 7 List of failure mode as a function of the driver and failure mode

Ethical Risk Failure Modes				
Failure Mode Driver	Failure Mode Family	Definition	Example	Recommended name
Data	Robustness	Failure to detect different body traits	Hand image recognition is strongly dependent on the hand positioning	Failure to robustness by poor human traits representability
Data	Bias	Failure to detect different race traits	Image recognition is strongly dependent on human traits	Failure to bias by poor representability of race traits
Data	Robustness	Failure to detect disruptive traits	Detection failed by devices or traits (e.g. tattoos) that alter the recognition process	Failure to robustness by disruptive traits.
Data	Robustness	Failure for quantity	Image failure to detect by lack of reverse / flipped image	Failure to robustness by poor representability.
Data	Robustness	Failure for quality	Data used for the training process show lower quality than the used for analyses	Failure to Robustness by a Quality discrepancy.
Data	Robustness	Failure for timeframe representability	The time frames used for training do not match the timeframes of analyses	Failure to robustness by timespan mismatch.
Data	Robustness	Timing gap	Distance between data points does not help to represent phenomena	Failure to robustness by timeframe granularity.
Data	Robustness	Timing	The algorithmic event happens too late or too early, or the control action mistimed	Failure to robustness.

Physical	Robustness	Timing gap	Lag or mismatch on timeframes between information capture and use of it	Failure to robustness by sensed timeframe mismatch.
Data	Robustness	Lack of Functionality	The algorithm provides no output or control action not provided when expected	Failure to robustness by lack of functionality
Internal Social	Robustness	Improper Functionality	The programmed control system software performs an unexpected action as defined by the user	Failure to robustness by improper functionality
User and System Interphase	Robustness	Improper software use	Requirements set by users are not achievable by the algorithm or its scope set for training	Failure to robustness by improper software use
Internal Social	Robustness	Lack of algorithmic corrective actions	Lack of identification and action associated with protecting the algorithmic robustness	Failure to robustness by lack of corrective actions
Algorithm	Robustness	Sequence	Algorithmic event occurs in the wrong order or control action with incomplete sequence concept error	Failure to robustness by sequencing actions
Algorithm	Robustness	False positive detection from alarm/action	The algorithm detects an error when there is no error or control action provided when not expected	Failure to robustness by algorithmic false positive
Algorithm	Robustness	False-negative detection from alarm/action	The algorithm does not detect an error when there is an error or control action provided when expected	Failure to robustness by algorithmic false negative
Algorithm	Robustness	Fault logic and Ranges	Concept error where the software or control actions contain incomplete or overlapping logic	Failure to robustness by incomplete logic actions
Algorithm	Robustness	Incorrect computation from recognised input	The software computes incorrectly based on some or all inputs or control actions. The potential source of error is identified.	Failure to robustness by incorrect computation from recognized input
Algorithm	Robustness	Incorrect computation from unrecognized sources	The software computes incorrectly. The potential source of error is NOT identified.	Failure to robustness by incorrect computation from unrecognized sources
Algorithm	Robustness	Memory Management	The algorithm performs actions that make the system run out of memory	Failure to robustness by excess memory usage
Physical driver	Robustness	Hardware requirement	The hardware is insufficient for the memory requirements of the algorithms	Failure to robustness by inadequate hardware
Physical driver	Robustness	Interface Failure	Software failure due to failure of hardware interfaces such as power supply	Failure to robustness by interface handling
User and system interphase	Security	Software virus	The software did not function on demand due to a software virus.	Failure to security by virus attack
Internal Social Driver	Societal wellbeing	Lack of social metrics	No features for tracking and reporting on social trends or impacts	Failure to Societal well-being by lack of tracking metrics
Internal Social Driver	Environmental wellbeing	Lack of environmental metrics	No features for tracking and reporting on environmental trends or impacts	Failure to environmental well-being by lack of tracking metrics
Algorithm	Societal Wellbeing	Lack of use or misuse of societal metrics	No use of features for tracking and reporting on social trends or impacts	Failure to societal well-being by misuse of metrics
Algorithm	Environmental wellbeing	Lack of use or misuse of societal metrics	No use of features for tracking and reporting on environmental trends or impacts	Failure to societal wellbeing by
Internal Social Driver	Data Governance	Lack of protective policies	No organizational policy for the protection of property, which is to prevent the theft of technical resources	Failure to Data Governance by lack of protective policies
Internal Social Driver	Bias	Human Rights communication and AI ethics	Lack of understanding of the importance of human rights in the organization	Failure to bias by a lack of definitions and understanding of human rights
Algorithm	Bias	Incomplete data sets	Lack of representability of clusters or groups by an uneven representation of data	Failure to bias by incomplete data sets
Algorithm	Bias	Lack of bias elimination	Lack of methods or approaches to eliminate biased data from data sources known to contain them	Failure to bias elimination by lack of methods
Data	Bias	Unrecognized bias	Lack of recognition or identification of bias from data sources	Failure to bias by undetected sources

Internal Social Driver	Societal Wellbeing	General communication problems	Lack of processes for resolving grievances from AI	Failure to societal well-being by lack of grievances resolving
Internal Social Driver	Societal wellbeing	Lack of regulation compliance	Conditions of work with the AI do not comply with local, regional, or national law	Failure to societal wellbeing by lack AI local compliances regulation compliance
Internal Social Driver	Societal wellbeing	Lack of fair operating practices	Error, no driver, or no methodologies to apply corrective actions related to fairness	Failure to societal well-being by lack of operating practices
Internal Social	Safety	Lack of security and safety corrective actions	Lack of identification and action associated with protecting the algorithmic robustness	Failure to safety by lack of corrective actions
Internal Social	Data governance	Lack of governance corrective actions	Lack of identification and action associated with securing data governance	Failure to governance by lack of corrective actions
Data	Data Governance	Lack of data protocols	Lack of protocols for data ownership and data responsibilities	Failure of data governance by lack of policies
Data	Data Governance	Lack of data usability	Data is not related or relevant for the problem to be solved	Failure to data governance ownership by data usability
Data	Data Governance	Lack of data format consistency	Data is supplied spread between formats that do not match	Failure to data governance by lack of format consistency
Data	Data Governance	Lack of data integrity	Data describe altered, unreal, or inconsistent trends in the information supplied.	Failure to data governance by lack of data integrity
Data	Data Governance	Lack of temporal data consistency	Data is supplied sporadically	Failure to data governance by lack of temporal consistency
User and system interphase	Data Governance	Lack of user responsibilities	Error applying Data management and designation of responsibilities from the user part, leading to poor data quality or quantity, miss direction of data, etc.	Failure to data governance by users lack of responsibilities
Data	Data Governance	Lack of external data management and responsibility	Poor Data governance from external sources that are dependent on supplied information from the AI	Failure to data governance by external sources
Data	Data governance	Lack of protocols for data validation	No protocols or poor application of them from data validation supplied to the system	Failure to data governance by lack of data validation and its protocols
Data	Data governance	Lack of protocols for data curation	No protocols or poor application of them from data curation supplied to the system	Failure to data governance by lack of data curation and its protocols
Data	Data governance	Lack of protocols for data tagging	Lack of methods to track data modifications, if allowed, by tagging and users identification	Failure to governance by lack of data tagging protocols.
Physical	Data governance	Lack of supporting hardware	Lack of protocols or physical components to secure data integrity and supporting track of information	Failure to governance by lack or failure from supportive hardware.
User and system interphase	Security & Data Governance	Lack of accessibility protocols	Lack of protocol for securing user access or user recognition	Failure to security & data governance by the lack or poor accessibility protocols
User and system interphase	Security	Over accessibility	Lack of control of the user and developers' access to restrictive information, source code, and algorithmic parameters	Failure to security by over accessibility
Data	Accountability	Lack of internal data or algorithmic responsibility	Poor or lack of designation of responsibilities for internal data sources management, quality, veracity, and quantity.	Failure to be accountable for the lack or poor internal data responsibility
Data	Accountability	Lack of external data or algorithmic responsibility	Poor or lack of designation of external data sources management, quality, veracity, and quantity responsibilities.	Failure to be accountable for the lack or poor external data responsibility
Internal Social	Accountability	Lack of accountability corrective actions	Lack of identification and action associated with securing data accountability for data and algorithms	Failure to accountability by lack of corrective actions
Internal Social	Transparency	Lack of Transparency in corrective actions	Lack of identification and action associated with securing system transparency in algorithms	Failure to transparency by lack of corrective actions
Internal Social	Societal Wellbeing	Lack of Societal well-being corrective actions	Lack of identification and action associated with securing societal wellbeing for data and algorithms	Failure to societal well-being by lack of corrective actions

Internal Social	Human Agency and Oversight	Lack of Human Agency and Oversight corrective actions	Lack of identification and action associated with Human Agency and Oversight	Failure to Human Agency and Oversight by lack of corrective actions
Internal Social	Privacy	Lack of privacy corrective actions	Lack of identification and action associated with data privacy	Failure to privacy by lack of corrective actions
Internal Social	bias	Lack of bias corrective actions	Lack of identification and action associated with a bias from data, developers, and algorithms	Failure to bias by lack of corrective actions
Internal Social	Users Values	Lack of User Values corrective actions	Lack of identification and action associated with users' values and its trends for data, developers, and algorithms	Failure to users' values by lack of corrective actions
Users and system interphase	Safety	Perturbation Attack	The attacker modifies the query to get an appropriate response	Failure to safety by perturbation attack
Users and system interphase	Safety	Poisoning Attack	Attacker contaminates the training phase of ML systems to get the intended result	Failure to safety by poisoning attack
Users and system interphase	Safety	Model Inversion	The attacker recovers the secret features used in the model	Failure to safety by model inversion attack
Users and system interphase	Safety	Membership Inference	Attacker infer if the given data record was part of the model's training data set	Failure to safety by membership inference attack
Users and system interphase	Safety	Model Stealing	The attacker can recover the model by constructing careful queries	Failure to safety by model stealing
Users and system interphase	Safety	Reprogramming ML system	Repurpose the ML system to perform a non-programmed activity	Failure to safety by the reprogramming ML system
Users and system interphase	Safety	Adversarial Example in Physical Domain	Attacker brings adversarial examples into the physical domain to subvert ML system	Failure to safety by adversarial example in the physical domain
Users and system interphase	Safety	Malicious ML Provider Recovering Training Data	Malicious ML providers can query the model used by the customer and recover the customer's training data	Failure to safety by malicious ML provider
Users and system interphase	Safety	Attacking the ML Supply Chain	Attacker compromises the ML model as it is being downloaded for use	Failure to safety by attacks over the ML supply chain
Users and system interphase	Safety	Backdoor ML	Malicious ML provider backdoors algorithm that does not work unless triggered	Failure to safety by backdoor ML
Users and system interphase	Safety	Exploit Software Dependencies	The attacker uses traditional software exploits to confuse ML systems	Failure to safety by exploiting software dependencies
Users and system interphase	Safety	Reward Hacking	Reinforcement learning systems act in unintended ways because of a mismatch between stated reward and true rewards	Failure to safety by reward hacking
Users and system interphase	Safety	Side Effects	System disrupts the environment as it tires of attaining its goal	Failure to safety by side effects
Algorithm	Robustness	Distributional shifts	The system is tested in one kind of environment but is unable to adapt to changes in other kinds of environment	Failure to robustness by distributional shifts
Users and system interphase	Safety	Natural adversarial examples	Without attacker perturbations, the ML system fails to owe to hard harmful mining	Failure to safety by natural adversarial examples
Algorithm	Robustness	Common corruption	The system is not able to handle common corruption and perturbations such as tilting, zooming, or noisy images	Failure to robustness by common corruption
Users and system interphase	Robustness	Incomplete testing or training	The Ml systems are not tasted or trained in realistic conditions that it is meant to operate	Failure to robustness by incomplete testing or training
Users and system interphase	Robustness	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by robustness.	Failure to robustness by users violation
Users and system interphase	Safety	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by safety.	Failure to safety by users violation
Users and system interphase	Transparency	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by transparency.	Failure to transparency by users violation

Users and system interphase	Accountability	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by accountability.	Failure to accountability by users violation
Users and system interphase	Societal Wellbeing	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by societal wellbeing.	Failure to societal well-being by users violation
Users and system interphase	Environmental Wellbeing	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by environmental wellbeing.	Failure to environmental well-being by users violation
Users and system interphase	Human Agency and Oversight	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by Human agency and oversight.	Failure to human agency and oversight by users violation
Users and system interphase	Privacy	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by privacy.	Failure to privacy by users violation
Users and system interphase	Bias	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by bias.	Failure to bias by users violation
Users and system interphase	Users Values	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by users' values.	Failure to users values by users violation

Again, and importantly, the previous failure mode effects should be analyzed under their local and higher-level effects, and they should be described as foreseeing and end effect that helps to evaluate and define the total effect and assume the accumulated effect on the operation, function, the status of the system, algorithms, the environment and users. The end effects may result from a combination of failure modes that can end in catastrophic end effects. If possible, these combined end effects should be reported if foreseen. Nevertheless, given the lower understanding of the impact of combined failure modes, prioritization should be given to those with individual higher impact, which could hive a high likelihood of occurrence.

The fourth step, named **Identify Failure Detection Methods**, is based on identifying, evaluating, and observing methodologies that can be used to detect the Failure Modes as the system runs. This step can be performed parallel to the fifth and sixth steps described next. Typically, the detection implies using visual or audible gears that sensor limits and warn users in the manufacturing sector. In the case of the AI component, this would imply that metrics and methods will have to be embedded or linked to inputs or outputs features and metaparameters to be linked to the risk conditions. Given some of the trustworthy requirements, some failure detection methods will be easier to implement than others. For example, methods for Failure to Robustness and Failure to safety already exist for AI components.

Nevertheless, evaluation methods for other requirements will have to be defined for online or offline detection. For example, independent of the nature of these methods (online or offline), adequate time must be available to react if operation actions are required to reach a safe state or prevent escalation (if the system's dynamic allows it; i.e. human-in-the-loop). This also implies that the own nature of the intrinsic risk level of the AI component (e.g. unacceptable or high risk) defines the detection process and involvement of human intervention.

Detection methods can be directly or indirectly linked to users' actions for recognition. At the latest, procedures are required (and should be defined) to detect the malfunction's specificity. These procedures could require other instruments, control devices, circuit breakers, or a combination of the previous for failure detection. The lack of these procedures is considered a failure mode to system robustness and security. In the case of direct method detection, they could be classified as abnormal or incorrect. Abnormal implies an indication that is evident to an operator when the system has a malfunction or failure, while the second

indicator is given by supporting devices such as instruments, sensing devices, visual or audible warning, etc.

The fifth step, named **Analyze Effects on parts, subparts, and globally**, implies that failure modes are analyzed in their consequences. The failure modes can significantly affect broad components of the system (i.e. cascade effect). Therefore, a comprehensive analysis of the interactions within AI components for software that involves complex architectures. Finally, the FMEA should identify the end effect of any failure in its impact. For the ethical Framework, this implies that focus should be driven on the impact of each of the possible effects over the trustworthy requirements, the regulatory considerations established so far, and the values set by the stakeholders involved in the design, implementation, and use decommissioning stages. In the end, the FMEA is a tool to evaluate whether enough measures are in place to prevent hazards.

In terms of the current framework, this is translated into setting enough safeguards, controls, and technical and non-technical approaches over the system that will comply with the legal, ethical, and value-based requirements established for the components (which are also dependent on the intrinsic risk level of the AI).

The sixth component, **named Identify Corrective Actions**, takes care in suggesting solutions or corrective actions for risk conditions. The corrective actions should comply with the system design philosophy [55], and therefore for the current framework, these imply to:

i. Actions that will reduce the likelihood of the failure mode (i.e. component within the ten families previously described).
ii. Actions that will reduce the intrinsic level of risk of the AI component
iii. Actions that will secure comply with legal requirements
iv. Actions that will secure the safe operation of the overall system
v. Define recovering actions for failing conditions
vi. Actions that will set humans as the centrepiece (i.e. human-centred focuses actions).
vii. Actions that will secure each requirement established by the EU (depending on the relevance for the intrinsic level of risk of the AI element).
viii. Others Relevant to values and ethical considerations (extend)

The corrective actions can be classified with different levels depending on the risk category involved in the AI component. We recommend the use of three levels of classification named **"For immediate Attention"**, **"For Serious Consideration"**, and **"For Future Improvement Decisions"**. One important consideration is that these levels would depend on the risk appetite established on the policies. Furthermore, none of these classifications is suitable for AI that presents unacceptable risk intrinsic levels since they should not be considered for any AI life cycles (i.e. development, deployment, use or decommission).

The missing components, named **Criticality Ranking, Tabulate and Report (or FMEA report)**, will be covered here and in the following sections since they involve using the risk register and can be linked to the critical matrix. The criticality ranking allows the set soft metrics to keep track of the failure more and thus can be used as representative base KPIs for estimating the state of ethical issues within the risk management process. The current framework proposes to use these and other KPIs to track the state of the AI component within a given system and ASSISTANT. Generally, the criticality ranking for FMEA analyses involves setting values for (1) the likelihood of a failure mode to take place (section 4.7), (2) the severity of the consequences in case the failure condition takes place (section 4.6), and (3) the capability of users and system to detect the failing condition (4.8). These metrics should ideally be based on historical information. Nevertheless, expert judgment can be used to rank each of these components.

The multiplication of each of the estimated indexes for each failure mode generates the Risk Priority Number ($RPN = S \cdot O \cdot D$; Where *S* is the severity, *O* is the Occurrence or likelihood, and *D* is the detection ranking). The RPN works as a fundamental index that merges these considerations in one metric. The higher the RPN, the higher the risk involved in the analysed failure mode, and thus, the item or component is the source of the failing condition. The RPN is not directly used in the criticality analyses but can be used as a source of information, independent of it, to give a sense of an item's global risk.

This could be achieved (as proposed here) by summing up the failure modes RPN multiplied by a Failure Mode Ratio. The failure mode ratio represents that, given a failing condition, what part could be explicitly attributed to the specific failure mode under evaluation. This failure mode ratio (α) is covered in section 4.5 since it is linked with the criticality analyses. Therefore, an item accumulated (RPN_{item}) can be estimated as shown in the following equation. In this equation, RPN_{item} is calculated as the sum of the RPN of item *i* and its corresponding i^{th} failure mode ratio. represent the RPN of failure mode I an

$$\boldsymbol{RPN_{item} = \sum_{i=1}^{n} RPN_i \alpha_i} \quad \textbf{(1)}$$

4.4.3 **Report:**

The FMEA report focuses on keeping a repository where enough information for the readers is used to understand the failure modes, effects, existing risk, control measures, safeguards, and related recommendations. As expected, their responsibility is dependent on the organisation's structure (i.e. the risk management architecture) and the domain and goals of implementation. One main document is described for internal reporting of the risk processes for the current framework. This document, named risk register, is described in other sections. Other documents that could be employed are (1) executive summary, (2) description of systems, (3) conclusions and corrective actions, and (4) referencing data [55]. Given the structure of ASSISTANT, the description of components (1) and (2) are defined throughout D2.1 and D2.2 up to D7.1 and D7.2. Therefore, the integration of reports and reviews in ASSISTANT will focus on using Risk Register (Section), Conclusions and Corrective Actions, and Referencing data summaries given by collecting these deliverables.

Significantly, the reliability block diagram can help construct the risk register (documental component) since it can withdraw corrective actions. Therefore, the risk register will condense the overall aspects of the FMEA analysis by including the RBD.

Even though a broad number of failure modes can be found, the information specified in the report should be good enough for performing critical item lists to settle what corrective actions should be prioritized.

4.4.4 **Review and Verification:**

The reviewing process of an FMEA is an iterative process that includes preliminary recommendations over the failure modes identified. Any technical issue or consideration of interest for the area of relevance (given the risk architecture) should be discussed among the FMEA team and stakeholders before delivering a final version for acceptance and sharing with another area of the architecture. Some of the scopes and pitfalls that should be considered in the reviewing process can be extracted from [63].

- Parts of the system or critical operations omitted in the analysis

- Incomplete failure list
- No consideration of common-cause of failures
- Global links and their effects not addressed
- No consideration of failures on existing safeguards and control
- Failure or delayed follow-through of corrective actions
- Insufficient descriptions in the worksheets to understand the failure scenarios
- Insufficient information in the FMEA report
- FMEA did not match the latest design or off-the-shelf FMEAs
- Submittals too late

Furthermore, the FMEA study alone could not provide enough assurance to achieve satisfactory levels of security over these systems. Therefore the purpose of the verification process is to secure that the conclusions reached in the FMEA study are verifiable. As described later, this process can be performed by external stakeholders (in an audit process) to check and test both correction actions and testing the FMEA study results.

The audit components should define supporting tools such as test sheets, checklists, and verification plans before implementing the verification process. For the current implementation in ASSISTANT, for each tested component, these checks should include:

- Hardware and software description
- The purpose of the test includes specification of the failure of concern
- Test methodology
- Procedures of testing
- Expected results
- Results section
- Comments section

The sheet proposed for ASSISTANT is included in the annexe section. As observed in the table, the sections include a description of the system (i.e. WP for ASSISTANT), sub-system, and component to be tested. The method and the expected results should be provided by the Risk Assessment team and agreed upon by the audit team. The other section should be filled as the failure modes, safeguards and components are tested. These sections can provide recommendations for component modifications or strategies to improve system performance regarding the probability of materialising risk components. If the system does not perform as expected, the system or its development should depend on the intrinsic risk level involved and the risk appetite. This is important for trustworthy considerations within enforced or driving regulation (e.g. a transparency component is not working as expected for a high-risk AI component – AI act).

The validation process should, ideally, be run over each failure mode. Nevertheless, the scope should be driven over those that will have a higher impact on the system's functionality and, at the same time, describe the higher risk (given as a combination of likelihood and outcomes – i.e. based on heat maps results, as defined later on). These tests should include at least:

- System confirmation to operate with/under failing conditions, following the design intent (safeguards and control approaches). These tests can include adversarial attacks guided by the trustworthy considerations imposed by the intrinsic risk level of the AI components and, at the same time, the values that want to be induced in the system.
- Confirmation, If possible, of the system response to common system failures
- Corrective actions implementation, as defined in previous stages.
- Confirmation of the correct process

As deduced from previous contexts, some tests would not be possible to be driven, given possible damage to hardware or inexperience in sources of information that could lead to ethical considerations. Nevertheless, if failure cannot be reproduced, the safeguards used to protect shall be evaluated in case of such failure. It should be tested their existence, specifications, functionality, maintainability process and needs, and methods that could lead to failure of the safeguard.

4.5 Criticality Analysis and Failure Mode, Effects, and Criticality Analysis (FMECA)

A Failure Mode, Effect, and Criticality Analysis (FMECA) extend the FMEA process by including a criticality assessment. Definitions of the criticality analyses are extracted from the: "Military Standard (MIL-STD-1629A), Procedures for Performing a Failure Mode, Effects and Criticality Analysis – Department of Defense, United States of America [54].

This process allows to bring to attention the most critical issues explicitly and, therefore, helps considerably in the decision-making process of what risk should be managed first or, depending on the use of supporting approaches (e.g. Pareto), what parts or components could be left for treatment or evaluation on consecutive applications of the risk management process. Generally speaking, the ranking constructed by this analysis helps allocate resources and effort to produce higher benefits.

The process uses a combination of the severity and the likelihood, highlighting those failure modes with the higher risk. Ideally, the estimates should be based on historically quantifiable data since the reliability of the collected information will strongly influence the curse of actions following the FMECA analysis.

The criticality analysis can follow both processes, qualitative and quantitative, based on the experience of the users, developers, and stakeholders that perform the analyses. The qualitative approach is appropriate when specific failure rate information is not available. When enough information is available, criticality numbers (i.e. that give the methods' name) should be calculated and incorporated into the overall analyses.

In the qualitative approach, failure sub-systems identified in the FMEA should be evaluated in terms of probability of occurrence when specific failure rate data is not available. In this case, individual failure mode probabilities of occurrence will be grouped by specifically defined levels that establish the qualitative failure probability level for entry into the critical analyses. The following three sections give the tables and corresponding probability of occurrence levels. Finally, the likelihood levels of occurrence (Frequent, Reasonably probable, occasional, remote, extremely unlikely) are given in section 4.7. Attention should be brought that the purely FMEA analyses also use these tables since they serve as a ranking strategy for the failure modes.

Once the probability numbers are transformed into categories based on the tables mentioned before or in the case pure expert knowledge is used for defining classification ranks for the failure modes severity and failure mode probability of occurrence, Figure 16 can be used to estimate the failure mode criticality number. This criticality number would follow here and after the same process of transformation and agglomeration as for the quantitative analyses. Furthermore, and as will be explained soon, this criticality number can be subject to different transformations based on specific parameters. These parameters are α, which is the failure mode ratio; β, which is the conditional probability of mission loss; γ which is the

part/component failure rate; and *t, which* is the duration of the applicable mission phase expressed in hours or several operating cycles.

The Failure effect probability represents the stakeholder judgment as to the conditional probability that the loss will occur as defined by the failure mode effect. In other words, it represents how accurate the expected effects of the failure mode are according to the pre-set analyses. The values can range from 1 to 0, following the scale defined in Table 8.

The failure mode ratio represents that, given a failing condition, what part could be explicitly attributed to the specific failure mode under evaluation. As a fraction, it can range from 0 to 1. As expected, If all potential failure modes of a particular part or item are listed, the sum of the α for that component will be equal to one. If information is not available for estimation (given historical records), the value should be represented as the expert judgment.

The part failure rate corresponds to the obtained failure rates multiplied by different factors that could alter the obtained failure rates. In other words, the Mathematical expression described in the following equation can be used to extract the final part failure rate. In this equation, π_I represent each of the *i* factors that will modify or alter the base failure rate γ_0.

$$\boldsymbol{\gamma = \prod_{i=1}^{n} \pi_i \gamma_0} \quad \textbf{(2)}$$

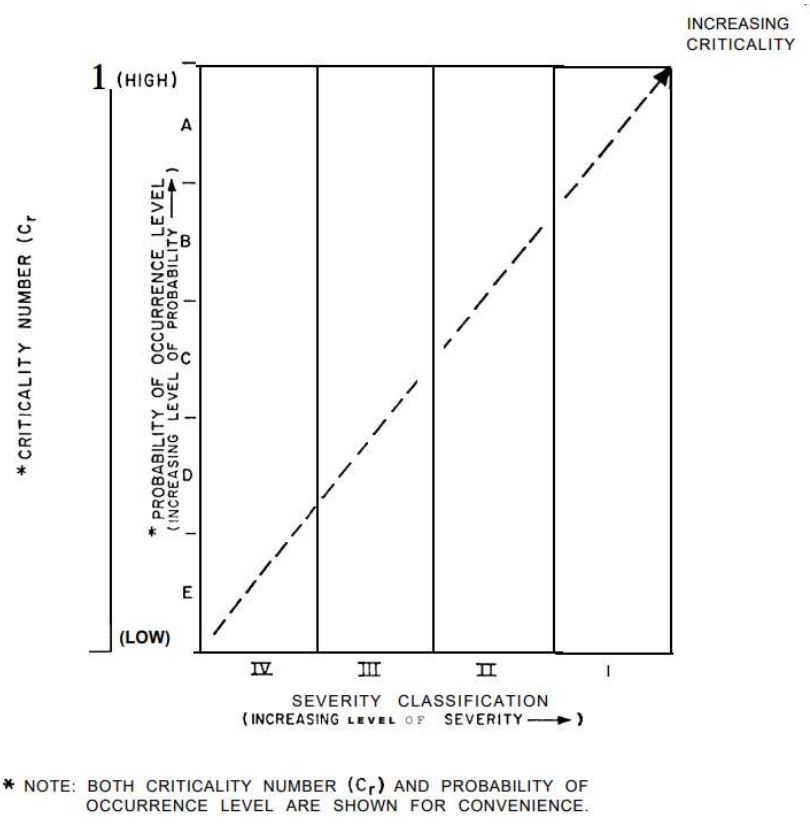


Figure 16. Critical Evaluation based on severity and likelihood (extracted from [54])

In the quantitative approach, ***failure rate*** data is used directly instead of an estimation that leads to a probability number (i.e. qualitative approach). Some failure rates and adjustment factors should be used when the running activity has no specified failure rate data source. These factors consider several external and internal factors that could affect the

provided and actual data (e.g. environmental conditions and quality factors). These factors should be developed through experience and are not necessarily readily available.

The criticality number for failure mode calculation is shown in the following equation. In it α, β, γ, and t have previously been defined.

$$C_n = \beta\alpha\gamma t \quad (3)$$

Table 8 Failure Mode Ratio in the function of the failure effect loss judgment

Failure Effect	Failure effect probability
Acutal / Expected	1
Probable	0.1 to 1
Possible	0 to 0.1
No effect	0

Finally, the part or component's overall criticality number, which represents the overall combination of the failure conditions that could take place for an item, corresponds to the sum of all the criticality numbers estimated by Equation 1. Importantly, if the part failure rate is not available, the likelihoods will be recommended as guidelines to estimate an overall component/item criticality number. Nevertheless, if likelihoods are used in the analyses, they should be used for the overall analysis and, therefore, each item evaluated under the risk assessment process should use a likelihood, as defined in Section 4.7.

Finally, qualitative and quantitative considerations for Criticality Analysis are made in the present framework. In fact, given the chance that the manufacturing sector (or any domain that would try to implement the present framework) would lack historical records or expert judgment to facilitate the implementation of quantitative analyses, the framework proposes the use of FMEA indexes as supporting driving an initial stage to support decision making. To be more specific, the FMEA index, named Risk Priority Number, which combines severity, detection, and rank of event occurrence, can be used at the initial stages to generate heat maps or risk matrixes that will support the decision-making process.

Similarly to FMEA, a worksheet is also used to track and analyse the process and, later on, construct and use a specific component named criticality matrix (also named risk matrix). Given the nature of the FMECA analyses extension over the FMEA, a unique worksheet is proposed to be used with a stakeholder decision to be made if it extends its use over the criticality analyses component.

4.6 Severity Classification and Ranking

A severity classification is helpful to provide a qualitative measure of the potential resulting consequences from a failing item. Therefore, a severity classification should be assigned for each failure mode and analysed component. As recommended by [54], if no categories can be defined, a similarity with loss statements based upon loss of system inputs or outputs shall be developed and included within the FMEA/FMECA ground rules. A multi-failure description should determine the severity level that considers each ethical-based FMECA issue. The table shows a proposition based on considerations of ethical and security-based concerns. This table will feed the critical matrix by setting severity categories with a critical number.

The severity code imposes the definition of low, minor, and significant injury. As specified in [64], these levels imply:

- Low-level exposure: An exposure at less than 25% of published Threshold Limit Value (TLV) or Short Term Exposure Limit (STEL).
- Minor Injury: A slight burn, light electrical shock, minor cut or pinch. First aid can handle these and are not OSHA recordable or considered as lost time cases.
- Significant Injury: Requires medical attention other than first aid. This is a medical risk condition.

Prioritization is given to the system based on the ES&H scale. If the levels based on this scale are acceptable under the institution's risk appetite, the severity ranking based on customer satisfaction could be used next. Table 9 and Table 10 show the severity ranking and classification in the function of the severity designation of the failure mode.

Table 9 Severity classification for Failure Modes Ranking based on ES&H severity code [64]

Severity designation of the Failure mode and its Effect Description	Severity Rank	Failure severity Classification
• Failure would cause loss of life or total disability to personnel • Failure would cause identifiably catastrophic damage to the system and repairs that are beyond the capability of the user or contractor to resolve the effects • Failure would lead to violating any regulatory consideration set as fundamental rights. • Failure would lead to violating principles that cause a non-recoverable and undermining of the users and environmental wellbeing	10	Catastrophic (A or I)
• Failure would cause severe disabling injury or severe occupational illness to personnel • Failure would cause identifiably critical damage to the system and extensive repairs to resolve the effects. • Failure would lead to violating principles that cause severe undermining of the users, and environmental wellbeing • can cause fire or environmentally adverse conditions.	8-9	Critical (B or II)
• Failure would cause a minor injury or minor occupational illness to personnel that may require hospitalisation, but failure is not disabling • Failure would cause identifiably marginal damage to the system an acceptable level of repairs and downtime to resolve effects • Failure would violate principles that will undermine the users and environmental well-being that could be managed with proper implementation actions. • The severity level is high and activates alarms, safeguards, and requirements of special system attention. • Can cause controllable environmentally adverse conditions.	6-7	Marginal (C III)
• Failure would cause minor injury to personnel, but those injuries would not require hospitalisation, or failure would cause minor occupational illness • Failure would cause identifiable minor damage to the system and minor repairs and short downtime to resolve effects • Failure would lead to violating principles that cause minor undermining of the users and environmental wellbeing • The severity level is high and activates alarms, safeguards, and requirements of special system attention.	3-5	Minor (D or IV)
• Failure would cause less than minor injury and no occupational illness • Failure would cause negligible damage to the system and insignificant or no downtime to resolve effects • Failure is not credible. • There is no impact on the environment.	1-2	Negligible (E or IV)

Table 10 Severity Ranking based on customer satisfaction qualitative information [64]

Description	Severity Level	Rank
Failure will result in significant customer dissatisfaction and cause non-system operation or non-compliance with government regulations	Catastrophic (A)	10
Failure will result in a high degree of customer dissatisfaction and cause non-system functionality	Critical (B)	8-9
Failure will result in customer dissatisfaction, annoyance and deterioration of part or system performance	Marginal (C)	6-7
Failure will result in slight customer annoyance and slight deterioration of part or system performance	Minor (D)	3-5
Failure is of such minor nature that the customer will not detect the failure	Negligible (E)	1-2

4.7 Likelihood Classification and Ranking

A likelihood or occurrence ranking metric helps measure how frequently an analysed failure mode could occur. The probability of occurrence (P_f) should be based on the failure mode's probability of occurring during operation time. The time frame's homogenisation should be based on an hourly or a 1E-6 hourly base for each failure mode considered. This designation of time frame is used since its commonly used also in criticality analyses. In case the process is sporadic, Table 12 could be used. Even though this distinction could obtain different rankings, as long as homogenization is performed for each classification, the shift in ranking would be homogenized in the end analysis. Table 11 shows the ranking and occurrence level in function of the description or designation of the occurrence.

