CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	A responsible AI framework: pipeline contextualisation
Authors	Vyhmeister, Eduardo;Castane, Gabriel;Östberg, P.-O.;Thevenin, Simon
Publication date	2022-04-19
Original Citation	Vyhmeister, E., Castane, G., Östberg, P.-O. and Thevenin, S. (2023) 'A responsible AI framework: pipeline contextualisation', AI and Ethics, 3(1), pp. 175–197. Available at: https://doi.org/10.1007/s43681-022-00154-8
Type of publication	Article (peer-reviewed);journal-article
Link to publisher's version	10.1007/s43681-022-00154-8
Rights	© The Authors 2022. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. - http://creativecommons.org/licenses/by/4.0/
Download date	2026-11-02 03:20:49
Item downloaded from	https://hdl.handle.net/10468/15435



University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

AI and Ethics (2023) 3:175–197
https://doi.org/10.1007/s43681-022-00154-8

ORIGINAL RESEARCH

A responsible AI framework: pipeline contextualisation

Eduardo Vyhmeister[1] · Gabriel Castane[1] · P.-O. Östberg[2] · Simon Thevenin[3]

Received: 17 December 2021 / Accepted: 10 March 2022 / Published online: 19 April 2022
© The Author(s) 2022

Abstract
Incorporating ethics and values within the life cycle of an AI asset means securing its development, deployment, use, and decommission under these perspectives. These approaches depend on the *market domain* where AI is operational – considering the interaction and the impact on humans if any process does not perform as expected – and *the legal compliance*, both required to ensure adequate fulfilment of ethics and values. Specifically, in the manufacturing sector, standards were developed since the 1990's to guarantee, among others, the correct use of mechanical machinery, systems robustness, low product variability, workers safety, system security, and adequate implementation of system constraints. However, it is challenging to blend the existing practices with the needs associated with deployments of AI in a trustworthy manner. This document provides an extended framework for AI Management within the Manufacturing sector. The framework is based on different perspectives related to responsible AI that handle trustworthy issues as risk. The approach is based on the idea that ethical considerations can and should be handled as hazards. If these requirements or constraints are not adequately fulfilled and managed, it is expected severe negative impact on different sustainable pillars. 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 in the manufacturing sector. The framework follows a pipeline structure, with the possibility of being extended and connected with other industrial Risk Management Processes, facilitating its implementation in the manufacturing domain. Furthermore, given the dynamic condition of the regulatory state of AI, the framework allows extension and considerations that could be developed in the future.

Keywords Responsible AI · Manufacturing · AI · Ethics · Framework

1 Introduction

The industry is becoming more automated in the Digital Era. The main focus has been acquiring, collecting, and managing all data produced intelligently and efficiently during the last decade. Current factories, and trends, 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]).

Generally speaking, the AI has been implemented in different industrial processes, including optimisation, quality thorough operational excellence, generative design, intelligent purchasing and supply management, supply chain risk assessment, robotics, and smart devices (that includes self-configuration, self-optimisation, self-protection, and self-maintenance) [3–5]. However, incorporating AI within the manufacturing sector comes with different challenges and risks that should be addressed. As identified by Fujimaki [6], these challenges include shortage of AI talent, technology infrastructure and interoperability, data quality, real-time decision-making, edge deployments, **trust**, and transparency.

✉ Eduardo Vyhmeister
eduardo.vyhmeister@insight-centre.org

Gabriel Castane
gabriel.castane@insight-centre.org

P.-O. Östberg
p-o.ostberg@biti.se

Simon Thevenin
simon.thevenin@imt-atlantique.fr

[1] Insight Research Centre, University Collage Cork, Cork, Ireland

[2] Biti Innovations AB, Umea, Sweden

[3] Institut Mines-Telecom, Palaiseau, France

Furthermore, as pointed out by Accenture and Deloitte [7, 8], challenges from a company perspective can include: security, budget, lack of talent to implement and run AI, big data and data analytics, integration with existing systems, and procurement limitations (e.g. on big data vendors and enterprise not ready for it).

According to [9], trust is a perceived benefit and perceived risk for users on the adoption of AI software components. In that work, challenges are classified, including those mentioned above, into ability, integrity, and benevolence. Other authors also consider the safety and security dimensions in addition to these three associated to market verticals such as medical or laws environments [10–12]

Furthermore, trust is vital for technology providers to be confident for consumers to use their products, independently of the market segment. Finally, trust is a basis for social organisation, civic democracy, and economic prosperity[13].

To ensure trust, the actions of any agent over humans must be reliable [14] or, in other words, with the lowest risk to produce adverse outcomes. This requirement depicts the linkage between trust and risk management.

Different organisations have developed various methods based on multiple ethical principles to facilitate practitioners in developing AI components. These organisations include academia, trade union, business, government, and NGOs. Examples include The Institute for Ethical AI and Machine learning [15], Microsoft's Responsible AI guidelines [16], UNI Global Union [17], the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [18] together with its newest release the IEEE Standard Model Process for Addressing Ethical Concerns During System Design [19].

However, considering authors' knowledge, neither of these works nor others have included a risk management approach incorporating the AI software life cycle and the ethical imperatives.

The current work presents a framework dedicated to managing AI components in the manufacturing sector. The general idea lies within the extension of well known and developed risk management approaches used in the industry, but with considerations of current trends in AI. In so doing, trustworthiness requirements and compliance with present regulatory conditions (and future ones - given the framework extension capabilities) could be achieved for AI assets and products. Current work continues previous work entitled "Risk as a Driver for AI Framework Development on Manufacturing". In it, the foundations for establishing an AI framework management that is rooted in responsible AI scrutiny was done. Furthermore, a link between ethical considerations, trustworthiness, and risk was presented to impose the validity of handling responsible AI requirements and ethical based considerations as a Risk Management Process (RMP). This work continues the development of the framework by introducing a well defined and structured approach for performing system contextualisation; step needed in any RMP and risk assessment. Future works will help define protocols (i.e. risk assessment) and metrics to perform and track ethical-based RMP. A thorough explanation of our previous, present and future contributions are encapsulated in the following section.

2 Contributions

The Responsible AI framework presented in this work provides a flexible methodology for the manufacturing sector to develop, implement, assess, and control AI assets with current trustworthiness considerations – combination of methods, frameworks and strategies including the ISO31000 [20], the trustworthy guidelines as ethical requirements [21], the white paper on artificial intelligence, the classification of AI elements based on the Artificial Intelligence Act [22], the charter of fundamental rights [23], Deloitte approach for management [24], the General Data Protection Regulation (GDPR) [25] and different techniques that support the framework use.

Thus, the framework objectives are:

- Support developers to incorporate ethical principles and values within the AI in product life-cycle processes. It is key that developers are familiar with the ethical principles at every stage of implementing/operating products with AI. Furthermore, it is critical to distinguish between requirements that 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 on the use of AI.
- Facilitate the combination of the framework with other approaches used to handle risks by industrial stakeholders. To enhance the adoption by companies that already have their own risk management process, the framework need to be design as a complementary asset to these 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 spiral/iterative development processes - e.g. agile techniques. Therefore, the framework should incorporate the benefits of these development cycles to ease the developers with its incorporation.

- Ensure that metrics and Key Performance Indicators (KPIs) can be tracked to register the evolution of the ethical based risks management. For many companies, specifically, the business units, tracking KPIs are essential for their daily operations. In addition, managerial levels must use this tool to have a broader understanding of 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 and developers (among other stakeholders) do not know AI 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 research advances and 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 ease its understanding and implementation.

As mentioned above, our approach currently focuses on weak AI components. Since there are no strong AI components, and in the foreseeable future, the lower chance of its incorporation in the imminent manufacturing sector, its impact on the current proposition should be minimal.

The current framework does not consider generating any specific tool or software to handle trustworthy components within the AI management process and their risk management. Instead, it use a some current applications (e.g. ALTAI tool [26]) to be combined with the RMP.

The framework does not intend to overlap regional regulatory conditions; they should be taken into consideration as part of the pipeline or considered before the implementation of the current framework. Furthermore, technical components, such as transparency, are not specified in the present framework. Inherit decisions based on data type, AI tool, and protocols for robustness evaluation are out of the framework's scope.

Figure 1 shows a high level overview of the framework. It has two primary components: (i) a Responsible AI Framework development - structured on four sub-components - and (ii) the validation of the framework – case studies evaluation and industrial partners at the bottom part of the Figure.

One of the sub-components, named Contextualisation, is the main focus of the present work. Within it is defining

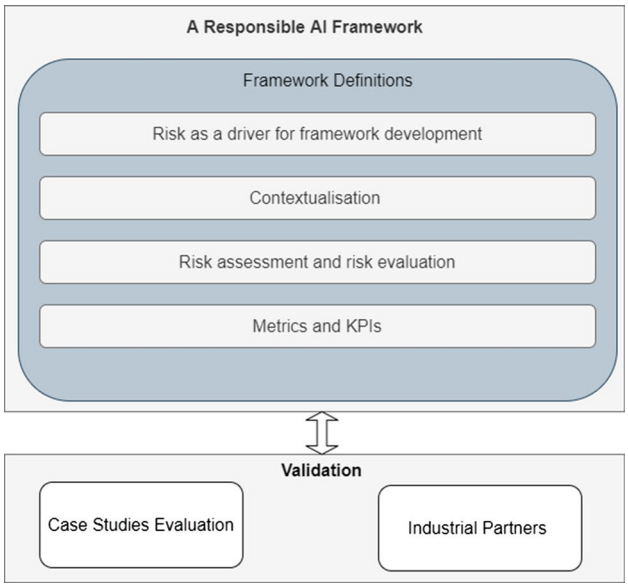


Fig. 1 Framework overall structure

and agglomerated different considerations before the risk assessment processes take place. To be more specific, this contextualisation can be linked to the ISO31000 step named Establishing the Context [20] that helps to define AI assets risk levels, trustworthy requirements, values (e.g. corporate values) and what elements should take further attention during risk management.

