CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Risk as a driver for AI framework development on manufacturing
Authors	Vyhmeister, Eduardo;Gonzalez-Castane, Gabriel;Östbergy, P.-O.
Publication date	2022-05-11
Original Citation	Vyhmeister, E., Gonzalez-Castane, G. and Östbergy, P.-O. (2023) 'Risk as a driver for AI framework development on manufacturing', AI and Ethics, 3(1), pp. 155–174. Available at: https://doi.org/10.1007/s43681-022-00159-3.
Type of publication	Article (peer-reviewed);journal-article
Link to publisher's version	10.1007/s43681-022-00159-3
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/. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-11-05 19:04:26
Item downloaded from	https://hdl.handle.net/10468/15436



University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

AI and Ethics (2023) 3:155–174
https://doi.org/10.1007/s43681-022-00159-3

ORIGINAL RESEARCH

Risk as a driver for AI framework development on manufacturing

Eduardo Vyhmeister[1] · **Gabriel Gonzalez-Castane**[1] · **P.-O. Östbergy**[2]

Received: 13 December 2021 / Accepted: 4 April 2022 / Published online: 11 May 2022
© The Author(s) 2022

Abstract
Incorporating ethics and values within the life cycle of an AI asset means to secure, under these perspectives, its development, deployment, use and decommission. These processes must be done safely, following current legislation, and incorporating the social needs towards having greater well-being over the agents and environment involved. Standards, frameworks and ethical imperatives—which are also considered a backbone structure for legal considerations—drive the development process of new AI assets for industry. However, given the lack of concrete standards and robust AI legislation, the gap between ethical principles and actionable approaches is still considerable. Different organisations have developed various methods based on multiple ethical principles to facilitate practitioners developing AI components worldwide. Nevertheless, these approaches can be driven by a self-claimed ethical shell or without a clear understanding of the impacts and risks involved in using their AI assets. The manufacturing sector has produced standards since 1990's to guarantee, among others, the correct use of mechanical machinery, workers security, and environmental impact. However, a revision is needed to blend these with the needs associated with AI's use. We propose using a vertical-domain framework for the manufacturing sector that will consider ethical perspectives, values, requirements, and well-known approaches related to risk management in the sector.

Keywords Responsible AI · Manufacturing · AI · Ethics · Framework · Risk Management

1 Introduction

The industry is becoming more automated in the Digital Era, with better sensors and captors, advanced planning systems, process controls, and supervisory control systems.

Although a direct adaptation of AI components could be made in the industrial sector, several challenges need to be overcome. A common denominator of these challenges is the need for trust. The concept of trust is key for technology providers allowing consumers to be confident in their use, independently of the market segment. Trust can be achieved as the combination of: (i) fulfilment of specific users interests - which can be translated into values, ethics, and technical requirements, (ii) a continual show of robustness over time under technical and social perspectives, and (iii) compliance with local and general regulations.

Risk is linked to trust. In fact, trust considerations are only relevant under conditions of risk and uncertainty [1]. By incorrectly placing trust, the results can lead to loss or harm [2]. To ensure trust, the actions of any agent—AI in our case—over other agents (including humans) must be reliable [3] or, in other words, with the lowest risk to produce adverse outcomes. This requirement further depicts the linkage between trust and risk and, therefore, to risk management - RM in short.

For industrial companies, the RM is an activity deployed on top of every production process. A key factor is incorporation of external and internal contexts to these, including human behaviour and cultural factors [4]. Incorporating external factors and internal factors linked to societal concerns in a RM activity can foster the incorporation of ethics and values.

To drive incorporation of reliable AI assets within the manufacturing sector, it can be foreseen the use of frameworks and standards based on risk management, promoting

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

Gabriel Gonzalez-Castane
gabriel.castane@insight-centre.org

P.-O. Östbergy
p-o.ostberg@biti.se

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

[2] Umeå University, Department of Computing Science / Biti Innovations, Umeå, Sweden

the incorporation of responsible-AI [5] considerations. This framework could allow upskilling of current technology trends and users understanding of AI. These skills could facilitate , among other benefits, a bias reduction, to improve monitoring, to improve performance, and to enhance systems robustness.

Even with the right skill-set, it is a challenge to select the standards and frameworks that suit a company's needs for risk management. In fact, different standards, frameworks, and structures can be used to handle risks.

The backbone framework used for AI ethical considerations in Europe is the Ethic Guidelines for Trustworthy AI [6]. 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 (i.e. life cycle) of AI elements (under the considerations of responsible AI [5]), 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, including the Ethic Guidelines for Trustworthy AI—e.g. ISO/IEC JTC 1 and the High-Level Expert Group on Artificial Intelligence.

The lack of applied standards, definitions, and regulations to AI and ethics create gaps that could lead to self-claimed implementations to have roots in the responsible AI considerations but in which these concepts are not adequately addressed. These implementations, or AI assets, could produce further misunderstanding within the business domains since they mislead how ethical concepts and moral values from the field's regulatory frameworks are integrated into the AI software components.

To facilitate the process of conceptualisation, planning, developing, deploying, tracking performance, use, decommissioning (i.e. its life-cycle), and continually improving the AI component on different systems, we propose the use of a vertical-domain framework for the manufacturing sector that will consider: ethical perspectives, values, requirements (as established by the European Union Ethic Guidelines for Trustworthy AI), and well-known approaches related to risk management.

More specifically, this work aims to set foundations for AI development and management, allowing stakeholders from the manufacturing sector to implement and assess AI assets. Furthermore, the proposed framework allows the use of management processes highly known in the manufacturing sector, facilitating its incorporation in that and possibly other domains. Thus, the management of AI assets can be provisioned by a proper definition and link between requirements and risk components. Therefore, by a proper definition and translation of responsible AI considerations into risk management, the trustworthiness requirements within AI assets can be secured. The following section gives a broader perspective of the scope by defining our work's current and future contributions.

2 Contributions

Figure 1 shows a schematic of the Responsible AI Framework contributions that our work seeks to achieve. The framework provides a methodology for RM of AI assets, within its life-cycle, with current and future (i.e. adaptable) trustworthiness trends. The framework is build as a combination of methods, frameworks and strategies including the ISO31000 [4], the trustworthy guidelines as ethical requirements [6], the white paper on artificial intelligence, the classification of AI elements based on the Artificial Intelligence Act [7], the charter of fundamental rights [8], Deloitte approach for management [9], the General Data Protection Regulation (GDPR) [10] and different techniques that support the framework use.

As observed in the figure, our contribution has two main components. One is dedicated to the Responsible AI Framework development (top box), and another is dedicated to framework validation (bottom box).