A critical characteristic of the occurrences ratios for some intrinsic algorithmic processes is that confusion matrixes estimations could estimate them. These numbers could be used directly for specific robustness-based failure modes. We, ASSISTANT, recommend its use if readily available. This approach is considered within the pipeline process of the proposed framework.

Table 11 Occurrence Ranking Criteria Likelihood or Level of Occurrence in the function of temporal probabilities as a single failure mode for quantitative analyses [54]

Description	Occurrence Level	Rank
Once a week. High probability is defined as a single $P_f > 0.20$ of the overall probability of failure during the item operating interval.	High probability (A)	10
Once every two weeks. Probability is defined as a single $P_f > 0.10$ but $P_f < 0.20$ of the overall probability of failure during the item operating time interval.	Probable (B)	7-9
Once a month. Occasional is defined as a single $P_f > 0.01$ but $P_f < 0.10$ of the overall probability of failure during the item operating time interval	Occasional (C)	4-6
Once every two months. Remote is defined as a single $P_f > 0.001$ but < 0.01 of the overall probability of failure during the item operating time interval.	Remote (D)	2-3
An unlikely probability of occurrence during the item operating time interval. Unlikely is defined as a single $P_f < 0.001$ of the overall probability of failure during the item operating time interval.	Unlikely (E)	1

*Each occurrence is considered during the system/part/sub-part/component operating time or the ratio of item build with failing conditions during the operating time. The normalisation comes from the operating time specification (e.g. hourly or 1E-6 hourly).

Table 12 Occurrence Ranking Based on Ratios [65]

Description of Ranking	Ratio	Rank
Very High (The failure is very likely to occur)	1 in 2	10
Very High	1 in 8	9
High (The failure will occur occasionally)	1 in 20	8
High	1 in 40	7
Moderate	1 in 80	6

Moderate	1 in 400	5
Moderate	1 in 1,000	4
Low (the failure will rarely occur)	1 in 4,000	3
Low	1 in 20,000	2
Remote (the failure is unlikely to occur)	< 1 in 10^6	1

4.8 Detection Classification and Ranking

The following tables cover the detection classification ranking. As observed, there are two tables in its use. The first defines the detection capacity of failing conditions based on the products (e.g. inspection of products), while the second is based on the detection of system control failures.

Table 13 Detection Ranking Criteria for products

Rank	Description	DetectionLevel
10	Very low (or zero) probability that the defect will be detected. Verification and controls will not or cannot detect the existence of a deficiency or defect	Very Low (A)
8-9	Low probability that the defect will be detected. Verification and controls are not likely to detect the existence of a deficiency or defect.	Low (B)
5-7	Moderate probability that the defect will be detected. Verification and controls are likely to detect the existence of a deficiency or defect	Moderate (C)
3-4	High probability that the defect will be detected. Verification and controls have a good chance of detecting the existence of a deficiency or defect.	High (D)
1-2	Very high probability that the defect will be detected. Verification and controls will almost certainly detect the existence of a deficiency or defect. (1 in 8)	Very High (E)

Table 14 Detection Ranking criteria based on design and control

Detection	Criteria: Likelihood of detection by Design Control	Ranking
Almost Certain	Design Control will almost certainly detect a potential cause/mechanism and subsequent failure mode	1
Very High	Very high chance the design control will detect a potential cause/mechanism and subsequent failure mode	2-3
High	High chance the design control will detect a potential cause/mechanism and subsequent failure mode	4
Moderately High	Moderately high chance the design control will detect a potential cause/mechanism and subsequent failure mode	5-6
Moderately	Moderately chance the design control will detect a potential cause/mechanism and subsequent failure mode	7
Low	Low chance the design control will detect a potential cause/mechanism and subsequent failure mode	8-9
Very Low	Very low chance the design control will detect a potential cause/mechanism and subsequent failure mode	10
Remote	Remote chance the design control will detect a potential cause/mechanism and subsequent failure mode	8
Very Remote	Very remote chance the design control will detect a potential cause/mechanism and subsequent failure mode	9
Absolutely Uncertain	Absolutely Uncertain chance the design control will detect a potential cause/mechanism and subsequent failure mode	10

4.9 Criticality matrix / Risk matrix.

The criticality matrix provides a means of identifying and comparing each failure mode with respect to severity and likelihood. Its construction depends on the followed process in the critical analyses performed. Independent of the approach, the matrix is constructed by incorporating the system, sub-system, or component criticality number in matrix locations representing the severity classification category. In case only the FMEA process was used, the analysis can still be performed by the agglomerated (and pondered) metrics (i.e. the Risk Priority Number). This agglomeration has previously been discussed and corresponds to the weighted Risk Priority Number of the failure modes corresponding to a specific component/AI asset. Nevertheless, this process is not mandatory since it can be used as a direct implementation of the failure modes Risk Priority Numbers

The critical matrix based on the Criticality Analysis distributes a "risk score" that sets the limits for defining the level of importance of the risk elements and, thus, the type of management that should be considered for such element or risk component. A representation of it is given In Figure 17.

Independent of the qualitative or quantitative nature of the criticality analysis, the pondered elements should be allocated over the critical matrix constructed based on the likelihood scale, the severity scale, and the risk score, which is a direct translation of the user risk appetite.

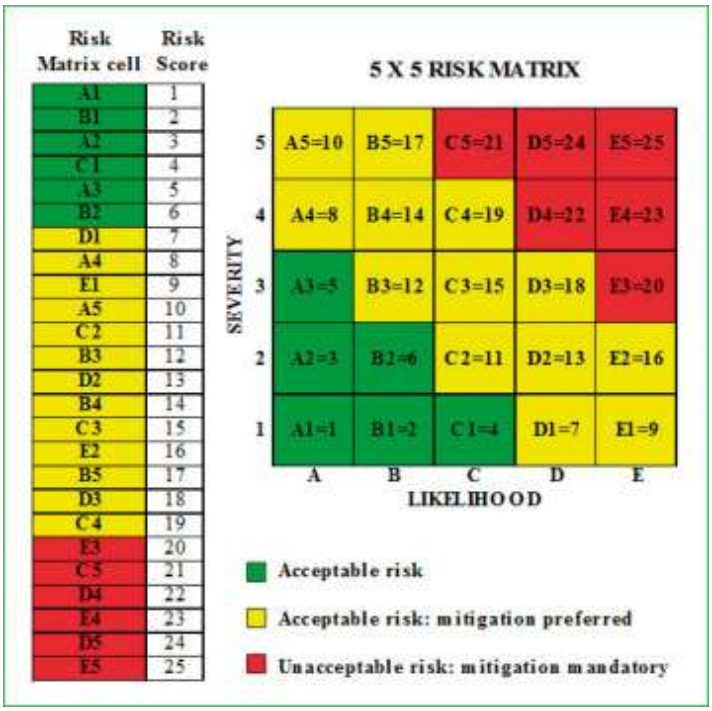


Figure 17 Critical Matrix representation based on generalizable risk score

The figure, as observed, does only consider 3 clusters of risks (acceptable risk, acceptable risk with mitigation preferred, and unacceptable risk with mandatory mitigation). In order to connect the risk matrix with the 4T's, it is required to set the risk score scale in 4 clusters. To set these 4 clusters, the risk appetite will establish the numerical risk scores for limiting the Tolerate, Transfer, Treat and Terminate processes. For example, it could be used a quartile distribution in order to set even distribution of the 4Ts.

Notably, the prioritization of what T use on specific ranges is connected to the risk matrix, as shown in Figure 18. As observed in it, the Tolerate range corresponds to those elements that possess low risk.

As observed in the figure, the elements to be transferred to another party (e.g. insurance company) possess a high impact but a low likelihood. However, the current status of enterprises dedicated to covering AI components has not been studied in the present work. Thus, it is not clear the absolute validity of considering transfer risk related to AI and, more specifically, e-risks. Based on this in ASSISTANT, there are no considerations of transfer and all the processes with transfer risk scores would be considered for Treatment instead (i.e. only three clusters will be used, and something similar to Figure 17 would be employed).

In a standard connection between the 4T's and the risk matrix, the risk Treatment is connected to risks with high likelihood and low impact; therefore, the prioritisation of the treatment should be on reducing the likelihood of the events. In ASSISTANT, since only three classifications would be employed, the treatment should be driven by reducing the impact and the likelihood of the events.

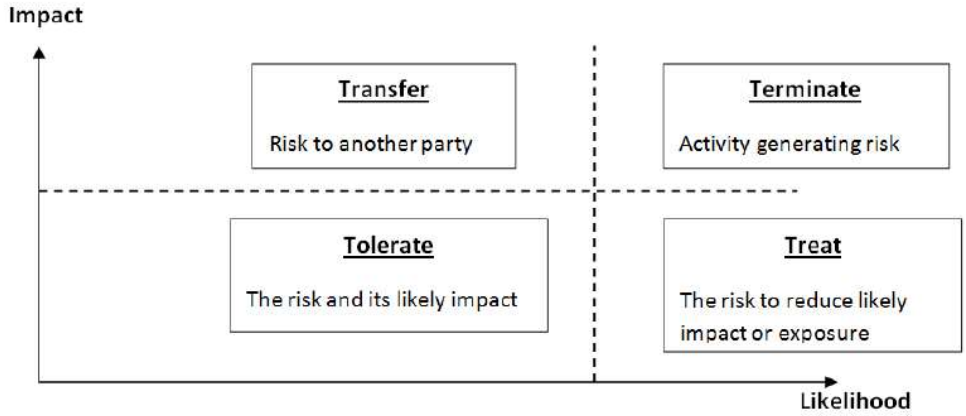


Figure 18 Connectivity between the risk matrix and the 4T's considerations

Finally, those processes, elements or failure modes with high impact and likelihood should be terminated. Again, the risk score that sets the initiation of the Terminate process is given by the risk appetite. A discussion of it is given in section 5.

The FMEA analysis is a valuable tool to quantify any risk associated with the failure of a system, subsystem or component. Regretfully, it does not provide a general framework to set acceptable levels of tolerance for risk nor agglomerate the information strategically. Nevertheless, the following procedure can help an FMEA be used as a risk management system for ISO to construct a risk matrix based on the Risk Priority Number.

1. Define how many percentiles can be used to cluster the information in case that is undecided, define the quartile, median, and third quartile (i.e. definition based on the risk appetite).
2. Construct Figure 19 as a rank matrix. This is built by putting severity on the y-axis, occurrence on the bottom x-axis, and the detection rate on the upper x-axis. Then, multiply the intersections to cover all possible RPN.
3. Use the percentiles derived from the risk appetite and the RPN range (from 0 to 1000) to define the risk level clusters. In the case of quartiles, the Q1 values will be considered tolerable (i.e. would not require any modification from the current condition).
4. Connect the clusters (e.g. quartiles) with the 4T's. For example, the Q3 implies that 25% of the risk is above this value and, therefore, can be considered High-Risk (i.e. should be terminated). Anything that falls in the interquartile would be considered a

risk condition that should be treated or transferred. These limits should be modified based on the percentiles used to construct the risk matrix and the institution's risk appetite. Figure 20 shows a risk matrix constructed by the previous methodology and the quartile considerations.

Severidad \ Deteccion	1	2	3	4	5	6	7	8	9	10
1	1	4	9	16	25	36	49	64	81	100
2	2	8	18	32	50	72	98	128	162	200
3	3	12	27	48	75	108	147	192	243	300
4	4	16	36	64	100	144	196	256	324	400
5	5	20	45	80	125	180	245	320	405	500
6	6	24	54	96	150	216	294	384	486	600
7	7	28	63	112	175	252	343	448	567	700
8	8	32	72	128	200	288	392	512	648	800
9	9	36	81	144	225	324	441	576	729	900
10	10	40	90	160	250	360	490	640	810	1000
	1	2	3	4	5	6	7	8	9	10

Ocurrencia

Figure 19. Risk matrix based on pure FMEA approach

Severidad \ Deteccion	1	2	3	4	5	6	7	8	9	10
1	1	4	9	16	25	36	49	64	81	100
2	2	8	18	32	50	72	98	128	162	200
3	3	12	27	48	75	108	147	192	243	300
4	4	16	36	64	100	144	196	256	324	400
5	5	20	45	80	125	180	245	320	405	500
6	6	24	54	96	150	216	294	384	486	600
7	7	28	63	112	175	252	343	448	567	700
8	8	32	72	128	200	288	392	512	648	800
9	9	36	81	144	225	324	441	576	729	900
10	10	40	90	160	250	360	490	640	810	1000
	1	2	3	4	5	6	7	8	9	10

Ocurrencia

Figure 20. Risk matrix based on quartiles for risk limits identification

4.9.1 Root Cause Analysis (RCA)

To understand root causes, the best way to do it is to analyse it as the analysis of a common problem. For example, if a business is underperforming, an effort to find the reasons for such underperformance will be made. Interestingly, to perform an analysis, one could define the symptoms to provide a situation remedy. However, these solutions only consider the symptoms and do not consider the underlying causes of those symptoms — i.e. the root of the problem.

Therefore, to solve or analyze a problem, we will need to perform a Root Cause Analysis (RCA) and find out precisely what the cause is and how to fix it. Even though there is comprehensive documentation that readers could use to get a deeper understanding of the RCA, we would perform a slight analysis here, outlining standard techniques and specifications of template methodology that can be used and therefore linked to the framework (covered later on).

RCA assumes that it is more effective to solve underlying issues rather than just treating ad hoc symptoms. RCA can be performed with a collection of principles, techniques, and methodologies that can be leveraged to identify the root causes of an event or trend (i.e. failure condition or failure mode). The goals of the RCA include (1) to discover the root cause of a problem or event; (2) to fully understand how to fix, compensate, or learn from any underlying issues; and (3) to apply what we learn from this analysis to prevent future issues. There are a few core principles that guide practical root cause analysis:

1. Focus on correcting and remedying root causes rather than just symptoms.
2. Do not ignore the importance of treating symptoms for short term relief.
3. There can be multiple root causes for one event.
4. Focus on how and why something happened; since accountability is a significant concern within ethical perspectives, who should also be included in the analysis.
5. Use cause-effect evidence to back up root cause claims.
6. Provide enough information to inform a corrective course of action.
7. Consider how a root cause can be prevented (or replicated) in the future.
8. Take a comprehensive and holistic approach to analysis.
9. Strive to provide context and information that will result in an action or a decision.

Different techniques can be used for RCA. Below are mentioned and defined the most widely used ones. Most RCA methods are top-down deductive analysis proceeding through successively, and a more detailed cause of the event is achieved.

4.9.1.1 5 Whys

A common technique in performing a root cause analysis is the 5 Whys approach. For every answer to a WHY question, follow it with an additional, deeper “Ok, but WHY?” question. The 5 Whys serve as a way to avoid assumptions. Finding detailed responses to incremental questions makes answers more precise and concise each time. Ideally, the last why will lead to a source of the failing condition.

4.9.1.2 Change Analysis/Event Analysis

This method is convenient when there are many potential causes with different time frames considerations. Instead of looking at the specific day or hour that something went wrong, we look at a more extended period and gain a historical context. To do (1) First, every potential cause leading up to an event is listed. These should be at any time a change occurs over the system condition. (2) It is categorised each change or event by how much influence we had over it. The categorization can include Internal/External, Owned/Unowned, or something similar. (3) Event by event is evaluated and decide whether or not that event was an unrelated factor, a correlated factor, a contributing factor, or a likely root cause. This evaluation can use the 5 Whys as support. (4) It is analysed how it can be replicated or remedy the root cause.

4.9.1.3 Cause and effect Fishbone diagram

Another common technique is creating a Fishbone diagram (also named Ishikawa diagram) to map cause and effect. It is similar to the 5 Whys. The diagram starts with the problem in the middle of the diagram, then brainstorm several categories of causes, which are then placed in off-shooting branches from the mainline. Categories are extensive. After grouping the categories, we break those down into smaller parts (e.g. for example, under

"trustworthy components", we might consider potential root cause factors like "robustness", "security", etc.

As the analysis goes deeper into potential causes and sub-causes, questioning each branch, it is a higher chance of getting the sources of the issue. Furthermore, this method eliminates unrelated categories and identifies correlated factors and likely root causes. For the sake of simplicity, carefully we recommend for ASSISTANT the following categories to consider in a Fishbone diagram:

Table 15. Recommended Cause and Effect Fishbone categorical base ordering for ASSISTANT

Skeleton	First Level	Second Level	Third Level
Failing Condition	AI	Robustness	Social
			Technical
		Safety	Algorithms
			Physical component
		Transparency	Traceability
			Explainability
			Open Communication
		Accountability	Auditability
			Accountability definitions
			AI Management processes
		Societal Wellbeing	
		Environmental Wellbeing	
		Human Agency and Oversight	Human Autonomy
			Human Oversight
		Privacy	Data
			Methods
		Data Governance	
		Bias	
		Values	
		Others	
	Model	Knowledge	
		Others	
	Schopfloor / Manufacturing system	Machine	
		Method / Process	
		Material	
		User	
		Knowledge	
		Management	
		Maintenance	
		Environment	
		Suppliers	
		Skills	
		others	

Further categories could be needed to understand the root causes better. Furthermore, this table can also be used by the 5why's analysis to facilitate the identification of root causes. If RCA were implemented within ASSISTANT use cases, proper extensions, as needed, and definitions of the "others" component would be presented as the risk management process validation product.

4.9.1.4 Fault tree analysis (FTA)

FTA is a diagrammatic analytical technique used for reliability, maintainability, and safety analysis. The FTA can be connected with qualitative or quantitative information (i.e. probabilities) to improve its analysis. Similarly to the fishbone, the outcome is taken as a logic tree's root (or top-level). Different logical operators, such as OR, AND, Exclusive OR, Priority AND, and IF, can be used to produce branching structures. If fault trees are labelled with failure

probabilities, overall failure probabilities and rates can be estimated by the trees. As expected, an event can have more than one outcome. Therefore, it could produce several tree extensions or have repetitive branches. The branches and operators can be connected with different events that give information about the activities that can take place if the branches follow the designated direction. These events can define (1) internal terminal even (e.g. root cause), external event, undeveloped event, conditioning event, and intermediate event. An exemplification of a fault tree analysis is given in the following figure.

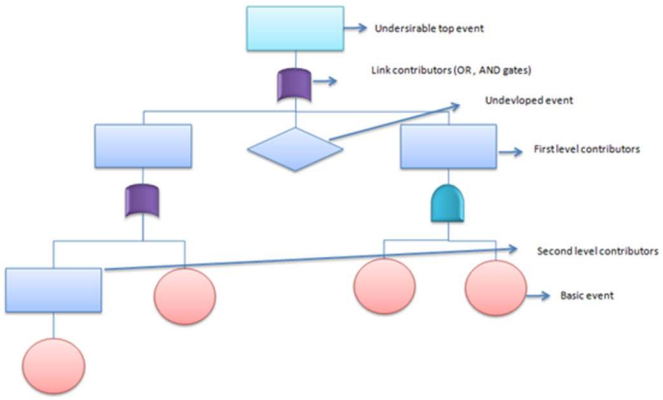


Figure 21. Fault Tree Analysis Exemplification

4.10 Implementation of AI-risk management

Three types of assessments are usually involved when evaluating conditions involved with risk management. First, a hazard assessment involves estimating how often events of various sizes occur. Second, it focuses on considerations such as what the probability of a given factor occurring under given circumstances is? Third, what is the variability of the hazard?. In other words focuses on understanding the hazard, its nature, and its variability.

An impact assessment involves focusing on the outcome of at least one single hazard scenario. It focuses on evaluating the impact on the system given that conditions exist for a risk to materialize. Finally, The risk assessment considers the full suite of hazard scenarios (i.e. the event sets) and understands the risk (probability and magnitude of loss) over the system.

The implementation of AI-risk management is based on the combination of the ISO 31000 (described in section 4.2.1) with requirements and restrictions imposed by the Trustworthy Guidelines, the White paper on Artificial Intelligence, The artificial Intelligence act, and several documents related to these previously mentioned guides (e.g. EU Charter of Fundamental Rights linked that is linked to the artificial intelligence act).

Even though the ISO 31000 can be seen as a general framework for risk management, a lower-level framework that complements the ISO 31000 was used to define the general structure and settle, in this way, the Risk Policy involved in ASSISTANT. In the following section (Section 5), each RASP component and a thorough definition of how to combine the tools defined in section 4 as part of the e-risk protocols are covered.

5. Ethical Risk Management (e-Risk) Framework

In this section, the Ethical Risk Management Framework is presented. First, a description of the general constituents of the ethical risk management framework (section 5.1) is

performed. Then, individual components linked to the e-risk management framework are presented. More specifically, in section 5.2, the Documentation and Instruments for Risk ASSESSMENT are presented. In section **Erreur ! Source du renvoi introuvable.**, the ethical FMEA and FMECA processes are better defined.

5.1 General Description

In order to consider all the components established in the ISO 31000 (or other standards relating to risk management that cover the same principles), its scope is covered by using the RASP approach. The Risk Architecture (RA) define the roles, responsibilities, communication and risk reporting structure (subsection 5.1.1). The strategy (S) defines the risk appetite, attitudes, and philosophy in risk management policy (subsection 5.1.2). The protocols (P) define the rules and procedures and the risk management methodologies, tools, and techniques that should be used (subsection 5.1.3).

Most of the specific definitions of each scenario are established in the risk management policy. However, since the relevance of the risk management policy does not overlap with a framework specification for generalised e-risk management, an exemplification of such information is incorporated within the Annex section as a specific risk management policy for ASSISTANT.

Another important consideration is that AI methods can be embedded within processes or be a stand-alone system used for mapping, prediction, forecasting, optimisation, and recommendations, among other tasks. A general definition of a system will be used in the framework to describe an agglomerate of AI (can be only one) that can be contained in integrative subsystems. Each AI is constructed or defined by different components or processing steps that would be denominated as Components.

This implies that a classification based on an **Architectural Definition** (which can be linked to technical architectures) can be established to define interdependencies between AI. A generative classification structure is given in

Figure **22.** The existence of the system, subsystem, and components, with a specification of each of these elements, is given.

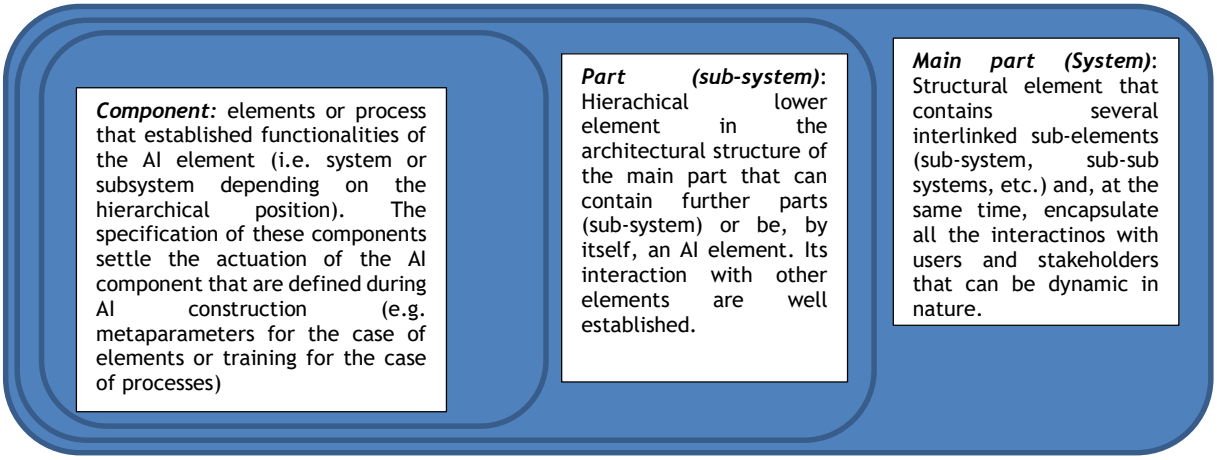


Figure 22 Arrangements for Incorporating risk management in ASSISTANT.

5.1.1 Risk Management Architecture:

The risk management architecture should define the committee structure in terms of reference, roles and responsibilities, internal reporting requirements, external reporting controls, and risk management assurance arrangements. This section will cover these topics except those not relevant and applicable to ASSISTANT (i.e. external reporting control and risk management assurance arrangements).

Figure 23 Ethical Risk Architecture Shows a base architecture that integrates the risk assessment process and the internal reporting channel specification about the committee structure. The figure also defines critical roles and responsibilities. This architecture can be modified depending on industrial complexity and its policy. ASSISTANT risk management architecture would be based on the structure specified in this figure and establish the minimum responsibilities set for the different risk owners, process owners, internal audit, risk management members, staff, contractors, and outsourced contributors (risk stakeholders). The definitions of roles for ASSISTANT are included in the appendixes. A detailed set of responsibilities will ensure that the roles of risk stakeholders are clearly defined and understood.

The first role (bottom right of the figure) is Divisional Management (DM). A division is a part of the system (process, business, or organization) that performs a critical role. Under the AI perspective, these divisions could be integrated as a whole main component (i.e. an AI division) or as sub-components that focus on a specific AI asset functionality (e.g. Training, data curation, optimization, etc.). The divisional management is incorporated by different stakeholders that perform the internal activities related to risk management. These include, for ASSISTANT considerations, performing the risk assessment process, preparing and keeping up the risk register (a document that will be explained later on), setting priorities for the division in terms of the focus on risk analysis and treatment, and keeping updated KPIs related to the AI elements. Information on each significant priority risk ownership should be included in the risk register.

If the risk management process is performed in parallel with processes, it can easily be seen that this simplistic architecture allows integration with other structures. The divisional management reports the risk assessment finding to the Management Committee. Documentally, the DM is responsible for the Risk Performance and Monitoring Reports that include the risk register as complementary information.

The second role is the Management Committee (MC). This role corresponds to a division's leader (in terms of e-risk management) who would integrate the reports from the different divisions. A general MC could integrate general risk management and incorporate those involved in ethics perspectives. Additionally, the MC provision and monitor the actions of the DM, securing that tasks involved in risk assessment are performed correctly. The MC also allocates responsibilities to its staff based on the definitions established by the e-risk board.

Within the processes involved in the divisions, the MC contemplates internal audits that are defined by an Audit Committee (AC). In this way, the MC is responsible for constructing complete documentation dedicated to the Events response and Recommendations.

The MC reports the accumulated information from the division to the Executive Risk Committee and proves to them with recommendations and suggestions provided by the DM and

AC that can alter or modify the risk management process or the actions implemented so far e-risks.

The third role is the Executive Risk Committee (ERC – led by the risk manager).

The risk manager is responsible for the corporate learning that has to take place so that the organisation can understand the benefit of risk management. As the person responsible for the RASP, the risk manager will be responsible for developing the strategy, systems and procedures by which the required risk management outcomes for the organisation are achieved.

These ERC should be covered by an expert (or group of experts) that will have a deep understanding of AI, the company objectives, and the present regulatory considerations, requirements and constraints of AI systems. These requirements are given since the ERC will have the primary responsibility of performing the analyses of the risk management processes and recommending actions to reduce the risk levels or to terminate or tolerate the AI's current conditions under the perspectives of the e-risk management process (based on the company risk appetite and regulations of AI – i.e. consider external materiality of information). Further responsibilities include ensuring that risk management is embedded within all AI systems or subsystems, independent of the innate level of risk (High, moderate or low risk – those classified as unacceptable should not be implemented) and reviewing the profiles of the DM groups in order to secure that a variety of experts with different expertise are integrated into the risk assessment process. Additionally, the ERC keep track of the whole process (i.e. internal materiality of information) and all the documentation generated during the risk management process. Therefore, it is necessary to maintain a range of risk management records that include details of various risk management activities, including administration, risk response and improvement plans, event reports and recommendations, and risk performance and certification reports. These documents correspond to:

- Risk management documentation manual and administration records and responsibilities – Responsibility of the E-risk board
- Risk response and improvement plans - Responsibility of the ERC
- Event reports, incidents, investigations and recommendations – Responsibility of the MC
- Risk performance and monitoring reports – Responsibility of the DM

The Risk Management Manual (RMM) contains details of all the responsibilities, procedures, protocols, Language and perception of risk in the organisation, framework for identifying significant risks, role of the risk manager and internal auditors, and guidelines regarding the risk management process and framework for the organisation. The manual should confirm the procedures for undertaking the activities and set out details of the systems and processes that will be put in place to monitor performance and the means for reporting and communicating on risk management. In addition, the risk management procedures will set out risk assessment processes, risk control objectives, risk resourcing arrangements, reaction planning requirements and risk assurance systems.

It is recommendable to update the risk management manual each year so that risk management activities and the overall risk management approach is in line with current best practice.

The fourth role is the e-risk bard (or team leader). The e-risk board has the overall responsibility for the e-risk management. These include allocating responsibilities for each principal component of the risk management architecture, making decisions regarding budgeting and effort specification, and making final decisions based on reports given by the ERC. Also, these roles define the level of participation of external component that includes insurance brokers, insurance companies, accountancy firms and external auditors, among

others involved in risk management, quality, regulation evaluations and social organizations that play an essential role in setting values and ethical considerations within the company.

There are no general specifications for the team leader's background, but it should be expected to consider executive or non-executive directors of the organisation. (Non-executive directors - role are restricted to audit, assurance and compliance activities, to assist with the formation of strategy and the monitoring of performance, and does not include the organisation's day-to-day management. Executive directors - involved in the management of individual risks and implementation of the strategy.)

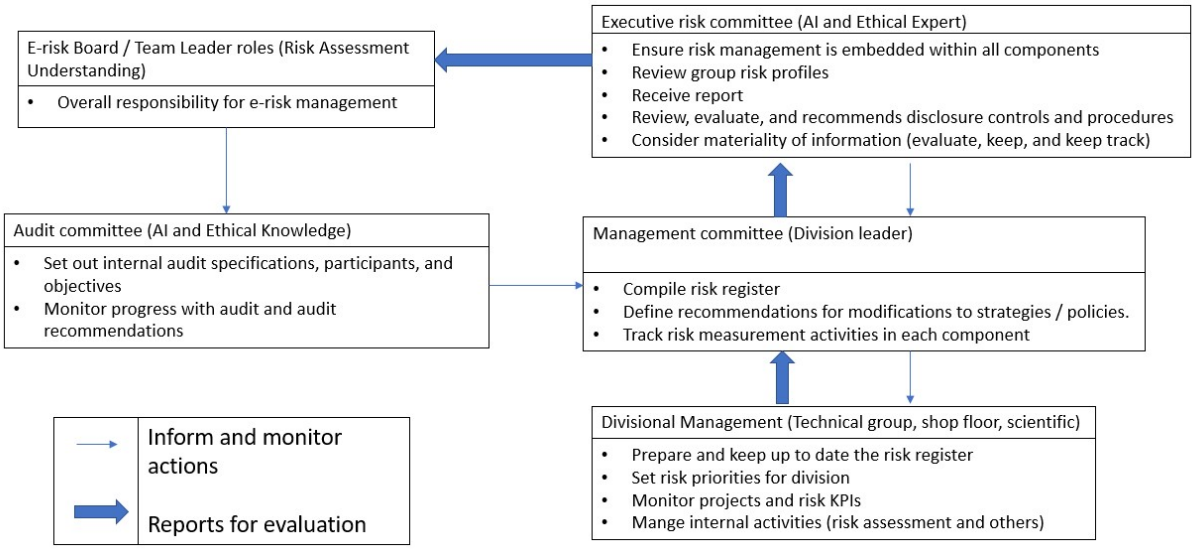


Figure 23 Ethical Risk Architecture

The final component is the audit committee. Internal audit has its expertise in the evaluation of controls and the testing of their efficiency and effectiveness. The RASP should set out the details of how this close cooperation will be achieved in practice (specified in the annexe section for ASSISTANT). The internal audit should be driven by experts that will cover the requirements of AI and, at the same time, those requirements established by Trustworthy AI definitions.

Even though the architectural components describe hierarchical structure and responsibilities, a decision-making process should follow the recommendations defined in section 3.4.2. Table 16 Decision-Making Considerations shows a general description of the decision making criteria involved in the definitions at each level.

The following Table describes an organizational recommendation level for decision-making based on the Architectural level. The first column capitalizes on the components that will lead the strategic, tactical, and operational decision-making. The second column capitalises on the goals of the decision-making approaches (that should not violate those defined by the previous hierarchical components). Finally, the third column defines the criteria for the decision making. The strategic, higher-level, will focus that the criteria in decision making follow the objectives of the policies, the risk follows the capacity established for the risk management (i.e. secure that risk management has an implementation level in accordance to the level of risk embedded on the AI elements), flexible to modifications depending on current tendency, and consistency over time.

Consistency, in terms of e-risks, implies that values and ethical components established during the decision-making process should be kept unmodified unless the company's risk policy is modified. The decision-making process is based on ANP and AHP tools and is described in section 5.1.3.

Table 16 Decision-Making Considerations

Management	Goals	Criteria
Strategic (e-risk board)	General Decision-Making	• Objectives • Capacity • Transparency • Budget • Flexibility • Consistency (value-based)
Tactical (Executive Risk Committee)	Final Analysis(Identification) Risk Analysis Rist Treatment Options	• Cost-Effective • Time-Effective
Operational (Divisional Management)	Implement Quality Control Program and products to reduce risk.	• Operational

5.1.2 Risk Management strategy

The risk management strategy should define the risk management philosophy, the arrangements for embedding risk management, the risk appetite and attitude to risk, benchmark test for significance, specific risk statements/policies, risk assessment techniques, and risk priorities for a given period. Most of these topics are described in the annexe section for ASSISTANT and will not be thoroughly covered here.

One crucial characteristic used for e-risk management is the risk management philosophy. This states how risk is considered in the organizations. For the case of AI, we recommend (and base on) the considerations of the risk management approach based on three fundamental principles.