In order to deep down into contextualisation and risk assessment and risk evaluation processes, some ethical frameworks and guidelines discussion, as a mode of introduction, is made next. Then the recap on RMP is performed.

A vertical-domain approach is proposed for the manufacturing sector that will consider its ethical perspectives, values, local requirements, and well-known approaches related to risk management. It is planned to be implemented on specific use cases through a European Project [intentionally removed the name for reviewing purposes].

3 Ethical frameworks and guidelines

Ethical frameworks and guidelines can be seen as an ethical-based generalisation or implementation approach for the AI life cycle. Different organisations have developed various methods based on multiple ethical principles to facilitate practitioners developing AI components worldwide [15–19, 21–23]. An extensive list of guidelines and strategies based on critical issues can be seen in [27]. As shown in the mentioned work, no approach covers all the 22 extracted issues.

Even though organisations show different interest on what issues and principles (ethical-based) should focus on, the most relevant concepts as seen by organisations

and companies includes privacy, fairness, accountability, transparency, explainability, and safety [27, 28]. These key concepts would have different importance and relevancy depending on the AI life cycle stage. Therefore, strategies for using ethical guidelines and general frameworks should be seen as a supporting approaches, independent of the AI stage.

Furthermore, ethical AI frameworks and AI implementation guidelines should consider the entire environment in which these components are developed and deployed [1]. Conditions could change over time as tools are integrated into dynamic environments and, therefore, challenges, concerns, and risks would not always be foreseen at the initial stages.

Improved risk identification requires advanced knowledge on the scope where AI will be developed and deployed. Thus domain environment should be considered in the contextualisation stages. 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 frameworks and tools as some ethical concerns, values (and their hierarchy), and decisions change over time. The sum-up of all these considerations and challenges makes the generation of frameworks a cumbersome task.

Even under those challenges, some higher-level frameworks and guidelines have set concepts that impact the ethical AI considerations, helping in the RMP contextualisation of AI assets. According to the authors at [22], understanding these approaches is essential since they can contribute to fostering trustworthiness and, at the same time, get a higher level of understanding of approaches that could be used for AI development. The relevant for the proposed framework are covered next.

3.1 Human-Centric

Human-Centric can be seen as a sub-class or a particular AI approach that focuses on the AI interaction and collaboration with human agents. Thus, the algorithms (and learning processes) can continually be updated and consider the human agents' state, needs, experiences, and human-AI physical component interactions. In addition, 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, the system could require an understanding of the environment and ongoing interactions depending on functionalities and objectives. 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 [29]), physical, collaborative, and integrative [30].

The perspectives in 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 component). These challenges could include, human factors and technical factors [28, 31, 32], some relevant are defined next:

- 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 components that, for example, are intrinsic in the case of image processing.
- Models or techniques to improve understanding 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.
- There is a suitable link between non-interpretable formalism (i.e., data and machine learning components) and interpretable formalism (symbolic models and specifications interpretable by human agents deployed by encoding processes).
- 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.
- interactions should have, developed, and refine a shared vocabulary of concepts and relations, agreements on interdependencies, knowledge models.
- Make the system reliable to the extent that critical applications (e.g., human surgery, automatic driving) would not produce erroneous estimates and be safe to a broad extend (including noisy information and cyberattacks) - in other words, verifiable to the extent in which performance surpass the current state.
- Defining standards and protocols to general/specific applications and domains for the AI components that will interact with human agents, independent of the method of communication (e.g., verbal).

Based on the previous list, developing a framework that provides a structured methodology for managing AI-based risks can contribute, among others, to the last three mentioned challenges. Furthermore, by securing a suitable

approach that continually assesses and improves AI assets' approaches, the challenge of eliminating or handling human biases could be fostered. Thus, technical components and a suitable framework secure better system conditioning in terms of diversity, non-discrimination, and fairness.

3.2 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 approach that extends on the autonomy and collaboration of AI in regards to human 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 oracles feedback. The benefits of HITL includes relative incorporation of transparency within the systems, incorporation of human judgment (i.e. accountability) and, among others, removing pressure from perfect algorithms [33]. As mentioned by Dignum [34], HITL is often the most appropriate since they allow for more clear responsibility attribution. Nevertheless, the decision made by the principal-agent is affected by societal, legal, and physical infrastructures.

Some important considerations within the HITL approach are that (1) it is dependant 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 can be linked to the perspective of human-centric approaches.

It can be seen that for HITL a human still has complete control over the 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. Thus, 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 intrinsic high-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 [35].

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 perceptions that humans and machines have different capabilities [36]. As expected, this process involves incorporating several human-related weaknesses that could derive from technical and non-technical failing conditions. Boredom at routine monitoring, bias incorporation, alert and fatigue, among other considerations, let to establish that humans perform poorly as supervisors of automated technical systems [36] (i.e. a restriction to be used on complex manufacturing systems).

A sound RMF contribute 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, a RMP oversees the overall system and secures intervention stages when needed.

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

3.3 In-, By-, and for-design (ethics) and design for values

AI components 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 components under different scopes. One scope involves that components can be built, deployed, and integrated under some pre-specified concepts or approaches (e.g., ethics) to embed the element with the 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). Readers are encouraged to check for a thorough understanding of these concepts [34, 37, 38]. 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 component 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).

To grasp a better understanding of these approaches, it is first required a good understanding of ethics.

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 values, principles, and purpose of an agent 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. 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 the concepts of "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).

The normative, virtue and applied subdomains of ethics can be readily implemented within the approaches of -in, -by, and -for-design. 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. As shown in the literature, different ethical frameworks have already been settled in their establishment as a solution for AI development and deployment [39–41]. It is also worth notice 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 optimized 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; benevolence and universalism over personal power or achievement enhancement [34, 42]. Therefore development and deployment stages could follow a structured methodology guided by hierarchically organized values [38]. The design-for-values approach allows incorporating values in a rational way guided by a process that involves: the identification of relevant values, generating a normative practice for the incorporation of such values, and linking such normative system with concrete functionalities [34].

As specified by Lason [43], AI components 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). In this work, the alignment based on morals and values is one of the most suitable alternatives to impress with ethical considerations different components.

Incorporating values (and ethics in that regard) can be done, from a technical point of view, integratively or controllably. The first can be considered as a parallel structure (e.g., merge it with current approaches such as architectures - see [34]). The second can be considered as a series structure (e.g., as part of a pipeline in which functionalities and normative sets are applied outside the own AI component, 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, values, even though generic, considerations such as wealth, health, safety, and others, can be linked to metrics representing the system's value state. This point is essential since they allow to monitor, given a pre-specification of suitable indicators, the state of a given condition. Finally, they allow the transformation of abstract concepts into norms that allows integration of such norms as specific requirements that different stakeholders can understand.

A clear example of implementing the Design-for-Values approach can be seen in [37, 38]. The Design-for-Values approach works as a filtering component around the developed AI component to map moral values into explicit, verifiable norms that constrain the system inputs and outputs.

3.4 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. 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. The system is evaluated as a whole and general complex in which specific components (Down) can interact in its definitions and analyses. 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. These Top-Down and Bottom-Up approaches have also been beneficial in designing indicators that help evaluate the systems' state [38].

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 implies deriving the intrinsic rules that will 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? [43].

4 Risk management as a source of trust

In order to have a sound understanding of risk management as a source of trust and have a better concept related to a base RMF, readers are encouraged to review a previous work [intentionally removed the name for reviewing purposes].

The present work continues the framework's development, focusing on different standards, guidelines, and frameworks (defined in the contribution section). For contextualisation purposes, some of the previous work most relevant contributions, considerations and definitions are summarised here.

Risk management can be seen as the process involved in identifying, assessing and controlling risks. The RMPes are well established, but they could be presented differently with different terminologies depending on the domain and the guidelines used in their implementation.

To use a RMF based on ethical scrutiny, it is first required to understand what implies an ethical risk. In terms of responsible AI, the ethical requirements imposed over AI, the values that would like to be branded on them, and the social, societal, legal, and environmental constraints should, among other considerations, be considered as ethical objectives of AI assets functionalities. In addition, several conditions, processes, and status with different probabilities or likelihood to materialize can damper or restrain the expected AI behaviours. We call the combination of these events' probabilities to materialize and the impact over the AI objectives as **ethical risks** or **e-risks**.

A RMF does require: (1) a clear structure and formality for performing communication and reporting, which is also denominated as Risk Architecture (RA); (2) a definition of the strategies for implementation set by the system/organization, denominated as Strategy (S); and (3) a set of guidelines and procedures for performing the process of managing risks, denominated as Protocols (P). The combination of these components is denominated in conjunction with the RASP strategy. However, the present work focuses on the protocol component since the RA and S were covered in the work entitled "Risk as a Driver for Ai Framework Development on Manufacturing".

One important consideration is that AI methods can be embedded within processes or be a stand-alone system used for the map, prediction, forecast, optimize, and recommend, among other tasks. A general definition as 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. A sub-system is hierarchically a lower element in the general architecture structure (i.e. system) containing different elements. This definition helps to establish interdependencies between AI and other elements. A generative classification structure is given in [intentionally removed the name for reviewing purposes].