The first component is subdivided into four subcomponents. Among these, the one named *Risk as a driver for framework development* is the main focus of the present work. In it, the focus is on: (1) a constructive explanation and model definition between the links of ethical imperatives, trust (and its trustworthiness considerations), and the risk considerations within AI components in terms of hazards; (2) general review and cover of the ethics and

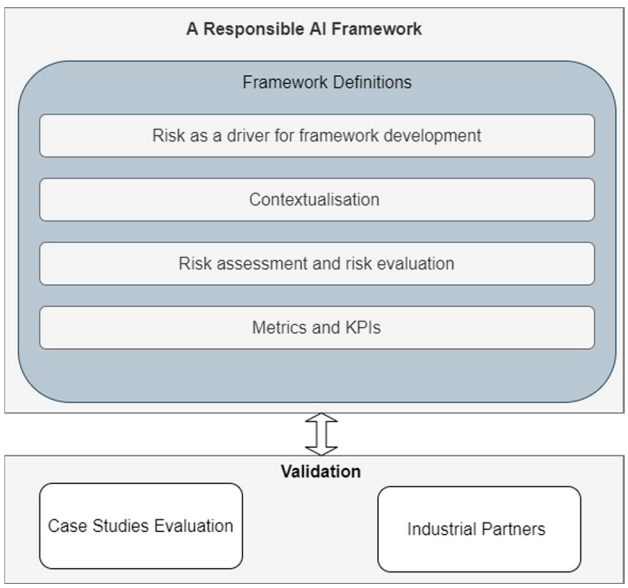


Fig. 1 Contribution overall structure and its strategy

risk considerations within the manufacturing sector; and (3) a discussion of standards and approaches for AI risk driven by known management processes.

Upcoming works will cover the contextualisation, risk assessment and risk evaluation processes, and useful metrics and KPIs that could be used to manage AI components in terms of ethical driven risk considerations. The approach proposed in this paper targets to:

- 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 key to distinguish between requirements—could be needed by law to acquire commercial certifications—and values—that are societal 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 on legislation to be applied by different countries on the use of AI.
- Flexibly adapt the framework to other approaches used to handle risks by industrial stakeholders. To enhance the adoption by companies that already have their own Risk Management Process (RMP), 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 a spiral/ iterative development processes - e.g. agile techniques. The framework should incorporate the benefits of these development cycles to ease developers their 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, this tool must be used by managerial levels to have broader understanding on the incorporation of ethical aspects into development in parallel to the existing process.
- Construct an architecture to support a better understanding on responsibilities and channels of communication between technical and non-technical stakeholders. For example, legal departments of many companies do not have technical knowledge to satisfy the legislations on some aspects of the AI life cycle. In the same manner, technical users - developers, architects—do not have the knowledge on 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 advance but also for companies to save revenues in future developments and to incorporate new processes into the existing ones. In addition, a well structured risk identification avoids to repeat failing conditions to similar AI components.
- Ease the ethical-based risk evaluation using a pipeline-based approach. Having flowcharts to model the framework ease its understanding and implementation.

The latest component of the schematic, validation, is performed by testing the framework in specific use cases. The case studies lie within an ongoing funded project ([intentionally removed the name for reviewing purposes]) that focuses on developing breakthrough solutions for the manufacturing industry, using artificial intelligence to optimise production systems.

Our contributions involves the full implementation of the RMP. Therefore, specifications and evaluations of each component (e.g. risk architecture, risk strategies, and protocols) will be described, tested, and implemented. For further information about the project and the manufacturing sector case studies, readers are invited to review [intentionally removed the name for reviewing purposes].

As mentioned above, our approach currently focuses on weak AI components. Weak AI corresponds to AIs that only perform a specific task. On the other hand, strong AI can perform several functions with the ability to learn by itself to solve new tasks; in other words, it understands and has other cognitive states [11]. Currently, there are no strong AI components, nor in the foreseeable future; therefore, its impact on the current proposition should be minimal.

The current framework would not generate any specific tool or software to handle trustworthy components within the AI RMP; instead, it would use the collaboration of current applications (e.g. ALTAI tool [12]). The framework does not intend to avoid or overlap regional regulatory conditions; they should be enforced to be taken into consideration as part of the pipeline process and considered before the framework implementation. Finally, technical components, such as transparency, are not intended to be 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.

3 Ethics, trust, and risk

This section presents a model for the link between ethics, trust, and risk, which can be considered a pillar for using the RMP for responsible AI considerations in a specific domain. First, some definitions are provided, followed by explanations on how ethical imperatives were used as a sustaining component for defining trustworthy requirements. Then, the latest is mapped to risk considerations, which should drive the discussion for RM as a solution for ethical 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 [13]. Several definitions exist for risk [4, 13–15]; nevertheless, they agree in considering uncertain phenomena that will impact the system goals. Notably, an event's outcomes can be positive or negative if materialised. Hence, a classification can be used by the nature of the outcome and occurrence.

Risks are classified as Opportunities 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 (i.e. related to budgets, timeframes, and management, where the outcomes can lead to different results).

Finally, risks are classified as hazards if the only outcome 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 materialisation of the risk condition. Given the nature and potential of materialising risk related to AI components, the present framework considers these risks as hazards. Nevertheless, we will interchangeably use the references hazards and risk here and after.

The European ethical principles for AI are based on ethical imperatives presented by the AI4People group in 2018. These imperatives define the approaches on which AI components should rely on [16]. 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 of people's goals and wishes, (4) Justice, that state that AI should act in a just and unbiased way, and (5) Explicability, that sates explanation on how an AI system arrives at a conclusion or result.

The ethical imperatives can be linked to seven requirements for AI systems that should be met to deem them trustworthy. These requirements, presented by the High-Level Expert Group on AI in 2019 [17], followed a reviewing process that allowed stakeholders (including those from the industrial sector) to revise and provide feedback to develop an initial integrative framework for AI management and development.

Importantly, linkage between ethical imperatives and trustworthiness is partially defined in the document [17] —e.g. "A crucial component of achieving Trustworthy AI is technical robustness, which is closely linked to the principle of prevention of harm." [17]).

These requirements set an initial framework for developing AI assets. Nevertheless, they do not define specific strategies for implementation nor secure that unexpected behaviours can materialise. As expected, depending on the impact or consequences of the possible unexpected behaviours, the severity of the measures integrated into the system should be more stringent. In other words, the AI assets **inherent risks**, which corresponds to a combination of an event probability and its consequence, should be minimised, controlled and tracked.

As specified by the European Commission, AI entails potential Risks. These risks must be addressed by regulatory frameworks that should concentrate on how to minimise the various potential harms, in particular, the most significant ones [18, 19].

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).