- **Proactive, innovative, and dynamic risk management process:** This implies creating a robust management process that can be used to identify, quantify, monitor, mitigate and manage e-risks. It is proactive in its philosophical perspective to use EC regulatory considerations to define benchmark approaches for global best risk management practices.
- **Corporate and societal Value-Based:** Promote the incorporation of values in the risk management process that is not contradictory to legal and ethical requirements established.
- **Not violate domain ethics:** Synergetically integrates the AI ethical considerations with the ethics intrinsic to the application domain (e.g. medical /healthcare ethics).
- **Technically considered:** Promote the integration of Trustworthy AI considerations with technical and functional requirements established by stakeholders, allowing the innovation of the AI element.

The risk appetite is a crucial set of statements that define the proactivity towards risk. For the ASSISTANT, the risk appetite statements are as follows:

- **EU Regulatory based approach:** ASSISTANT is committed to delivering value to our AI elements by securing the use of risk strategies established by the EC. We will obey the spirit and the letter of the laws and regulations that apply to the EU and those established regionally.
- **Operational Challenge:** We recognize the complexities of integrating AI elements with an agnostic specification of ethical considerations. Even though we are fully dedicated to

dedicating the development of AI elements with innovative functionalities, the integration of ethical risk considerations will be promoted at all levels of the project.
- **Industry Risk:** The industry is constantly changing, and sector developments, and mandatory industry changes are not correctly implemented. We will always seek to remain current and adhere to regulations uncles prevented by our infrastructure (all partners considered).
- **Ethical Risk consideration:** Even though ASSISTANT recognize its intention to produce an innovative approach that combines state of the art component to improve processes of product design, process design, and process control, our risk appetite seeks to optimize a high level of performance while achieving the fulfilment of ethical risk considerations

- **Third-Party Risk:** We are willing to consider working in parallel with our partners to conceptualise and integrate e-risks into their current risk management processes.
- **Value Risks:** We are not willing to accept the incorporation of values contradictory to regulations and ethical requirements established by the EC.
- **Domain ethics:** We are considering the integration of ethical considerations that are intrinsic to the domain of application or the domain in which the AI element is embedded.

5.1.3 Risk Management Protocol

This section presents the risk management protocol (i.e. the framework pipeline). Each component explanation would be presented first to facilitate these critical approaches, followed by the whole set of figures representing the overall framework (from page 94).

5.1.3.1 Benchmark e-Risk Management Process

Figure 24 provides a general level of detail for the benchmark framework for e-risk management. This benchmarking framework extends the ISO Risk Management Process by incorporating several supporting tasks that secure ethical and regulatory considerations implementation processes parallel to the classical ISO process (Figure 5). The analogical components of the presented framework with the ISO process are as follows:

- All the boxes, except for the box named "*Execute e-Risk Management Process*" (EeRMP), correspond to the component of "*Establishing the Context*" of the ISO process. A more detailed process has been developed here to incorporate current and future regulations that could be defined for AI assets.
- The box EeRMP also contains an "establishing the Context" component, but it only performs the accumulation and use of the context defined in previous steps.
- The box named EeRMP contains all the iso processes, except the communication and consultation. This is done since the combination of the architecture and policies should impose the frequency and channels of communication over the risk management process.
- The box named EeRMP does contain the ISO-defined "Monitoring and Review" process. However, in order to improve the pipeline flow process, it was defined as a central component after the ISO-defined "Risk Evaluation" and "Risk Treatment" process only (i.e. is not connected directly to the context or the risk identification process). Furthermore, framework updating is enforced in the pipeline structure if new regulations or considerations are required; therefore, the reviewing process has partially been integrated within the framework itself.

The Figure, and the general structure of the benchmark process, are based on a UML recursive pipeline approach. In this figure and the others in this section, the white boxes represent activities to be performed by the stakeholders involved in the AI risk management

process. The diamond boxes correspond to check components, while the blue boxes correspond to a whole process described by another UML benchmark process. The extensions of these last ones are given in the same order as those processes described in the figure. The circles with a number within the figure are used as reference points in the text for better explanation. Finally, The Black dots correspond to an initial point of the pipeline, while the circle with a cross in it represent the endpoint and termination.

Following Figure 24, the first process confirms that AI elements are considered within the evaluation system, subsystem, or component. This process implies understanding and differentiation between AI and other algorithmic processes that should not be classified as AI.

An AI is a system designed by humans that, given a complex goal, acts in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimisation), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems). This definition implies that data is used for both learnings and acting upon. Therefore, those algorithms that have not included these processes should not be considered AI.

If an AI algorithm is considered for evaluation or is embedded within the system, the **e-Risk Identification and Classification** process occurs. This process is defined in Section 5.1.3.2, which focuses on defining the AI elements' intrinsic level of risk under regulatory conditions. This classification is based on the AI act [8] and includes modification if new regulations are defined for the AI elements.

As defined in the AI act, the AI element can be classified as **Unacceptable**, **High**, **Limited** and **Minimal** Risk. After the classification and identification, if the AI element has an unacceptable risk (i.e. yes, in Diamond 1), the AI element's modifications can be done to secure that a lower level of risk is achieved for the AI element. These modifications are based on the idea that they can affect the technical considerations that make the AI element unacceptable.

Nevertheless, if the domain and scope of implementation give the limitation of unacceptability, the AI would not be able to be modified to reduce its level of risk. If these modifications are possible (yes, in diamond 2), the modifications have to be implemented (the process named AI modifications). In it, the required modifications of the scope, data managed, or other conflicts that limit the AI element are re-defined. Otherwise (no in diamond 2), the risk management process is terminated since the AI element cannot be developed, implemented or used. If the AI asset is already underused, decommissioning should be considered.

Following the figure, if the AI risk component has an acceptable intrinsic level (no in diamond 1), the process of **AI Scope Definition** occurs. This process establishes what components, based on the trustworthy requirements, the AI-act, and other regulations should be considered during the risk assessment processes. This component, which is covered in Section 5.1.3.4 and follows the ISO 31000 Risk Management Process, this individual process component is part of Establishing the Context within the risk management process (see Figure **5**).

After establishing an initial context regarding the requirements of the trustworthy guidelines, a secondary context related to values integration within the system is performed.

This process, covered in Section 5.1.3.5 and named **Analysis of Values Definition**, involves establishing what values and requirements can be incorporated that are sound to regulations. In case contradictory values exist, this process involves using decision-making processes using ANP or AHP tools depending on the interdependencies that could exist between values components and criteria. One of the criteria that the ANP and AHP process should be considered are those related to social and legal compliances and, of course, the regulatory and ethical considerations of the AI domain of implementation (e.g. medical ethics). Finally, since these tools allow the evaluation from several stakeholders' perspectives, they can be used to homogenize the perspectives, generating the most suitable combinations of values and the hierarchy they should be integrated into the systems.

After the context of the risk management process is done, the risk assessment, risk treatment, and risk monitoring and review take place. All these previously mentioned components directly specified in the ISO 31000 are encapsulated within the Execute e-risk management process (in section 5.1.3.6).

The ISO 31000 framework established that the risk management process should be dynamic and continual. The endpoint shown in the Figure only helps visualize the process as a pipeline system. Nevertheless, the benchmark ethical framework's idea is periodically used for risk management. Recursive processes are included within the previously described steps, and they should be reinitiated as impactful modifications are made to the system, the company's policies, or the regulations.
Under the current framework, considerable system modifications imply, as MINIMUM, one of the following:

- An additional part has been added to the overall system architecture
- Modifications have been made on the interdependencies of the system's parts that are hierarchically high, forcing other components or subsystems to modify their connectivity or data usage.
- The data source or type is modified
- New functionalities have been added to the AI (e.g. automatization of training processes)
- Interfaces are modified
- The scope of usage or deployment changes.
- Regulations are modified that affect the risk level of the systems or their parts

5.1.3.2 e-Risk Identification and Classification

Figure 25 shows the e-Risk Identification and Classification pipeline. The general approach focuses on setting trustworthy requirements based on: (1) the scope of the AI elements, (2) the domain in which they are involved, and (3) their functionalities. The first process in the pipeline involves analyzing the components under the Artificial Intelligence Act. If this has not been performed yet, the stakeholders involved in the risk management process can proceed independently of this since the pipeline force an initial early e-Risk assessment in case these considerations have not been implemented or the AI element possesses intrinsic functionalities that could lead to behaviours that are contradictory to corporate values and policies.

After performing this analysis, a question addressing if new regulations (or corporate considerations) should be integrated into the framework is done (diamond 1). For the framework and definition of new regulations, we define two types of modifications that could impact and therefore be considered in the pipeline. These modifications (or incorporation of requirements) are over:

- The risk classification and identification processor of the different risk levels (i.e. a new risk level is defined in addition to unacceptable, high, limited, and minimal risk or the regulations and identification process of AI components within these risk levels is modified).
- The regulations level (Higher or lower levels) over AI assets, enforcing them to change their functionality, data usage, security, or other operational considerations.

This list could be extended in the future; the objective is to provide approaches to update the pipeline process, specifically the ***Early e-Risk Assessment*** and the ***e-risk identification and classification*** pipelines. Additionally, if new regulations are required, the pipeline checks if these modifications impact the risk levels defined for AI components (diamond 3).

If answered yes to the previous question (i.e. first item of the previous list), a new cluster(s) (or modification of them) should be incorporated within the risk evaluation process and the pipeline. If so, a whole process named ***Define New e-Risk Assessment*** takes place.

If no new regulations are required, it is evaluated if the user could classify the component according to its risk level (diamond 2). If this was not possible (or was not performed), The pipeline enforces the risk evaluation stakeholders to **Initiate an Early e-Risk Assessment**. This assessment is specified in Section 785.1.3.3 and, as observed in the Figure, corresponds to a hierarchically lower pipeline. Finally, if new regulations are required to be implemented and incorporated into the framework, a distinction of the regulation scope must be defined (diamond 3).

In this process, a clarification based on well-established questions that settled the domain, functionalities, and approaches of this new cluster should be defined. Readers are encouraged to wait until the early e-risk assessment process is covered to understand this point better. The new cluster will imply the definition of requirements (or risk concepts) derived or extended from the trustworthiness requirements. These considerations are evaluated in this new class later in this same pipeline (diamond 7 as reference).

If answered no to the previous question (diamond 3), an internal AI asset process takes place that analyses if the regulatory modifications or AI constraints could lead to a different intrinsic risk level classification. As shown in the figure, the AI asset modification is evaluated to secure the fulfilment of regulations; otherwise, consider the assets as one with unacceptable risk. It is convenient to highlight that these modifications could be derived from the RMP and thus, should consider alternatives to risk treatment given by the new regulatory conditions.

Following the overall pipeline, the process of risk level identification throughout the Artificial Intelligence Act is confirmed (diamond 2). If it was not performed or was not possible to achieve a classification, the pipeline will enforce to perform a process named ***Early e-Risk Assessment***; covered in other sections. Furthermore, the same process is performed if modifications were performed over the AI regulatory conditions or new risk classification levels were incorporated.

Independent of the case of modification, the ***Early e-Risk Assessment*** process will be initiated, considering that the AI asset possesses an unacceptable risk level if there is any violation of the new regulations. If under current use, the AI asset should be modified to achieve a tolerable level of risk before being considered for decommissioning.

After performing an ***Early e-risk assessment*** or having defined the risk level of the AI assets (Yes in diamond 2), a pipeline evaluation for setting minimal trustworthy requirements, and thus risk attention, is done. Further requirements can be added depending on the

companies' policy interests; this framework only helps set minimal considerations, as risk components, for users and developers regarding AI assets.

Following the pipeline, if the risk level of the AI component is defined as Low Risk (yes in Diamond 4), the MINIMAL consideration to be implemented in the AI development involves Societal and Environmental Well Being (as mentioned in the process named Consider Low/Minimal Risk). Implementing environmental and societal reflections in any AI components does help the company to execute the process with a perspective on sustainability but is not enforced under current regulations (e.g. Europe's current regulatory conditions).
There is no need to specify the economic perspectives associated with economic benefits since they are enforced by companies' interest in an external e-risk management approach or can be incorporated as values, if necessary, for evaluating possible discrepancies between all the users' points of view.

It is essential to mention that for the current framework, the considerations established based on the risk levels define the analyses that will be taking place during the RMP. Nevertheless, the processes of treatment, tolerate, transfer, or terminate of the AI assets or their functionalities will be dependent on: (1) the likelihood of an event to occur, (2) the outcome that could take place if these events materialise, and (3) the risk appetite established by regulations and the companies policies and interest.

Some exemplifications of companies' policies were presented previously. Therefore, this consideration should be considered for this and upcoming intrinsic risk levels.

In the case that the AI asset possesses an intrinsic limited-Risk (no in Diamond 4 and yes in Diamond 5), the MINIMAL set of requirements established for the AI components are: (1) Societal and Environmental Well Being, (2) Transparency, and (3) Technical Robustness and Safety. The need for transparency is based on the AI act; the need for societal and environmental well being follows here and after the same consideration as that established for low/minimal risk AI assets; The need for Technical Robustness and Safety are included to foster quality and efficiency in the manufacturing sector (as described in the introduction).

In the case that the AI asset possesses an intrinsic high-risk (no in Diamond 5 and yes in Diamond 6), the MINIMAL set of requirements established for the AI components include, in addition to those requirements established for the Limited-Risk: (4) Human Agency and Oversight and (5) Accountability. The difference between previous risk classification and this one is that there are obligations on adequate risk assessment and mitigation systems as established in the AI act. This implies that the risk appetite should be more severe and thus, secure appropriate human oversight, high level of robustness, security, accuracy, and minimisation of risk derived from biased information. Furthermore, the increase in risk appetite will define lower tolerance on AI assets and, therefore, will foster the implementation of treatment or terminate conditions, if needed, during the e-risk management process.

After these evaluations, the possibility of extending the classification, and its test, is done throughout a specific evaluation (Diamond 7). As mentioned in the diagram, this new class should consider, as MINIMUM, the previous reflections established by the corresponding intrinsic level defined by the Artificial Intelligence Act or that extracted from the Early e-risk Assessment step. In addition, further considerations could be included in this new class that should not contradict those established by local and global regulatory conditions (e.g. Charter of the EU concerning fundamental rights).

Finally, suppose any of the previous stages did not classify the risk level of the AI asset. In that case, the AI is considered an unacceptable risk, leading to a restriction to its

development, a decommissioning if currently used, or a considerable modification of the AI scope that could secure the intrinsic level of the AI component to a lower one.

5.1.3.3 Early e-Risk Assessment

Figure 26 Early e-risk identification shows the pipeline process to define the intrinsic level of the AI element. The pipeline is constructed so that the intrinsic risk level under evaluation decreases (from higher to lower risk considerations). This consideration should be taken into account when new regulations or classes need to be incorporated into this framework and, therefore, the new classes identifications component should be placed between intermediate risk level classes. Furthermore, if new identification components are placed over a risk class, they should be placed as an evaluation component (i.e. diamond structure) at any position within the blocks that define a specific risk class.

The first part involves evaluating and understanding the Human Rights Considerations. This implies, for one part, an understanding of the type of information handled by the system, its goals, objectives, and possible deviations that could have over expected functionalities. On the other hand, it requires a complete understanding of the Human Rights requirements [66]. These considerations are assumed to be known by the frameworks' users, and therefore, Human Rights considerations are left as a checking process. After performing this step, eight questions extracted from the Artificial Intelligence Act are used to define, under the human rights considerations, if the AI elements should be considered as one with an unacceptable risk level. These questions include, for example, understanding if the system is contravening human dignity, freedom, democracy, equality, the rule of law, solidarity, justice, and the right to life. For a complete understanding of these concepts, readers are encouraged to review [66].

After the Human Rights Considerations, the AI functionalities are evaluated. As observed in the figure, four questions (diamonds) are used to evaluate if the AI functionalities and domain of applications are considered unacceptable and, therefore, with an unacceptable level of risk.

Again, if other considerations are defined in the future over regulatory conditions that make an AI have an unacceptable risk level, the pipeline process can be extended to accommodate these new considerations further. A box named "Do you add new considerations for unacceptable Risk?" was added to secure the extension of unacceptable risks.

The following risk level considered for implementation and classification is the **High**-Risk Level. Since, at this level, no Human Rights elements should be vulnerable under the AI functionalities, the evaluation focuses on the domains and functionalities of the AI. These questions involve social scoring, law enforcement, human resources, etc. The list of questions is based on the EU AI act and should incorporate local regulations or future AI constraints in the future that will signify the higher level of risk AI and, therefore, full consideration of the requirements of the trustworthy guidelines. A set of 15 questions is currently used in the AI high-risk level identification.

Nevertheless, further questions can be included to consider local, sustainable, or corporate definitions. Two boxes (17 in total – last two) define corporate responsibilities and values to restrict them to ethical requirements. Finally, an extra box is added (box 18th in the high-risk process) that helps to remember to incorporate new considerations or regulations concerning High-Risk Considerations.

The next level of risk is the Limited-Risk Evaluation. This risk level focuses on the AI impact on the system and environment in which they are or will be implemented. This level should consider local and sustainable considerations or additional corporate definitions that can be **downgraded** (not violated) by the AI component. Therefore, similar conditions as those

established for **High-Risk** considerations could be placed, with the distinction that these conditions can be downgraded instead of violated. In total, nine considerations (boxes) are included in the current status of this framework.
The consideration of downgrading establishes the need for values measuring and, therefore, KPIs are required to measure the level of downgrading of acceptable conditions. Similarly to previous clusters, the **Limited-Risk** cluster includes the possibility of incorporating further considerations or regulations by extending the box named "Do you add new consideration for **Limited-Risk**".

Finally, If the AI does not belong to any of the previous risk clusters, the AI component is considered Low-risk for the risk management exercise.

5.1.3.4 AI scope Definition

Figure 27 AI Scope DefinitioFigure 27 shows the AI scope definition pipeline. This component focuses on extending the e-Risk identification and classification pipeline by analyzing to greater detail considerations based on information used by the AI component and the possible interactions that could be defined between AI and agents. In other words, this pipeline evaluates the possibility of biases and personal information usage in greater detail, which impacts the requirement of DnDF. In terms of AI and agents interactions, it focuses on the level of automatization left over by the AI component that translates on requirements over the agency of humans on decisions made by the AI component.

These agency components are connected to the Human-Centric perspective and, depending on the risks considerations of the AI application, are linked to human-in-the-loop and human-on-command needs. As observed in the Figure, the pipeline starts by performing the **Data consideration structure** process. This process focuses on analyzing and developing a complete understanding of the type of data that the AI will be managing, what type of transformation, if any, would be made by the system, and considering which of the sources or outcomes of the AI assets could have biases that could derive on issues related to DnDF. Additionally, given current trends, ask for analyzing under the GDPR regulatory framework the type of data that will be managed and kept, interacting and managing by the AI asset.

After the users perform the previous process, questions (5 in total) are used to analyse the MINIMAL conditions to establish the need to incorporate DnDF requirements. These questions are linked to historical records, output information, disabilities, and other considerations. A specific question (Diamond 1) allows extending the analysis with greater detail and, therefore, allows the extension of the current frameworks as further definitions are constructed with greater detail over the capabilities of AI components and their impact on DnDF topics. A specific question related to disabilities is used to check the impact of AI on the disability and vice versa. This question allows establishing if AI would impact disability or if the disability restricts the use of the AI assets.

After the DnDF definitions, a definition step over a MINIMAL analysis over Data Privacy and Governance. Even though there is only one evaluation performed over these considerations, the framework allows, as observed in the figure (diamond 2), extensions to easily be incorporated depending on new definitions or companies' interests to include these considerations within the risk management analysis.

Immediately after the Data Privacy and Governance analysis, a step dedicated to Human Agency and Oversight. In order to do that, a process named **AI Scope** is performed first. An internal analysis has to be performed over the AI assets, all the interactions, and between agents and AI components. These interactions can be direct, such as a user-UI interaction, and indirect, such as patient-AI predictions components that could substantially impact decision-

making (e.g. AI-image cancer prediction). In general, all these processes would require consideration from the Human Agency and oversight. Therefore, they would require some specific definitions, depending on the AI behaviour, the responsibilities that will lay over humans, the control the AI will have over human decisions, and to define until what point human-centric considerations will apply over the AI asset.

If more than one type of agent is under the AI's approach, the analysis should be driven on a per-user base. Similarly, if more than one interaction with the same AI tool but under different UI interfaces, a differentiated analysis should be driven based on each UI interface's functionalities (based on the hierarchical structure definition previously established – i.e. see Figure **22**).

The final process used to define what requirements should be established to be included within the AI management process is the **ALTAI tool**. The Assessment List for Turtworthy Artificial Intelligence (ALTAI) for self-assessment tool [67] supports the actionability of the critical requirements outlined by the Ethical guidelines for Trustworthy AI [5].

The ALTAI tool aims to provide a basic evaluation process for Trustworthy AI. First, it helps users understand what Trustworthy AI is and what risks an AI system might generate. Second, it raises awareness of the potential impact of AI on society, the environment, consumers, workers and citizens. Third, it promotes the involvement of all relevant stakeholders. Finally, it helps gain insight into whether meaningful and appropriate solutions or processes to adhere to the requirements are already in place or need to be put in place.

This step aims to bring further awareness to the current framework users about what other requirements could be considered (not MINIMAL) to be incorporated within the e-Risk management process. The MINIMAL requirements were established in previous stages, and thus, the use of the ALTAI tool is complementary to the current framework but not enforced.

5.1.3.5 Analysis of Value Definition

Figure 28 Defines the pipeline to check if values could be incorporated within the AI risk management framework. In order to do that first, as observed in the figure, a checking process is performed that the value(s) to be incorporated should not be, first, contradictory to the values established by the EC. These values correspond to Human Dignity, Freedom, Democracy, Equality, Rule of Law, and Human Rights, and they should be checked [66] and understood if there are at any stage some values to be incorporated that could contradict them.

As observed in the pipeline, the first question does address the topics related to adding values additional to those established by the EU since those, under the current framework, is the MINIMAL condition established as values defined by the AI act [8] (i.e. established as those with unacceptable conditions). Even though the AI that contradicts the EU values should have been screened out at these stages, given that their nature is of unacceptable risk, an additional check is made to see if these values contradict the EU fundamental rights (Diamond 1). As expected, these values cannot be incorporated or considered for analysis and the AI development or use should terminate, leading to a decommissioning process if underuse. If the value(s) are not contradictory, a value hierarchy should be defined to address their incorporation relative importance, especially if the values to be incorporated are contradictory. If the hierarchy has not been defined, a process named **Define Hierarchy** is initiated. In this process, the hierarchy is recommended to be driven first by weighting all the desired values to be incorporated within the risk management process, followed by a discretization process that will define the most relevant values to be incorporated first.

Based on the current framework, the recommendation is to use a decision-making-based approach. More precisely, the ANP or AHP processes are used to define the hierarchy of the values and the definition of what method to use strictly depends on the possible inter-correlations between criteria and values defined in the analysis. The criteria that should be considered within this approach are listed in Table 16 Decision-Making Considerations. This does not imply that other criteria could be added to the analyses.

In the case of ASSISTANT, the list of criteria would be analysed by the case study to reflect and propose the MINIMAL criteria in the manufacturing sector. Additionally, further reports about ASSISTANT will be dedicated to evaluating what other values could be interested in being incorporated into such analyses.

After performing the ANP or AHP processes for hierarchizing the values, a discretization process to define the most relevant values to be included is recommended. To do that, methods well-known, such as Pareto 80/20 or others could be used to eliminate those that would have a relatively low impact on the system's functionality or that are highly contradictory to those that are hierarchically relevant.

The final consideration of the current pipeline involves defining metrics to track the values desired to be incorporated into the system. These metrics can be qualitative or quantitative, and, as will be seen later on, both can be used for the e-risk management process. If several values will be considered incorporated in the framework, it is recommended that these metrics be normalized or managed to analyse the effect over the different e-risk treatment processes comparable between values effect. This is important, especially in cases where contradictory values are incorporated, and therefore can help measure the relative inverse effect of the values to be incorporated.

The normalization process can be made by considering “best scenario” and “worst scenarios” considerations, allowing fixing the cap value of the metric. This will allow standardizing each value measure into a quantitative form using the following equation.

$$\frac{V_a - V_w}{V_b - V_w} \tag{1}$$

In the previous equation, V represent the qualitative or quantitative estimates for the values-based variables under the actual state (*a*), worst state (*w*), or best state (*b*). The incorporation of these standardized metrics would allow evaluation of each state modification based on previous conditions (η_p) or its best (η_b) scenarios (see equations 2 and 3, respectively).

$$h_p = \frac{V_{a,new} - V_{a,old}}{V_{a,old}} \tag{2}$$

$$h_b = \frac{V_{a,new} - V_{a,old}}{V_b} \tag{3}$$

In these equations, the “new” and “old” captions describe the previous and current states of the values-based variables. This is a helpful index to evaluate the effect on the system when modifications are performed.

5.1.3.6 Execute the e-risk Assessment Process

Figure 30 shows the e-Risk Management process Pipeline. As observed in the figure and previously defined, a small process of **Establishing Context** is performed before processes commonly linked to ISO 31000 steps (i.e. Risk analysis, Risk evaluation, Risk Treatment, and monitoring and review) are run. The figure shows that two processes are extended

hierarchically in other pipelines (Risk analysis and evaluation – see section 855.1.3.7 and Risk Treatment, Transfer, Tolerate, or Terminate – see section 5.1.3.11). The first process in the pipeline (**Establishing Context**) allows incorporating all previous requirements definitions and establishing the interconnections that AI will have with users and other AI components. More specifically, this process looks at:

1. **Connectivity with other components and subsystems**: This allows a clear understanding of how AI affects internal and external parts of the overall system. This is relevant, especially if there are several processes or AI that are orchestrating at the same time in order to perform systems outputs. Additionally, having a clear understanding of the connectivity with other components and subsystems (including visualization and interfaces) allows understanding the secondary effects that if a risk will materialize (i.e. failure mode), what would be the impact on the own part and parts connected to it and therefore a more accessible establishment of accountability on more complex systems. This connectivity is typically established in software development systems by establishing software architectures that define software elements, relations among them, and properties of both elements and relations.
2. **Dependencies**: Dependencies are the hierarchical extension of previous definitions. This allows specifying what parts and components will describe a cascade effect on risk analyses and help understand what parts should drive the greater attention in terms of risk analyses.
3. **Constrain and Context**: Constrains are delimitations of the parts' functionalities, inputs behaviours, systems outcomes, and components' values. If relevant, these constraints can directly or indirectly relate to physical values (i.e. its physical context), given a higher degree in considerations of systems security, especially for those cases related to AI-user interactions.
4. **Diagrams**: System representation linked to numerals 1 and 2.
5. **Requirement's definitions**: Agglomerate all the previous requirements that would be important to include within the risk management process (MINIMAL and additional ones). As specified in the diagram, these are obtained by the previous analyses performed by the e-Risk identification and Classification and AI Scope Definition processes.
6. **Values**: The values to be incorporated into the framework after performing the Analysis of Values Definition process.

The following two components (Diamond 1 and 2) are helpful to analyse whether the risk management process would be run in parallel or not with other risk management processes. Specifically, Diamond 1 focuses on the design process, which will involve performing in parallel the well-known DFMEA process (where D stands for design FMEA), while Diamond 2 involves performing in parallel the well known PFMEA process (where P stands for Process).

The PFMEA is a helpful tool run by institutions to identify and evaluate the potential failures of processes, which involves the possibility of already under use processes. At the same time, the DFMEA is a systematic group of activities to recognize and evaluate potential systems, products, or processes failures and, therefore, involves a more comprehensive analysis of systems part at the early stage of systems development. Figure 31 describes a structural diagram in which the current framework is merged with other risk management processes based on FMEA approaches to understanding better the approach used for merging risk management processes.

As observed in the figure, each component evaluated under the risk assessment process could follow an ethical base analysis (ethical failure modes in the figure) and, if considered, run the DFMEA and PFMEA processes. For simplicity, the figure only describes the context in general perspectives from an FMEA-based analysis. Before using the FMEA/FMECA, approaches such as contextualization should be run before considering merging the ethical-based and non-

ethical-based FMEA/FMECA. This should be done to secure that the scope of the risk assessment process allows for the running of both processes simultaneously. In other words, this implies that AI ethical based failure modes would consider as another item or component failure mode, but with the definition that they could have different severity, causes, mitigation, and likelihoods.

Following the bottom pipeline, different failure modes will be identified and analyzed, depending on: (1) the risk considerations constructed during scope definition and (2) all the items, interfaces, and components considered in the analysis. The analysis will follow the structure of the FMEA approach described in Figure 13. Therefore different stages will be run, such as Define the Analysis, Identify Failure Mode, etc. For simplicity, only a few components are included in this figure.

Independent of the ethical requirements established for analysis during the contextualization process, the FMEA will analyse the causes of the ethical failure modes. The likelihood of their events, their severity of materializing, the detection capacity of the failing conditions, and alternatives for mitigation will be estimated. As described in the FMEA process, Risk Priority Numbers will be estimated for each failure mode (RPN_i) that could be combined using Equation 1.

The combination of the DFMEA RPNs and ethical based RPNs could follow the same technique described in this equation. Nevertheless, we recommend (in ASSISTANT) to address a differentiated weighting factor (w_e) that would represent the relative importance of purely ethical based failure modes concerning those based on technical components (i.e. technical robustness should be considered as a technical component). This weighting factor could range between 0 and 1, obtaining a combined RPN ($combined\ RPN = w_e \sum RPN_{i,eth} \quad \alpha_{i,ethic} + (1 - w_e) \sum RPN_{i,technical} \alpha_{i,technical}$). The combined RPN could be used as a first-hand metric for item analyses. Nevertheless, Critical Analysis should be performed for such. Critical numbers will be built if criticality analyses are performed on each process (after the e-FMEA and the DFMEA/PFMEA) (section 4.5). Similarly, the criticality numbers could be combined for items using the expression used in the corresponding section or the weighting factor $w_e . \boldsymbol{RPN_{item}} = \sum_{i=1}^{n} \boldsymbol{RPN_i \alpha_i}$ (1

Following the pipeline, the **Risk analysis and Evaluation Process** takes place. Different tools are used in it depending on the AI functionalities, the type of information collected, and the pre-specification of the system. Specifically, at this stage, a Failure Mode and Effects Analysis is proposed as the main component to understanding the system risk, and the implicates of its failing conditions. This failure mode that focuses on e-Risk, here and after also named e-FMEA, complements other failure modes and analyses all the conditions linked to trustworthy requirements, ethical and values considerations. A thorough description of this process will be made in the corresponding section.

Given the iterative nature of the Risk Management Process, as Failure Modes are specified for the system (or its parts and components), they can be integrated and kept for posterior analyses from the same system or for being considered on other systems that describe similar functionalities, interactions, or data usage. To keep this information, after performing the Risk Analysis and Evaluation Process, a question (Diamond 3) define if the Update Definitions process can be run. If run, the new failure modes should be included within the following analyses and documented (Risk Register and other channels defined in the Communication Protocols) with enough specificity and scope for a possible extension to other parts and components.

After the Update Definitions or if there is no requirement to perform this process, the Risk Treatment, Transfer, Termination, or Tolerate Process is performed. In this step, different

actions can be taken depending on the risk appetite, the risk levels, the probabilities of events occurrence, and the chance of detection of the risks to be materialized. In other words, it is defined what part, component, or system should be:

- Treated: modified, upgraded or include enough safeguards to reduce the risk level of the failing condition happen
- Transfer: if the risk level allows it, use external safeguards approaches such as insurances that allow transferring the responsibility of the events if they materialize)
- Terminate (i.e. stop the development and use stage of the AI part, component, or system. Proceed with decommissioning if necessary)
- Tolerate (i.e. Do not perform any part or component modification, keep the analysis of it and continue updating the status of the elements under evaluation with the frequency established in the risk management protocols)

A thorough description of this process will be made in the corresponding section.

The following process corresponds to Estimate KPIs. KPIs linked to the Risk Management Process, each Risk component, and KPIs directly linked to the methodologies used for risk assessment should be estimated. The section dedicated to this corresponds to Section 5.3 and will not be covered here.

The following process corresponds to the update of the Risk Register. The Risk Register corresponds to a risk management tool that acts as a repository of the risk identified throughout the risk management process. It includes diverse information that helps keep track of the propositions made for risk management, KPIs and, among others, relevant information related to the methodological methods used for evaluating risks (e.g. FMEA/FMECA). An entire section is dedicated to the Risk Register (Section 5.2.1.1) and describes the proposed Risk Register for the Current Framework.

The following process, Monitor, involves the internal evaluation and comments on the risk management process. This involves evaluating the correct application of the risk management processes and, at the same time, generating feedback that would allow improvements over the protocols used. The E-risk board should define the implementation of these management processes after being collected and reported by the Executive Risk Management Committee.

After the Monitor processes, different questions (Diamonds 4, to 7) are used to evaluate modifications related to the AI Interaction with other components, the data structure managed by the AI components, the incorporation of other AI components or functionalities that were not foreseen to be implemented, the incorporate further functionalities of current AI that can impact the trustworthiness of the system. Depending on the response nature, updates must be performed over the Establishing Context process and are required to rerun the risk management process. All the previous possible modifications should be derived from the Risk Treatment components since the rest of the T's (in the 4T's of risk management) would not affect the architecture or functionality of the system. The implementation of treatment under the current framework is not proposed until a complete understanding of the implications is performed. To do this, and as observed in the Figure, several updates or restarts of the framework analyses (**Update Interactions**, Update of **Requirements**, or **Restart.**) force to analyze the implications of the proposed modifications with the new considerations. These propositions will be kept track under the risk register, and therefore, a binnacle will be kept for understanding the process behind the risk treatment propositions.

Once there are no further updates on the risk treatment components (i.e. No in Diamond 7), the review, update, and implementation process occurs. These steps strategically

correspond to defining what strategies for risk treatment will be implemented based on the performed risk assessment processes. This process is directly connected to Ethical Risk Architecture (Figure 23) since it involves the interactions between the Management Committee, which will comply with the risk register, the Executive Risk Committee, which will review and report the findings, and the E-Risk Board that will define, and inform back, the strategies of implementing the risk 4T's. For a complete understanding of how the recommendations and strategies should be followed for defining what processes of the 4T's will be followed, readers are encouraged to wait until the complete Risk Protocol process is presented.