5 A responsible AI framework: contextualisation

Figure 2 provides a general level of detail for the benchmark framework for e-risk management. This framework extends the ISO RMP by incorporating several supporting tasks that secure an implementation process of ethical and regulatory considerations parallel to the classical ISO process. The contextualization process, which is the main component of the present work, corresponds to the components labelled **e-risk identification and classification**, **AI Scope definition**, and the **Analysis of values definition**. Each of these components corresponds to a pipeline process similar to that shown in Fig. 2.

In the figure, and the others in this work, the white boxes represent activities to be performed by the stakeholders involved in the AI RMP. 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 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.

For a better grasp of the link with the ISO standard, an analogical analysis per components of the presented framework is as follows:

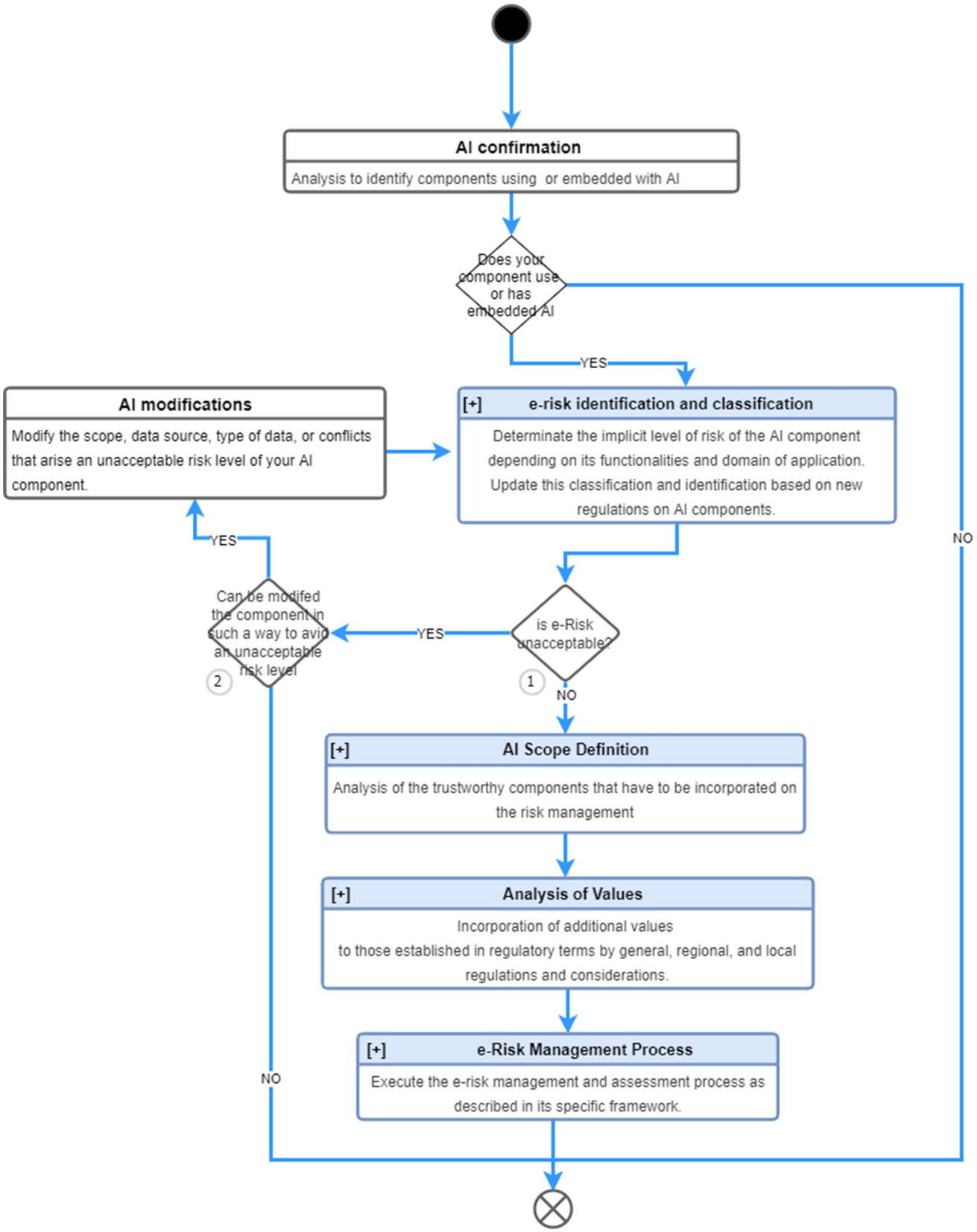


Fig. 2 Benchmark e-risk Management Process

- 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. This is presented here too.
- The box named EeRMP contains all the iso processes within it, except for 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 RMP.
- The box named EeRMP does contain the ISO-defined "Monitoring and Review" process but in order to improve the pipeline flow process, it was defined as a main 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 to be imposed; therefore, the reviewing process has partially been integrated within the framework itself

Following Fig. 2 (covered thoroughly in [intentionally removed the name for reviewing purposes] , the first process identifies or confirms that AI elements are considered within the system or subsystem under evaluation. If an AI algorithm is considered for evaluation or is embedded within the system, the **e-Risk Identification and Classification** process takes place (covered next). This process focuses on defining the AI elements' intrinsic level of risk under regulatory conditions. This classification is based on the AI act [44] and includes modification if new regulations are defined for the AI elements.

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 acts, and other regulations should be considered during the risk assessment processes.

After establishing an initial context regarding the requirements of the trustworthy guidelines, a secondary context regarding values to be integrated within the system, if any, is performed (named Analysis of Values Definition in Fig. 2). In case contradictory values exist, this process involves using decision-making processes (e.g., ANP or AHP tools) depending on the interdependencies between values components and criteria. For example, one of the criteria that the ANP and AHP process should consider is a social and legal compliance and, of course, the regulatory and ethical considerations of the AI domain of implementation (e.g. medical ethics).

After the context of the RMP 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.

The ISO 31000 framework established that the RMP should be dynamic and continual. The endpoint in the figure only helps visualize the process as a pipeline system. Nevertheless, the idea behind the benchmark ethical framework is to be periodically being used, and updated, for risk management.

As defined in the ISO 31000, processes should be dynamic and readily available for changes. This reflection will be relevant for recursive processes in which system and assets modifications can take place. 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 have hierarchically big enough that force other parts (i.e. subsystems) also to modify their connectivity or data usage.
- The data source or type are 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 e-risk identification and classification

Figure 3 shows the **e-Risk Identification and Classification** pipeline. The general approach focuses on evaluating 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, named **Level of Risk**, analyses the components under the Artificial Intelligence Act.

After performing this analysis, a question addressing if new regulations (or corporate reflection) 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, on the pipeline. These modifications (or incorporation of requirements) are over:

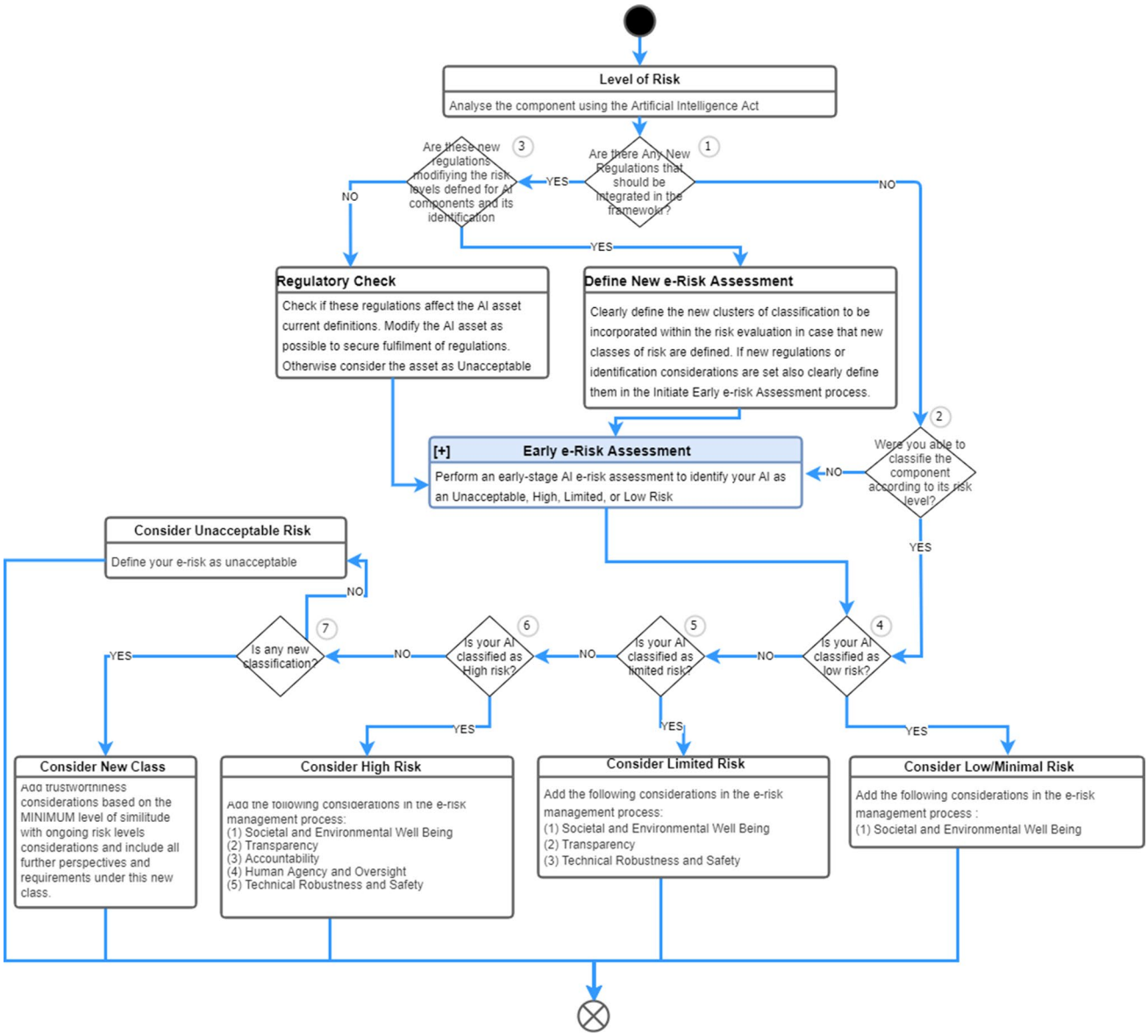


Fig. 3 e-Risk Identification and Classification pipeline

- 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, more 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 full process, named **Define New e-Risk Assessment** takes place.