This list is driven by objectives set by the European Commission regarding a regulatory framework on Artificial Intelligence. The first objective corresponds to "ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union Values" [19]; Thus, a linkage with (1) and (2) of the previous list. The second objective is to "ensure legal certainty to facilitate investment and innovation in AI" [19]; which can be linked to (3) of the previous list. The third objective is to "enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems". [19]; this objective can be linked to (1) and (3) of previous list. Finally, the fourth objective is to "facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation" [19]; which is directly related to liability (3) and system consistency (2). Other risks could be intrinsic to the system, including technical and non-technical, and the values imposed by non-regulatory stakeholders (e.g. companies values). Therefore the fourth source of risk was included in our analysis, which can be further analyzed (and split in a more specific source of trust. But for simplicity,

this risk was not further extended nor directly connected to trust considerations since it involves system dependencies.

The EU has issued liability guidelines [20] that establish some applications that could be warranted strict liability. Liability does distinguish cases in which the existing uncertainty that could come from a person's wrongful conduct can dictate no blaming of any persons [21].In the case of AI this could be translated on the undermining effect of AI over humans' actions. These considerations should be included on the perspectives of AI human interactions and, thus, on the human agency and oversight.

The European Commission has defined in its Artificial Intelligence Act [19] 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 fines. Artificial Intelligence components can be categorised as unacceptable, high, limited, and minimal risk [19]. This categorisation is based on the AI functionality (i.e. what the AI component does) and the implementation domain. Even though this first approximation to cluster AI based on their risk is helpful, further classification, given future trends, could be given.

It is plausible that the European Commission has settled risk levels for AI components. Nevertheless, a framework that deals with risks in AI component should follow both general and specific RM approaches. General, in a sense that should be able to be adaptable with the consideration of different domains. Specific, in a sense that the AI managing should be differentiated depending on the possible adverse outcomes, application domain, and its functionalities.

Even when ethics is essential in developing AI assets, since it allows the inclusion of a general management approach and covers a component consideration to achieve trust, the ethical imperatives do not encompass all considerations involved in developing technically sound components. Responsible AI does not only imply the consideration of ethical use and development 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. Foremost among these are: agents' roles, societal issues (inclusion, diversity, and universal access), prediction, reflection, analyses (positively and negatively), and given their discrepancy, the analysis and consideration of legal concerns. Responsible AI is a pillar that should be considered incorporated at every stage

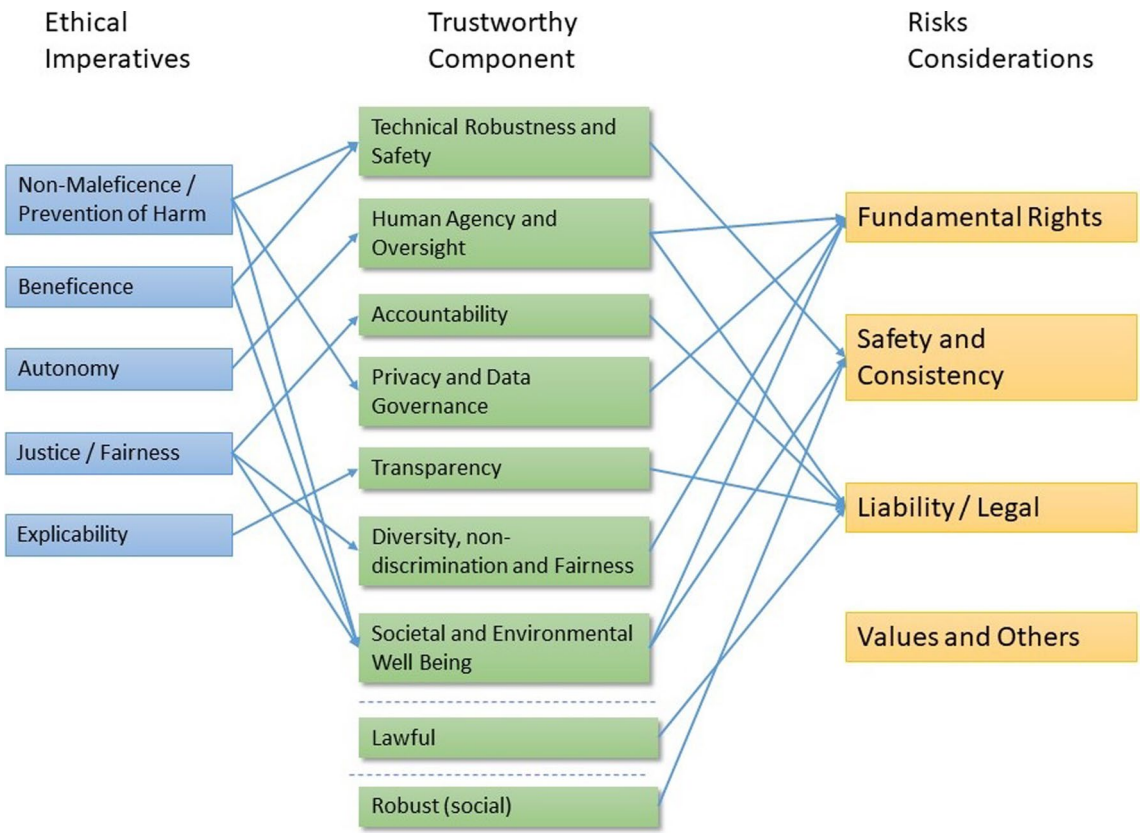


Fig. 2 Ethics, Trust, and Risk Pillars Link

in the production of AI assets, and therefore, any framework used for AI assets life cycle.

It is clear up to this point that the consideration of risk, and risk management, is fundamental for the AI assets life cycle. Figure 2 describes the link between trust, ethics, and risk from a broader perspective. In the figure, the ethics and trust links are based on the trustworthy requirements definitions of the European Commission, while the trust and risk column are connected based on the impact that the materialisation of events could have over the components of the central column. It can be argued that several connections are possible (e.g. non-maleficence and prevention of harm, ethical imperative should be connected to all trust components); nevertheless, these are defined as main contributors to support the ethical imperatives or trust components with an intrinsic perspective of the manufacturing sector. Further explanation of Fig. 2 will be given in Sect. 7.

As represented in the figure (middle column), trust can be achieved by securing that the AI component fulfils lawful, ethical, and robust components throughout its entire life cycle. First, it should be ethical, ensuring adherence to the ethical imperatives and values that govern local and regional social behaviours and regulations. Second, it should be lawful, securing compliance with local and global laws and regulations. Finally, it should be robust in a technical perspective (embedded within the trustworthy requirements) and a social perspective. At the same time, these three components are linked to ethical risks for AI technologies. Even though there could be many ethical concerns, they can be encapsulated within the critical risks associated with (1) fundamental rights, (2) safety and consistency risks, or (3) legal and liability risks. If they cannot be encapsulated within these critical risks, they could be considered within those named as values and others. Those encapsulated as values and others could include systemic risks, functional risks, and reputational risk, among different business and value-derived risks [3]. The fundamental rights encompass all the base rights established for human agents locally (e.g. EU Charter of Fundamental Rights [8]). Furthermore, the legal risks should not only encompass restrictions and regulatory frameworks for AI systems; they should also encompass a situation where a system becomes too successful and performs as an anti-competitive environment [3]. Safety and consistency risks are associated with those that can harm agents and the environment, causing catastrophic loss.