Finally, The last question (Diamond 8) checks if the risk implementation processes considerably modify the AI components' functionalities or interactions. If so, a process of re-evaluation of the risk levels of the AI components (based on the early risk assessment process, see section 5.1.3.3) should be defined.

Users can extend these considerations to incorporate different conditions that will force the re-evaluation of the whole risk assessment process.

Additionally, the E-risk board should coordinate an audit process to evaluate the implementation, in due time, of the corresponding strategies implemented throughout the 4T's considerations. However, this does not imply that internal audits should be applied only when substantial modifications are performed over the AI components or the general architecture.

5.1.3.7 Risk Analysis and Evaluation

Figure 31 shows the pipeline used within the e-Risk management process dedicated to Risk Analysis and Evaluation. This pipeline focuses first on defining what instruments, between the FMEA, FMECA, or the RCA (Root Cause Analyses) approaches will be used for the risk assessment to force the running of the selected ones. If FMECA is done, the Critical Analysis is performed after the FMEA process.

As observed in the figure, thirteen questions define what instrument(s) will be used for risk assessment. The first question, Diamond 1, directly tests if a predefinition of the instrument is done. This is important, especially in the iterative process of risk assessment. In order to keep the integrity of the metrics used for dynamic evaluation, the same instruments should be used. If a modification of the instruments is needed, framework users are encouraged to include new and old methodologies until enough records for risk assessment exist. Other instruments outside the scope of the current framework can be implemented, but given the well the nature well documented and the possibility to be linked to the users' ongoing risk management process, the present framework encourages the use of the FMEA/FMECA approaches.

Starting from question number 2 (Diamond 2), eleven questions are used to see if the most convenient approach is the FMEA. These questions include (following the same order as shown in the figure):

- If interested in identifying all failure modes
- If the top events (risk conditions) can be defined or limited to a few events. This consideration implies similar agglomerate events in a cluster that share commonalities and correspond to the same failure mode. (e.g. two different tanks of water and oil both will have the same failure mode if they run out of fluid, in other words, an empty tank failure mode).
- If the AI includes human or software updates. The nature of the FMEA process has shown great applicability and effectiveness in such fields.
- If the system is in the early stage of development
- If the system is going to be modified considerably, that will imply several functionalities or interaction modifications.

- If the system is required to be evaluated by its robustness, which implies a thorough analysis under critical analyses and relevant use of metrics to keep track of the system robustness.
- If there are considerable modifications to the system (based on the description given in Section 5.1.3.1 as the MINIMAL conditions for strong modification considerations).
- If it is required to quantify the risk levels.
- If there is considerable human intervention over the system that can cause erroneous functionality of it
- If there is some need to have a deep understanding of the cascade consequences of events.
- If there is a need to completely understand events, consequences, and impacts.

Immediately after the set of questions, the most suitable approaches (between FMEA and RCA) should be run. In case that FMEA is run, if enough information regarding qualitative and quantitative information (i.e. likelihood and severity – for quantitative and competent expert knowledge – for qualitative information of the e-Risk component to materialize). Then, two questions are used to define if a qualitative Criticality Analysis or quantitative Criticality Analysis should be run (diamonds 3 and 4).

In both cases, there is the consideration of constructing heat maps and performing analyses, but with the differentiation that quantitative information allows the incorporation of numerical analysis values based on the probabilistic information collected about the failure modes.

If a Criticality Analysis is not performed, the FMEA can also be used to construct heat maps. Nevertheless, these heat maps are based on intrinsic e-FMEA KPIs (Risk Priority Number - as specified in the implementation of FMEA and the KPIs section). In such cases, precaution should be used to define managing strategies since they are not entirely based on quantitative information.

Finally, Heat Map Construction and Perform Analyses are made as observed on the pipeline. The heat mat construction can generally be linked with Criticality Analysis, but since they can be constructed based only on FMEA information in the present Framework, they are left as a separate process (covered in section **Erreur ! Source du renvoi introuvable.**). Description of the Heat Map construction is left in other sections that describe the implementation of the FMEA and FMECA approaches

5.1.3.8 Use FMEA

Figure 32, Figure 33, and Figure **34** Show the FMEA approach based on the description shown in Section 4.4. As described in the pipeline, the whole process has been divided into three main steps. These steps are: (1) defining if merging with other risk management processes will take place (and executing it - Figure 32), Identifying the failure modes and determining the rates for the failure modes (Figure 33), and performing the post-analysis together with placing the information in the risk register (Figure **34**). Since most of the previous information related to process, interconnectivity, diagrams, and scope should have been defined in previous stages of the framework, the transfer of that information for the merging process should be direct. The merging process was previously explained in section 5.1.3.6 using Figure 29 and will not be thoroughly covered here.

As observed in the pipeline, the first question is if PFMEA or DFMEA analyses are used parallel with the current framework (diamond 1). Suppose the answer is yes; another question (Diamond 2) checks if the scope of the FMEA and FMECA approaches allows extension and or merging with the current framework. In other words, what components/items are analyzed by the previously mentioned processes and the policies established by the organisations running

them. If the current focus does not allow it, the pipeline will focus only on the e-FMEA process. If it does, depending on the level of implementation, a stage of “define and merge DFMEA/and PFMEA” occurs. It defines the strategy to follow for the scope and functional block constructions - i.e. extends the functional blocks if the process is running or create and extend the incorporation of AI functionalities on the functional blocks if the process is starting.

On the other hand, Dimond 2 also forces the e-FMEA analysis's execution since, regardless of the answer to its question, it redirects to diamond 4. In Diamond 4, it is asked if the whole life cycle of the AI components is considered for the analyses. If it does, an e-FMEA approach should be considered for each stage of the AI life cycle. This consideration is implemented since the risks involved during development, use, or decommissioning could be considerably different. Furthermore, by considering in advance processes that have not started yet, a better implementation of the following steps of the life cycle could be achieved. For example, by considering AI decommissioning from the beginning, policies could be placed to eliminate relevant information and transfer such approaches with the same level of security during the AI use stage.

After defining and merging the process of the DFMEA and PFMEA with the e-FMEA takes place, a question (diamond 3) is used to check if failure modes have been previously identified for the design and or process analysis. It should be taken into consideration that by integrating the AI component in the analysis, different functionalities and dependencies could be incorporated in the DFMEA and PFMEA, and, therefore, new failure modes could occur. Therefore, a well-defined interdependencies link should be generated to rationalise such failure modes during the merging process. If there is a need to define new failure modes for the process or design FMEA, a step named Identify Process and/or Design Failure Modes occurs.

Going back to Diamond 4, after its execution, the last question in the figure checks if the AI is used within maintenance or operational processes. If it does, it is executed a process named PFMEA that allows the user to extend the risk management with the PFMEA process. If it does not (or is not considered the execution of the PFMEA process, or has not defined the) the identification of failures modes is started.

Independent of the processing direction used in the analysis, two identification processes of failure modes can take place (see Figure 33). The first, named Identify Ethical Failure Modes defines, based on the requirements established in the ***e-risk identification*** and the ethical considerations to be performed in the FMEA analyses, the possible failure modes that could take place. For ASSISTANT, this stage will be exploratory, and thus several failure modes will be proposed that could extend, modify, or eliminate (due to lack of relevancy) those previously stated. The second process corresponds to the identifying process and defines failure modes in cases that have not been recognized before (or require an update). In such a step, based on the approach (process or design analysis), the failure mode of the same system, sub-system or components under the analysis of the present risk assessment are identified.

For each of the AI components, each of the system life stages should be considered, and for each ethical consideration, a definition of the Failure Modes (FM) should occur. If this stage has not been addressed before or is required to incorporate new ones, they should be thoroughly defined, explained, and recorded within the risk register. Consider the different life stages, if applicable, to the process and design of the system, sub-system and components.

The dashed arrows included in these figures help as a visualization tool in the pipeline process but do not involve executing any task on the part of the framework users. As observed in the pipeline, identifying ethical failure modes will produce a series of failure modes “classes” linked to the trustworthy components and values set during the contextualization of the risk assessment. Eight different encapsulated areas (human agency and oversight failure mode,

transparency, etc.) are shown for simplification, linked to the trustworthy component. Nevertheless, the failures modes should be based on the **Ethical-Based General Failure Modes Families** described in previous sections.

After identifying the failure modes, ranking the occurrence likelihood takes place. As observed in the pipeline, for each failure condition on each component failure mode/root, the occurrence rank (O) should be estimated. For the case of robustness analyses, confusion matrix analyses from the AI component can easily be linked to the probability of occurrence. Therefore we recommend that ASSISTANT use this approach as the first source of rank information. For further understanding of this process, please check the dashed box in the diagram). If no probability numbers or historical information is accessible (i.e. quantitative information), expert judgment should drive the estimation of the occurrence level and, thus, the ranking.

Next, the failure modes ranking based on the severity analysis takes place (process name Severity Analysis in the figure). For each failure mode, evaluate the consequence/severity of a failure mode in terms of the operation, function, or system status. Consider that a failure condition may be caused by one or more failure modes and could be linked to the PFMEA and DFMEA (if considered). Furthermore, consider if its effect is local (within the same subsystems) or global (other subsystems and systems). Finally, consider each of the remarks established in Section 4.6.

Similarly to the severity ranking, a dashed box references the ranking system. Nevertheless, the severity ranking cannot be linked to probability ranges since it will be a situation dependant. Nevertheless, the definitions used before for Catastrophic, Critical, Marginal, Minor or Negligible should be used.

Finally, the failing detection raking (D) takes place. Evaluate the way the failure mode/root cause is detected and the means the user is made aware of the failure conditions and system status. The detection can include messages, identifiable metrics, alarms, human perception, process stops and process halts. Notably, some traits linked to ethical concerns could not be immediately recognizable. Thus approaches should consider the time frames required, methods, and impacts that a failure mode could take over the system.

If the occurrence was estimated using a confusion matrix and algorithmics, the same approach or methods used to estimate the matrix (not its probabilities) could be linked to the detection method. If detection is performed not locally (i.e. through external subsystems or components external to the local subsystem), consider making the risk appetite more stringent to provide for the failures involved in the communication channels used for information transfer.

The final steps of the FMEA process are shown in Figure **34**. These processes involve:

1. Analyse the causes of potential failures, define recommendations for the failure modes identified (i.e. how to reduce its likelihood or its severity)
2. Define detection methods if not specified (i.e. reduction of the detection-ranking component).
3. Populate the risk register with the collected information
4. Specify any recommendation about the failure modes (not the FMEA analysis or the framework; these processes are performed later on).

When specifying how the failure could occur, we recommend performing the description of something that can be corrected or controlled. Additionally, perform the decryption of the cause following an "if X occurs, then Y happens" where X is the cause, or RCA root and Y is the failure mode or the origin of the risk condition in the RCA approach specification if performed

the RCA. Use the RCA information only if the FMEA was not performed (as described in the Risk Analysis and Evaluation pipeline).

When specifying the recommendations (i.e. Define recommendations and estimate the Risk Priority Number process in the figure) related to failure modes incorporation, failure compensating previsions, or modifications that could prevent, reduce, or increase the detection of the failure mode effects, in other words, perform recommendations that will reduce the risk priority number (RPN = O * S * D).

Finally, as observed in the figure, the Populate FMEA forms process occurs. In other words, populate the risk register components related to the FMEA (i.e. the risk register, presented later on, contains the FMEA and the CA components together).

5.1.3.9 Use RCA

Figure 35 defines the RCA pipeline process. As observed in the figure, the RCA involves four processes named: Scope of the RCA, Root Cause Analysis, Analysis and recommendations and Record Results. The first one involves studying each of the considerations established during the process named "Establishing the Context" of the risk management process pipeline (previously defined). Differently to the FMEA process, the RCA process is a Top-Down approach; thus, by establishing the context of what trustworthy components are the most relevant ones for the risk management process, the root causes analyses methodologies should be used to extract the causes (expected more than one) that could produce the undesired event (i.e. diminishing on the trustworthy component). This step also involves collecting all the relevant data that would help to drive the RCA analysis. Even when the process could be performed qualitatively, quantitative approaches should be fostered (e.g., to implement FTA). Therefore, in ASSISTANT, we would, if possible, use quantitative analyses, mainly if RCA approaches are applied.

After the scoping process, the application of the RCA methods is performed. Depending on the information collected and the type of analysis required, the 5why's, change analysis /event analysis, fishbone diagrams, or FTA should be performed. The 5why's can be used with other methods to facilitate their implementation. Further definitions of applying them have previously been covered (see section 4.9.1).

The following step involves the analysis and recommendation based on the RCA analyses. This specification should (1) explain the root cause claims and (2) provide a corrective course of action for the roots of the problems detected.

The final step involves the recording of the results using the risk register. The risk register defined in ASSISTANT is based on the FMEA approach; nevertheless, the areas related to failure modes identification, recommended actions and explanation of the causes can be populated with the analysis performed in the previous process. Furthermore, as seen in the risk register, each field with a "*" can be populated if an RCA is performed. Finally, it should be clarified in the notes that register that the observations are based on the RCA analysis.

5.1.3.10 Use Critical Analysis

Figure 36 shows the Critical Analysis pipeline. The pipeline follows the specifications established in section 4.5. As observed in the figure, the first step involves defining if the criticality analysis would be performed qualitatively or quantitatively.

If partial information is gathered (i.e. some failure rates would be obtained), those failure modes in which failure rates are obtained should follow the quantitative branch of the

pipeline, while those that only probability information or expert judgment would be used would follow the qualitative analyse. Independent of the path followed and, as observed in the figure, all processes will end up in an agglomeration stage in which the critical numbers will be accumulated per item.

Following the pipeline, if the quantitative analysis is performed, the first process is to secure or estimate the failure rates for the failure mode under the quantitative analysis. Then, parameters such as the conditional probability are estimated to estimate the failure mode critical number (in the process with the same name). The pipeline adds the equation to facilitate the readers' understanding of the involved stage.

If a qualitative analysis is performed, it is essential to define if the process will use any probability number or will be used only expert knowledge. The occurrence level category process would occur for those failure modes based on expert knowledge (NO to the question "Do you have probability numbers to use?"). It estimates the occurrence level category (from A to E) based on expert judgment regarding system/sub-system/component fixed operational times. The operational time should be the same for each failure mode under consideration. If probability information is used, the probability numbers would be used to estimate the category levels (from A to E).

Once the category levels are defined for all failure modes with qualitative information, the critical number would be constructed based on the specifications mentioned in section 4.5 (i.e. use of the referencing figure and multiplier parameters to set the criticality number).

Once all the criticality numbers are generated, they can be agglomerated on the process named "Agglomerate". Again, the equation for agglomerating the criticality number is shown in the pipeline process for reader understanding. Immediately after, a process named "Ethical KPIs" takes place. Ethical Criticality Number and the Ethical Relative Criticallity Numbers can be estimated that instead of agglomerating over a component, they agglomerate over specific, trustworthy considerations. Further information on KPIs will be covered in the following sections.

Once KPIs are estimated, the final analysis and hierarchisation processes can occur. Based on the agglomerated critical numbers, it defines the components that have the intrinsic higher level of risk. Prioritize later on the corrective actions on these items. Use tools based on decision making or discretisation approaches (e.g. Pareto 80/20) to define the most relevant ones.

5.1.3.11 Risk Treatment, Transfer, Termination or Tolerate

Figure **37** shows the last pipeline process related to the 4T's of risk management. As observed in the figure, the first component is a question to check if failure modes, treatment, transfer, terminate or tolerate approaches were defined before. Furthermore, this question also checks If a new failure mode is identified. If yes to this question (diamond 1), it is implied that a recursive risk management process is involved. Therefore, first, a process named New Failure Modes/Failing Condition Analysis will separate those new conditions and send them through checking alternatives for the 4T's (i.e. Diamond 2). Diamond 2 is also reached if this is the first time the 4T's process is implemented.

For the 4T's processes designated as "Old Ones", a question (Diamond 3) checks if there have been any modifications to the risk appetite or new KPIs are required to be set for the risk management process for the system sub-system or components. If NO modifications exist, the following questions (diamond 4) ask if there have been any improvements based on previous historical records (i.e. risk register) when a change was expected. A modification was expected

if treatment was performed over the components to improve their performance under the e-risk considerations.

No further processing is required for treatments or system modifications that have shown improvements under the KPIs analyses (i.e. end pipeline). For those that did not, a process named Risk Appetite is run. In this process, it is defined, updated, or check if needed modifications the risk appetite is based on the policies established for the risk management process. These modifications could imply, in the end, further and stringent conditions for the risks under consideration and, thus, higher incentives to define several 4T's or safeguards that could guarantee improvement of the management KPIs. This process is not mandatory since the 4T's process will be independently run in the pipeline process. This process is only connected with diamond 5, which represents the initiation of analyses and definitions of the 4T's.

Before continuing with the 4T's process, the path from diamond 2 has to be explained. Diamond 2 asks if the ALTAI tools have been used to define some general ideas on the risk treatment defined during the FMEA or RCA analyses. This question checks if the user has defined recommendations of treatment for each of the failure modes or root cases established in the risk analysis and, at the same time, if the user has used the ALTAI tool before to establish some considerations to include for risk treatment. The diamond 5 (previously mentioned) is reached if it is answered yes to both questions. In case of any no, first, the ALTAI tool process is run to, later on, connect to the Risk Appetite process defined above. The ALTAI tool process seeks that the user utilizes the ALTAI tool to extract considerations (not foreseen by the framework user) that could be helpful at the moment of treating risk considerations.

From Diamond 5 up to diamond 8, different questions are used to propose some base recommendations that the framework user could consider for implementing on their AI assets in the case that the corresponding e-risk component was identified during the ***e-risk identification and classification***, ***AI scope definition***, and ***Analysis of Values*** processes. The first of these diamonds checks if there are requirements related to Environmental wellbeing. If so, different processes are used to check time frames, meta parameters and re-estimation, dimensionality, and other recommendations concerning AI. These recommendations include: (1) modifying the updating processes timeframes of metaparameters of your AI component to reduce the energy use during configurational processes while securing system robustness, (2) securing the saving of runs results (especially in highly energy-intensive processes such as optimization) and a method to provision reuse of these results, (3) perform dimensionality analyses of your data to secure a repetitive task over variables that are statistically the same and eliminate the recording of unnecessary information, and (4) incorporate novel processes and methods to secure reduction of the computational burden of your system with proper testing that does not impose a robustness burden. The pipeline follows two paths after the last of these processes (i.e. Others process). The ones to the left continue with the recommendations steps (i.e. diamonds 6 to 8), while the ones to the right go to Risk Treatment.

In the Risk Treatment process, for each element with risk conditions of likelihood and impact of treatment level conditions that define to treat of the risk (based on the risk matrix and the risk appetite), consider the implementation of recommended approaches from ALTAI and this framework (i.e. coming from FMEA/FMECA/RCA or this pipeline). Secure the implementation of alternatives that reduce the likelihood of the risk condition or the impact on them (i.e. safeguards)

Dimond 6 performs similarly to Dimond 5, except for focusing on transparency. If Yes to diamond 6, a process of defining technical components to improve transparency (dedicated explicitly to explainability) is used. The explainability recommendations are based on the data type of the information handled by the AI assets. As observed, these can be Image, Text and Tabular data. For the first, it is asked to Define metrics to evaluate the explainability

capabilities of the design and use one of the most suitable generic approaches for the system, e.g.:(1) Saliency Maps (Include consider Digital-Twins-approach). (2) Concept Attribute (3) Counterfactual (4) Prototypes. For the second, it is asked to Define metrics to evaluate the explainability capabilities of the design and use one of the most suitable generic approaches for the system, e.g.: (1) Sentence Highlight (2) Attention-Based Method. Finally, for the third, it is asked to Define metrics to evaluate the explainability capabilities of the design and use one of the most suitable generic approaches for the system, e.g.: (1) Feature importance (Include consider digital-twin-approach) (2) Rule-Based (3) Prototypes (4) Counterfactual.

One recommendation defined for ASSISTANT is to use their digital twin as an approach to produce explainability. This idea es based on the neighbourhood exploratory approach of explainability that in order to produce transparency, optimal solutions (from the AI components) are evaluated on the neighbourhood space (i.e. solutions with input conditions or features close to the optimal, as long as they are feasible) in order to "explain" why the AI produced the before mentioned outcome. Since the digital twin represents real systems, neighbourhoods could also be linked to positionings' physical conditions. Even though this recommendation is defined for ASSISTANT, the approach of generating specific tools for explainability is not part of the WP2 objectives. Thus, it is recommended that the technical WPs (WP3-WP5) use the digital twins constructed to test this approach.

Following the pipeline process, similar to the previous recommendation process (the one started in Diamond 5 or 6), all the recommendations follow the same structure. i.e. once terminated the recommendation analysis, the following recommendation block is followed, and, at the same time, a link to the risk treatment process is defined. This structure allows expansion of the current framework by extending the internal recommendations (i.e. add further process in case an ethical requirement has been set during the scope definition process) or its external component (i.e. add diamonds that would open the possibility of giving recommendations based on the ethical requirements that have NOT been covered yet). As seen in Diamond 8, it asks if the user has specified any other requirement not stated before or explicitly covered by the ALTAI evaluation or scope definition. If not, a process named **Framework Construction** takes place. This process requires a general framework that will define an approach to handle the system, sub-system, or component in a continual mode. This definition cannot be contradictory to trustworthy requirements.

Finally, the last question that has not been implemented so far in the framework (this will be expanded after its implementation stage on the technical component – i.e. starting from M18 of ASSISTANT) corresponds to the accountability and human agency component (diamond 7). Specifically, it is asked if the user has a requirement of Accountability or human agency or any other process that requires HITL, HIC or HOTL that requires establishing responsibilities and obligations. If there is a need to check such requirements, three processes have been defined to give recommendations regarding AI considerations.

The first, named **Responsibilities and Obligations**, Establishes responsibilities for AI failing conditions. It defines that developers should be accountable for faults due to the product's design, while users are accountable for faults resulting from the product's specification, their actions with the AI asset, and the design requirements. In ASSISTANT, this specification translates into technical WPs as developers, while use cases are seen as users. Furthermore, this process mention setting condition in which responsibilities do not lie over users or designers (i.e. system errors) when unforeseen interactions between the system, sub-system or components and the components compromise the product's parts. Independent of the lack of accountability in such cases, it is required to establish clearly, the obligations under such events and corrective actions based on the risk assessment approach (from users and developers).

The second process, named HITL –HIC- HOTL, defines that if decision-making is involved, or any other process of presenting alternatives, the final decision could have a considerable impact on the system, subsystem, components, users, and ambients; it should be secure to:

1. Include a Human in control (HIC), Human on The Loop (HOTL), or human in the loop for accountability considerations
2. Include an explainability process that allows, if possible, to visualize other alternatives and why the system recommendation was chosen
3. Establish users' responsibilities based on previous considerations. Furthermore, as mentioned before, this process would depend on system dynamics and thus include more human intervention only when the dynamics allow it.

The final process is named Interfaces. It helps to specify that if it is unclear to the user that it is interacting with an AI (which supplies/help decision-making processes), there should be enough information to clarify this. This could be more relevant as third-party stakeholders are involved (e.g. AI shopfloor interactions or shop floor managers).

The process named Risk Treatment is only one of the 4T's alternatives to manage risk. Depending on the risk appetite and policies, what approach to managing the AI should be defined is established. This process directly uses the combination of the FMEA, FMECA, CA, and the Risk Matrix (which also uses the risk appetite) tools. The following process, named Risk Terminate, established that for each AI asset or process with risk conditions of likelihood and impact of terminate level conditions, consider a total modification of the system, sub-system, or component to avoid high-risk conditions. If not possible, the AI asset or process should be avoided.

The final pipeline process contains the missing elements of the 4T's (Named Risk Transfer and Tolerate). It specifies that for each element with risk conditions of likelihood and impact of transfer level, if possible, perform a transfer to cover that conditions. For those not handled by treatment, terminate, or transfer, keep them evaluated by the corresponding KPIs to check progress. In ASSISTANT, the Transfer alternative would not be considered, and, thus, these elements would be moved hierarchically as those of higher risk. This implies that it would be considered to be Terminated or Treated.