In this process, a clarification based on well-established questions that settled the domain, functionalities, and approaches of this new cluster should be defined. To better understand this point, readers are encouraged to wait until the **early e-risk assessment** process is covered. The new cluster will imply the definition of requirements (or risk concepts) derived, or extended, from the trustworthiness

requirements [21]. 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 with 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 classifications levels were incorporated.

Independent of the case of modification, the **Early e-Risk Assessment** process will be initiated with the consideration 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**). The idea of 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 current regulatory conditions [22]). 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 the AI assets or their functionalities will be dependant 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 exemplification of companies policies were presented previously [Name has been removed for reviewing process]. Therefore, this consideration should be taken into account 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 [22]; 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 as established in the AI act [22], there are obligations on adequate risk assessment and mitigation systems. 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 take into account, 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. Further considerations could be included in this new class that should not contradict those established by local and global regulatory

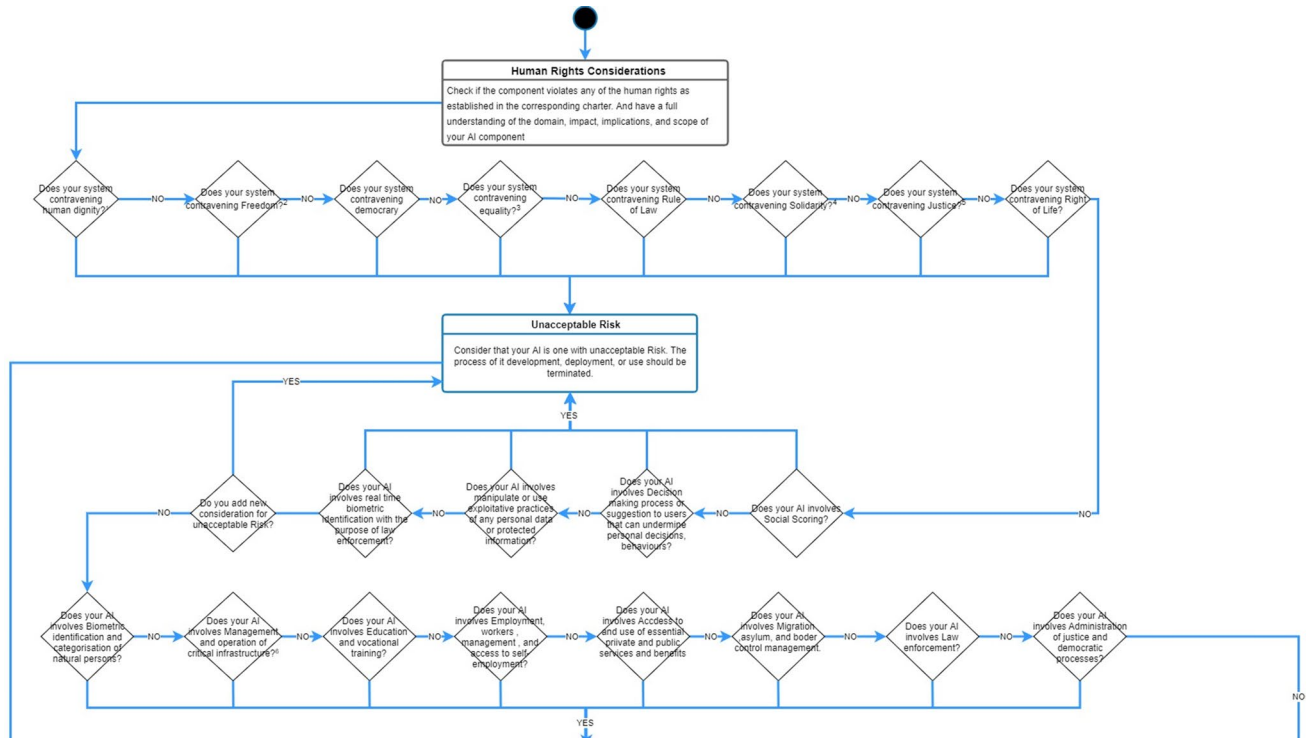


Fig. 4 Early e-risk Identification

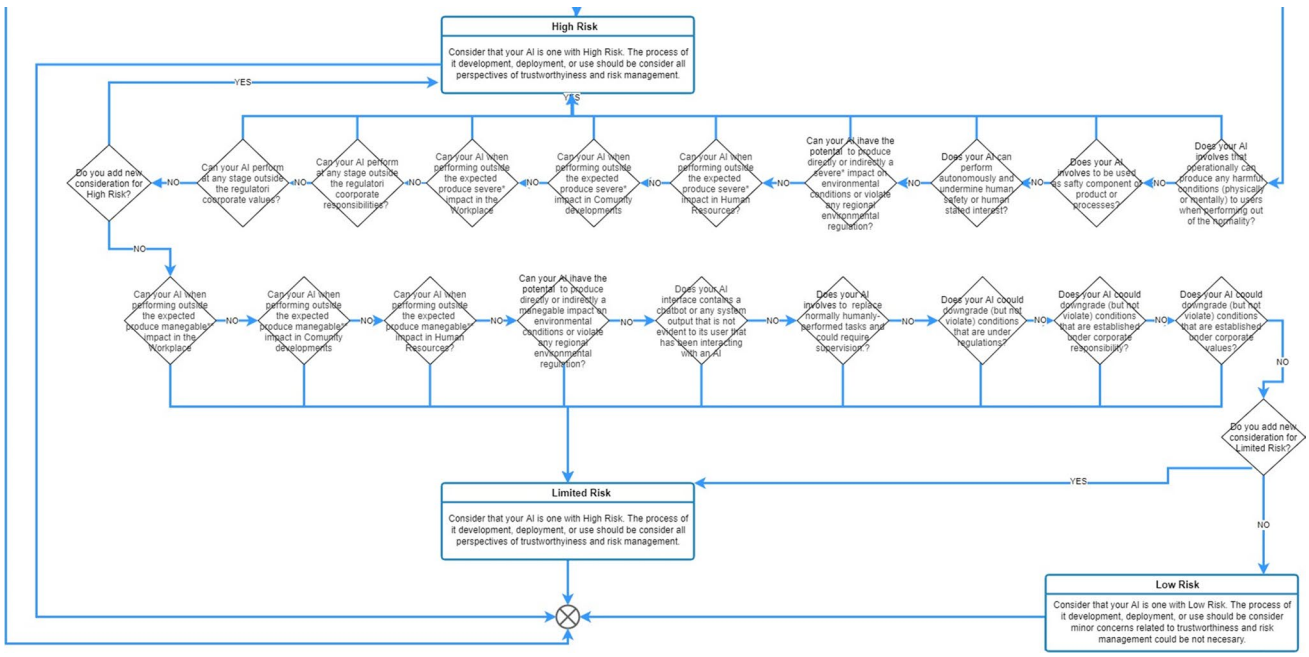


Fig. 5 Early e-risk Identification

conditions (e.g. Charter of the EU respect to 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 a lower one.