4 Ethical risks

After describing the link between ethics and risks, the present section focuses on defining an ethical risk helpful definition for the current framework.

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

Finally, even though ethics imperatives define the need for beneficence on AI assets, they do not define the negative impact of a e-risk to materialise over AI assets objectives; therefore, hazard consideration is reinforced.

5 Frameworks, guidelines, and other approaches for responsible AI implementation

Before presenting the responsible AI framework, this section covers some general approaches that can help to contextualise the present framework proposition. Furthermore, it allows readers to have a wider perspective over the considerations that could be transferred into a RMP of AI ethical-based assets.

The 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 incorporation of flaws embedded by their creators [22]. This antecedent means that humans have a concrete view and responsibility of designing a software component to develop, test, and use until 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. Moreover, these decisions and predictions are guided and influenced by personal biases and values developers and data inscribe into the applications.

Therefore, AI management regulations, frameworks and guidelines should contain technical and non-technical considerations. These considerations 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), human resources definitions, and/or social, ethical and values considerations that will come from stakeholders involved in the AI life cycle.

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 generalisation approach for the AI life cycle. An extensive list of guidelines and strategies based on critical AI issues can be seen in [23]. Nevertheless, considerations of the latest approaches to set, for example, integration of values within the development stages [24], should also be included in these analyses. As presented in [23], no approach covers all the issues related to responsible AI considerations. The most mentioned include privacy protection, fairness, non-discrimination, justice, accountability, transparency, openness, safety, and cybersecurity. Even though organisations show different interests in what principles should focus on, the most relevant concepts seen by organisations and companies includes privacy, fairness, accountability, transparency, explainability, and safety [23, 25]. 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 with the end goal of generating suitable frameworks.

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 [26], Microsoft's Responsible AI guidelines [27], UNI Global Union [28], the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [29] together with its newest release the IEEE Standard Model Process for Addressing Ethical Concerns During System Design [24].

Controversy has been generated related to specificity and industry influence in frameworks and guidelines [23, 30]. Even though some could agree with these statements, industrial values should be considered within the life stages of AI component. Furthermore, the tools specification, domain-specific impact and requirements will be better understood as use cases are performed with the current approaches. 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. Finally, 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).

It is essential to highlight that ethical AI frameworks and AI implementation guidelines should consider the entire environment [30] in which these components are developed and deployed. 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. Risks identification can be performed by understanding the environment where AI will be developed and deployed. Furthermore, given the dynamic nature of the systems and current regulatory frameworks in which AI could be deployed, monitoring the application throughout its life cycle could be mandatory. This implies that frameworks should ideally include or facilitate KPIs and metrics that would help track the impact of implementing different approaches or the status of a system under a continual risk evaluation, with the end goal of improving performance and trustworthiness. Finally, as new legislative endeavours emerge or ethical and values concerns rise from users and developers, it might be essential to update the application and frameworks.

Therefore, the establishment of ethical AI frameworks in the industry sector requires considering the domain complexity; otherwise, they will fail in the continuously changing environment [30]. Therefore the industry requires generalisable frameworks with enough specificity and aided tools to prevent struggling during its implementation. Additionally, approaches similar to those already run by the sector regarding RM would facilitate the actionability of ethical risks.

Independent on how well-developed guidelines and frameworks are constructed and used, failures will raise by unintentionally negative consequences (i.e., when AI are developed and deployed without sufficiently robust governance and compliance [23]) 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 poor or lack of company management, regulatory incentives, manufacturing practices, employee training, and quality assurance, in addition to the e-risks (e.g., lack of understanding, biased information, improper combination and managing of data, and misuse of algorithms).

6 Responsible AI considerations in the manufacturing sector

Since the current framework focuses on the manufacturing sector, some considerations regarding AI implementation in this domain are covered in this section.

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. The incorporation of AI in manufacturing expects to value U$ 16.7 billion by 2026 [31]. This trend is driven by an increasing number of large and complex data sets, the revolution of interconnectivity

and sensing (provided in Industry 4.0 and IoT), and improving automation and computational power.

The manufacturing sector is highly competitive and in constant need of automation and efficiency. To foster the achievement of such objectives, the sector incorporates AI technology with different objectives. This incorporation, in the manufacturing, is characterised 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 [32]. Therefore, incorporating AI in the manufacturing sector should consider both the technical challenges and the risks involved within responsible AI considerations. In fact, most of the critical factors involved in AI implementation are not driven purely by technological considerations [30]. Instead, companies' fundamental lack of responsible AI considerations (including poor requirements, governance, and processes) leads to adverse outcomes.

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 a risk to materialise in the manufacturing sector can be equally dangerous for the users, workers, and companies. The trends show that around 85% of the AI projects could deliver erroneous outcomes due to bias in data, algorithms, or poor managing by the teams involved in their development and implementation [25].

A significant requirement to incorporate AI in the factories is to keep processes robust and low risk (optimally free). Thus, AI techniques that enable the scenarios exploration without interrupting the production, future actions, and complex trade-offs in real-time are highly beneficial for the production and processes [33].

It can be foreseen that the relevance of incorporating responsible AI would be seen during failing conditions and their different outcomes with different severities (including production halt, safety, loss of revenue, demands and brand name damage). Independent of the results, acting reactively over failing systems does not prevent future problems since they tend to be systemic and could involve cascading or multi-system failures [30]. Therefore, applying methodological approaches that are proactive in reducing failing conditions and evaluating or analysing cascade considerations should be the primary driving condition in the AI assets life cycle.

By themselves, the failing considerations put a general framework in which analysis and implementation of ethical considerations could be driven. The combination of methodologies and frameworks previously established 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 does require (1) an accurate definition of the problem, (2) identification of potential root causes, (3) objectively evaluate, quantitatively or qualitatively, 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, prioritising those with higher impact and likelihood.

Furthermore, responsible AI considerations within the manufacturing sector should be focused on two different approaches depending on the nature of the goods produced. First, the goods could or could not have included AI elements interacting 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 the generalised guidelines and frameworks should be possible. As expected, this would impact, among others, the considerations of users' accountability, agents privacy and data governance, and possibly system transparency (depending on the user, the domain impact, the data types, and the level of detail on the system output). Of course, depending on the intrinsic risk level of the AI component, trustworthy components can be included or excluded leading to a broad range of risks to be handled.