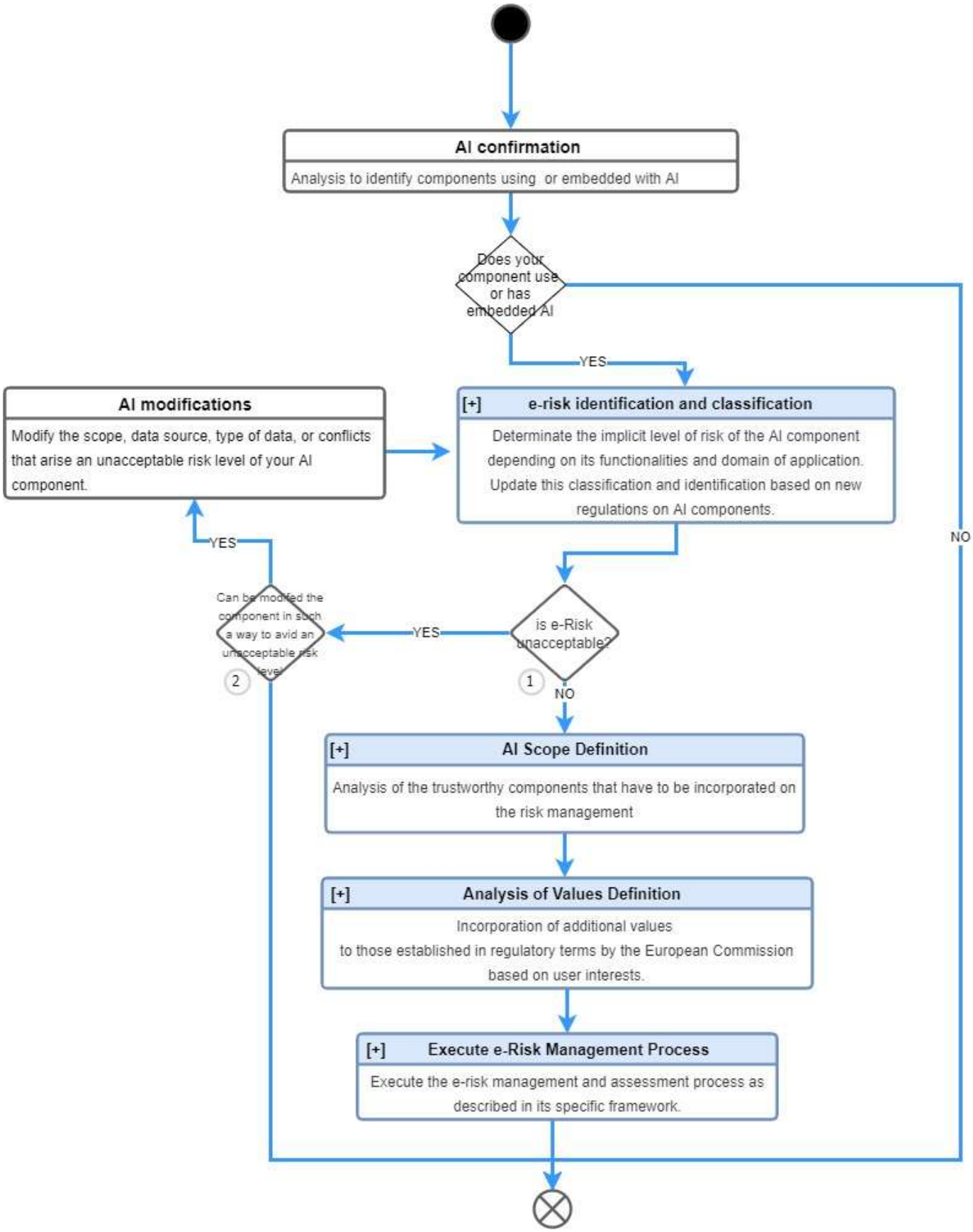


Figure 24 Benchmark e-risk Management Process

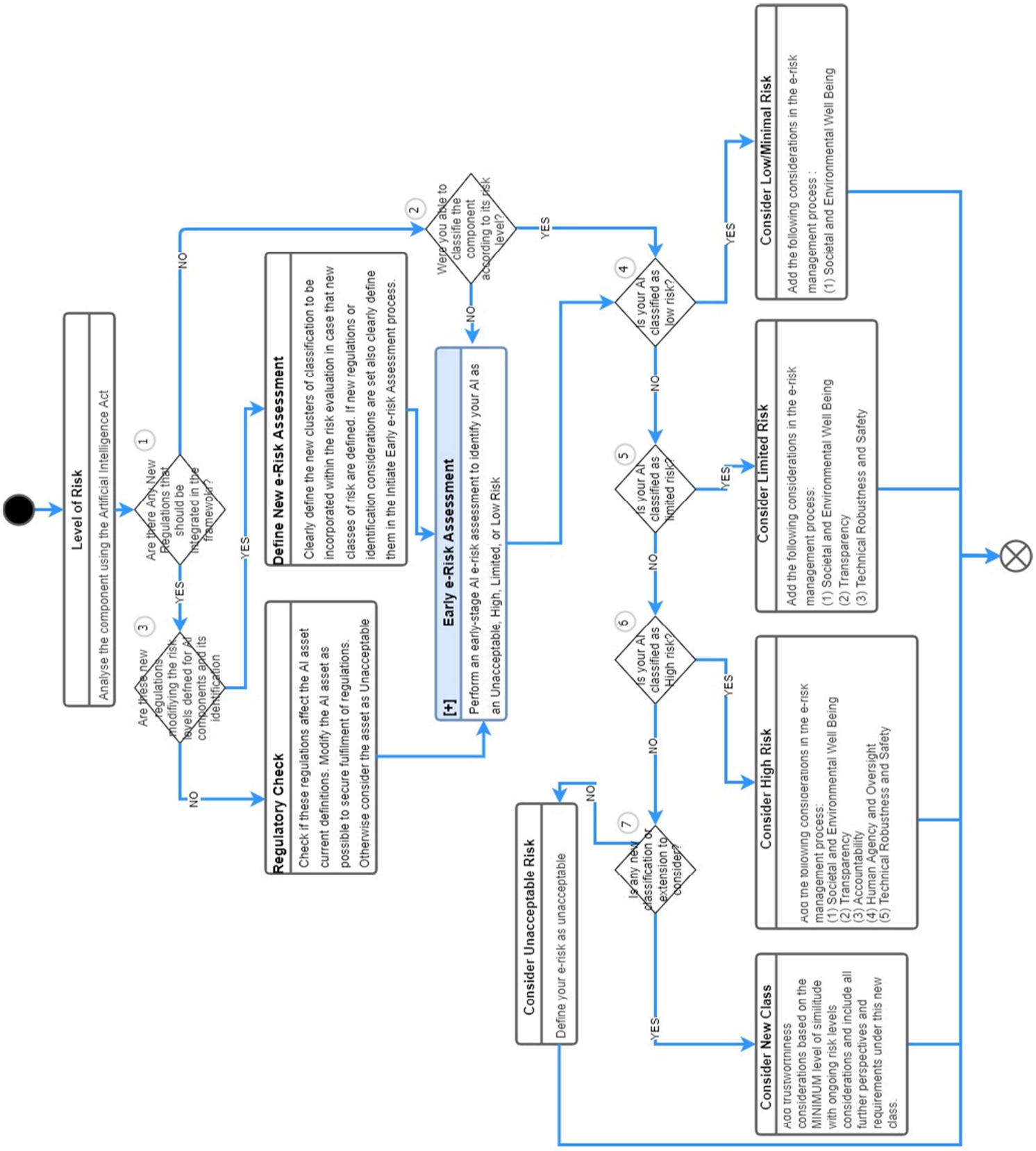


Figure 25 e-Risk Identification and Classification Pipeline

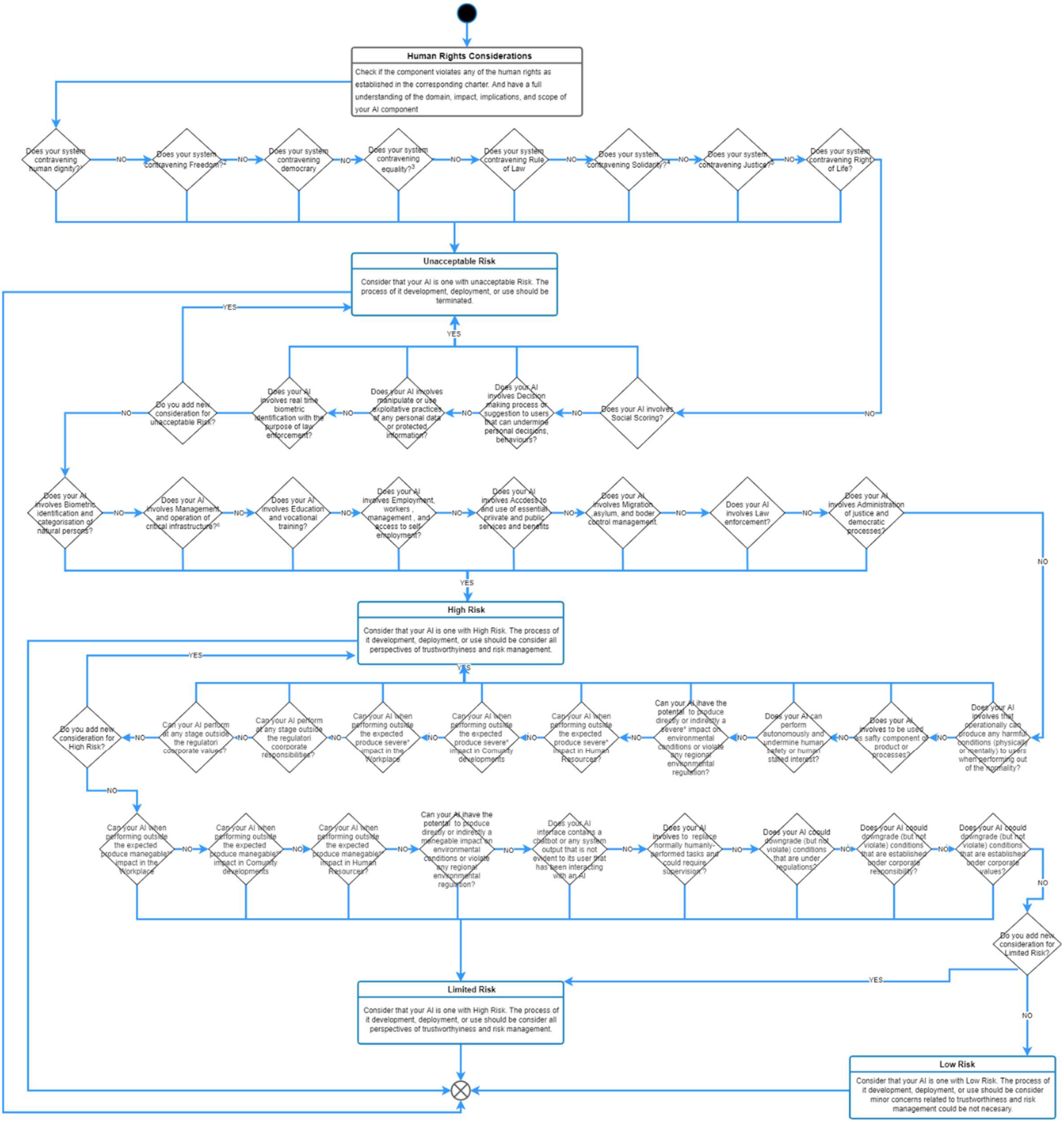


Figure 26 Early e-risk identification

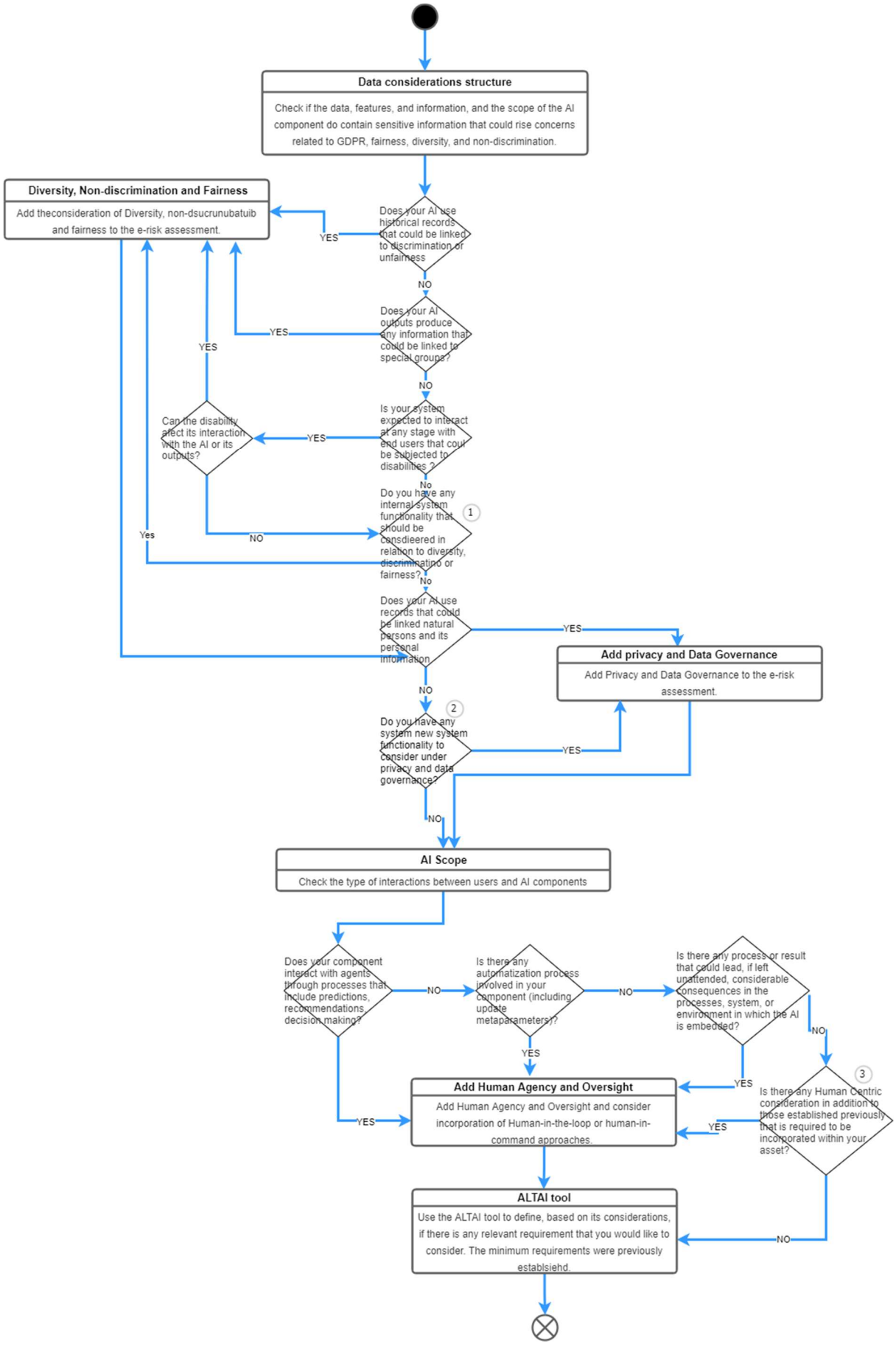


Figure 27 AI Scope Definition

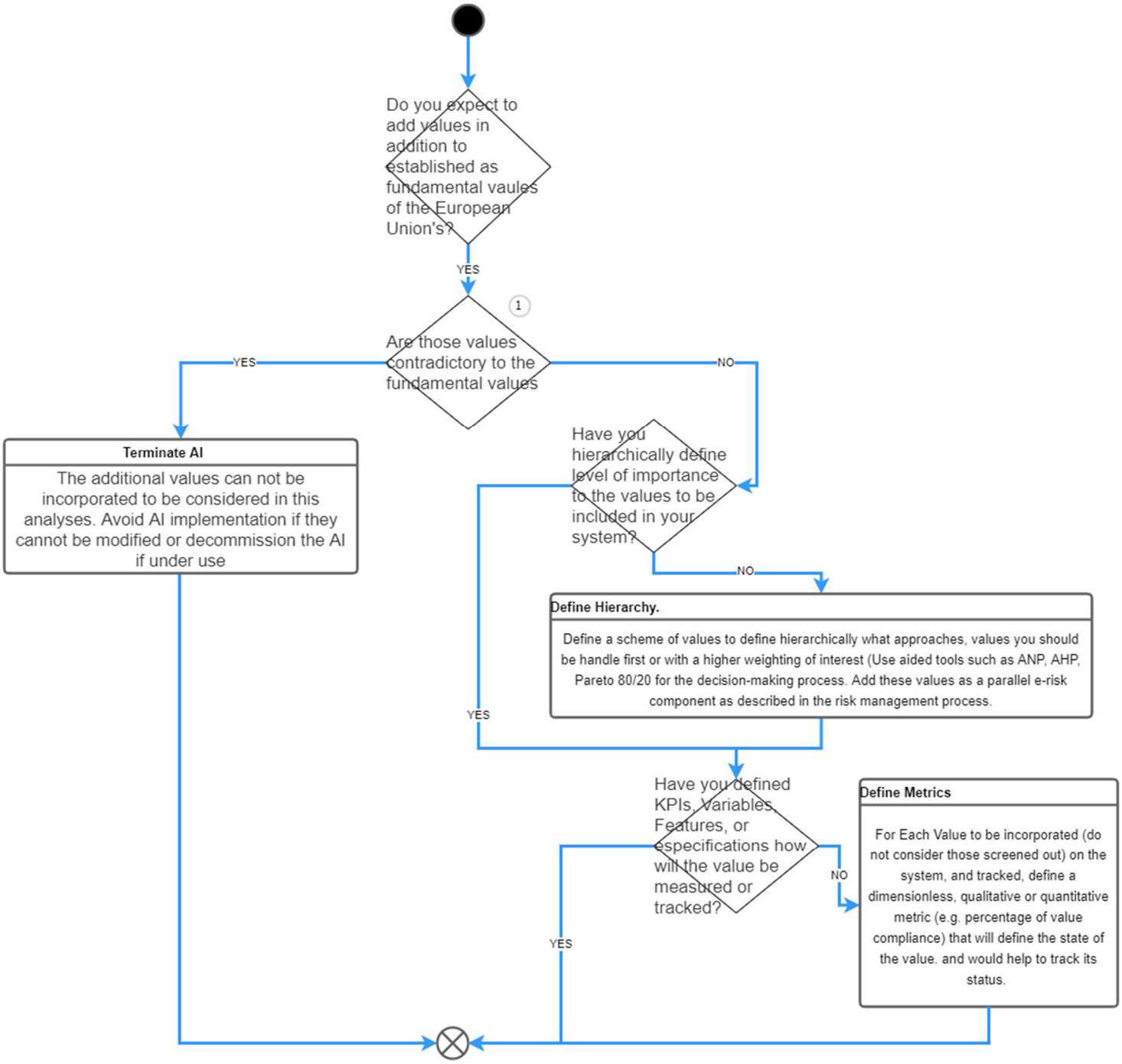


Figure 28 Analysis of values and definitions

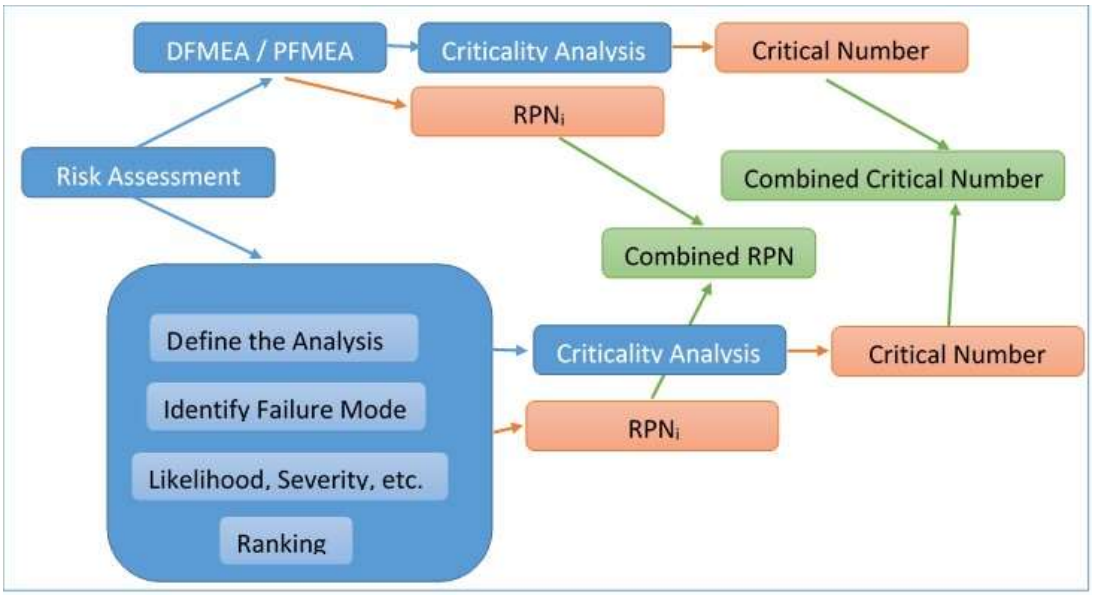


Figure 29 Combination of ethical based FMEA with DFMEA or PFMEA processes.

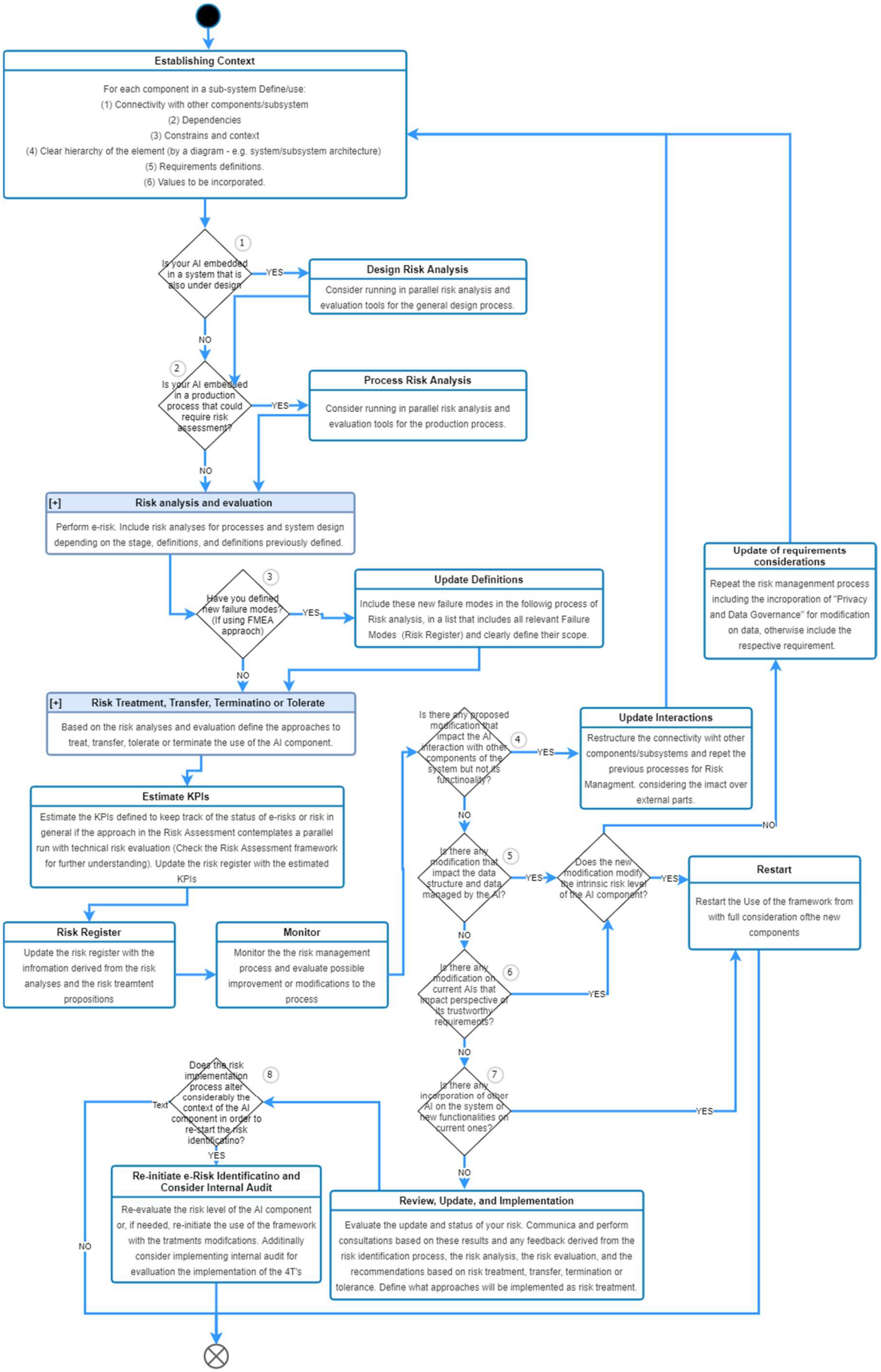


Figure 30 E-risk Management Process

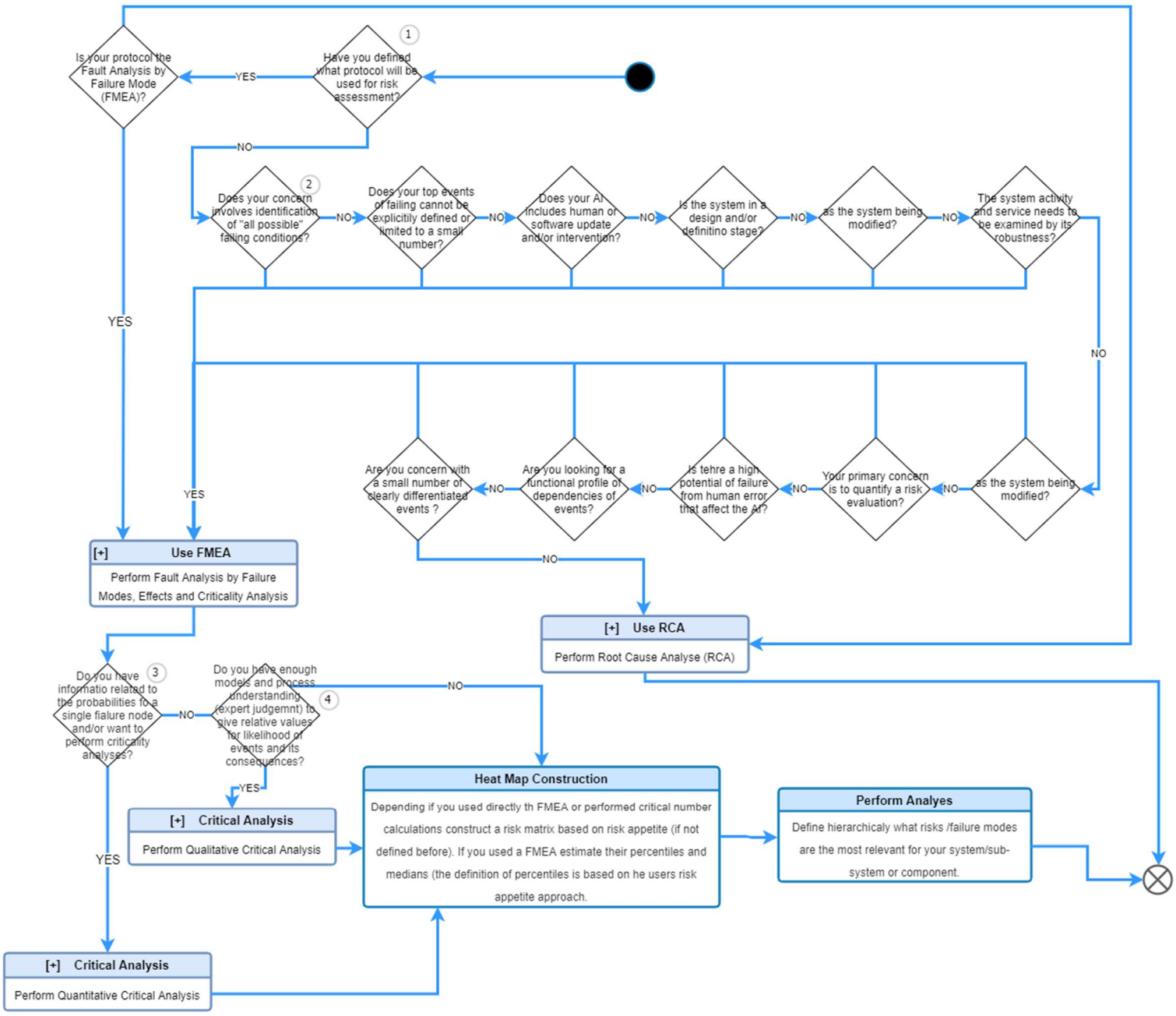


Figure 31 Risk analysis and evaluation

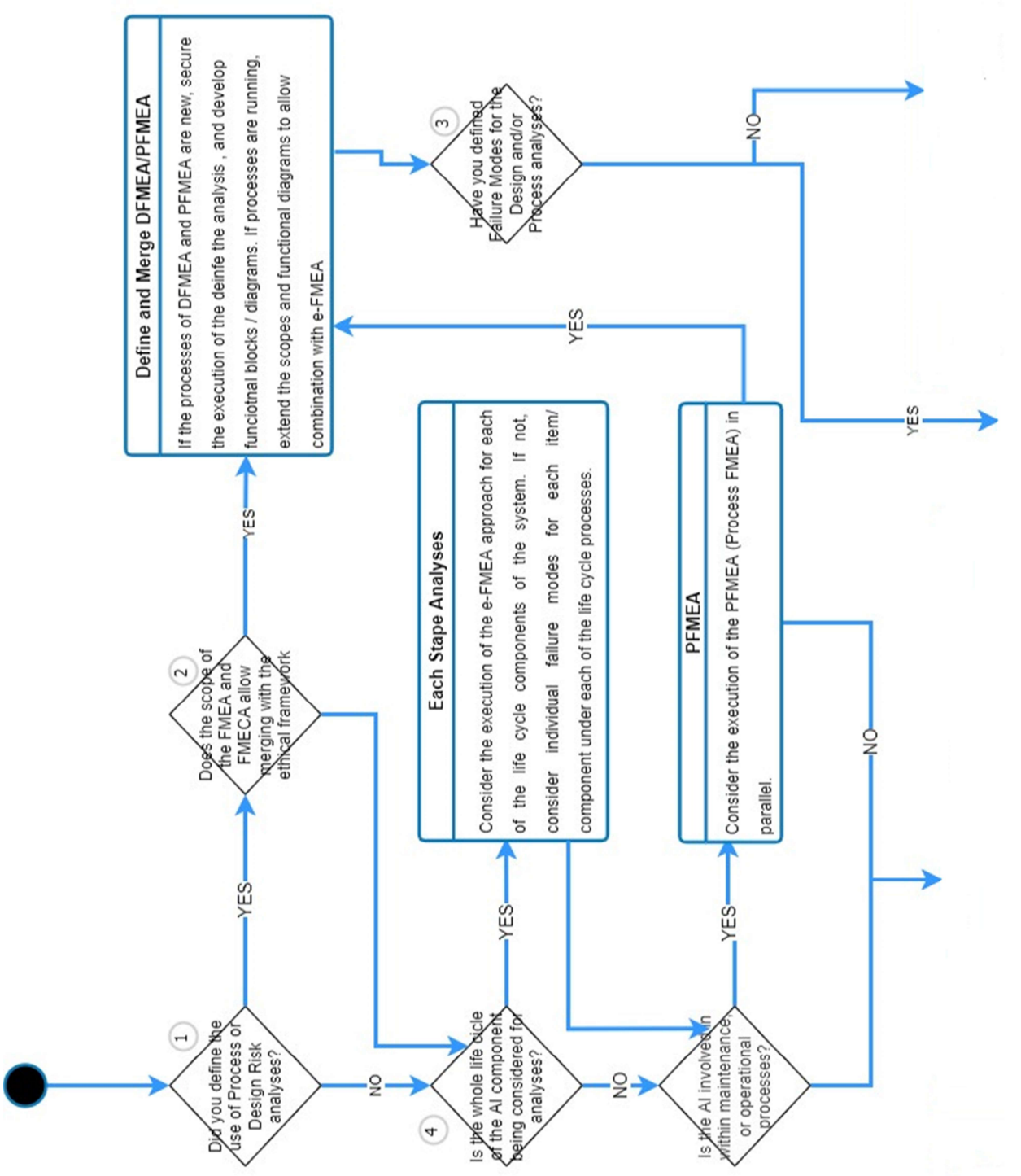


Figure 32 FMEA – Part I – Define if merging with other risk management approaches and execute

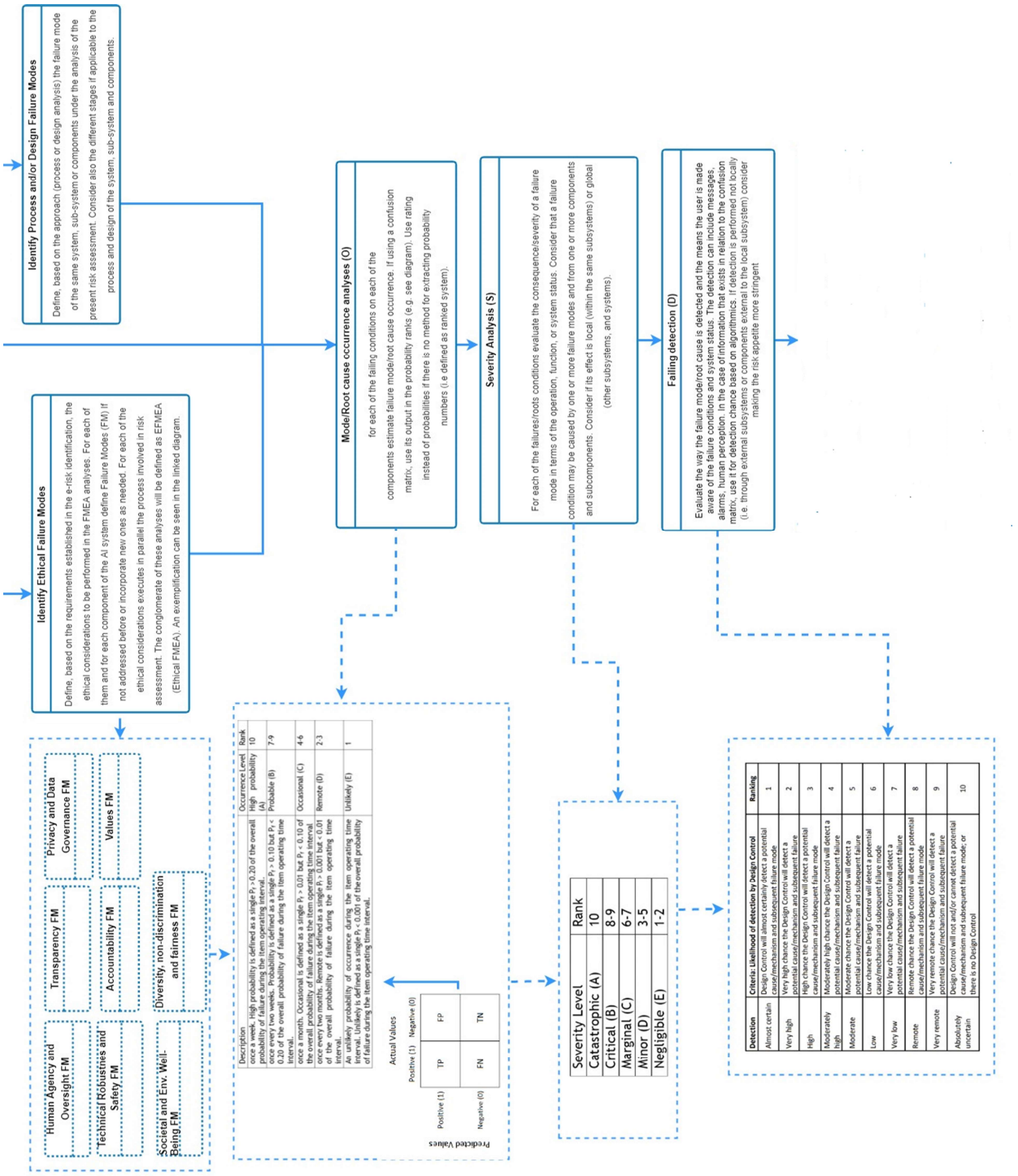


Figure 33 FMEA – Part II – Estimating FMEA indexes

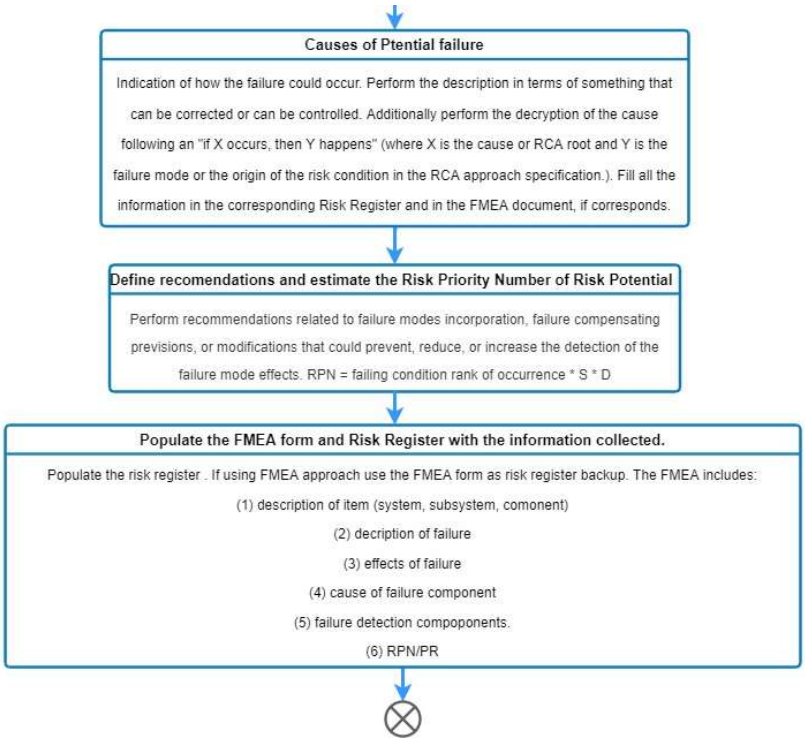


Figure 34 FMEA – Part III – Analysis of the FMEA process

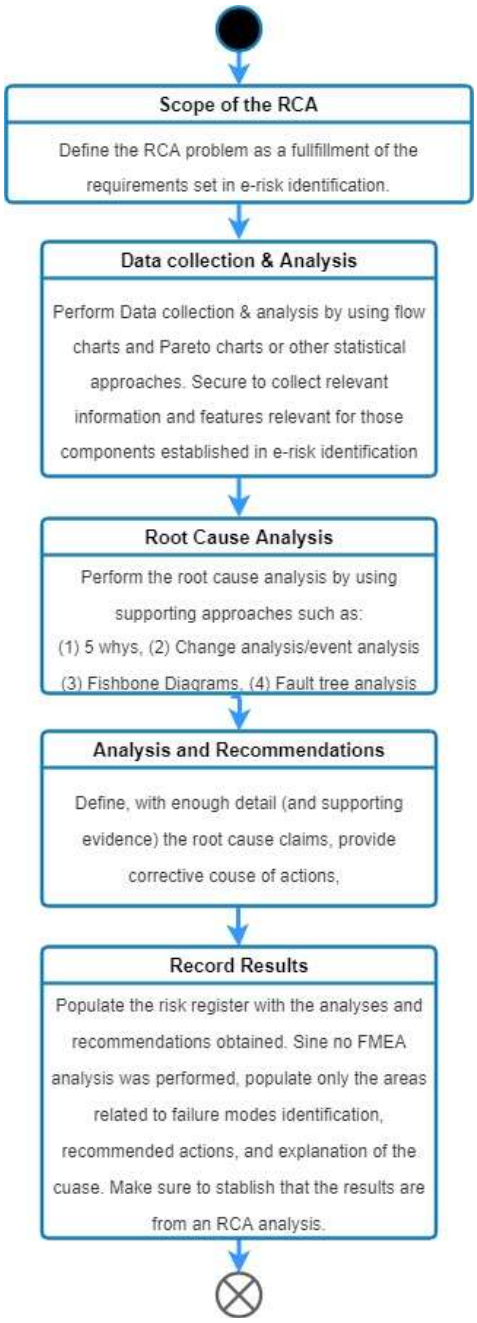


Figure 35 RCA

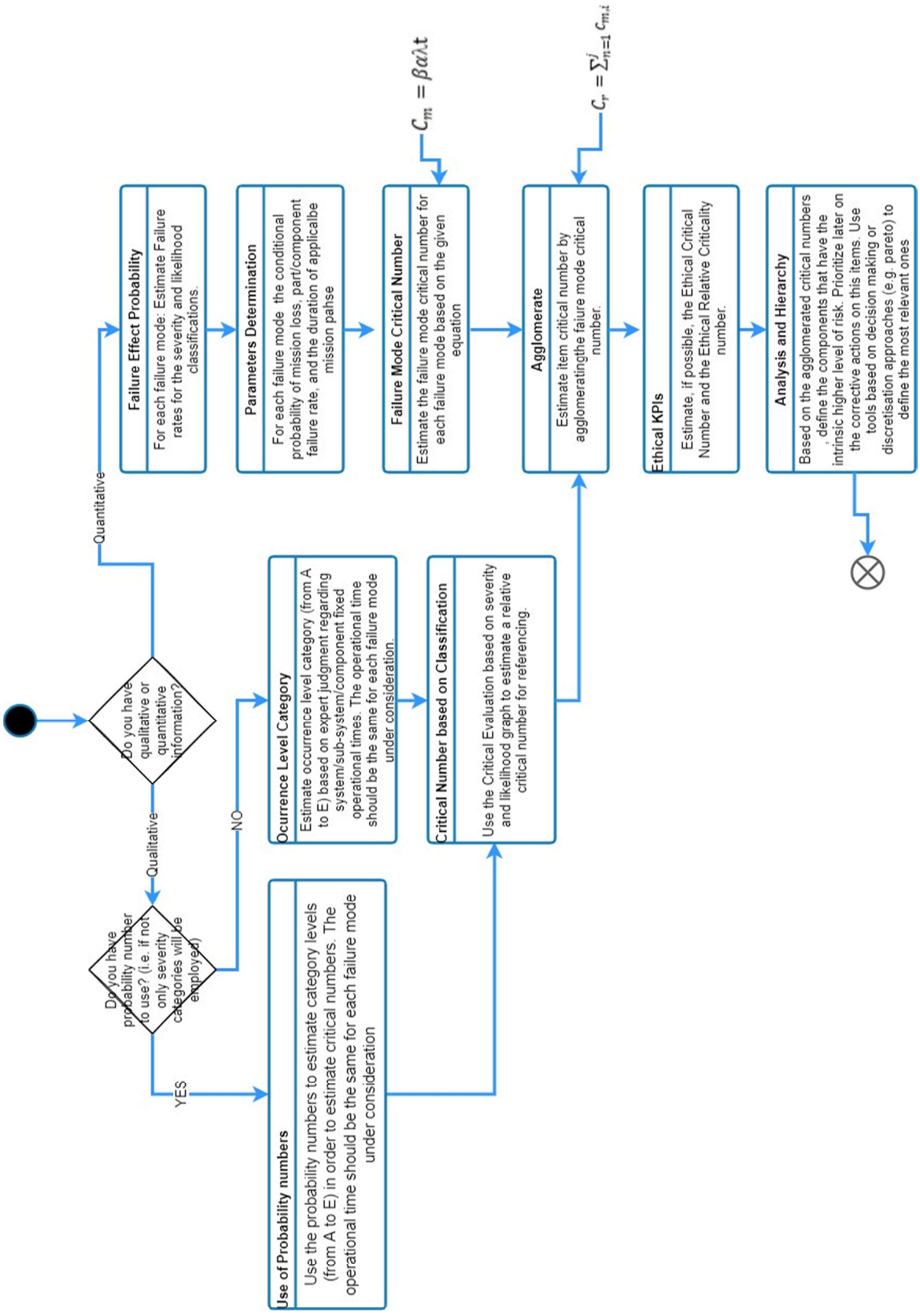


Figure 36 Critical Analysis.