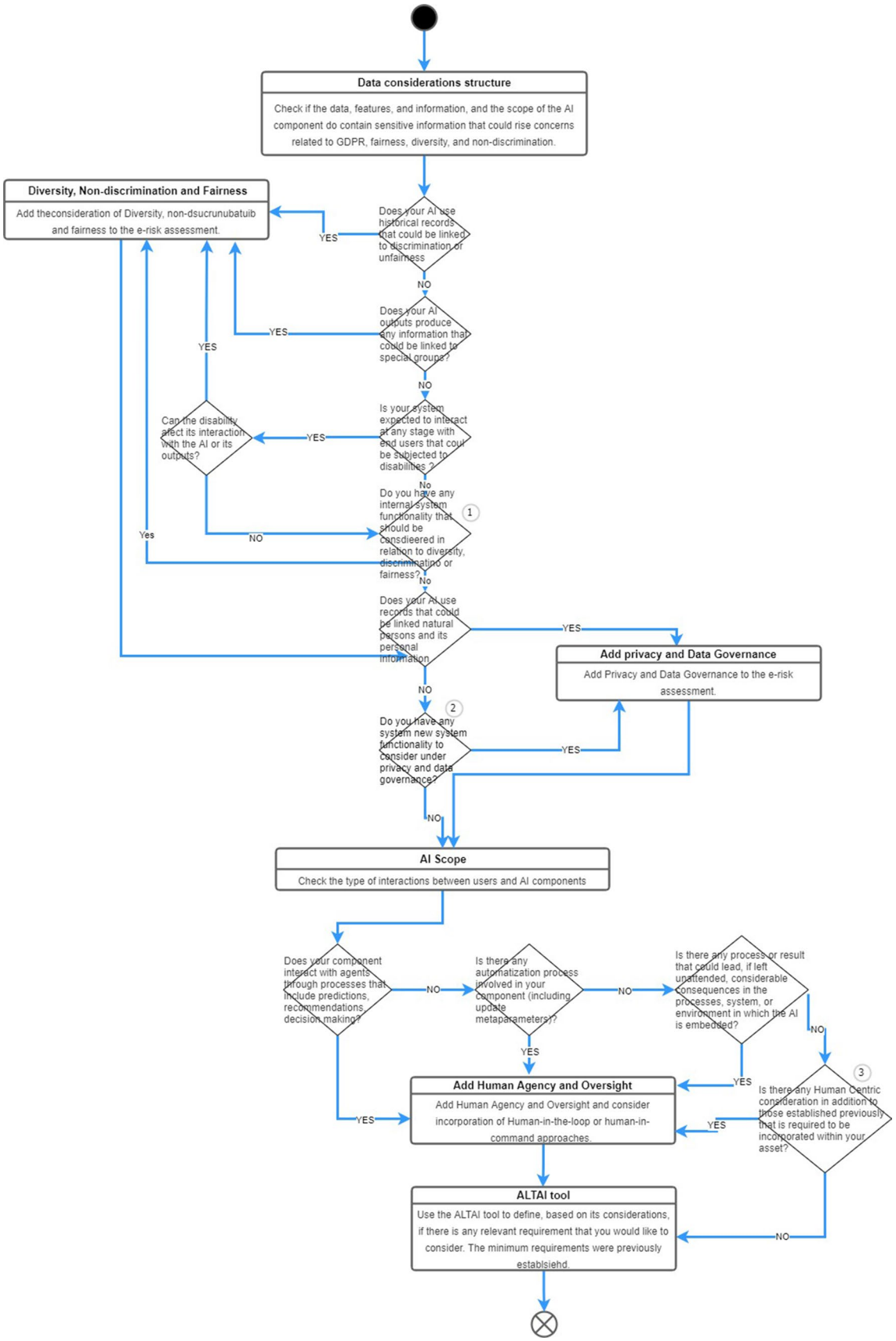


Fig. 6 AI Scope Definitions

5.1.1 Early e-risk assessment

Figures 4 and 5 shows the **Early e-risk identification** pipeline. This pipeline focuses on defining the intrinsic level of the AI element in case that is unknown by users, or the analyses has not been performed based on the AI act.

The pipeline is constructed so that the intrinsic risk level under evaluation, and therefore its identification, decreases (from higher to lower risk). This format should be taken into account when new regulations or classes (by the users) need to be incorporated into this framework. Therefore, the new classes identifications processes should be placed between intermediate risk level classes. Furthermore, if new identification processes 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 intrinsic risk class.

As observed in Fig. 4, the first pipeline part involves evaluating the AI assets to understand 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 [23]. These scrutiny process are assumed to be known by the frameworks' users, and therefore, Human Rights evaluation, depicted in the process named **Human Rights Considerations** are left as a checking process.

After performing this process, a set of eight questions extracted from the Artificial Intelligence Act [22] are used to define if the AI asset has an unacceptable intrinsic level of risk. These questions include, for example, understanding if the system is contravening human dignity, freedom, democracy, equality, the rule of law, solidarity, justice, and right of life. For a complete understanding of these concepts, readers are encouraged to review [23, 45, 46].

After the Human Rights scrutiny, the AI functionalities are evaluated. As observed in the figure, four questions (diamonds) are used to evaluate if the AI functionalities and application domain enforces the AI asset to be considered with an unacceptable intrinsic level of risk.

Immediately after these four questions, there is an additional one, *Do you add new consideration for unacceptable risk?*, that allows incorporating further regulatory conditions derived from stringent regulatory scrutiny given by the risk appetite of the companies, or the modification of regulatory conditions. The objective is to define further reflections that could make an AI have an unacceptable intrinsic risk level. As expected, several questions can be incorporated, allowing the dynamicity of the framework for current and future trends.

The following risk level scrutiny 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 pipeline evaluation focuses on the domains and functionalities of the AI. To do it, several questions derived from the artificial intelligence act [22] are used. Nevertheless, local regulations or future AI constraints should be considered for incorporation of the current pipeline process.

The questions involve issues such as social scoring, law enforcement, human resources, among others. Currently, a set of 15 questions are used in the AI high-risk level identification (the last 8 in Fig. 4 and the first 7 in Fig. 5).

Further questions can be included to scrutiny local, sustainable, or corporate definitions. In this regard, two additional boxes immediately after the last ones in Fig. 5 define corporate responsibilities and values to make them entirely restricted to ethical requirements. Finally, an additional question is added that helps to incorporate new considerations or regulations about High-Risk scrutiny. The process of incorporating should be similar to that discussed for unacceptable risk (concerning regulations, risk appetite, companies policies, and regulatory conditions). The objective is to define further considerations that could make an AI have a High-Risk intrinsic risk level.

The next level of risk is the Limited-Risk Evaluation. The focus of this risk level relates to the AI impact on the system and environment. This level should attend local and sustainable reflections or additional corporate definitions that can be downgraded (not violated) by the AI assets and their functionalities. Therefore, similar conditions, as those established for the High-Risk scrutiny cluster, could be placed here (with the distinction that these conditions are downgraded instead of violated). In total, nine considerations (see Fig. 5) are included in the current status of this framework.

Similarly to the previous level of risks, there is an additional question, *Do you add new consideration for Limited-Risk* , to extend the functionality of the current framework by allowing incorporating the recognition of AI with limited risk.

The reflection of downgrading establishes the need for defining metrics, which helps to determine the level of acceptable downgrading conditions. Although, as shown in Fig. 1, Metrics and KPIs are considered to be defined within the current framework proposition, they are not presented within the scope of the present work.

Finalising with the present pipeline, 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.

5.1.2 AI scope definition

Following with the contextualisation process, and as seen in Fig. 3, the upcoming process that should be performed if the AI assets do not possess an unacceptable intrinsic level of risk (screened out by diamond 1) is the **AI scope definition**. Figure 6 shows this processing pipeline. The main objective of this process is to extend the **e-Risk identification and classification** step by analysing to greater detail considerations based on information used by the AI component and the AI - agents interactions.

More specifically, this pipeline evaluates the possibility of biases and personal information usage, which impacts the requirement of Diversity, non-Discrimination and Fairness (DnDF). In terms of AI and agents interactions, it focuses on the level of automatisation left over the AI component that translates on requirements over the agency of humans on decision-making by the AI component. These agency components can be connected to the Human-Centric perspective

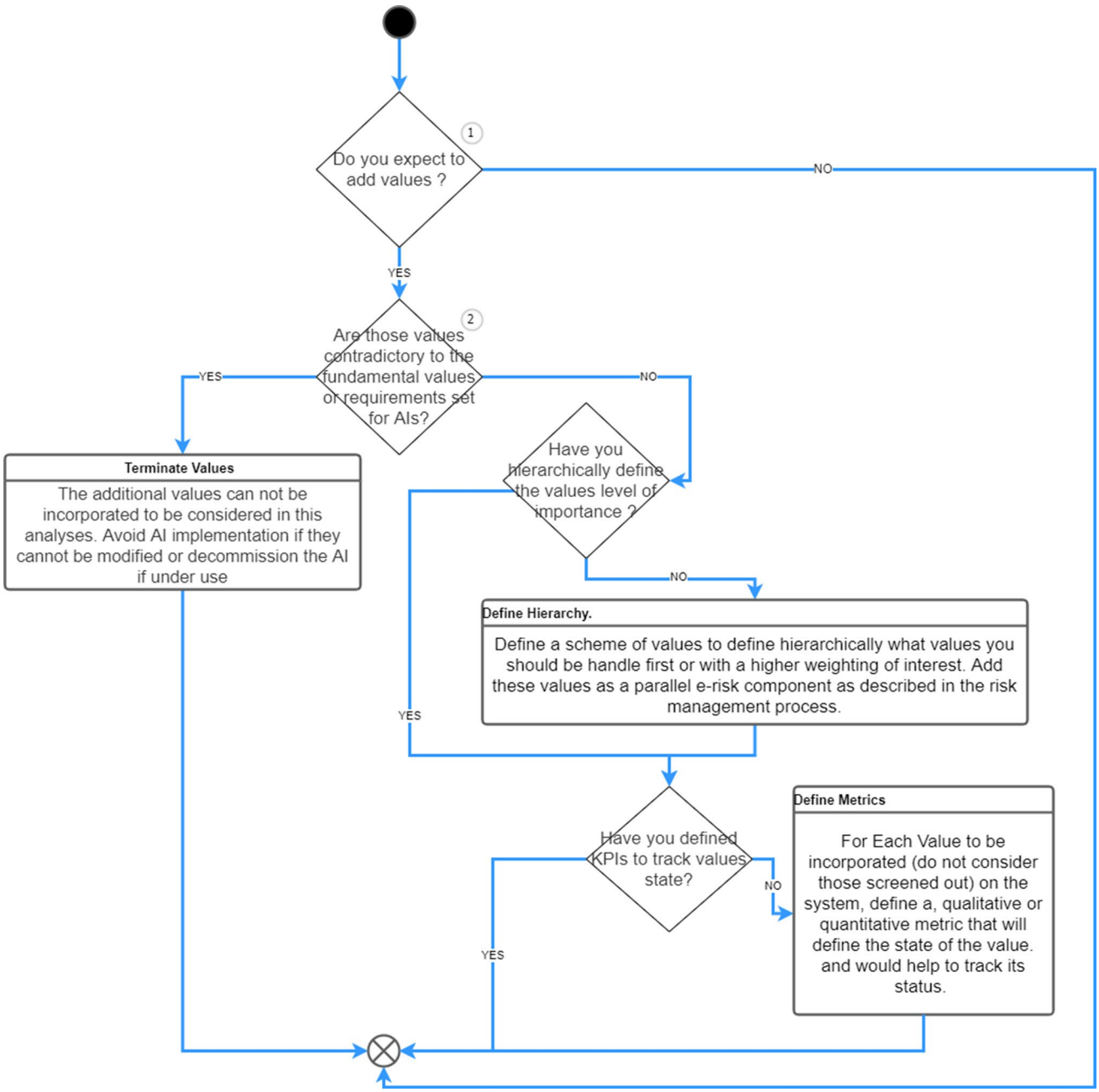


Fig. 7 Analysis of Values and Definitions

and, depending on the risks considerations of the AI application, linked to human-in-the-loop or human-on-the-loop approaches. Human-on-command has been left outside the current framework applications since it does not solve problems related to biases and efficiency, which could be considered necessary in the manufacturing sector.