All the previous considerations (ethical considerations, trust, risk, root causes, among others) could be linked together by a suitable management process of the risk involved in responsible AI; in other words, e-risk management. To facilitate such a task, the following sections present, in general terms, definitions and approaches that could be used to describe ethical risk and its management process.

7 Risk management as a source of trust and its link to risk considerations

As previously described, ethical risks correlate to trust through ethical, lawful, and robustness considerations. This definition implies that a suitable RMP that contemplates each of these components will improve the AI assets perception from their users and, therefore, an improvement in trust. Even though legal components could be handled based on a RM framework, legal constraints should not allow uncertainties in their range of AI applications. Therefore, their considerations should not be perceived with soft boundaries that allow relative violations.

In our proposition for the e-RMP, legal issues should be considered in the early stages of AI development. Nevertheless, the incorporation of an early e-risk

management should be fostered independently. In general, RMP , are fostered to be executed before major resources are committed into any project [34]. This statement is also transferable to e-RMP, since the definitions of requirements could impact functionalities or avoid pitfalls during its life-cycle.

Current and future regulatory considerations of AI elements could be tested initially following a pipeline evaluation structure that secures that the minimal constraints are fulfilled before starting the development stage of any AI element.

Further constraints could be over imposed over the technical components depending on technical and non-technical requirements, as long as they are not contradictory with regulatory considerations (including those imposed by ethics and values - e.g. fundamental rights).

Next, a discussion based on Fig. 2 and its link as a source of trust for RM point of view is performed.

Robustness: Robustness considerations, as seen in Fig. 2, follow two categories. one related to technical robustness and another related to social robustness. The first, part of the trustworthiness requirements [6], will be covered together with the other ethical-based requirements. The 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)* [35]. However, the definition of present and future points of view consideration place a heavier burden on social robustness since, as explained by Beumer et al. [35], change of strategies tend 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 [36]. 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 (e.g. 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 proposed e-risk management framework could possess two characteristics that allow the fulfilment of present and future points of view.

First, the framework incorporates ethical requirements and values within the e-RMP. Therefore, a comprehensive framework modification can easily incorporate present and future perspectives or requirements derived from social considerations, ethical concerns, or regulatory conditions. 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 be incorporated as long as they do not contradict the local and global legal and ethical requirements established for the AI asset.

Ethics and values often involve managing tradeoffs that cannot be satisfied simultaneously. For example, some fatality rates are acceptable within most risk works environments with the beneficence of higher-paying loads (i.e. a tradeoff between beneficence and non-maleficence). Therefore, processes to define an optimal set of values derived from several stakeholders' perspectives could be used. Well-known industrial decision-making tools could be integrated for defining these sets from the manufacturing sector.

Second, the framework developed can be based on well-known domain RM strategies. Therefore incorporation of them in the manufacturing sector can be considered acceptable since it would not impose a considerable change of stakeholders methodologies.

Accountability: In terms of AI, the accountability place and distribute responsibilities between AI developers, deployers and users. It is reasonable to take into account that if a clear uncertainty and responsibilities definitions are provided (or clarified) to the AI users, the AI developers could be considered to act in an accountable manner. In the end, these links (risk of failing accountability given lack of definitions) established that the risks verification process and its management, derived from the understanding of the system uncertainties, would increase the assets trust.

Accountability can be seen as a mandatory requirement for AI elements with high inherent risk; a well developed and structured definition of responsibilities would also facilitate defining legal responsibilities. This consideration can be essential and more complex as AI assets are Incorporated within the business products since it involves incorporating further stakeholders. For example, those assets classified as limited risk or minimal risk [7] are not explicitly required to provide accountability specifications. Nevertheless, they should cover any legal requirements, including those defined for their accountability.

Human agency and oversight: Oversight over the AI elements does help in the accountability process and improves 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 [37]. One of the significant concerns about human oversight is that people's growing dependence on algorithms. This consideration is crucial in fields in which the outcomes of potential risk are considerable (e.g. healthcare). A RMP could help directly 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.

Like accountability, human agency and oversight are relevant to AI assets defined as High-Risk [7]. In this case, the consideration should be equally crucial for products embedded with AI and AI processes since the over relay on algorithms and methods could occur in both cases. A RMP would also allow improving human-AI oversight since instead of considering it as an agent that assumes responsibility for the AI outcomes, it will establish strategies for users (e.g. error process protocols) on the how and when to interact with AI assets.

Technical robustness and safety - TR&S: TR&S are crucial for ensuring that fallback plans exist if something does not go as intended with the AI assets. As mentioned in [6] TR&S secure the minimisation of harm. As previously mentioned, RM does involve the managing of hazards. Thus a consistent implementation of risk assessment, which includes the definitions of treat and terminates with risk conditions, would secure a constant improvement of the AI assets regarding their TR&S.

Since TR&S are crucial elements in the manufacturing domain, their relevancy should encompass both high risk and limited risk AI assets. Those with minimal risk would not require robustness consideration. This proposition would not contradict [7] since it establishes only transparency requirements for limited risk AI assets.

As mentioned in [38] safety and reliability can be achieved if challenges of data reliability, human-machine interactions, security, transparency, and explainability are addressed.

Privacy and data governance: Regulatory requirements regarding data managing and handling have already been set (e.g. [6]). 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 RMP can secure better security over data access and incorporate fallback planes in case that 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 the AI.

Privacy and data protection would be crucial for AI assets embedded on products that handle user information and link them to different databases. Even though there is no specific requirement on privacy and data governance for limited and low-risk AI assets, these considerations are imposed indirectly by other legal requirements (e.g. GDPR regulations). Similar to the previous cases, well-performing management of the risk involved in managing data will improve the security of the data managed, if relevant, in the manufacturing sector.

Transparency: Indicates the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions. Transparency dependends on the type of human-AI interaction (i.e. the user type) [39]. Thus, transparency is a precondition to determine responsibilities and to hold the responsible people accountable. Therefore, a e-RMP would allow to increase and secure trust. As could be observed in literature, there is a boom in technical explainability methods for AI; therefore, the future from a technical perspective of this requirement looks promising.

Based on Diversity, non-Discrimination and Fairness (DnDF): Bias is defined as the risk of a systematic error or deviation from the truth [40]. 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 [6], but it is essential to recognise those linked explicitly to the marginalisation 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 the biases.

Since it is well documented that biases can exist in complex historical data, AI-based risk scoring systems could perpetrate such biases. In addition, 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% [41]. This consideration has been foreseen by the European Commission and has defined some High-Risk or prohibitive AI applications that include scoring systems.

Even though a e-RMP would not eliminate the biases, the own nature would allow setting mechanisms to prevent the second and third numerals previously mentioned. The e-RMP will be considerably helpful for securing standards for High-Risk AI elements.