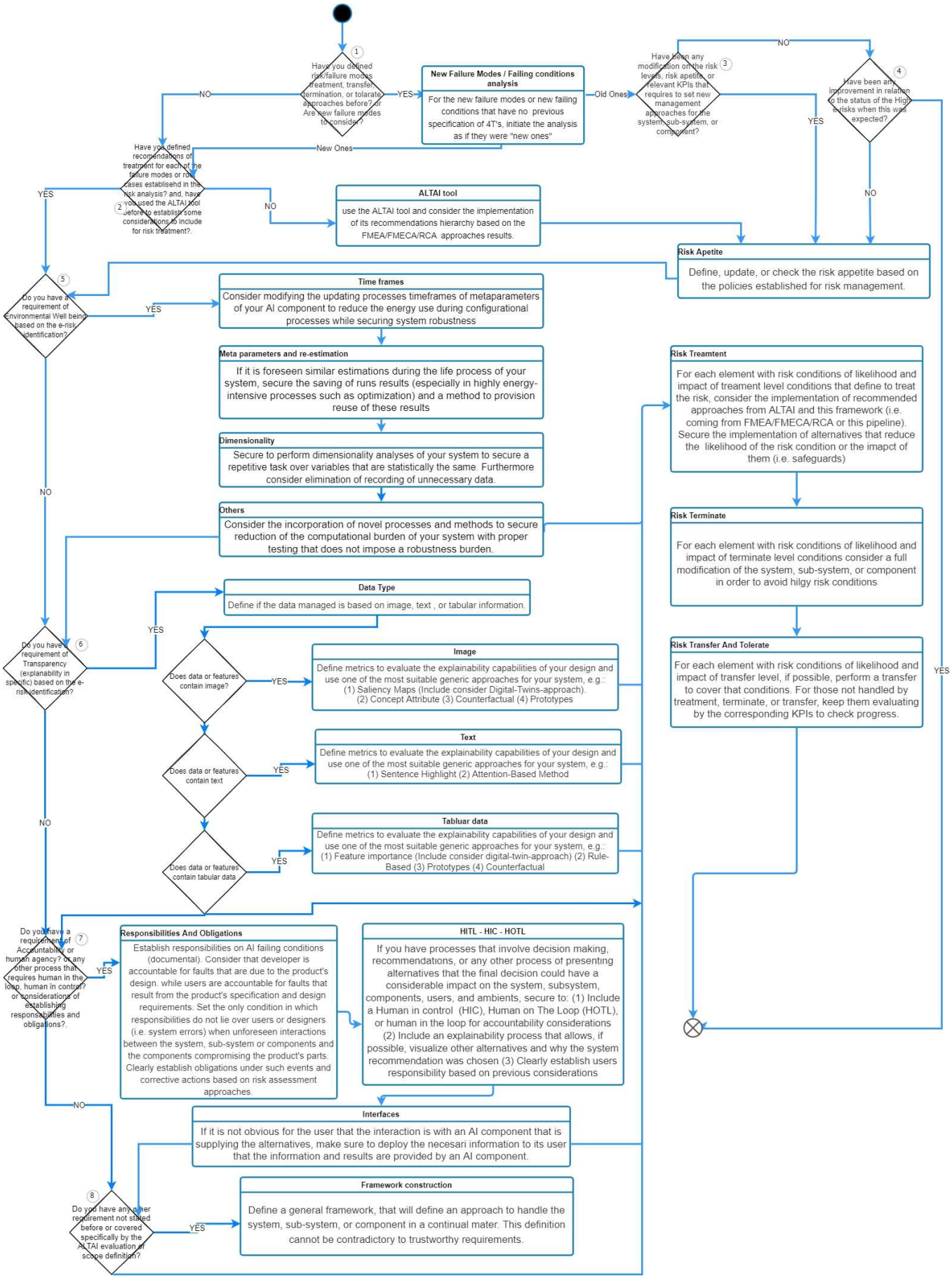


Figure 37 Risk treatment, transfer, terminate or tolerate

5.1.4 **Risk Appetite**

The risk appetite imposes the corrective actions for the risks (i.e. treat, transfer, terminate or tolerate) in the function of the risk scores and the desire of the user to deal with their risk components. Therefore, these specifications should be defined in the risk policies and easily translated to functional risk matrixes.

For ASSISTANT and the framework, the risk appetites should be considered depending on the intrinsic risk considerations derived from the trustworthy requirements and the values incorporated into the system. In other words, if the AI elements correspond to a high-risk element (see Figure 1), the risk appetite should be more stringent in its considerations, while the opposite should be the truth for low-risk AI elements. Based on this approach, we propose testing in ASSISTANT a four-level risk range categorization described in the following tables.

The first of these tables is directly linked to the Criticality Analysis, while the second is linked to the RPN numbers (i.e. purely FMEA based approach). The reflection and analysis of the applicability of these approaches would be evaluated during the validation step of ASSISTANT (i.e. T2.5), thus would initiate in M18.

Table 17 Risk Score Ranges in function of the intrinsic risk level

Risk Level	Tolerate Risk Score Range	Treat Risk Score Range	Terminate Risk Score Range
Unacceptable Risk	-	-	1-25
High Risk	1-5	6-20	21-25
Limited Risk	1-10	11-25	-
Minimal Risk	1-20	21-25	-

Table 18 RPN Ranges in function of the intrinsic risk level

Risk Level	Tolerate RPN Range	Treat RPN Range	Terminate RPN Range
Unacceptable Risk	-	-	1-1000
High Risk	1-200	201-800	801-1000
Limited Risk	1-400	401-1000	-
Minimal Risk	1-800	801-1000	-

5.2 Documentation and Instruments for Risk ASSESSMENT

This section discusses the documents used for the risk management process within the current framework. Different documents could be used to track information, keep the risk assessment status, and update recommendations and modifications over the system and management process, among other tracking documents. We hold the risk register as the primary document to track and keep risk assessment information within the current framework. This register, together with other general management documentation, is described next. The other management documentation instruments include (1) a request and checklist for initiating FMEA processes, (2) a recommendation and modification document for the framework (i.e. updated if needed), and (3) an audit form that specifies what component or sub-system should be evaluated. All these instruments will be used within ASSISTANT. Nevertheless, given the research nature of the current framework, different instruments could be added in the future.

5.2.1.1 Risk Register:

The risk register is an instrument used for recording the risk management process dedicated to identifying risks'. Its purpose is to facilitate ownership and management of each risk. At the same time, it is used to generate a record of the risks that have been identified. Furthermore,

it serves as a record of the control activities that are currently undertaken or will be used to improve the control of the particular risk. The risk register will cover the significant risks facing the organisation/project and record the results of the system risk assessment.

A different form of tracking, such as any information system, could be used to record the information held in the risk register. Therefore is out of the scope of the framework to specify such tools. In terms of ASSISTANT, the information will be tracked within the same repositories used within the project.

A risk register is also a fundamental tool used in internal audit since it allows track modifications from the past, allowing go back on more suitable modifications. Risk registers also help support strategic decisions and relate to the routine operations undertaken. The risk assessment of the strategy (i.e. 4T's of risk management) could include both the risks of undertaking and not undertaking the proposed strategy. Importantly, when considering reputational issues (i.e. social perspectives), the required level of control should be evaluated together with the responsibility for managing the brand. The risk register should be a dynamic document seen frequently and kept up to date.

Table 19, Following Erreur ! Référence non valide pour un signet., the first column is the Failure Modes Name. This section corresponds to the descriptive name of the failure mode detected. The column Description of Failure covers a thorough description of the failure mode that can be used to understand the different stakeholders involved in the risk management process. The previous columns are used for a general description of the failure mode.

The following columns are used to understand the effects of failure. The local, subsystem, system, and global columns are used to understand the effect over each component (local), subsystem, and system linked hierarchically to a failure mode (i.e. bottom-up). The severity is the metric used for the severity measure based on the severity levels described in Table 9. Finally, the last column allows the incorporation of further comments to understand the failure mode's effects better.

Table 20 to Table 22 show the risk register content with an exemplification for a component withheld within the ASSISTANT linked system (WP4). These tables have been separated since the first is linked to a general description of the component to be analysed, the second table is linked to the failure modes description, and the last table focuses on the FMEA/FMECA/RCA analyses (i.e. post-processes after a definition of the failure mode). In each table field, an asterisk (*) can be used to identify those fields that the RCA analyses can populate.

Following Table 19, The first section is the date, which helps keep the information updated. Next, a serial number ID is used for tracking in the Description of Item section. This ID should be linked to the system/subsystem identification to help track information (e.g. 4-1-25-2, system 4, subsystem 1 and ID of it 25 with a version 2; a component can have more than one failure mode).

The Linked ID helps to understand that this is an update or some linked information between IDs. The item is a soft description for understanding the component; the Mode/Phase/Process is a description of the system functionality before the failure mode takes place; the component, subsystem, and system are linked to the description of components previously mentioned in this document (see Figure 22 Arrangements for Incorporating risk management in ASSISTANT.).

Table 19 Risk Register – Part Description

Update	Description of Item

Date*	ID*	Linked ID*	Item*	Mode/Phase/Process*	Component & Ethics*	Subsystem*	System*
2021-22-9	4-1-25-2	4-1-23-1	AI-tolerance limit	Operational	DNN predictor - Robustness	Modeler	WP4

Following **Erreur ! Référence non valide pour un signet.**, the first column is the Failure Modes Name. This section corresponds to the descriptive name of the failure mode detected. The column Description of Failure covers a thorough description of the failure mode that can be used to understand the different stakeholders involved in the risk management process. The previous columns are used for a general description of the failure mode.

The following columns are used to understand the effects of failure. The local, subsystem, system, and global columns are used to understand the effect over each component (local), subsystem, and system linked hierarchically to a failure mode (i.e. bottom-up). The severity is the metric used for the severity measure based on the severity levels described in Table 9. Finally, the last column allows the incorporation of further comments to understand the failure mode's effects better.

Table 20 Risk Register – Failure Mode and Effects Description

Description of Failure		Effects of Failure						
Failure Modes*	Description of Failure*	Local*	Subsystems*	System*	Global*	Severity (S)	System status at the failing condition*	Additional Comments on effects*
Side Effects	Tolerance auto settings do not reach a value within the expected one for safety considerations	X System will not be able to predict a correct model	X Error will be passed to Y components that will produce …	N.A.	N.A.	Moderate (7)	Idle	

Table 21 focuses on a post-analysis of the description of the failure mode. It includes the Causes of Failure that describe, under the analyses performed, a broad understanding of the failure mode and its root causes (e.g. use root cause analysis tools). The item causing is a soft description of the processes or conditions that links to the failure mode. Finally, the occurrence is linked to the table used for specifying the occurrence level based on **Erreur ! Source du renvoi introuvable.**.

Then, a description of the current control for prevention and detection is registered. The first column, detection methods, describes the process or action linked to detecting the failing condition.

Then a thorough description of the detection is made. Finally, a detection probability is used based on the detection ranking (main document section 4.1.7). The risk priority number gives a sensation of the risk level of the failing condition based on the information collected. Nevertheless, this number can be corrected based on the Criticality Analyses correction factor (if needed). This number corresponds to the multiplication of the S (severity), O (Occurrence), and D (Detection) components of the table.

The recommended actions and correction responsibility correspond to the analyses stage that will come with recommendations of actions to perform. Therefore, these recommendations could be linked to the current framework (as seen in the treat, tolerate, terminate, or transfer components pipeline).

There are also columns to specify the corrective action responsible (i.e. division). Furthermore, it is required to specify if the action was implemented. Finally, the last columns refer to the status of KPIs used for AI management purposes.

These KPIs can help to visualize the status of the components based on previous and after corrective actions, if possible to be measured.

Table 21 Risk Register – Failing effects, KPIs and Actions

Cause of Failure			Current controls for prevention / detection			Risk Priority Number	Recommendations and actions			KPIs	
Failure Causes*	Item Causing	Occurrence (O)	Detection Method*	Detection description*	Detection Probability (D)	S*O*D	Recommended Actions*	Correction Responsibility*	Actions made?*	Previous	Actual/during failure mode
Data quality / data curation / incorrect data source	Poor control	Moderate (6)	Model Performance RMSE	The models RMSE over periodic runs	Almost certain (1)	42		WP4	Yes	RRS=0.85	RRS = 0.4

Table 22 Risk Register – Criticality Analysis and Remarks (Optional based on FMEA/FMECA definitions to be used), describe the content of the risk register to be used in case a criticality analysis is desired to be performed. The first component is the severity class that follows the classes defined in Table 9 or Table 10. The failure probability of failure rate data source would correspond to the likelihood of failure if no historical records were used in the analysis (i.e. failure rates per hour basis). Different methods could be used in the analysis (e.g. per hour basis or per million hours basis); nevertheless, once defined for one failure component, it should keep the same basis for each item or component.

The β, α, γ and t columns were previously defined in section 4.5., therefore, they will not be covered here. The Failure mode critical number is estimated using Equation 1, while the Item Critical Number is estimated by summing up all the failure mode Critical Numbers for the same item (i.e. several failure modes will have the same number in the column). The item failure modes, and their Critical Numbers, are linked (and reported) in the column named Linked ID, which could include the ID of the own failure mode under consideration.

The first of the last two components are left for remarks about the Criticality Analysis, If performed. The last one is used for overall remarks for the FMEA or FMECA process performed.

Table 22 Risk Register – Criticality Analysis and Remarks (Optional based on FMEA/FMECA definitions to be used)

Criticality Analysis										Remarks	
Severity Class	Failure Probability or Failure Rate Data Source	Failure Effect Probability (β)	Failure Mode Ratio (α)	Failure Rate γ	Factors Under Considerations for Failure Rate (πi)	Operating Time (T)	Failure Mode Critical Number	Item Critical Number (summation of the item overall critical numbers)	Linked ID	Remarks on Criticality Analysis*	Overal Remarks for FMEA/FMECA

5.2.1.2 FMEA requirement or FMEA Definition

ASSISTANT FMEA/FMECA REQUIREMENT AND DEFINITION	
System/sub-system/Component:	Name of the system for which the FMEA is required
Requirements:	Guide indicating the need for FMEA/FMECA
Purpose of FMEA/FMECA:	The purpose of an FMEA is to demonstrate compliance with the design philosophy for failure situations. The specified undesired event is typically

	one of the three listed in the section below. As part of the FMEA process, corrective action should be proposed.
Undesired Events:	This section specifies the system and global consequences after a failure. These undesired consequences of interest or events fall within 11 broad categories defined for the failure modes.
Interdependencies:	This section lists systems and subsystems whose failures must be addressed in the FMEA to determine their compliance with design philosophy.
Modes of Operation:	A system has typically multiple modes of operation, and each mode can present distinct failure scenarios.
Typical Failures:	This section illustrates the types of failures expected to be analyzed in the FMEA. The list is not comprehensive. All foreseeable failures must be considered in the FMEA, even if not listed in this section.
Timeline:	This section suggests the optimal time in the system life to conduct the FMEA.
Lifecycle Management:	This section indicates how the FMEA will be used and updated during the asset's operational life.
Additional Comments:	comments and notes not fitting in previous categories

5.2.1.3 Recommendations and Modifications for Risk Management Process

ASSISTANT Risk Management Process Recommendations	
Division / WP	This section corresponds to the division and WP that is generating the recommendations to take into account for Risk Management Process Modifications
Date	Date of the recommendation
Reasons for Modification	Describe under what context is necessary to modify the risk management process and provide enough information (diagrams, results, KPIs, or other supporting data) to define the priority of this modification
Recommended Actions	Define what strategies or protocol modifications could be performed to facilitate or improve the use of the risk management process.
Priority	Define the priority and urgency to implement the risk management process recommendations based on the following levels **High** – It is impossible to perform and use the risk management process correctly unless the recommendations are not included. **Medium** – The risk management process is possible to be executed. Nevertheless, the recommended actions would improve metrics and processes, reducing times and costs. **Low** - General recommendations that do not fit in previous levels
Timeline:	This section suggests the optimal time for activating the recommendations.
Additional Comments:	comments and notes not fitting in previous categories
Final Status:	This line defines if the recommendations are accepted or rejected after being analyzed by the E-risk board or Executive Risk Committee. The final status is accepted and implemented, accepted and implemented with modifications, or rejected.
Date of analysis:	Date of analyzing the recommendations
Reasons and Comments:	Comments that describe reasons for acceptance or rejection of the recommendations. Furthermore is described the modifications to be executed if the final status is accepted and implemented with modifications.

5.2.1.4 Audit Initiation Form

AUDIT FORM for	
System/sub-system/Component:	Name of the system for which the FMEA is required

Requirements:	Guide indicating the need for FMEA/FMECA
Purpose of FMEA/FMECA:	The purpose of an FMEA is to demonstrate compliance with the design philosophy for failure situations. The specified undesired event is typically one of the three listed in the section below. As part of the FMEA process, corrective action should be proposed.
Undesired Events:	This section specifies the system and global consequences after a failure. These undesired consequences of interest or events fall within 11 broad categories defined for the failure modes.
Interdependencies:	This section lists systems and subsystems whose failures must be addressed in the FMEA to determine their compliance with design philosophy.
Modes of Operation:	A system has typically multiple modes of operation, and each mode can present distinct failure scenarios.
Typical Failures:	This section illustrates the types of failures expected to be analyzed in the FMEA. The list is not comprehensive. All foreseeable failures must be considered in the FMEA, even if not listed in this section.
Timeline:	This section suggests the optimal time in the system life to conduct the FMEA.
Lifecycle Management:	This section indicates how the FMEA will be used and updated during the asset's operational life.
Additional Comments:	comments and notes not fitting in previous categories

5.3 KPIs

KPIs are measurable values that demonstrate the state of a given system condition. For enterprises, they are generally liked to evaluate the success (or failure) of achieving business objectives. For the present framework, the KPIs should be settled based on the objectives within the Risk Management Protocols objectives (i.e. internal objectives) and those settled by the current interest in Trustworthy AI (i.e. external objectives).

As stated at the begging of the document, the Framework objectives include “Develop a framework that will contain metrics that will allow tracking improvements of ethical-based risks”. Therefore, the KPIs should be intrinsically correlated to evaluate the system state (i.e. condition based on the most relevant intrinsic risk defined for the AI component) on a scale that allows an easy understanding of the level of improvement, risk state, or performance of actions taken (e.g. percentual scale).

Even though the current framework focuses on ethical considerations and has metrics for each of the ten failure mode families (or its specific components), general management KPIs have been included to foster the general system evaluation and its robustness. To be more specific, we recommend using four classes of KPIs. These classes are based on:

1. They depend directly on the current framework approach and the tools proposed here (i.e. Based on the FMEA, FMECA, and CA).
2. They are based on framework KPIs and can be used to link or help to track the state of ethical considerations.
3. They are proposed for AI management at a higher level but are not linked directly to AI ethics.
4. They are linked to specific industrial recognized ratings that could be linked to the current framework (e.g. the MSCI ESG rating).

The following subsections and their tables cover these KPIs in the function of the numeral stated before. For readers, Table 23 Nomenclature for KPIsTable 23 includes the referencing nomenclature used for the equations.

Table 23 Nomenclature for KPIs

Symbol	Name	Units (if any) and Comments
▫	Failure Effect Probability	
▫	Failure Mode Ratio	
▫	Failure Rate	Two failure rates are specified the pondered λ and the base λ_β The last needs multiplication from the different corrective factors
P_i	Corrective Factors for the Failure Rate	
C_m	Failure Mode Criticality Number	
C_r	Item / Component / system criticality number	
t	Time /number of activities	This variable reflects the number of activities that the system/AI has performed in its functionalities; different subindexes are used to define the referencing time (e.g. for maintenance, $t_{maintenance}$)
TP	True Positive	A correct classification (e.g. an apple classified as an apple)
TN	True Negative	A correct negative classification (e.g. a pear is not classified as an apple)
FP	False Positives	An Incorrect classification (e.g. an apple is not classified as an apple)
FN	False Negatives	An Incorrect negative classification (e.g. a pear is classified as an apple when asking for apples)
KPI_i	Domain KPI	KPI used for comparison of AI state or AI with human performance. These KPIs are linked to the domain (e.g. Incidents per month, tickets resolved per month, etc.)

5.3.1 Framework specific KPIs (FMEA / FMECA / CA)

Quantitative Approach: When specific failure rate data is available, direct use of it will be used for the critical analyses. In this case, the natural process is the construction of the criticality matrix. If quantitative information has been provided, the calculation of a criticality number or assignment of a probability of occurrence level and its documentation are performed by providing the following information.

Table 24 Framework dependant KPIs based on Quantitative Information

Name	Definitions	Comments
Failure probability	The probability of the failure mode to take place. The probability does not involve time and thus is not considered a rate.	If available, the failure probability of occurrence shall be listed within the risk register. The change of the Failure Probability could give a metric for reducing the likelihood of events.
Failure effect probability	The Failure Effect Probability is the conditional probability that the failure effect will result in the identified criticality classification, given that the failure mode occurs. The values of it depend on the analyst's judgment and the conditions that correspond to the actual loss (P=1), probable loss (0.1<=P<1), possible loss (0<P<0.1), No effect (P=0).	

Failure mode ratio	This index corresponds to the fraction of the system, sub-system, or component failure rate related to the particular failure mode. The Failure mode ratio shall be expressed as a decimal fraction that the system, sub-system, or component will fail in the identified mode.	
Failure rate	The DART failure rate. DART is an acronym for Days Away, Restricted, or Transferred. Moreover, it is a measure of impact on the workplace. On the other hand, in technical approaches, a base failure rate is weighted by factors that modify a base failure rate. In case those factors are unknown. Consider only an essential part failure rate. The general formulation implies $\lambda = \lambda_B \prod p_i$ λ_B represents the base failure rate, and p_I are the factors (e.g. manufacturing process correction factor, package type correction factor, temperature conditions correction factor, etc.) that weight the base failure rate depending on the system's state.	
Failure mode criticality number	This KPI represents the critical effect of the system, sub-system or component individual failure. They are estimated based on previous numbers and correspond to $C_m = \beta\alpha\lambda t$, where t corresponds to the operating time generally expressed in hours or number of system cycles and therefore depends on the way features are expressed.	
Item criticality number	correspond to the accumulated critical numbers over the same system, sub-system, or component. This accumulation could be driven by failure modes of different nature (e.g. Human Agency and Oversight vs Accountability). Therefore, the same scaling system should be used if accumulated over all possible sources of critical numbers (including system and sub-system failing conditions). (j). $C_r = \sum_{n=1}^{j} c_{m,i}$	
Ethical critical number (ECN - ASSISTANT	This metric corresponds to the total item criticality number given by ethical considerations that would be produced by the system, sub-system, and components. Its estimation corresponds to $ECN = \sum_{n=1}^{k} c_{r,i}$ Where k correspond to all item criticality number related to ethical failure modes.	This number can be used as an index of the current status of ethical concerns.
Ethical relative criticality number (ERCN - ASSISTANT)	This KPI corresponds to the ratio of item criticality number given by ethical considerations to the total critical numbers produced by the system, sub-system, and components. Its estimation corresponds to $ERCN = \frac{\sum_{n=1}^{k} c_{r,i}}{\sum_{n=1}^{j} c_{r,i}}$ Where k correspond to all item criticality number related to ethical failure modes, and j correspond to all criticality numbers. This number can be used as an indication of how technical failure modes.	This number can be used as an index of the current status of ethical concerns

In case information for performing the critical analyses is not available but the FMEA approach was performed, the fed information from FMEA could be used as an independent failure mode evaluation tool as long as it is not merged with criticality analyses. The main difference is based on the idea that the different failing conditions of a component work independently, and therefore, the rates of occurrence should not be combined. In order to perform this approach. Each failing rate should be based on the same time frame, cycles, or produced items consideration:

Table 25 Pure FMEA based approach (proposed for ASSISTANT) KPIs

Name	Definition

Total probability failing rate (Ca - ASSISTANT)	Sum up the failure rates that correspond to the same failure mode ($C_a = \sum_{n=1}^{j} r_i$). This KPI will provide a total probability rate of the failure mode produced by any failure conditions. Secure that each rate (r_i) is over the same time frame or cycle considerations. Otherwise, normalise them to a daily process running.
Ethical critical number (ECN - ASSISTANT)	This KPI corresponds to the total item criticality number given by ethical considerations that would be produced by the system, sub-system, and components. Its estimation corresponds to $ECN = \sum_{n=1}^{k} c_{a,i}$ Where k corresponds to all item criticality numbers related to ethical failure modes, this number can be used to index the current status of ethical concerns.
Ethical relative criticality number (ERCN - ASSISTANT)	This KPI corresponds to the ratio of item criticality numbers given by ethical considerations concerning the total Critical Number produced by the system, sub-system, and components. Its estimation corresponds to $ECN = \frac{\sum_{n=1}^{k} c_{a,i}}{\sum_{n=1}^{j} c_{a,i}}$ Where k corresponds to all item criticality number related to ethical failure modes, and j correspond to all criticality numbers.

5.3.2 Framework general and ethical based KPIs

Table 26 Framework General and Ethical Based KPIs

Name	Definition	Further Information
The likelihood ratio of risks	Number of risks with the likelihood of occurrence over the stated limits / Number of Risks identified	Metric depends on the limit defined to the risk score level (i.e. risk appetite) that is acceptable (Tolerate from the 4T's)
The managed ratio of likely risks	1 – (Number of risks with a likelihood of occurrence over the stated limits - Number of risks managed with a likelihood of occurrence over the stated limits)/ Number of risks identified	Same as before
Severity ratio of risks	Number of risks with severity over the stated limits /Number of risks identified	Same as before
The managed ratio of severe risks	1 – (Number of risks with the severity of occurrence over the stated limits - Number of risk-managed with the severity of occurrence over the stated limits)/ Number of risks identified	Same as before
Item Average Detection Capacity	Average detection capacity of failure modes with risk levels over those in the stated limits.	Same as before

5.3.3 Framework independent and for general AI management

Table 27 Framework Independent and for General AI Management

Name	Definition	Further Information
AI Effective Capacity (AI_{EC})	Measured as the AI usage without modifications over the overall time the AI has been implemented/run. If the AI process is not continual, instead of referring to time, refer to the number of users of the AI. $AI_{EC} = {}^{t_{used}}/_{t_{total}}$	KPI that could be linked to the number of times AI predictions are considered by users (linked to accountability).
AI Downtime Rate (AI_{DR})	The ratio of unscheduled downtime. The schedule includes training/parametrization and AI maintenance. All the other processes or times the AI	KPI that provides insight on AI stability. The time implemented to

	was idle would be considered an unscheduled downtime period. $AI_{DR} = {}^{t_{unscheduled.downtime}}/_{t_{total}}$	correspond to the functional time
AI Percentage Planned Maintenance (AI_{PPM})	The ratio of scheduled downtime. The schedule includes training /parametrization and AI maintenance. $AI_{PPM} = {}^{t_{scheduled.downtime}}/_{t_{total}}$	Can be estimated as 1-AI downtime Rate
AI Capacity Utilization (AI_{CU})	This production KPI measures the amount of capacity utilised as a function of the total capacity available. $AI_{PPM} = {}^{t_{scheduled.downtime}}/_{t_{implemented}}$	KPI is similar to AI_{EC} with the difference that it considers only the functional time (i.e. total time – downtime times)
AI performance indicator (AI_{PI})	A comparison performance metric in which AI is compared to human intervention capacity. $AI_{PI} = {}^{KPI_{AI}}/_{KPI_{human}}$	It can give a metric on system robustness (and comparison). These performance metrics would depend on the implementation domain and could translate into increased trustworthiness of the systems. (e.g. TP_{AI}/TP_{human})
AI correction indicator (AI_{CI})	Percentual reduction of tickets or events about a specific previous condition. $AI_{CI} = {KPI_{initial} - KPI_{AI,new}}/{KPI_{initial} - KPI_{AI,old}}$	It helps to measure how effective the AI makes corrective problems before they occur compared to previous conditions of implementation or AI modifications. For its estimation, use relevant KPIs (e.g. $KPI_{fail\ per\ month}$)
Overall AI Effectiveness (OAE)	This key performance indicator is based on a gold standard for measuring manufacturing productivity: the higher the OAE, the more influential the AI. OAE = Availability * Performance * Quality	
AI Work-in-Process	This KPI metric measures the time requirements for the AI element to produce an outcome in function of the overall production time. It helps manufacturing companies understand how much AI processing information is required for processing information and, at the same time, linked to energy consumption in the case of highly computational required algorithms. AI Work-in-Process = AI requirements for settling a result in hrs (algorithm dependent - e.g. optimisation time, NN training) / 24hrs	
AI accuracy rate (AI_A)	Correct estimation from the AI. $AI_A = {TP + TN}/{TP + TN + FP + FN}$	Only for classification
AI error rate (AI_{ER})	Incorrect estimation from the AI. $AI_M = {FP + FN}/{FP + FN + FP + FN}$	Only for classification
AI precision rate (AI_P)	How often the estimation of positive estimations are correct $AI_P = {}^{TP}/_{TP + FP}$	Only for classification

AI F1 score (AI_{F1})	A balanced metric for predicting AI classification performance $AI_{F1} = {}^{2TP}/_{2TP + FP + FN}$	Only for classification

5.3.4 Based on Environmental, social, and governance (ESG)

As specified in [68], Environmental, Social, and Governance (ESG) is a broad field with many different investment approaches addressing various investment objectives that cover three areas. The first is the ESG integration, which improves the risk-return characteristics of investment, and the second is the values-based investing, in which the investor seeks to align his investment with his norms and beliefs. Finally, impact investing seeks to trigger change for social or environmental purposes.

A (Morgan Stanly Capital International) MSCI ESG Rating is designed to measure a company's resilience to long-term, industry material, environmental, social and governance (ESG) risks. They use a rules-based methodology to identify industry leaders and laggards according to their exposure to ESG risks and how well they manage those risks relative to peers. Our ESG Ratings range from the leader (AAA, AA), average (A, BBB, BB) to laggard (B, CCC). They also rate equity and fixed income securities, loans, mutual funds, ETFs and countries. Even though these risks are not directly linked to AI, they could serve internally to the manufacturing company in the long term to reference their ESG status. Even though this could facilitate tracking some of the general perspectives of the manufacturing sector, they will not foster a direct impact on the internal AI components; therefore, these KPIs could be used generally but not for tracking the internal risk management process. Further information on these KPIs can be found in (https://www.msci.com/our-solutions/esg-investing/esg-ratings - accessed 2022-1-25).

6. Implementation within ASSISTANT

This section covers some specificity regarding implementing the ethical frameworks and tools in ASSISTANT. Furthermore, this section includes a small discussion of the specific considerations that should be taken for ASSISTANT and could be translated to the manufacturing sector.

Figure 38 shows the specific classification of system, subsystem, and components. Given the specific organization of ASSISTANT, the definition of a system is given to each work package that has the participation or development of an AI element (WP3- WP7). A subsystem corresponds to each division or independently functional system within each WP. Based on the D3.1 to D5.1, these subsystems correspond to WP3: The process manager UI, The Process Designer, The process Predictor, and the Process Optimiser components. For WP4: The simulation, The Production Planner, The Model Acquisition for Scheduling, the Scheduler Optimisation, and the Production Manager UI. For WP5: The Streamhandler, The Execution Control and Reconfiguration, The Digital Twin for Execution, The Human Body Detection and Human Task Prediction, and the Human Side Interfaces. For WP6: The Data Fabric in a general sense.

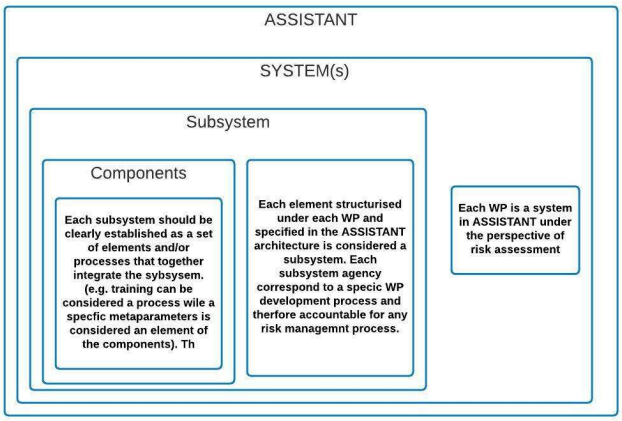


Figure 38 Arrangements for Incorporating risk management in ASSISTANT.

Most of the documentation related to ASSISTANT is presented as annexes. This approach avoids overlapping concepts between the framework and the ASSISTANT use case. Annex A gives a broader description and Considerations for Framework Implementation within ASSISTANT. Annexe B presents a Safeguard proposition based on fuzzy logic. This tool is not considered a product of WP2 but is given as a research proposition for developing and improving AI ethics-by-design considerations. Finally, Annex C presents the ASSISTANT Ethical Risk Management Policy. The policy sets binding rules within ASSISTANT to implement the proposed framework and define approaches to initiate its activities.