As observed in the figure, the pipeline starts by performing the **Data consideration structure** process. This process focuses on analysing and developing a complete understanding of: (1) the type of data that the AI will be managing, (2) what type of data curation, transformation and features generations, if any, would be made by the system, and (3) which of the AI assets inputs and outputs could be related to biases, private information, fairness, diversity, and/or discrimination.

Additionally, it is important to consider under the GDPR regulatory framework that if the AI asset will manage personal information, how secure will be kept any information made and used by it.

After this analysis, a set of questions (5 in total) are used to analyse the MINIMAL conditions to establish the need to incorporate DnDF requirements within the RMP. These questions are linked to historical records, output information, disabilities, among other considerations. A specific question related to disabilities is used to check the impact of AI on the disability, allowing to establish if AI would impact disability or if the disability restricts the use of the AI assets. The last of these questions (Diamond 1) allows extending the analysis with greater detail and, therefore, allows the extension of the current frameworks as further requirements are established about DnDF topics by regulations or users considerations.

After the DnDF definitions, a process named **Add privacy and Data Governance** takes place in case that records and information can be linked to natural persons (directly o throughout the combination of data sets) or some of their personal information. Only one question is used for such a process; nevertheless, the framework allows, as observed in the figure (diamond 2), extensions to easily incorporate new regulations or definitions based on companies interests.

Immediately after the Data Privacy and Governance analysis, a step dedicated to Human Agency and Oversight (HAaO). In order to do that, a process, named **AI interactions** is performed. In this process, an analysis is performed over all possible AI assets - human interactions. These interactions can be direct, such as a user-UI interaction, or 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 some scrutiny of Human Agency and oversight. Therefore, they would require some specific definitions, depending on the AI behaviour, on the responsibilities that will lay over humans, the control the AI will have over human decisions, and to define until what point human-centric, HITL, and HOTL considerations will apply over the AI asset.

If there is more than one type of agent under the approach the AI fundamentally is based on, the analysis should be driven in a per-user base (e.g. patient and medic). 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. This last point implies that the analyses has to be linked throughout the whole scope of the system; this implies that if an AI feeds information to another algorithm, the new element interactions are also important to be known.

After performing the **AI interactions** process, three questions are used to determine the incorporation of HAaO as a trustworthiness considerations within the RMP. Importantly, AIs that were already classified with a high intrinsic level of risk was already enforced to consider HAaO, so this stage considers extensions and incorporation of further assessment (throughout diamond 3) not defined during the **e-risk identification and classification** process.

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 [26] supports the actionability the key requirements outlined by the Ethica guidelines for Trustworthy AI [21].

The ALTAI tool aims to provide a basic evaluation process for Trustworthy AI. First, it helps users to 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

Table 1 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)	Analysis, Treatment Options	Cost Effective, Time Effective
Operational (Divisional Management)	Implement, Quality Control, Program and products to reduce risk	Operational

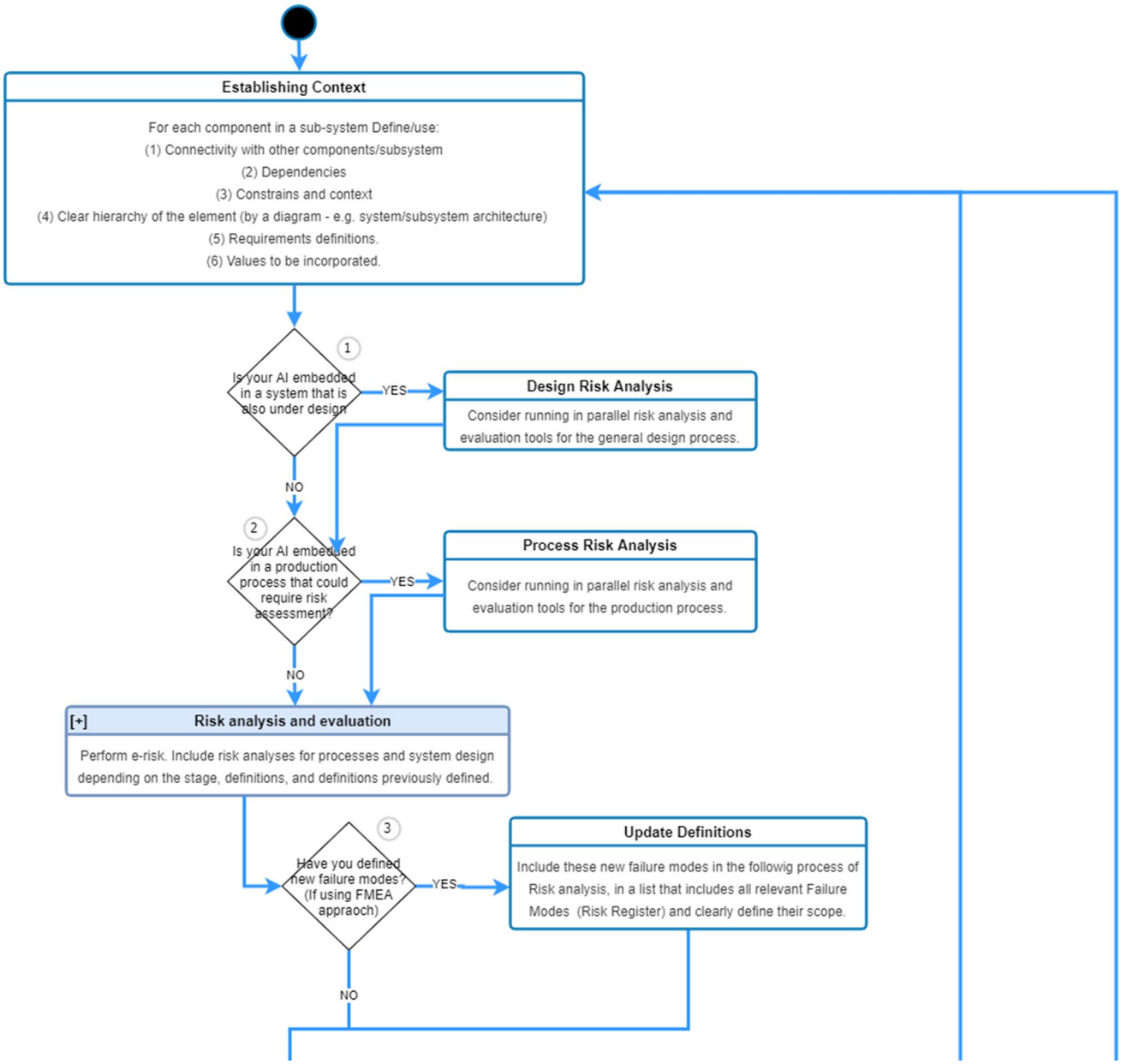


Fig. 8 Risk Management: Part 1

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 users of 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 Analysis of values

Figure 7 shows the pipeline to check if values could be incorporated within the AI RMF. To do that, a checking process that evaluates if any value would be incorporated is performed (diamond 1). If the answer is *NO*, the **Analysis of Value Definition** process is terminated. On the other hand, if the answer is *YES*, a question regarding the fundamental rights (diamond 2) is performed. In it, the value(s) to be incorporated are checked that they are not contradictory to the values established by general, regional, and local