The AI act defines transparency considerations imposed over AI assets with higher risks than low/minimal risk.

Societal and environmental well-being: As mentioned by [41], 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 scale of implementation. Therefore evaluating and managing the risk related to 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 on 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, there is no clear identification of the process recommended to manage and ensure sustainable protocols in that regard. This implies that the manufacturing sector should consider the societal and environmental well being at every level of risk consideration. Including these perspectives within their RMP can produce an overall impact on the manufacturing, including brand recognition.

8 Standards and approaches for risk management

The Australian Standards Body developed the first standard by 1995 [42]. That work, was followed by other countries later, until this was withdrawn by the International Organization for Standardization, a worldwide federation of national standard bodies, to use the ISO 31000 [13]. Since then, several RM standards have been released depending on the updating needs and implementation domain. The latest version of the ISO 31000 provides principles, definitions, framework and processes, encompassing: RM - Guidelines [4], IEC31010: 2019 Risk Management - Risk Assessment Techniques[43], and the ISO GUIDE 73:2009 Risk Management - Vocabulary [44]. The ISO is currently developing standards with a specific focus on AI. These standards are encapsulated within ISO/IEC JTC 1 family, which is developing 31 standards with ten published; being only one (ISO/IEC TR 24028:2020 [45]) focused on trustworthiness. One of the standards under development – ISO/IEC DIS 23894 – focuses specifically on the RM of AI. Regretfully it is not clear, at this stage, if such standards will include ethical considerations and values within the RMP.

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 has expanded on their applicability to improve performances by enhancing internal systems control, risk management, system governance, and fraud deterrance [46]. Additionally, the British Standards BS 31100:2011 "Risk management - code of practice". The latest version of these standards explains how to develop, implement, and maintain risk management processes.

Even though there are different frameworks, guidelines, and approaches to manage responsible AI considerations, the focus of the proposed work is based on: 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 supporting the use of the framework. Given the importance of the ISO standard as a base for risk management, a description is made next.

8.1 ISO31000

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 RM approaches applied to different types of risks (i.e. hazards, control risks, and opportunity / speculative risks).

The general RM framework is based on the principles, a framework, and a process. The principles guide effective and efficient RM characteristics, communicating its values and explaining its intention and purpose. As described, the principles should be (1) integrative, (2) structured and comprehensive, (3) customized, (4) inclusive, (5) dynamic, (6) Use best available information, (7) consider human and cultural factors, and (8) perform under continual improvement.

It should be integrative to other activities and, therefore, for the case of AI, should be integrative to AI life cycle processes in the manufacturing sector.

It should be structured and comprehensive, implying that it should contribute to consistent and comparable results and, therefore, should include metrics that will allow measuring its integration in the processes or systems. Furthermore, it should be customized to the proportional level of risk, securing that the costs involved in the RMP are level to the possible consequences. Thankfully, the

European Commission has already settled the first layer of risk levels [19] and therefore, as described in the framework, the effort and considerations for the RMP 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 integrations and, therefore, there is still considerable space for generating historical information that could allow improvement on the RMP.

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.

The ISO 31000 framework helps assist in integrating RM into organizations activities and functionalities. The leadership and commitment focus on ensuring that RM is integrated into organization activities. The integration framework specifies that RM relies on the organizational structure and, thus, should be dynamic.

In general, the framework constructed under the ISO31000 should consider the internal and external context. External does refer to social, cultural, political, legal, financial, technological, national, regional, and environmental factors. Internal does refer to vision, mission, values, strategy, policies, organizations culture, standards, capabilities, data, relationships with stakeholders, contractual relationships, and interdependencies, including communication, commitment, roles, resources, and accountabilities.

Furthermore, it states that both the risk and the framework should be measured to check the suitability of the approach used. This 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 RM framework performance against its purpose, implementation plans, and expected behaviour. These KPIs are also involved in the improvement of 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 RMP within an organization and, therefore, suitable for different processes involved in the life cycle of an AI. Importantly 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.

9 E-Risk management

In this section, the Ethical Risk Management Framework is presented. First, a description of the general constituents of the ethical RM framework is performed. Then, individual components linked to the e-risk management framework are presented.

9.1 General description

Systems and organisations can face different risks that could impact their operations with a wide range of outcomes. These events can inhibit what is desired to be achieved, enhance that aim, or incorporate uncertainty into their outcomes. RM can be seen as the process involved in identifying, assessing and controlling such risks. The RMPes are well established, but they could be presented differently with different terminologies depending on the domain and the guidelines and frameworks used on their implementation (such as different standards - covered in the following section).

The RMP cannot be a separate system component and require the intervention, communication, and definition of the different areas involved in the systems under consideration. Independent of this, a RM framework 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/organisation, 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 [47] (see Fig. 3).

RASP provides details of the RM framework, which helps to define the RM context. The essential component to " glue all together " is the RM policy statement (which