7. Conclusions

This document presents a framework for developing and designing AI components within the Manufacturing sector under the Trustworthy AI scrutiny (i.e. framework for developing ethics in/by design). We are proposing a well-structured approach based on risk management that would allow implementing ethical concerns in any life cycle stages of AI components (named development, deployment, use, and decommission). However, there are still considerable areas in which further definitions are required to generate a global approach for AI management under risk assessment. These areas can include technical and non-technical considerations and functionalities given by the algorithms used. Furthermore, depending on their nature, values defined by the users can be challenging to track and measure, which can limit the implementation of the current framework. The framework does not give this limitation; it is given just by a poor or incorrect definition and representation of the system's ethical considerations and values, which are not covered nor intended to be covered in the present work.

Future works will expand missing topics (e.g. different failure modes found during implementation and additional items to observe during risk analyses) that will help settle the approaches for risk management with the final goal of securing the development of AI components under the Trustworthy AI perspective in the manufacturing domain. The goals would be achieved by using ASSISTANT use cases as test scenarios. Furthermore, expert judgment from internal and external stakeholders would be considered for modifications or improvements of the current framework.

8. Bibliography

[1] D. Lauer, "you cannot have AI ethics without ethics," *AI and ethics,* pp. 21-25, 2021.

[2] P. Brosset, S. Patsko, A. Khadikar, A.-L. Thieullent, J. Buvat, Y. Khemka and A. Jain, "Scaling AI in manufacturing Operations: A Practitioners' Perspective," Capgemini Research Institute, 2019.

[3] M. Hengstler, E. Enkel and S. Duelli, "Applied Artificial Intelligence and Trust - the Case of Autonomous Vehicles and Medical Assistance Devices," *Technological Forecasting and Social Change,* vol. 105, pp. 105-120, 2016.

[4] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke and E. Vayena, "AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines,* vol. 28, pp. 689-707, 2018.

[5] E. Commission, "Ethics Guidelines for Trustworthy AI," European Commission, Brussels, 2019.

[6] V. Dignum, Responsible Artificial Intelligence, Springer, 2020.

[7] E. Commission, "White Paper On Artificial Intelligence - A European Approach to Excellence and Trust," European Commission, Brussels, 2020.

[8] E. Commission, "Regulation of the European Parliament and of the Council - Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," European Commission, Brussels, 2021.

[9] R. Corvet, C. Rooney, O. Mulloly, I. Duffy, C. Anderson, C. Stafford, A. Coll, A. Peate, S. O'Shea and R. Benson, "European Union: The EU's New Regulation on Artificial Intelligence," Mondqaq, Dublin, 2021.

[10] D. E. C. o. t. F. Government, "Opinion of the Data Ethics Commission," Data Ethics Commission of the Federal Government, Berlin, 2019.

[11] C. Bartneck, C. Lutge, A. Wagner and S. Welsh, An Introduction to Ethics in Robotics and AI., Springer, 2020.

[12] R. Bengamins, A. Barbado and D. Sierra, "Responsible Ai by Design in Practice," *arXiv,* vol. arXiv:1909.12838, 2019.

[13] C. Ebell, R. Baeza-Yates, R. Bengamins, H. Cai, M. Coeckelbergh, T. Duarte, M. Hickok, A. Jacquet, A. Kim, J. Krijger, J. MacIntyre, P. Madhamshettiwar, L. Maffeo, J. Matthews, L. Medsker, P. Smith and S. Thais, "Towards Intellectual freedom in an AI Ethics Global Community," *AI and Ethics,* vol. 1, no. 1, pp. 131-138, 2021.

[14] B.-h. Li, B.-c. Hou, W.-t. Yu, X.-b. Lu and C.-w. Yang, "Applications of Artificial Intelligence in Intelligent Manufacturing: a Review," *Frontiers of Information Technology & Electronic Engineering,* vol. 18, no. 1, pp. 86-96, 2017.

[15] P. Brey and H. Soraker, "Philosophy of Computing and Information Technology," *Philosophy of Technology and Engineering Sciences: Handbook of the Philosophy of Science,* pp. 1341-1407, 2009.

[16] H. Thilo, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds and Machines,* vol. 30, no. 1, pp. 99-120, 2020.

[17] R. Eitel-Porter, "Beyond the Promise Implementing Ethical AI," *AI and Ethics,* vol. 1, no. 1, pp. 73-80, 2021.

[18] D. Lauer, "You Cannot Have AI Ethics Without Ethics," *AI and Ethics,* vol. 1, no. 1, pp. 21-25, 2021.

[19] V. Dignum, F. Dignum, J. Vazquez-Salceda, A. Clodic, M. Gentile, S. Mascarenhas and A. Augello, "Design for Values for Social Robot Architectures," *Envisioning Robots in Society - Power, Politics, and Public Space,* no. 978-1-61499-931-7, 2018.

[20] J. Knight, Fundamentals of Dependable Computing for Software Engineers, Boca Raton: CRC Press, 2012.

[21] V. Unhelkar, C. Perez-D'Arpino, L. Stirling and J. A. Shah, "Human-robot co-navigation using anticipatory indicators of human walking motion," *2015 IEEE International Conference on Robotics and Automation (ICRA),* 2015.

[22] J. Mainprice, R. Hayne and D. Berenson, "Predicting human reaching motion in collaborative tasks using Inverse Optimal Control and iterative re-planning," *2015 IEEE International Conference on Robotics and Automation (ICRA),* pp. 885-892, 2015.

[23] R. Freedman and S. Zilberstein, "Safety in AI-HRI: Challenges Complementing User Experience Quality," *AAAI Fall Symposia,* 2016.

[24] G. Want, "Humans in the Loop: The Design of Interactive AI Systems," Standford HAI, 2019. [Online]. Available: https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems. [Accessed 17 11 2021].

[25] M. Vierhauser, M. A. Islam, A. Agrawal, J. Cleland-Huang and J. Mason, "Hazard analysis for human-on-the-loop interactions in sUAS systems," in *ESEC/FSE '21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Athens Greece, 2021.

[26] R. Koulu, "Human control over automation: EU policy and AI ethics," *European Journal of Legal Studies,* no. 1, pp. 9-46, 2020.

[27] A. Tubella, A. Theodorou, F. Dignum and V. Dignum, "Governance by Glass-Box; Implementing Transparent Moral Bounds for AI Behaviour," *arXiv:1905.04994,* 2019.

[28] V. Dignum, "Ethics in Artificial Intelligence: Introduction to the special issue," *Ethics and Information Technology,* vol. 1, no. 1, pp. 1-3, 2018.

[29] C. Closse, "The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism," *AAAI Fall Symposium,* p. 38, 2005.

[30] M. Saidani, B. Yanno, Y. Leroy and F. Cluzel, "Hybrid top-down and bottom-up framework to measure products circularity performance.," *International Conference on Engineering Design, ICED 17,* 2017.

[31] O. Evans, A. Stuhlmuller and N. Goodman, "Learning the Preferences of Ignorant, Inconsistent Agents," *arXiv:1512.05832v1,* 2015.

[32] G. Lason, "Artificial Intelligence, Values, and Alignment," *Minds and Machines,* vol. 30, pp. 411-437, 2020.

[33] O. Evans, A. Stuhlmueller, J. Salvatier and D. Filan. [Online]. Available: https://agentmodels.org/. [Accessed 2021 4 2021].

[34] M. a. Markets, "Artificial Intelligence in Manufacturing Market by Offering (Hardware, Software, and Services), Technology (Machine Learning, Computer Vision, Context-Aware Computing, and NLP), Application, End-user Industry and Region - Global Forecast to 2026," Markets and Markets, [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-manufacturing-market-72679105.html. [Accessed 7 4 2021].

[35] Synced, "AI-Powered 'Genderify' Platform Shut Down After Bias-Based Backlash | Synced," [Online]. Available: https://syncedreview.com/2020/07/30/ai-powered-genderify-platform-shut-down-after-bias-based-backlash/. [Accessed 7 4 2021].

[36] K. Lyons, "Clearview's facial recognition tech is illegal mass surveillance, Canada privacy commissioners say - The Verge," [Online]. Available: https://www.theverge.com/2021/2/4/22266055/clearview-facial-recognition-illegal-mass-surveillance-canada-privacy. [Accessed 7 4 2021].

[37] C. Stupp, “Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case - WSJ,” [Online]. Available: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402. [Accessed 7 4 2021].

[38] D. Wakabayashi, “Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam - The New York Times,” [Online]. Available: https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html. [Accessed 7 4 2021].

[39] J. Vincent, “Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day - The Verge,” [Online]. Available: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. [Accessed 7 4 2021].

[40] A. Jobin, M. Lenca and E. Vayena, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence,* vol. 1, no. 9, pp. 389-399, 2019.

[41] C. Beumer, L. Figge and J. Elliott, “The sustainability of globalisation: including the 'social robustness criterion',” *Journal of Cleaner Production,* vol. 179, pp. 704-715, 2018.

[42] S. Vallance, H. Perkins and J. Dixon, “What is social sustainability? A clarification of concepts,” *Geoforum,* vol. 42, pp. 342-348, 2011.

[43] D. Leprince-Ringuet, “AI{s big problemÑ Laxy humans just trust the algorithms too much.,” ZDnet, 2020. [Online]. Available: https://www.zdnet.com/article/ai-needs-to-be-controlled-but-lazy-humans-may-not-be-up-to-the-job/. [Accessed 13 08 2021].

[44] J. Higgins and S. Green, Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0, The Cochrane Collaboration, 2011.

[45] H. L. E. Group, “Ethics Guidelines for Trustworthy AI,” European Commission, 2019.

[46] M. Chui, M. Harrysson, J. Manyika, R. Roberts, R. Chung, P. Nel and A. van Heteren, “Applying artificial intelligence for social good,” McKinsey & Company, [Online]. Available: https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good. [Accessed 16 08 2021].

[47] I. 2. R. Management, “ISO31000:2018,” 2018.

[48] J. T. C. OB/7, “AS/NZS 4360:1995 Australian/New Zealand Standard; Risk Management,” Melbourne, 1995.

[49] P. Hopkin, Fundamentals of Risk Management, London: Kogan Page, 2010.

[50] J. SPeer, “Why FMEA is Not ISO 14971 Risk Management,” Greenlight Guru, 2016. [Online]. Available: https://www.greenlight.guru/blog/fmea-is-not-iso-14971-risk-management. [Accessed 8 17 2021].

[51] S. Sholla, R. Naaz Mir and M. Ahsnat Chisti, “A Neuro SYstem for Incorporating Ethics in the Internet of Things,” *Journal of Ambient Intelligence and Humanized Computing,* 2020.

[52] T. Saaty, Decision Making with Dependence and Feedback: The Analytic Network Process, Pittsburgh: RWS Publications, 1996.

[53] T. Saaty, “Decision Making - The Analytic Hierarchy and Network Processes (AHP/ANP),” *Journal of Systems Science and SYstems Engineering,* vol. 13, no. 1, pp. 1-35, 2004.

[54] D. o. D. o. t. U. S. o. America, “Procedures for Performing a Failure Mode, Effects and Criticality Analysis,” Department of Defense, Washington, DC, 1977.

[55] I. E. C. (IEC), “IEC 60812:2018 - Failure Modes and Effect Analysis (FMEA and FMECA),” IEC, 2018.

[56] H. Pentti and H. Atte, “STUK-YTO-TR 190 - Failure Mode and Effects Analysis of Software-BasedAutomation Systems,” STUK, Helsinki, 2002.

[57] A. B. o. Shipping, “Failure Mode and Effects Analysis (FMEA) for Classification,” American Bureau of Shipping, Houston, 2015.

[58] D. o. Defence, “Military Standard: Procedure for performing A Failure Mode, Effects and Criticality Analysis,” Department of Defense, Washington, DC, 1980.

[59] R. S. K. Shankar, D. O'Brien , K. Albert, S. Viljoen and J. Snover, "Failure Modes in Machine Learning," *arXiv:1911.11034,* p. 12, 2019.

[60] J. Millar, "Social Failure Modes in Technology and THe Ethics of AI," in *Oxford Handbook of Ethics of AI*, OUP USA, 2021, p. 896.

[61] P. Habek and M. Molenda, "Using the FMEA Method as a Support for IMproving the Social Responsibility of a Company," *Proceedings of the 6th International Conference on Operations Research and Enterprise Systems (ICORES),* pp. 57-65, 2017.

[62] A. B. o. Shipping, "Failure Mode and Effects Analysis (FMEA) for Classification," American Bureau of Shipping, Houston, Texas, 2015.

[63] A. B. o. Shipping, "Failure Mode and Effects Analysis (FMEA) for Classification," American Bureau of Shipping, Houston, Texas, 2015.

[64] M. Villacourt, "Failure Mode and Effects Analysis (FMEA): A Guide for Continuous Improvement for the Semiconductor Equipment Industry," SEMATECH, 1992.

[65] B. DHillon, "Maintainability Tools," in *Engineering Maintainability*, Elsevier, 1999, pp. 50-81.

[66] E. Commission, "Charter of Fundamental Rights of the European Union," Official journal of the European Union, Brussels, 2012.

[67] E. Commission, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment," [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment. [Accessed 21 9 2021].

[68] G. Giese, L.-E. Lee, D. Melas, Z. Nagy and L. Nishikawa, "Fundations of ESG Investing: How ESG affects Equity Valuation, Risk, and Performance," *The Journal of Portfolio Management,* vol. 45, no. 5, pp. 1-15, 2019.

[69] H.-L. E. G. o. A. Intelligence, "Ethics Guidelines for Trustworthy AI," European Commission, Brussels, 2018.

[70] A. Schuhly, F. Becker and F. Klein, Real Time Strategy, Bingley: Emerald Publishing Limited, 2020.

Annex A. ASSISTANT description and Considerations for Framework Implementation

ASSISTANT is a funded European Project ASSISTANT 2021) that aims to create intelligent DTs through the joint use of machine learning, optimisation, simulation, and domain models. The results (and advanced tools)would help design and operate complex collaborative and reconfigurable production systems. This project target a significant increase in manufacturing sector flexibility by incorporating AI components within the generative design, production planning, and control approaches.

Trustworthy AI is a fundamental pillar on which ASSISTANT focuses. It is considered to be applied vertically to each of the approaches (Work packages), tools, architectures, and frameworks that would have embedded AI components. The embedded AI components would participate in several tasks that include, among others: decision-making, control (including Process and robotic control), data cleaning, and modelling tasks.

ASSISTANT would allow a better understanding of the limitations on strengths of different approaches that implement ethical considerations within AI development and

deployment, but with a focus on the manufacturing domain. The following bullets give insight into ASSISTANT essential factors to consider when generating and implementing a framework.

- **Domain**: AI components are already in various manufacturing processes that, in general, include: (1) modelling and simulation, (2) predictions and estimations, (3) decision-making and optimisation, and (4) robotics. Contrarily, even though AI is already used to make low-level decisions in the manufacturing sector, such as automated machine tuning or predictive quality, there is still significant room for improvement and extension into higher-level manufacturing decisions. Incorporating data-driven and AI approaches would facilitate the design, planning, control, testing, management, and integration of the product and processes. A significant application of AI elements involves computational design. Specific elements are optimised for some criteria or technical specifications from a design perspective. Depending on the nature of the variables selected and the space of solution, the designing process can be classified as parametric (use of physical meaning parameters that are optimized during the design specification process) and generative design (space of solution is less institutive than parametric, given the problem to be optimized and the specification of the variables to be searched as an objective function). The generative design approach will be used in ASSISTANT for both process and production planning. Therefore, its implications in data managing and processing should be included in the different analyses performed under Trustworthy AI considerations.
- **Stakeholders**: ASSISTANT will be developed and implemented with a perspective in the manufacturing sector. The developed components could be further validated under three different industrial partners' requirements. Their interest in ASSISTANT includes reducing costs, improving quality, and gaining insight into specific operational units to secure processing conditions, using generative design approaches as well as human-robot cooperation components inside different operations (that could be dynamically modified/reconfigurable, and therefore controlled, depending on actual production requirements), data orchestration by using data fabric as a backbone module for connectivity between legacy, IoT, Industry 4.0, and modules, and using generative design approaches for manufacturing schedules and using data fabric for managing heterogeneous information. Additionally, academic and other industrial partners would develop the ASSISTANT approach, allowing different perspectives to be incorporated from different domains. Primary industrial stakeholders that will benefit by using the ASSISTANT approach include, but are not limited to: process planning engineers, employees responsible for designing and deploying a production line, production line operators, budget planners, managers (of different areas), shop floor planners, control and automation engineers, and other experts from the manufacturing domain that are a participant on production, scheduling, logistics, process control, and maintenance.
  Secondary stakeholders can be classified into two types. Those indirectly benefited by the improvement over the goods produced and those that benefit from improving the goods built-in with AI elements. This difference is essential since there is a broad difference in the responsibility involved concerning Trustworthy AI. ASSISTANT focuses primarily on developing AI elements that will be used within the manufacturing sector and, therefore, secondary and excluded stakeholders are not considered in the development and deployment of AI elements (i.e., under the umbrella of Trustworthy AI).
- **Physical environments**. Combining the different validation scenarios allows testing ASSISTANT approaches in a broad gamut of conditions and types of production. The productions cover low and high ranges of mixing components produced simultaneously at different production volumes. Independent of the specificity of a given scenario, since ASSISTANT focuses on implementing generative design approaches for the process, production, and scheduling steps, through the use of digital twins for systems representability, a commonality in the physical environments could drown:

- Shop floor: Since ASSISTANT output involves process design, It can directly affect workshop layout and the full considerations (including materials and products transportation, ergonimics, safety, and risk management (e.g., hazard, hazan), among others).
- Operational units: The critical components of transforming raw materials or parts by physical, chemical, thermal, or assembling processes. These components are connected to generate a whole operational process. The operational units correspond to a specific component within the shop floor.
- Robots: Since robots constitute a part of processing lines in several technified manufacturing industries, their considerations and interactions with humans and their environment should be considered in ASSISTANT. Robots are heavily being intensified in the manufacturing sector to work collaboratively. This consideration is based on the consideration that robots and humans perform better in different tasks (e.g., robots perform repetitive tasks that include physical actions).
- Wearables, interactive gadgets, and other communication components: Interaction with the different stakeholders involved in the manufacturing sector can be made through different channels. These could include wearables, computers, and gadgets that allow communication through natural language, programmatic languages, and interactive visualization channels. The physical components would imply interaction by input and output channels to allow the systems to feed and retrieve information. In ASSISTANT, the communication intends to be at a level where users would not need to understand the components behind the estimations deeply but provide enough information to be self-explanatory.
- Products and parts. Even though products and inputs materials could be foreseen as external components of manufacturing processes since they contain the capabilities to be fed with AI elements, considering them as an environment relative to an AI element should be considered.

- **Digital environment**. In ASSISTANT, different components will be used to process and handle information that could have embedded AI elements. Hierarchically, they cannot be organized since they could perform separately or under other components' specifications or functionalities. The following list describes them from a general perspective.
  - *Data fabric*: A data fabric is a system that provides a unified architecture for managing and providing data. Data fabric can provide flexibility and scalability to the system by providing service-oriented distributed systems for accessing and storing data. While data fabrics can be seen as a data managing and communication system, they can also be integrated with analytical processes (including AI elements). The ASSISTANT data fabric will be used to coordinate and provide the system with the data requirements.
  - *Digital twins*: A digital twin is a higher-level model corresponding to a digital representation of physical objects or processes. A digital twin is encompassed by three main components that include the physical component (e.g., shop floor, robots, and operational units), the digital representation (encompassed by domain models, metamodels, and constraints), and its communication (e.g., the data fabric). In ASSISTANT, there will be independent interactive digital twins: a process planning digital twin, a production planning and scheduling digital twin, and a reconfigurable manufacturing execution digital twin
  - *Models*: Models are an abstract representation of system conditions used for higher-level decisions (e.g., optimization, decision making, simulation, forecast, control). In ASSISTANT domain models (representations specific to a domain), metamodels (simplification models from other models), state models (models representing the dynamic variation of state variables), and different data-driven models (models derived from data, including AI). Models can be obtained by AI using different approaches (e.g., classification, regression, time series)

- *Algorithmics*: Different processes, including optimization, data curation, data elimination, data managing, and algorithmic processes, would be implemented in ASSISTANT. These can be embedded within AI processes to learn from modified environments and conditions with different autonomy levels.

In assistant AI, management would be done by considering a suitable management structure that defines policies, architectures, and strategies, among others and, at the same time, considers the Trustworthy AI concepts. Therefore, the settlement of risk management processes and the ART principles are the most suitable approaches to handling AI components. Based on these considerations, the following sections do comment on these perspectives.

Since ASSISTANT is based on the Ethics Guidelines for Trustworthy AI, a description of considerations based on the framework requirements is given next. In an industrial case, an AI element could supply a production plan based on optimization components that could dynamically check the state of materials, labour, workforce, and machinery availability. This approach implies two significant concerns about human agency.

First, until what point autonomy should be given to the AI elements. This question implies that general frameworks should secure the incorporation of human agency in the manufacturing sector by considering the use of human-in-the-loop (HITL) human-on-the-loop (HOTL) or human-in-command (HIC). The choice of which approach would be more suitable depends on the complexity of the decision to be made and the user's level of expertise involved in the decision-making process. In producing goods with embedded AI elements, a continual agreement with the end-user should be embedded to supply enough information to make its own decision. Pre-stated options should be avoided, given human bias tendencies to consider those machine-based decisions are more suitable than pure human-based approaches.

Second, until what point should workforce information be provided to AI elements. This implies that tracking behaviours (through gadgets and wearables) should be specified to users. Their use should be deployed with different constraining conditions that could keep personal information with limited access. For example, anonymous data should be analysed with a specific area of applicability within the environments (e.g., limit tracking to specific shop floor areas). Similar conditions should be made about human augmentation devices in which HITL, HOTL, or HIC approaches should be implemented.

In terms of safety, the focus should be made on the principle of prevention of harm. One considerable difference between both cases is that manufacturing safety procedures can be easily implemented in manufacturing sectors to facilitate fallback plans and general safety procedures. Given the p probability of interaction with robots on the shop floor, several technical approaches linked to human-centric verifiable, integrative, and physical approaches should be implemented. Autonomous vehicles are a hot topic that could be seen as a good in which AI elements are embedded within it. There is a broad discussion in the literature about safeties of autonomous vehicles (which could be correlated to the safety of the produced goods).

Nevertheless, this topic is out of the scope of ASSISTANT. ASSISTANT focused on AI within the manufacturing sector, not those embedded in products. One characteristic of ASSISTANT is that safety considerations can be considered directly in the analyses since the risk analyses' background will be implemented.

In terms of robustness, if the end goal is to increase accuracy and reproducibility, different algorithmic techniques that evaluate the sense of perturbation and the definition of specific metrics can be implemented. Furthermore, different approaches linked with explainability that focus on neighbour analyses of fed imputes (e.g., see) can also be used to evaluate the

robustness of systems based on perturbation analyses. Approaches similar to those will be implemented/developed on ASSISTANT.

Since digital twin components would be the primary driver used in generative design approaches in ASSISTANT, it is required to consider that, independent of the level of detail imposed over them, accuracy would always be subject to a lack of precise system representability. The simulation does not model the entirety of physics and, therefore, there is always a possibility of a mismatch between the simulation (performed by the digital twin) and the physical environment.

In terms of privacy and data governance, information within the manufacturing sector should follow privacy regulations. In that sense, personal information would be managed separately with security access to a specific person within the industry (differential privacy and access control). Furthermore, algorithms deployed within the manufacturing sector will avoid using specific identifiers for customers, workers, and definitions based on preferences (e.g., given customer records and loyalty) should be left outside the scope of AI decision-making components. This does not imply that algorithms could not include such information, but definitions as such should be established by users and the company and not defined or predicted by algorithms.

In terms of transparency and accountability, the processes, algorithms, and approaches to secure traceability, explainability, and communication should not be different from those applied outside the manufacturing sector's scope. This is because even though the level of expertise could be higher within the manufacturing sector, that does not imply the need for persons outside the scope of it would require an understanding of the methods involved within the AI element (especially in cases of malfunctions and safety problems that could lead to external process scrutiny - linked to auditability). Notably, the digital twin, which can be seen as a virtual and interactive representation of the shop floor or components to be modelled, given its direct representability in the physical domain, provides a degree of explainability. This implies that neighbourhood approaches could be easily implemented if alternative solutions are constructed (interactively if necessary) to supply "reasoning" on how the algorithm derived the end conclusion. Of course, these options imply that visualization processes will have to be constructed to secure understanding to the user of the discarded alternatives (i.e., alternatives that fulfil the system constraints but are not optimal).

ASSISTANT focuses on the use of generative design approaches. This implies that human decisions can intervene/affect the system output by specifying associated bound or defining different fitness functions. This consideration put a clear responsibility framework to be considered in the development and deployment of generative design components. Unless a fault results from unforeseen interactions between the materials of the final goods and the design (via computational design), the human user or developers are wholly accountable for faults in the final design. This implies that responsibility should be stated and linked from the beginning of the AI development based on expected AI-human interactions and behaviours and AI functionalities.

Finally, ASSISTANT will achieve a general sense of transparency and oversight since, as stated in it, tools and components developed in it will be shared through the AI4EU portal. Even though this would make the algorithms developed scrutinized by the scientific community, the provision of explanation for non-technical will also be made. Additionally, the validation of the tools will be made over manufacturing sector partners that will provide additional robustness and reliability on the final components developed.

In terms of DnDF, society, and democracy, it is expected that most of the AI elements within the manufacturing sector would not involve direct managing of information that could lead to

biases or misuse of information related to these topics (if previous definitions on anonymous are preserved). Contrarily, sustainable and environmental concerns do still apply to the industrial sector. Therefore they should drive the application of algorithms with efficiency concerns to reduce the repetition of unneeded processes (e.g., repetitive algorithmic training) when unneeded (i.e., include metrics to track a fundamental need of an update or reconfigure AI elements). Concepts of beneficence and non-maleficence should not be a primordial focus in AI elements within the manufacturing sector since a sense of regular safety protocols that generally exist in the sector should support such concepts.

A combination of values rooted in computational science, business, and AI elements would be evaluated for implementation in the design-for-values. The values rooted in computational science: Autonomy, Accountability, Access, Authority, Consent, Democracy, Freedom (from bias), Justice, Privacy, Power, Service, Sociability, Transparency, Trust, The world community, The future, and Usability; business values that should include elements of quality, safety, efficiency, costs, sustainability, that could be easily tracked by specific KPIs (e.g., client satisfaction, decarbonization, energy efficiency, increase recycling rate, material efficiency, minimize emissions, accident rate, waste reduction). Additionally, specific ethical and business values include integrity, honesty, openness, respect, fairness, responsibility, customer service, quality, innovation, reliability, efficiency, and value for money (wealth). Not all these values should be incorporated into the definitions; instead, a implementation hierarchy will be used. The list will be constructed based on ASSISTANT partners based on industrial interest and ethical considerations.

In terms of Top-Down and Bottom-Up approaches, it has been argued that Top-Down approaches can hardly be used for principles formulation and implementation into practice. This could be considered valid when more robust social components are embedded into the system, and a Bottom-Up or hybrid system could be a better option (which does not apply to ASSISTANT or with a bit of consideration AI elements within the manufacturing sector). On the other hand, at the current stage of technological development, the extraction of desires and behaviours is in a development stage, and considerable care should be placed on the quality and quantity of information supplied to extract principles, values, and concepts helpful in implementing in ethical frameworks. Furthermore, in the manufacturing sector, intrinsic values that could be extracted for information should be driven in terms of those previously stated, and therefore the consideration is not about extracting what the manufacturing is seeking but instead about how to achieve it. In this regard, we think that a Top-Down approach that settles a combination of ethical values and business values could produce more straightforward and effective incorporation of ethical considerations. It should be noticed that there could be several types of explanation approaches (trace, justification, and strategy). The justification approach has been mentioned as the most effective one. Such an approach would be sought in ASSISTANT implementation.

Annex B. Safeguard based on fuzzy-logic

The safeguard construction based on fuzzy logic has previously been analysed [51]. Even when the approach covered in the mentioned manuscripts is based on neural networks, the approach can easily be transferred to any fuzzy-logic based tools (i.e. pure fuzzy logic construction or neural-based approach). The work focuses on implementing ethics in the context of a smart component and learning appropriate ethical behaviours. To further understand this approach, an explanation of the mentioned manuscript is first given to describe the main differences that can be defined for ASSISTANT.

The work refers to collecting moral, ethical, religious, legal, cultural, regional, or management policies as of Ethics of Operations (EOP). The ethical behaviours are expressed in

terms of fuzzy rules and for each fuzzy rule also specifies its ethical desirability (i.e. weight factor). Furthermore, rules are connected by inputs and outputs that are critical in determining ethical behaviours (i.e. values variables as inputs and KPIs as outputs – ASSISTANT). These variables are called Fuzzy Ethics Variables (FEVs). And the extent to which a given mapping between input FEV and output FEV is ethically desirable is reflected in a Scaled Ethics Value (SEV – i.e. Membership Functions -MFs).

The membership functions for ethical representation have five regions. They also define four ethical modes named mild, default, strong, and stringent to decide how to comply with ethical outcomes. This classification is considerably important for the current conditions of AI ethics since these ethical modes can be correlated directly with the intrinsic risk levels of the AI (Unacceptable, high, limited, and minimal risks).

To illustrate and connect the approach proposed in the manuscript, we will be guided by the same implementation case they defined.

They consider a smart healthcare product used to monitor patient condition that informs close family and doctor about the patient's well-being. The EOPs are:

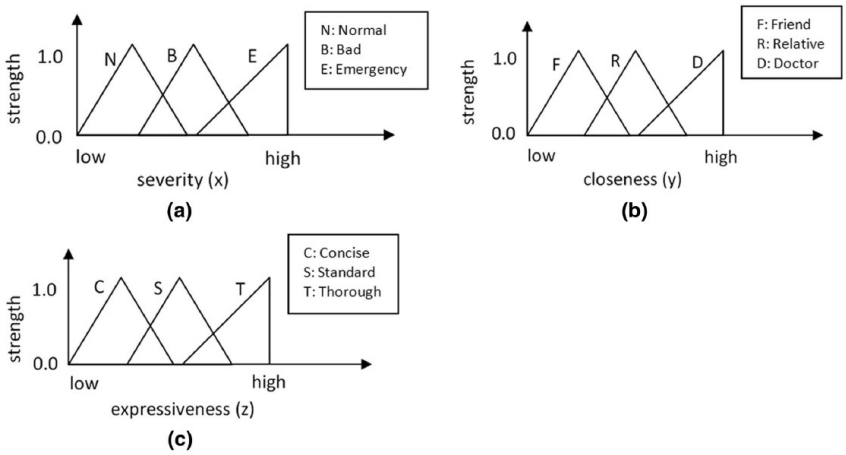
- If the patient's health is critical, it is obligatory for the smart device to inform the doctor.
- If a patient is mildly un-well close family should be informed but not friends.
- Also, the healthcare device may not disturb the doctor by sending any notifications if the patient is fine.
- It is permissible to share health updates with close family.

These EOPs translate into five ethical rules, as observed in the following table:

#	Ethical Rule	Meaning	Ethical Status	Range
1	If severity is Normal and closeness is Doctor Then Expressiveness is Concise	Inform the doctor if health is ok. Do not disturb the doctor.	Forbidden	0 - 0.2
2	If severity is BAD and closeness is FRIEND then expressiveness is CONCISE	Inform friend in case of bad health	Disliked	0.2 - 0.4
3	If severity is NORMAL and closeness is RELATIVE then expressiveness is STANDARD	Inform relative if health is ok	Permissible	0.4 - 0.6
4	If severity is BAD and closeness is RELATIVE then expressiveness is STANDARD	Inform relative in case of bad health	Recommended	0.6 - 0.8
5	If severity is EMERGENCY and closeness is DOCTOR then expressiveness is THOROUGH	Inform the doctor in case of emergency	Obligatory	0.8 - 1

As observed in this table, the ethical rules translate the meaning in which the ethical output FEV can be linked to the decision-making process based on the defuzzification process (which translates on the given ranges).

The following diagram shows the membership functions used in the system for the different FEVs. The severity of conditions and the closeness correspond to input variables. As observed, each MFs is represented through three linguistic labels



As previously mentioned, the authors used a neural structure to do the whole representation. They used a five-layer structure that can be linked directly to Figure 7 of the main document. The first layer corresponds to the input layer, where no transformation of the input information takes place. The second layer performs the fuzzification process, transforming the input data into the linguistic variables defined in the previous figure (e.g. normal, bad, emergency for the severity). The third layer has inputs for each linguistic variable depending on the activation functions, and only a few will be activated for evaluation. Therefore, the third layer performs the ethical rules described in previous tables, being the output of each EoP matric to be evaluated on the output FEV (which corresponds to layer 4. Finally, layer 5 merges the evaluated rules and performs the defuzzification process.

As defined here, the application of fuzzy logic to incorporating ethics in the IoT is not different to the approach defined in the main manuscript. Nevertheless, sound definitions of the FEVs and the ethical rules are required for implementation.

For the case of ASSISTANT, we propose to use such an approach as a safeguard system. The FEVs should be connected to relevant metrics (FEVs) that could help define the violation (or agreement) of different ethical rules extracted from Trustworthy AI considerations. In order to do it, we recommend the following considerations:

- The input FEVs should be defined by users considering the main decision-making process to be solved.
- The fuzzification process should be based on expert judgment with enough MFs as needed for a sound SEV representation.
- Ethical rules should be based on trustworthy requirements, and thus, the ethics rules should be defined in function of the intrinsic level of risk of the AI. To be more concise, we recommend using the framework scope definition to clarify the most relevant trustworthy components or values to be implemented as a safeguard.
- As a safeguard, the defuzzification process, and therefore the decision-making process, should define whether or not to allow the execution of the AI activity. If the process is allowed, the normal execution of the AI component can be run. Contrarily, the execution should be run, halted, or stopped, depending on the user's definitions.

The approach proposed here is not part of the ASSISTANT expected products; nevertheless, its implementation could be seen as a sound actionability to implement ethical safeguards on AI systems.