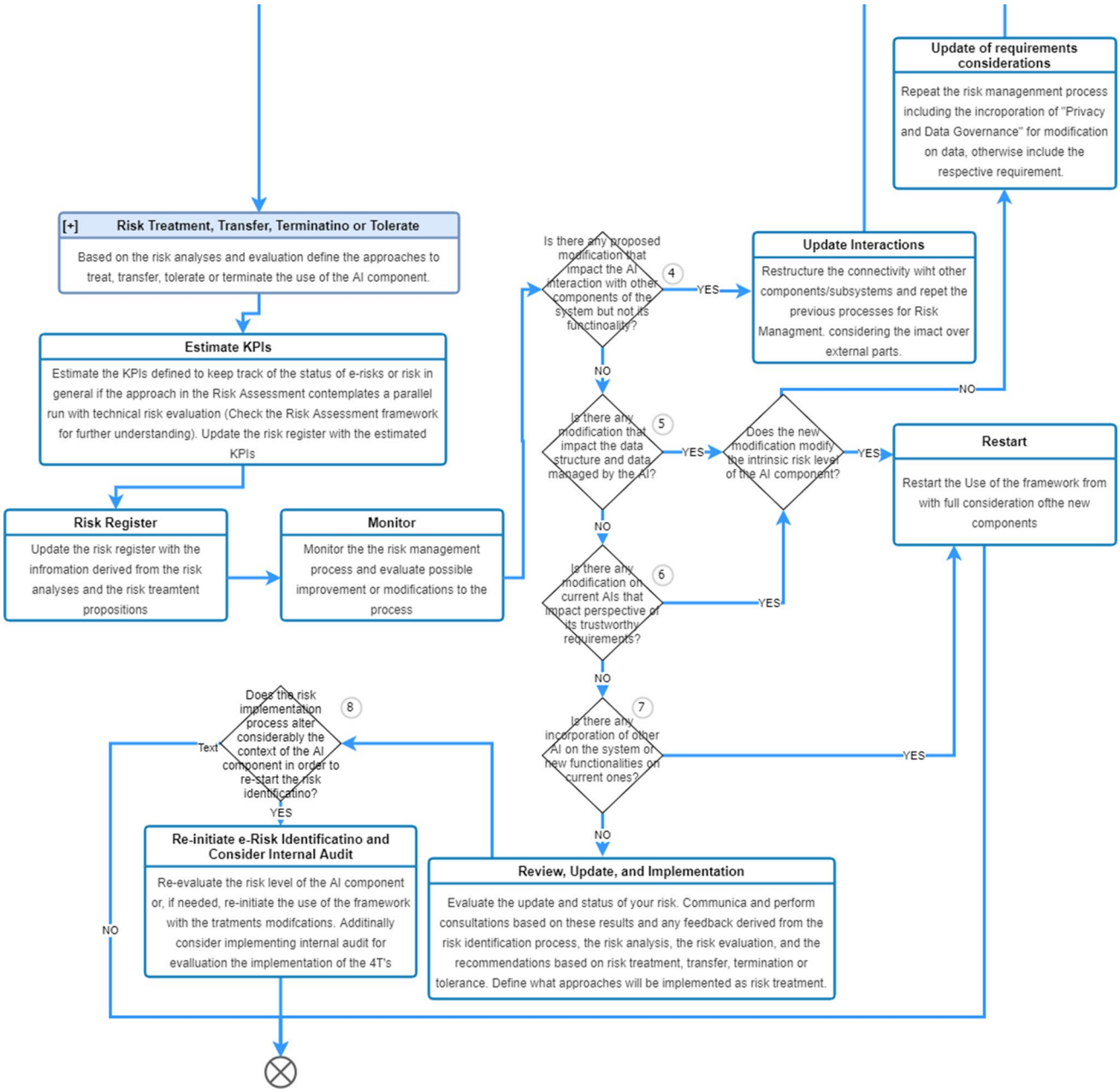


Fig. 9 Risk Management: Part 2

regulations (e.g. Human Dignity, Freedom, Democracy, Equality, Rule of Law, and Human Rights in Europe [23]).

Even though the AIs that contradicts general, regional, or local values should have been screened out at these stages, given that their nature is of unacceptable risk, an additional check is made to secure that users do not incorporate values that could be contradictory to them.

If the value(s) to be incorporated are not contradictory, a value hierarchy should be defined in order to address their incorporation relative importance, especially if the values to be incorporated are, between them, contradictory. In case that the hierarchy has not been defined, a process named **Define Hierarchy** is initiated. In this process, the hierarchy is recommended to be driven by weighting the desired values in the RMP.

Based on the current framework, the recommendation is to use a decision-making-based approach for the weighting process. We recommend the use of approaches such as ANP or AHP to define the hierarchy of the values. The definition of what method to use among them depends on the possible inter-correlations between criteria and values defined in the analysis. The criteria used for these analyses will depend

on users goals; therefore, based on the risk management architecture. An extract from previous work [intentionally removed the name for reviewing purposes] of the decision making considerations based on the recommended architecture are listed in the following Table 1.

After performing the weighting processes, a discretisation step to define the most relevant values to be included could be used. To do that, methods well-known such as Pareto 80/20 or others could be implemented to eliminate those values that would have a relatively low impact on the system's functionality or that are highly contradictory to the hierarchically relevant ones.

The final consideration of the current pipeline involves defining metrics to track the values desired to be incorporated into the system. To do that, specific questions are addressed if KPIs have already been set to measure the values state within the AI system. If yes, the **Analysis of Values process** is terminated; otherwise, a **Define Metrics** process takes place.

These metrics to be specified can be qualitative or quantitative. If several values will be considered incorporated in the framework, it is recommended that these metrics be normalised or managed to make the analyses of 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 values inverse effect.

The normalisation process can be made by considering "best scenario" and "worst scenarios", allowing fixing the cap for values metrics. Furthermore, by using the following equation normalisation into a quantitative form of each value-KPI is possible.

$$(V_a - V_w)/(V_b - V_w) \tag{1}$$

In equation 1, 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 evaluating at each state modification based on previous conditions (μ_p) or its best μ_b scenarios (see equations 2 and 3, respectively).

$$\mu_p = (V_{a,new} - V_{a,old})/V_{a,old} \tag{2}$$

$$\mu_b = (V_{a,new} - V_{a,old})/V_b \tag{3}$$

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.

5.2 e-Risk management process

As specified in the contribution section, the RMP is not thoroughly covered in this work. Nevertheless, as previously mentioned, some contextualization takes place that merges previous analysis. Furthermore, some definitions and approaches are covered for readers to have a wider understanding of how the ISO standards could be used to define an ethical-based RMP. Some of the techniques mentioned here are well known in the industry, allowing AI practitioners of the corresponding domain to understand how to merge or use these approaches with the current framework. Furthermore, for stakeholders that already run RMPes (e.g. dedicated to product, processes or design), the current framework can be used collaboratively, allowing full integration in the manufacturing sector. It is worth noting that this is part of further development that the authors are focusing on and will contribute to the definition of a generative process that could be implemented independently of the domain. Finally, comprehensive coverage of all these topics will be done in upcoming works.

Figures 8 and 9 show the e-Risk Management process Pipeline. The pipeline initiate with a small process of Establishing the general system Context (named **Establishing Context** - Figure 8). This process is performed before those steps commonly linked to the ISO 31000 risk assessment process (i.e., risk identification, Risk analysis, Risk evaluation, Risk Treatment, and monitoring and review). The **Establishing Context** process 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:

- *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. These connectivities are typically established in software development systems by establishing software architectures that define software elements, relations among them, and properties of both elements and relations.
- *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 greater attention.

- *Human-AI and Human-UI interactions*: This context allows us to extract and analyze the interactions that will happen between the AI elements and their UIs with humans. Furthermore, reflections of data types and the methods used to request and give information to users should be considered. Previous analyses performed in the **AI Interaction** process (in the **AI scope** section) analyses can be integrated here.
- *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 be connected with 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.
- *Diagrams*: Diagrams should be constructed based on the information collected in *Connectivity with other components and subsystems*, *Dependencies*, and *Human-AI and Human-UI interactions*. These diagrams can be used to track the impact of risks to materialize and check interdependencies. As mentioned in the figure, a clear hierarchy of the system elements should be defined. Combining these diagrams with the feedback of previously mentioned contextualization steps would improve system security, accountability, transparency, and robustness, among other requirements.
- *Requirements definitions*: Agglomerate all the previous requirements that would be important to include within the RMP (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.
- *Values*: The values to be incorporated in the framework after performing the *Analysis of Values* process.

The following two components (Diamond 1 and 2) are helpful to analyse if the RMP would be run in parallel or not, with other RMPes. Specifically, Diamond 1 focuses on the design process, while Diamond 2 involves parallel risk management dedicated to processes. Without further specification, the current framework focuses on using Failure Mode and Effect Analysis (FMEA) as the driver of risk assessment processes. A failure mode is how an item or operation could fail to meet its objectives. Therefore the FMEA process focuses on determining these failing conditions and the effect and impact on the system and processes. If the system or processes involve design or process, extensions to the FMEA are denominated DFMEA and PFMEA, respectively, are used.

The DFMEA is a systematic group of activities to recognise and evaluate potential systems, products, or processes failures and, therefore, involves a more comprehensive analysis of systems part at the early stage of systems developments. The PFMEA is a useful tool run by institutions to identify and evaluate the potential failures of processes, which involves the possibility of extension to processes currently used.

The focus of the present work is not on risk assessment, and therefore topics related to FMEA and how to use it for e-risk management is not covered here. Nevertheless, this information is valuable for setting approaches that could be foreseen for risk assessment implementation beforehand.

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 (made in **Establishing Context** process). Specifically, at this stage, a Failure Mode and Effects Analysis is proposed as the main component to have a high understanding of 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 focuses on the analysis over all the conditions linked to trustworthy requirements, ethical and values considerations. A thorough description of this process will not be covered here. This process involves another pipeline process that will be covered in future works.

Given the iterative nature of the RMP, 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 adequately documented.

After the **Update Definitions** process (or if there is no need to run it), the **Risk Treatment, Transfer, Termination or Tolerate** Process is performed (See Figure 9). In this step, depending on the risk appetite, the risk levels, the probabilities of events occurrence, and the chance of detection of the risks to be materialised, a part, component, or system should be:

- *treated*: modified, upgraded or include enough safeguards to reduce the risk level of the failing condition to happen
- *Transfer*: if the risk level allows it (given the risk appetite and the intrinsic risk level of the AI component), use external safeguards approaches such as insurances that allows transferring the responsibility of the events if materialized.