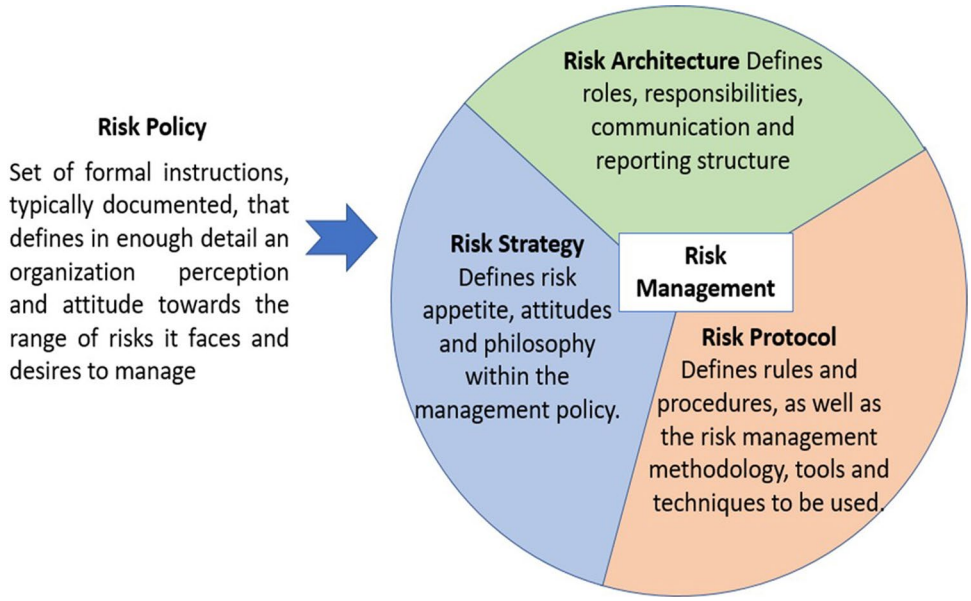


Fig. 3 Risk Management structure

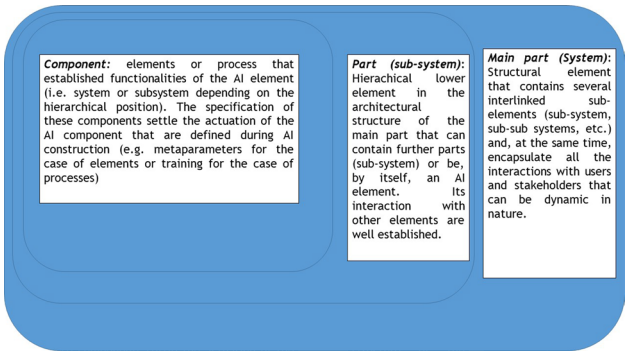


Fig. 4 Hierarchy for Incorporating Risk Management

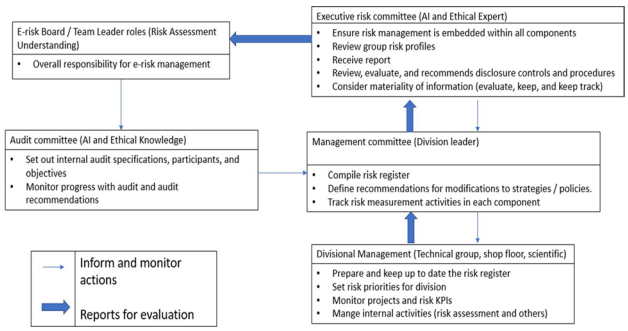


Fig. 5 Ethical Risk Architecture

sets out the organisation's overall strategy towards risk management). The risk policy contemplates the set of formal instructions (well documented) that defines in enough detail an organisation perception and attitude towards the range of risks it faces and desires to manage. Since this policy, and in general, the RMP, is related to a specific environment, the definition of the policies should be sound to the objectives established by the organisation and should not give generalities. Nevertheless, since the consideration is based on fulfilling at least requirements established by different regulatory bodies, some commonalities could be derived.

AI methods can be embedded within processes or be a stand-alone system used for the map, prediction, forecast, optimize, and recommendations, among other tasks. Therefore, a general definition as a system should be used in the framework to describe hierarchy of AI assets (can be only one). In the present framework, 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 general classification structure is given in Fig. 4, which shows the existence of the system, subsystem, and components, with a specification of each of these elements.

9.2 Risk management architecture

The RM architecture defines the committee structure and its: roles and responsibilities, internal reporting requirements, external reporting controls, and RM assurance arrangements. In this subsection, an exemplification that could be modified to implement e-risk processes, in function of industrial complexity and its policy, is given. This architecture is based on the implementation scenario previously described (i.e. [intentionally removed the name for reviewing purposes]).

Figure 5 shows this base architecture, which also describes the risk assessment process, the internal reporting channel, and critical roles and responsibilities.

The first role (bottom right of the figure) is the Divisional Management (DM). A division is a part of the system (process, business, or organisation) that perform 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, optimisation, etc.). The divisional management is incorporated by different stakeholders that perform the internal activities related to risk management. These could include performing the risk assessment process, preparing and keeping up the documentation, setting priorities for the division in terms of the focus on risk analysis and treatment, and keeping updated KPIs related to the AI elements.

In case that the AI e-Risk management is performed in parallel with other RMPes, this simplistic architecture allows integration with other structures. The divisional management reports the finding on the risk assessment to the Management Committee. Finally, 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 could integrate the reports. 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 full 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 from the DM and AC that can alter or modify the RMP or the actions implemented so far e-risks.

The third role is the Executive e-Risk Committee (ERC - lead by the risk manager). The e-risk manager is responsible for the corporate learning that has to take place to understand the benefit of e-risk management. In addition, as the person responsible for the RASP, the e-risk manager will be responsible for developing the strategy, systems and procedures by which the required e-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 on 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 e-RMPes and recommend actions to reduce the e-risk levels or to terminate or tolerate the AI current conditions under the perspectives of the e-RMP (based on the company e-risk appetite and regulations of AI). Further responsibilities include ensuring that RM 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 to secure that a variety of experts with different expertise are integrated into the process of risk assessment. Additionally, the ERC keep track of the whole process and all the documentation generated on the RMP. Therefore, it is necessary to maintain a range of e-risk management records, including details of various RM activities, including administration, risk response and improvement plans, event reports and recommendations, and risk performance and certification reports. These documents, following regular RMPes, could consider the following documentation and risk architecture responsibility:

- Risk management documentation manual and administration records and responsibilities—E-risk board
- Risk response and improvement plans—ERC
- Event reports, incidents, investigations and recommendations—MC
- Risk performance and monitoring reports—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 RMP 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. The RM procedures will set out risk assessment processes, risk control objectives, risk resourcing arrangements, reaction planning requirements and risk assurance systems.

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 main component of the RM 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 organisations that play

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

an essential role in setting values and ethical considerations within the company. There are no general specifications for the team leader 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.)

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 co-operation will be achieved in practice. The internal audit should be driven by experts that will cover the requirements of AI and, at the same time, those requirements established by responsible AI definitions.

Table 1 Decision-Making Considerations shows a general description of the decision making criteria involved on the definitions involved at each level. The table describes an organizational recommendation level for decision-making based on the Architecture. The first column capitalizes on the components that will lead the strategic, tactical, and operational decision-making. The second column defines the goals of the decision-making approaches (that should not violate those defined by the previous hierarchical components). The third column defines the criteria for the decision making. The strategic, higher-level, should secure that decision making processes follow the policies objectives, 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), and is flexible to modifications depending on current tendency and consistency. This table has been constructed on the based on proposed risk architecture and following descriptions defined in literature [47, 48]. The risk architecture and the description given in this table for decision-making considerations should be treated, as recommended on the mentioned literature, as a case-base scenario. We are defining this strategy for implementation within our case-study scenario.

Importantly, the previously described architecture is intended to define the approaches for managing, treating, transferring, terminating, or tolerating risks within the AI components and is not intended for real-time decision-making processes. The definitions established over the risk components, such as defining technical safeguards, will be determined based on the AI components' risk analyses.

9.3 Risk management strategy

The RM strategy should define the RM philosophy, the arrangements for embedding RM, 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.

One crucial characteristic used for e-risk management is that the RM philosophy. This states how risk is considered in the organizations. For the case of AI we recommend (base on case study—[intentionally removed the name for reviewing purposes]) the considerations of RM approach based on four key principles.