Annex C. ASSISTANT Ethical Risk Management Policy

Purpose

This document aims to provide a risk management policy for the ASSISTANT project. We focus on providing guidance regarding risk management that has its **origins in Trustworthy AI** considerations. This document intends to support the achievement of the consortium objectives by giving a structured approach to handling ethical considerations that could have their origins during the different parts of the life cycle of an AI element. Even though the approach seeks to help incorporate ethical considerations during the development of AI elements within the technical components of the ASSISTANT project, this approach can easily be used and extended for processes that involve deployment, decommission, and continual improvement of AI tools applied in different sectors.

The roots of this approach are based on the idea that ethical concerns raised by the current trends in AI can be handled as risks (and more specifically, hazards) that can produce, among others, adverse outcomes to operations, costs, processes, and safety and brand recognition. Given the experience that the manufacturing (and industrial sector in general) has in managing hazard conditions, the extension of existing approaches (given by the considerable gamut of standards and methods related to risk management) can be seen as a plausible alternative to incorporating ethical considerations during the development of AI elements and, in general, an alternative that the manufacturing sector can use to deploy within their dependencies specific values and ethical concerns.

Scope

The Policy is applied to l elements of the ASSISTANT project in which AI elements are involved. In addition, the methodologies and framework can be merged with other frameworks as long as they fulfil the requirements established in section “Integration with other systems and processes” are met. Finally, the approach used here can be combined with other risk Management processes to unify risk assessment approaches and produce a global evaluation of each sub-system and component that constitute the overall system in which ethical/values-based hazards are considered.

The current risk management process is intended to be tested in each element of the ASSISTANT project in which AI elements are involved. Nevertheless, responsibility for using or correct use of the presented framework does not lie within the work package dedicated to constructing this proposition (i.e. WP2). The responsibilities of each WP are presented in the Governance section of the present document, and they should be fulfilled as long as the technical components (i.e. technical WPs) are willing to evaluate the present framework. The present framework could be used alone or in conjunction with other risk management processes, but the responsibilities regarding external approaches (e.g. use cases risk management process) are not within the scope of the present approach. Therefore, any metrics, processes, and considerations concerning external factors will be the responsibility of the corresponding owner of these factors and methodologies. Independent of the previous section, a description of the methods that could be used to merge with other ISO31000 based approaches are presented here. The merging with other frameworks can be performed as long as they fulfil the requirements in the “Integration with other systems and processes” section. This merge would unify risk assessment and management approaches and produce a global evaluation of each sub-system and component that constitute the overall system in which ethical/values-based hazards are considered.

Risk Governance

An overview of the ethical-AI risk governance is described in the following table. This risk governance is only related to ethical-based risks for the ASSISTANT project and is not intended to compete or collaborate with the risk governance that considers management and control processes involved within the general framework of the ASSISTANT project. Furthermore, the governance presented here only considers application and considerations related to AI elements developed within ASSISTANT that could require particular concern about Trustworthy AI considerations.

Structural Component	Contributor	Responsibilities
Board	WP2 – WP7	Overall Responsibility for e-risk management
Audit Committee	WP2	Set out internal audit specifications, participants (i.e. WPs audit other WPs), and objectives. Monitor progress of audits and audits recommendations
Executive e-risk committee	WP2	Ensure e-risk management is embedded within ASSISTANT, receive reports, review e-risk profiles, review, evaluate, and recommend control and procedures, evaluate and keep track of the materiality of information.
The management committee (one per WP)	WP2	Compile risk register, define recommendations for modifications to strategies and policies, and track risk measurement activities.
Divisional management	WP3 – WP6	Prepare and keep up to date on the WPs risk register, set risk priorities for the WP, monitor projects and risk KPIs, and Manage internal activities related to risk and risk assessments (i.e. execute risk management processes).

Nomenclauture and definitions

This section presents a list of definitions for setting a common language for AI management. This will allow establishing, within the assistant, a more straightforward communication process within the different components that are built within ASSISTANT.

Concept	Definition
Artificial Intelligence	
Artificial intelligence (AI)	systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimisation), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).
Autonomous systems	can independently solve complex tasks in a specific application domain despite varying objectives and initial situations. Autonomous systems must independently generate an action plan, depending on the current task context, with which an overall goal specified by the operator of the autonomous system can be achieved without remote control and, if possible, without the intervention and assistance of human operators within the framework of legal and ethical requirements. If individual actions of the autonomous system fail during the execution of the plan, the system must be able to carry out a plan revision on its own to achieve the specified objective by adapting the original plan in another way.
Back-box models	The performance of AI systems in prediction, recommendation, and decision making support is generally reached by adopting complex Machine Learning models that “hide” the logic of their internal processes. Examples include deep learning models and machine learning ensembles (bagging and boosting models).
Ethical requirements	

GDPR	A set of clauses for collecting, storing, and using information that could correspond to personalised data or indirectly used to identify agents. (Data protection - Better rules for small business (europa.eu)). Data protection – by – design would also involve automatisation processes to secure the data governance of the system.
Agent	An agent is a system component (person or not) that acts on behalf of the system or another component. Additionally, an agent can be a Moral agent that can discern right from wrong and, therefore, be accountable for his/her actions. In the case of AI, since the AI buy itself should not be accountable buy itself, an accountability definition description should be stated at the beginning of the system/component development.
ANP and AHP	The Analytic Hierarchical Process (AHP) and the Analytic Network Process (ANP) measure intangibles using human judgment. These are structured techniques for organising and analysing complex decisions based on mathematics and psychology. They have particular application in group decision making and are used worldwide in a wide variety of decision situations, in fields such as government, business, industry, healthcare, shipbuilding and education. Rather than prescribing a "correct" decision, these methods help the decision-makers find a solution that best suits their goal and understanding of the problem. It provides a comprehensive and rational framework for structuring a decision problem, representing and quantifying its elements, relating them to overall goals, and evaluating alternative solutions.
Explanation–by–design	The own model can produce an explanation of the resulting outcome of the algorithm
Model explanation	It aims to explain the whole logic of a model
Outcome explanation	Understand and give a reason why an algorithm produces a specific outcome
Black-box inspection	Retrieve a visual representation for understanding how the black – box works.
Model-specific / vs model agnostic	If the approach to be used are applicable to a specific case or can be applied in general to all different solutions.
Explanation	An explanation answers a why question justifying an event.
Interpretability (in explainable AI)	The requirement describes the internals of a system in a way that is understandable to humans.
Metrics	
Functionally-grounded	Metrics that aim to evaluate the interpretability by exploiting some formal definitions that are used as proxies. They do not require humans for validation. For example, the interpretability of a model can be validated by showing the improvements wrt another model (or another solution) already proven to be interpretable by human-based experiments.
Application-grounded	Require human experts able to validate the specific task and explanation under analysis. They are usually employed in specific settings (assistant in the decision making).
Human-grounded	Metrics evaluate the explanations through humans who are not experts. The goal is to measure the overall understandability of the explanation in simplified tasks.
fidelity	It aims to evaluate how good is an explanation method at mimicking the black-box decisions. For example, if a surrogate model exists, the fidelity compares the prediction of the surrogate model on different instances used to create the surrogate model.
stability	A measure of how consistent is the explanation for similar records. The higher the value, the better the model. It can be evaluated by exploiting the Lipschitz constant, considering the neighbourhood of instances x's similar to x. then L = ||ex –ex'||/||x-x'||, where ex is the explanation.
accuracy	Metric to test the method performance.
Precision	Metric to test the method performance.
recall	Metric to test the method performance.
Feature importance	Each feature is assigned with an importance value representing how much that particular feature was necessary for the prediction under analysis. To understand each feature's contribution, the magnitude and the sign of each value of explanation are considered. ei<0 explanation contribute negatively to the outcome (opposite if ei>0).
LIME	Local Interpretable Model Agnostic Explanations. A model-agnostic explanation which returns explanations as feature importance vectors. The idea is that the explanation can be derived locally from records generated randomly in the neighbourhood of the instance that has to be explained. The key factor is that it samples instances in the vicinity of x and far from it. The explanation model is then constructed as a sparse linear model on the perturbed samples. The local feature importance consists of the weights of the linear model. IN ASSISTANT, we can use this approach but use the same digital twin to construct the system. (M. T. Ribeiro, S. Singh, and C. Guestrin. " why should I trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135-1144, 2016.
Saliency Map	It is an image in which a pixel's brightness represents how salient the pixel is. Formally, an SM is modelled as a matrix S with dimensions depending on the image pixels. The value of this matrix specifies the saliency values. A positive value means a pixel contributes positively to a classification, while a negative one, blue, means that it has contributed negatively.
Explainability	
Explainability	Active characteristic of a model, denoting any action taken to clarify or detail its internal functions.
Explanation by design	Explainable methods that return a decision and the reasons for the decision are directly accessible because the model is transparent.
Complexity (in explainable AI)	Degree of effort required by a user to comprehend an explanation. The complexity can consider the user background or time limitations necessary for the understanding

Logic Gate	Gates used in Fault Three analyses. The most common gates are the Or gate and the AND gate. Both gates produce one output.
PDDL (Planning Domain Definition Language)	The standardisation provided by PDDL can make research more reusable and easily comparable, though at the cost of some expressive power, compared to domain-specific systems. Planning tasks specified in PDDL are separated into two files: 1. A domain file for predicates and actions. 2. A problem file for objects, initial state and goal specification. Objects: Things in the world that interest us. Predicates: Properties of objects that we are interested in - can be true or false. Initial state: The state of the world that we start in. Goal specification: Things that we want to be true. Actions/Operators: Ways of changing the state of the world. https://arxiv.org/abs/1804.08229
HDDL (Hierarchical Domain Definition Language)	An extension to PDDL, the description language used in non-hierarchical planning, to the needs of hierarchical planning systems. https://arxiv.org/abs/1911.05499
Kappa statistics	Standard variable to define the quality of the training data sets. Thus as one of the criteria for the quality of the annotated training data, the reliability of the annotations for certification of an AI system based on ML becomes operationalisable and comparable through a standard metric. In order to generate a kappa statistic is necessary the evaluation of the same output features of systems for more than one agent. The kappa value range from 0-1. One is a perfect agreement between two observers. The kappa value considers the statistical chances of producing similar outputs by both observers and is estimated as Kappa k =(P_o –P_e) / (1-P_e) Where P_e=P+_1*P+_2+P-_1*P-_2 P_o=(P+_1+P+_2) /(p+_1+p+_2+p-_1+p-_2) P+_1= correct assessment by observer 1 P+_2 = correct asseesment by obersver 2 p-_1 = incorrect assessment by observer 1 p-_2 = incorrect assessment by observer 2
GDPR	
profiling	Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, including the natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;
pseudonymisation	means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject
controller	'' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data;
processor	means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller;
Risk, Risk assessment, and Risk Management	
Risk	Effect of uncertainty on objectives An effect is a deviation from the expected. It can be positive, negative or both and can address, create or result in opportunities and threats. Objectives can have different aspects and categories and can be applied at different levels. Risk is usually expressed in terms of risk sources, potential events, their consequences and their likelihood
Risk Management	coordinated activities to direct and control an organisation about risk
Stakeholder	coordinated activities to direct and control an organisation about risk
Risk Source	an element that alone or in combination has the potential to give rise to a risk
Event	occurrence or change of a particular set of circumstances. An event can have one or more occurrences and can have several causes and several consequences. An event can also be something that is expected which does not happen, or something that is not expected, which does happen. An event can be a risk source.
Consequence	the outcome of an event affecting objectives A consequence can be certain or uncertain and can have positive or negative direct or indirect effects on objectives. Consequences can be expressed qualitatively or quantitatively. Any consequence can escalate through cascading and cumulative effects.
Likelihood	chance of something happening In risk management terminology, the word "likelihood" is used to refer to the chance of something happening, whether defined, measured or determined objectively or subjectively, qualitatively or quantitatively, and described using general terms or mathematically (such as a probability or a frequency over a given period). The English term "likelihood" does not have a direct equivalent in some languages; instead, the equivalent of the term "probability" is often used. However, in English, "probability" is often narrowly interpreted as a mathematical term. Therefore, in risk management terminology,

	"likelihood" is used with the intent that it should have the same broad interpretation as the term "probability" has in many languages other than English.
Control	measure that maintains and modifies risk Controls include but are not limited to any process, policy, device, practice, or other conditions and actions that maintain and modify risk. Controls may not always exert the intended or assumed modifying effect.
Safeguard	Control action that is not continual and is only activated when specific conditions are met. These are used in extreme situations that minimise the impact of the risk conditions.
Corrective Actions	A documented design, process, procedure, or change implemented and validated to correct the cause of failure or design deficiency
criticality	A relative measure of the consequences of a failure mode and its frequency of occurrences
Severity	A measure of the degree of failure consequences that can be determined by the level of injury, damage, degree of rules violation
Detection Mechanism	Means of methods by which a failure can be discovered by an operator or system under regular operation.
Dignity	dignity, right to live, integrity, degrading or punishment, forced labour)
Freedom	liberty and security, respect for private and family life, protection of personal data, right of education, freedom of thought, conscience, expression, information, assembly, association, arts, science, and religion, right to property, right to asylum, freedom to choose an occupation and right to engage in work, and protection in the event of removal expulsion or extradition
Equality	equality before the law and non-discrimination, equality between men and woman, child rights, elderly rights, and integration of persons with disabilities
Solidarity	Right of collective bargaining and action, right of access to placement services, protection in the event of unjustified dismissal, workers' right to information and consultation within the undertaking, fair and just working conditions, child labour and protection of young people, family and professional life, social security and social assistance, environmental protection, consumer protection, access to services of general economic interest
Justice	Right to an effective remedy and to a fair trial, presumption of innocence and right of defence, principles of legality and proportionality of criminal offences and penalties, right not to be tried or punished twice in criminal proceedings for the same criminal offence
Critical Infrastructure	Energy, Information, Communication Technologies, Water, Food, Health, Financial, Public & legal order and safety, Transport, Chemical and Nuclear Industries, Space and Research
Severe*	Severe conditions imply those conditions that violate regulations can produce a considerable economic and environmental loss. Severe conditions also violate regional regulations, which could negatively impact legal and social components. The limit established for considering an economic loss (under severe conditions) should be defined by each company depending on their risk appetite.
Manageable**	Manageable conditions imply those conditions that could, if unattended, violate regulations and can produce economic and environmental losses. Manageable conditions also violate and contradict societal regulations, which could negatively impact legal and social components. The limit established for considering an economic loss (under manageable conditions) should be defined by each company depending on their risk appetite.
Failure Mode	How a component fails to perform its functions or achieve its objectives.
Failure Modes and Effects Analysis (FMEA)	A systemic process to identify potential failures to fulfil the intended function, identify possible causes, determine approaches to eliminate failing conditions and locate the failure impacts on reducing the impacts.
Failure Modes, Effects and Criticality Analysis	An extension of FMEA. It includes a consequences severity estimation of the failure (or a combination of the failure likelihood and the severity of the consequences).

Risk Management Architecture

A general scheme of the management architecture that will inform and monitor the management of the different e-risk components is presented in Figure 1. There are five main management components (also listed in section 3), with specific responsibilities that will secure correct risk management within ASSISTANT. In addition, a description of the main general and individual responsibilities, communication processes, and reporting structure are described in the following subsections.

Additionally, a framework describing the global process in which AI elements will follow their development and deployment in ASSISTANT is presented in D2.4. This framework has its root in both a definition of general frameworks used in risk management with the addition of the architecture describing the framework for AI development under Trustworthy AI considerations.

Finally, Figure 2 describes the system architectural link (i.e. the hierarchy of the AI assets) within ASSISTANT and their descriptions within their respective systems (i.e. work package) and subsystems (i.e. each element structured within each technical work package).

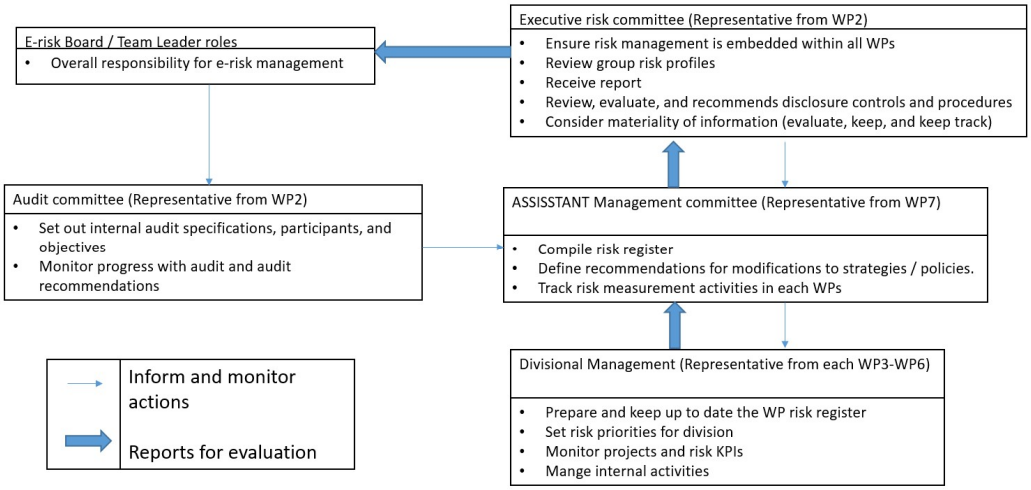


Figure 1. E-Risk Management Architecture

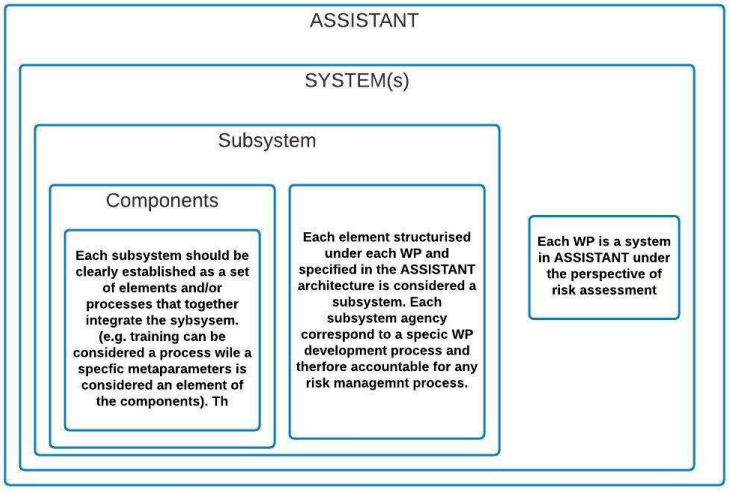


Figure 2. System architectural link to the risk assessment

Roles and Responsibilities

Each WP within ASSISTANT will contain a **divisional management**. Its responsibilities involved preparing and keeping updated the risk register and performing all activities related to the risk management process. The risk register would be populated based on the internal risk assessment (i.e. manage internal activities). Furthermore, based on the e-risk assessment result and risk management process, the KPIs used to track the performance of implemented e-risk managing approaches will be monitored and reported. All the reports from the divisional management will be given to the management committee. Finally, the Divisional management has the responsibility to set the risk priorities for the internal WPs, ensuring that the hazards/e-risk that require the earliest attention are managed within each WPs. In general terms, The divisional management will specify the team members that will participate in the risk assessment process that should involve, among others, individual closes to the event and

component, sub-system manager or supervisors, and other experts qualified for the expertise required in the definitions of the sub-system and its components under evaluation.

The Management Committee is responsible for accumulating the information generated from the different WPs and audit processes related to risk management activities. More specifically, the management committee will compile the individual risk registers from each WPs in a unified risk register, define, if possible, an early stage of recommendations for modifying the current strategies and policies established within the divisional management and audit processes (i.e. work as an informer of issues described by the different WPs and the audit committee), and finally track risk measurement activities in each WPs (i.e. compliance based on a recommended scheduled process defined for risk measurement activities). The management committee will work as a link between the risk committee and the divisional management by securing an expedited flow of information and communication (see section 4.2).

The Risk Committee has the general responsibility of reviewing, evaluating, and recommending the process involved in the risk management and the status of the risk management and audit processes. The risk committee will use the complete information reported by the management committee and perform strategic recommendations that will be informed to the assistant board or sent back to the divisional management depending on the scope it desired to be covered in this strategic step modification.

The **E-risk board** will take the overall responsibility for the e-risk management process and, if required, inform the status of the E-risk management process to the ASSISTANT board. Additionally, the E-risk board will decide on the audit process to communicate to the Audit Committee when necessary to perform one a WP. The audit could be related to a specific AI element of the processes involved in the development and deployment of AI elements but always with the perspective of Trustworthy AI (i.e. within the scopes set in the policies). In general, the E-risk bard responsibilities involve, among other, coordination of the risk management process, assembling, encouraging, and supporting a proactive team, assigning implementation tasks to team members (i.e. specifying roles and responsibilities for the rest of the components in Figure 1), be involved in the analysis and action plan implementation processes (Treat, Transfer, Tolerate or terminate – known as 4T's), communicate the progress of institutional barriers, monitor goals and progress towards completion of action plans and submit the finalize action plan register, help break down barriers to implementing the action plans, and define audit processes to evaluate the status of implementation within the system, and encourage the carried out on the time of actions within the definitions established in the cycling process of continual evaluation and improvement of the risk management process.

Communication

The communication and consultation between the different risk management bodies will follow the structure established in the Risk Architecture. Furthermore, this communication would follow the formality established within the ASSISTANT project and would be performed promptly and ensure that relevant information is collected before submitting any report. All communication channels should ensure that the relevant information shared should be synthesised, appropriate, and sound to the requirements established in the report format. Since the risk management process will be considered an activity within the ASSISTANT project, the same communication channels could share information and conduct a consultation. Furthermore, monthly meetings within WP2 would be used as a communication driver between the different bodies, ensuring that participants are aware of the risk management process statuses and all the activities required to be performed.

Management Process

The management process within ASSISTANT is considered an independent management process sustained and fostered by WP2. Nevertheless, the primary responsibilities in executing risk assessments and using the proposed tools are from the corresponding WPs, including AI assets. To initiate activities of the present framework and continue with a dynamic execution of the framework, the following figure shows a pipeline that describes the constitution of the different bodies and the members of ASSISTANT that will participate in it. Furthermore, the figure shows the recursivity needed to perform modifications in the framework as the current approach is implemented. Following the figure, the first process corresponds to setting the Risk Management Policies Agreement (i.e. current document). The agreement of the different technical WP to implement the framework is shown in the following table. Even though some technical components were not set to implement and test the framework, the different ASSISTANT partners were willing to participate and test the current framework.

Table 28 Agreement by technical WPs

Step	Alternatives	WP3	WP4	WP5	WP6	WP7
Implementation of the Ai framework?	Yes/ No					
Policies agreement (as set in D2.3)	Agree / Disagree & why					
Architecture Definition	Agree on an architecture proposition or define a new one					
Set of internal Responsibilities within WP for risk management process	Names of participants based on architecture					

After the different technical WPs agree with the use of the present framework, the definition of the E-Risk board members is performed. Each participant WP will require such set members to participate in this board, including WP2. Then, further specification of the participants that will be driving the different activities of the risk management process takes place together with the specification of the scopes of analysis. These specifications are based on the ASSISTANT technical architecture to establish the main components that will require the tool implementation. After setting the definitive members of the risk management process, the ERC and MC settle the documents used to start the process. Finally, a list of different documents is provided in the main body of the deliverable D2.4.

The dashed box represents the continual process of the framework use, and, as observed in the figure, it is initiated with the pipeline processes shown in the D2.4 sections 5. As expected, different reports will be produced by the framework that will contain the FMEA/FMECA analysis (i.e. risk register), recommendation, and different analyses. The MC will compile these reports and extend the recommendations to the E-risk board or the executive risk committee to determine if these recommendations are over the framework or the risks. In the first case, modifications are defined by the risk board. For the second, the definitions are settled by the management committee. The E-risk board can require audits to check the risk management assessment and the correct implementation of risk treatments depending on the results. If the recommendations are over risk, the divisional management will integrate the recommendations based on the 4T's (Treat, Transfer, Tolerate, Terminate) before reinitiating the whole process.

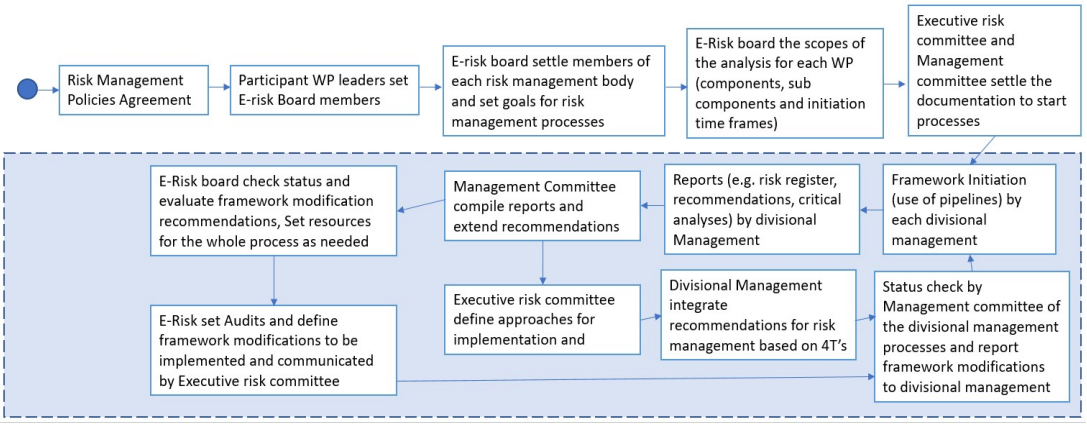


Figure 39 Framework Initiation and Workflow in ASSISTANT

The workflow of activities can be redesigned and restructured as needed. The responsibility for these changes will lay over the E-Risk board. The activities' duration will be bound to the corresponding formal noticing of the different participants. Even though there is no current expected time for execution of tasks, given the condition of framework evaluation, we encourage the participants of the risk management process to consider the following table for executing the activities. These times could be modified as a broader experience is obtained from the framework evaluation.

Recommended time for activities execution (per division - per WP in ASSISTANT)	
ACTIVITY	Time (in days)
MANAGEMENT	
Set the Management Committee and the Divisional Management	5
Set documentation in the function of the scope defined by e-risk board	10
Reports Generation	2
Compile reports and extend recommendations	2
E-Risk board evaluate modifications and recommendations	5
Audits	10
Risk ASSESSMENT (based on Bechnmar e-risk management process - Section 5)	
AI Confirmation	0.5
E-risk identification and classification	1
AI scope definitions	1
Analysis of values	5
Execute e-risk management process	
Execute e-risk management process	
Establishing Context	10
Risk Analysis and Evaluation	5
Risk Treatment, transfer, termination or tolerate	N.A.
Reviews, Updates, and implementation	5

Risk Strategy

The process of Identification of potential failures In ASSISTANT would be rooted in the combination of the problem definition, the use of approaches used in system failure analyses, and the use of specific frameworks. The frameworks would allow tackling specific components that could produce failing conditions under the umbrella of ethical considerations and values.

Since the purpose of the approach presented here focuses on the considerations and regulations generated by the EC regarding **Trustworthy AI**, the approach is based, in addition to standards such as the ISO31000, ISO/IEC TR 20547-1:2020, ISO/IEC TR 20547-2:2018, ISO/IEC

20547-3:2020, ISO/IEC TR 20547-5:2018, and the ISO/IEC TR 24028:2020, to the Ethical Guidelines for Trustworthy AI produced by the High-Level Expert Group on Artificial Intelligence, the White Paper on Artificial Intelligence, and The Artificial Intelligence Act. Further standards are under development for different recognized bodies, and we seek to consider their approaches as much as possible within the current risk management process.

Since ASSISTANT goal does not impose a specific scenario in which its approaches will hierarchically be implemented (i.e., there is no preference between focusing on process design, production, scheduling, or control), the likelihood considerations of event would be driven by expert judgment. Nevertheless, the relative occurrence of adverse events triggered by the activity or activation of AI components could be studied based on the confusion matrix and statistical methods (if possible).

The backbone frameworks and guidelines for implementing ethics in ASSISTANT include the ART principle, the Ethics Guidelines for Trustworthy AI, the white paper on artificial intelligence, the artificial intelligence act, and several standards and approaches that support risk management implementation. These documents were selected based on geographical considerations and trends in relation to legal concerns in the region.

In terms of Top-Down and Bottom-Up approaches, it has been argued that Top-Down approaches can hardly be used for principles formulation and implementation into practice. For assistant implementation, the focus is not on extracting what the manufacturing seeks but on how to achieve it. Thus a Top-Down approach that settles a combination of ethical values and business will be more straightforward and effective to incorporate for analysis. Contrarily, several risk management techniques are based on both approaches. This implies that the method specification would be goal-oriented.

Risk Management Process

The overall Risk Management Process is presented in D2.4. Section 5 covers the methodology that should be used to incorporate the risk management process. The approach is based on the ISO 31000 with modifications to define sound scopes for AI ethical considerations. Furthermore, different pipelines are defined to allow a more effortless experience from the user's point of view.

Integration with other processes

The following figure shows a schematic of the plan in which ethical perspectives are embedded within the ASSISTANT project. As observed in the figure, an ethical work package is dedicated to analysing and fostering the incorporation of Trustworthy AI-driven concepts within ASSISTANT. To do it, two branches of development (as seen in the figure) are used. The first branch (top of the ethical work package – Figure 1) focuses on analysing and implementing responsible

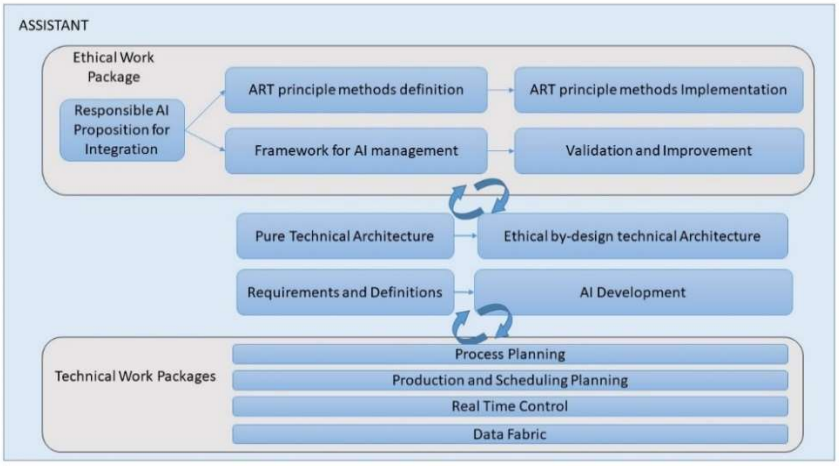
AI concepts within the technical architecture in which different AI components would be deployed within the project. The technical architecture that will settle the interconnectivity, interactions, rules, and specification of the components within the developed tools within ASSISTANT, including visualisation components that will lead to interactions with different agents. The ART principle allows us to integrate ethics-by-design into the project architecture.

The second branch involves the definition of management strategies of the AI component for the manufacturing sector. This implies generating a framework to consider current and future trends in Trustworthy AI, standardisation, and AI risk management. The components

development will check the compliance with the trustworthy guidelines with considerations based on the risk analyses assessment that describes the most relevant risks that could be foreseen to produce undesired outcomes regarding the Trustworthy Guidelines requirements.

As observed in the figure, a step of framework development is followed by a validation step. The validation step involves ASSISTANT internal validation and a manufacturing sector review. Since a management plan based on risk management involves a definition of risk management architecture, strategies, and protocols, the first primary consideration within the project development is the agreement of such concepts (policies) by the different stakeholders involved.

The Policy is applied to all elements of the ASSISTANT project (i.e. work packages) in which AI elements are involved. Furthermore, the policy involves the methodologies developed can be merged with other frameworks as long as they fulfil the specific requirements that settle the approaches used for risk assessment and management (i.e. ISO31000 based approaches and top-down or bottom-up based analyses).



Further integration specification would be made during task 2.5 since it would depend on the use cases involved and the expectations (and implementation motivation) of the AI framework proposed here. Independent of this statement, a proper implementation with the stakeholders would require defining:

- An appropriate plan including time and resources.
- Identifying where, when and how different types of decisions are made across the organization and by whom (i.e. implementation of risk architecture)
- Ensuring that the organization’s arrangements for managing risk are clearly understood and practised. (i.e. evaluation of further workshops within ASSISTANT).

Furthermore, an understanding of its external and internal context is required; therefore, Examining WPs and use cases' external contexts may include:

- Whether international, national, regional or local, social, cultural, political, legal, regulatory, financial, technological, economic and environmental factors are involved.
- Key drivers and trends affecting the objectives of the WPs and use cases.
- Users' perceptions, values, needs and expectations.
- User's case vision, mission and values.
- Users governance, organizational structure, roles and accountabilities;
- Users strategy, objectives and policies;
- Current standards, guidelines and models related to Trustworthy AI.

- Understanding resources capabilities in terms of capital, time, people, intellectual property, processes, systems and technologies.
- Understand linkages of data, information systems and information flows (i.e. technical architecture).
- Interdependencies and interconnections.

Risk categories

The risk categories are specified in D2.3 and would not be covered in the risk policy specification. The categories are based on ethical concerns and can be classified by their intrinsic ethical risk level (i.e. Unacceptable, High, Low, Minimal)

Risk Register

The risk register is specified in D2.3 and would not be covered in the risk policy specification.

Risk Reporting

The risk categories are specified in D2.3 and would not be covered in the risk policy specification.

Risk Management Performance

Different KPIs are specified in the D2.3 and would not be covered in the risk policy specification. The definition of the most suitable KPIs to track both the AI performance and the ethical framework performance would be dependent on the use case interest and information that exist for their tracking (e.g. pure qualitative vs pure quantitative).

Risk Appetite

Risk appetite should be settled based on the tables defined in D2.4. However, the user's interest could impose a more restrictive risk appetite on their AI and thus impose further restrictions over the risk scores settled for the risk analysis. Further information on settling the risk matrix and the risk appetite combination can be found in the deliverable.

Review and approval

This document has been reviewed internally by ASSISTANT members. Moreover, the policies will be presented in ASSISTANT general assemblies to have a general review and approval from the different ASSISTANT participant institutions.