- *Terminate*: stop the development and use stage of the AI part, component, or system. Proceed with decommissioning if necessary
- *Tolerate*: 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

These four possibilities are well documented and known in RMPes and receive as a whole the name of 4Ts of risk management.

The following process corresponds to Estimate KPIs. KPIs linked to the RMP, each Risk component, and KPIs directly linked to the methodologies used for risk assessment should be estimated.

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. It includes diverse information that helps to keep track of the propositions made for risk management, KPIs and, among other, relevant information related to the methodological methods used for evaluating risks (e.g. FMEA/FMECA) and those descriptions specified in the **Update Definitions** process.

The following process named Monitor Involves the internal evaluation and comments of the RMP. This involves evaluating the correct application of the RMPes and, at the same time, generating feedbacks 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 (for further understanding of this process, please review [reference removed for the reviewing process].

After the Monitor processes, different questions (Diamonds 4, to 7) are used to evaluate modifications related to the AI Interaction with: (1) other components, (2) the data structure managed by the AI components, (3) the incorporation of other AI components or functionalities that were not foreseen to be implemented, or (4) to incorporate additional functionalities of current AI that can impact the trustworthiness of the system.

Depending on the responses, updates must be performed over the Establishing Context process and is required to rerun the RMP. All the previous possible modifications should be derived from the Risk Treatment components since the rest of the T's of the 4T's of risk management would not affect the system functionality architecture. 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 of requirements considerations**, **update interactions**, and **restart** processes) force to analyse with the new proposed modifications. These propositions should be kept track. Thus a binnacle should be kept for understanding the process behind the risk treatment modifications.

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 since it involves the interactions between the different bodies involved in the RMP. 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 review [intentionally removed the name for review purposes] .

Finally, The last question (Diamond 8) is used to check if the risk implementation processes modified the AI components functionalities or interactions considerably (as specified at the beginning of the current section). If so, a process of re-evaluation of the risk levels of the AI components (based on the early risk assessment process) should be defined.

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

Additionally, the audit process should be coordinated to evaluate the implementation, in due time, of the corresponding strategies implemented throughout the 4T's scrutiny. 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.

6 Conclusions

This document presents a framework for developing and designing AI components within the Manufacturing sector under the responsible 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 both any of the 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 the perspective of risk assessment. Future works will expand missing topics that will help to settle the approaches for risk management with the final goal of securing the development of AI components under the responsible AI perspective.

Ethical imperatives - which are also considered backbone structures for legal definitions - together with standards and

frameworks can drive the development of new AI assets for industry. The ethical imperatives covered in this work are related to Risk Protocols of the RASP approach (Risk Architecture, Strategy and Protocols) the main focus of this work. Further definitions related to the RASP was covered before [referenced not mentioned for reviewing process]. Upcoming works will cover risk assessment and risk evaluation processes and valuable metrics that could be used to manage AI components in terms of risk considerations.

Funding Open Access funding provided by the IReL Consortium.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Lauer, D.: You cannot have AI ethics without ethics. AI and Ethics **1**, 21–25 (2021)
2. Brosset, P., Patsko, S., Khadikir, A., Thieullent, A., Buvat, J., Khemka, Y., Jain, A.: "Scaling AI in manufacturing operations: A practitioners' perspective. Capgemini Research Institute"
3. Integra N.: "Artificial Intelligence: the driving force behind industry 4.0," Aug. (2020)
4. Xu, L.D., Xu, E.L., Li, L.: Industry 4.0 state of the art and future trends. Int J Prod Res **56**, 2941–2962 (2018)
5. Zheng, T., Ardolino, M., Bacchetti, A., Perona, M.: The applications of Industry 4.0 technologies in manufacturing context: a systematic literature review. Int J Prod Res **59**, 1922–1954 (2021)
6. Fujimaki, R.: "The 6 Challenges of Implementing AI in Manufacturing," (2020)
7. Accenture, "Big Success With Big Data - Executive Summary," p. 12 (2014)
8. Deloitte, "Industry 4.0 Challenges and solutions for the digital transformation and use of exponential technologies," tech. rep., Deloitte (2015)
9. Bedué, P., Fritzsche, A.: "Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption," J Enterp Inform Manag vol. ahead-of-print (2021)
10. Pieters, W.: Explanation and trust: what to tell the user in security and AI? Ethics Inform Technol **13**, 53–64 (2011)
11. Quinn, T.P., Senadeera, M., Jacobs, S., Coghlan, S., Le, V.: Trust and medical AI: the challenges we face and the expertise needed to overcome them. J Am Med Inform Assoc **28**(4), 890–894 (2021)
12. Devitt, S.K., Horne, R., Assaad, Z., Broad, E., Kurniawati, H., Cardier, B., Scott, A., Lazar, S., Gould, M., Adamson, C., Karl, C., Schrever, F., Keay, S., Tranter, K., Shellshear, E., Hunter, D., Brady, M., Putland, T.: "Trust and Safety," 2021. Publisher: arXiv Version Number: 1
13. Tonkiss, F., Passey, A.: Trust, confidence and voluntary organisations: between values and institutions. Sociology **33**, 257–274 (1999)
14. Bartneck, C., Lütge, C., Wagner, A., Welsh, S.: An Introduction to Ethics in Robotics and AI. SpringerBriefs in Ethics, Cham: Springer International Publishing, (2021)
15. T. I. f. E. A. . M. Learning, "The Institute for Ethical AI & Machine Learning."
16. Microsoft, "Home."
17. U. N. I. Global, "10 Principles for Ethical AI."
18. IEEE, "IEEE Global A/IS Ethics Initiative Newsletter."
19. IEEE, "IEEE SA - Standards Store | IEEE 7000-2021," (2021)
20. ISO, "ISO 31000 - Risk management," (2018)
21. H.-L. E. G. on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," Eur Comm (2019)
22. E. Commission, "Regulation of the European Parliament and of the Council; Laying Down Harmonised Rurles on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," (2021
23. E. Commission, "Charter of Fundamental Rights of the European Union," p. 17, (2012)
24. Bigham, T., Tua, A., Mews, T., Nair, S., Gallo, V., Fouche, M., Soral, S., Lee, M.:"AI and risk management," Centre for Regulatory Strategy EMEA, Deloitte, p. 32, (2018)
25. O. J. of the European Union, "on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," (2016)
26. I. C. of Data Analitics, "Home page - ALTAI," (2020)
27. Hagendorff, T.: The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines **30**, 99–120 (2020)
28. Eitel-Porter, R.: Beyond the promise: implementing ethical AI. AI and Ethics **1**, 73–80 (2021)
29. Knight, J.:Fundamentals of dependable computing for software engineers. Chapman & Hall/CRC innovations in software engineering and software development, Boca Raton: CRC Press, (2012). OCLC: ocn668197239
30. AI4EEU, "Tech Gaps | AI4EU."
31. Sheridan, T.B.: Human-Robot Interaction: Status and Challenges. Human Factors: The Journal of the Human Factors and Ergonomics Society **58**, 525–532 (2016)
32. van den Bosch, K., Schoonderwoerd, T., Blankendaal, R., Neerincx, M.: "Six Challenges for Human-AI Co-learning," in Adaptive Instructional Systems (Sottilare RA and Schwarz J eds.), **11597**, pp. 572–589, Cham: Springer International Publishing, (2019). Series Title: Lecture Notes in Computer Science
33. Wang, G.: "Humans in the Loop: The Design of Interactive AI Systems," (2019)
34. Dignum, V.: Responsible artificial intelligence. Place of publication not identified: Springer, (2020) OCLC: 1129396014
35. Vierhauser, M., Islam, M.N.A., Agrawal, A., Cleland-Huang, J., Mason, J.: "Hazard analysis for human-on-the-loop interactions in sUAS systems," in Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, (Athens Greece), pp. 8–19, ACM (2021)
36. Koulu, R.: "Human control over automation: EU policy and AI ethics," Eur J Legal Stud pp. 9–46, (2020)
37. Dignum, V.: Ethics in artificial intelligence: introduction to the special issue. Ethics and Information Technology **20**, 1–3 (2018)

38. Tubella, A.A., Theodorou, A., Dignum, V., Dignum, F.: "Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour," arXiv:1905.04994 [cs], (2019). arXiv: 1905.04994
39. Cloos, C.: "The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism," p. 8
40. Saidani, M., Yannou, B., Leroy, Y., Cluzel, F.: "Hybrid top-down and bottom-up framework to measure products' circularity performance," in International Conference on Engineering Design, ICED 17, (Vancouver, Canada) (2017)
41. Evans, O., Stuhlmueller, A., Goodman, N.D.: Learning the Preferences of Ignorant, Inconsistent Agents arXiv:1512.05832 [cs] (2015). arXiv: 1512.05832
42. Schwartz, S.H.: "An Overview of the Schwartz Theory of Basic Values," Online Readings in Psychology and Culture, **2** (2012)
43. Gabriel, I.: Artificial Intelligence, Values, and Alignment. Minds and Machines **30**, 411–437 (2020)
44. European Commission. Directorate General for Communications Networks, Content and Technology. and High Level Expert Group on Artificial Intelligence., Ethics guidelines for trustworthy AI. LU: Publications Office (2019)
45. Aizenberg, E., van den Hoven, J.: Designing for human rights in AI. Big Data & Society **7**, 205395172094956 (2020)
46. Gordon, J.S.: ed., Smart technologies and fundamental rights. No. volume 350 in Value inquiry book series, Leiden ; Boston: Brill-Rodopi (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.