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

The risk appetite is a crucial set of statements that define the proactivity towards risk. Some of the examples that could be useful for ethical RMPes are provided as follows (based on case study - [intentionally removed the name for reviewing purposes]):

- Regulatory based approach: developments should be committed to delivering value to our AI elements by securing the use of risk strategies established by the local and general regulations (e.g. European Commission). It will be obey the spirit and the letter of the laws and regulations that apply regionally.
- Operational Challenge: We recognize the complexities involved in the integration of 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, mandatory industry changes that are not correctly implemented. We will always seek to remain current and adhere to regulations uncles prevented by our infrastructure.
- Third-Party Risk: We are willing to consider working in parallel with our partners for the conceptualization and integration of e-risks on their current RMP.
- Value Risks: We are not willing into accepting the incorporation of values that are contradictory to regulations and ethical requirements established by the local and regional bodies.
- 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.

9.4 Risk management protocol

Figure 6 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 an implementation process of ethical and regulatory considerations in parallel with the classical ISO process (as shown in Fig. 6). The analogical components of the presented framework with the ISO process is as follows:

- All the boxes with the exception of 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 does also contain 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 within it, exception of the communication and consultation. This is done since the combination of the architecture and policies should impose over the RMP the frequency and channels of communication.
- The box named EeRMP does contain the ISO-defined "Monitoring and Review" process but 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.

The Figure, and the general structure of the benchmark process, is based on a UML recursive pipeline approach. In this figure 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 colored boxes correspond to a whole process that should be described by another UML benchmark process (future works). Finally, The Black dots correspond to an initial point of the pipeline, while the circle with a cross in it represent the end point and termination.

9.4.1 Benchmark e-risk management process

Following Fig. 6, the first process identifies or confirms that AI elements are considered within the system, subsystem, or component under evaluation. 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, 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

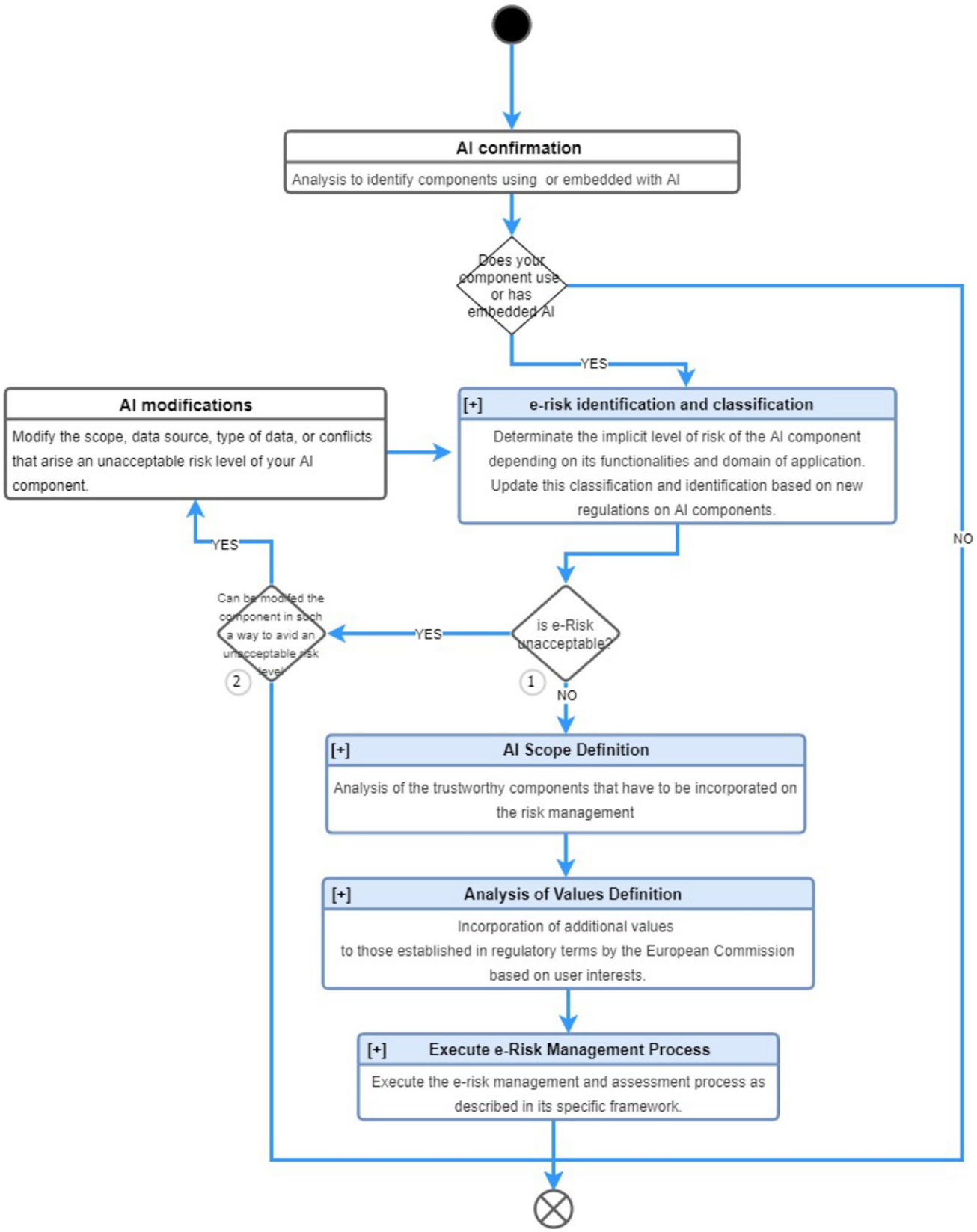


Fig. 6 Benchmark e-risk Management Process

of all other techniques into cyber-physical systems). This definition implies that data is used for both learning and act 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 focuses on defining the AI elements' intrinsic level of risk under regulatory conditions. This classification is based on the AI act 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 on 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 should 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 RMP is terminated since the AI element cannot be developed, implemented or used. If the AI asset is already under use, the considerations for decommissioning should be made.

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 trustworthy guidelines requirements, a secondary context regarding values to be integrated within the system, if any, is performed. This process 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 depending on the interdependencies between values components and criteria. Finally, these tools should allow the evaluation from several stakeholders' perspectives. Therefore, they can be used to homogenise the perspectives, generating the most suitable combinations of values and the hierarchy they should be integrated into the systems.

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 is shown in the figure only helps to visualise the process as a pipeline system. Nevertheless, the idea behind the benchmark e-risk framework is to be periodically used for e-risk management. Recursive processes are included within the previously described steps, and they should be reinitiated as impactful modifications are made over the system, the companies policies, or the regulations. We specify these impactful modifications as at least one of the following causes:

- 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. automatisation 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 its parts

10 Conclusions

In the present work, an effort has been made to settle the foundations for establishing an AI framework management rooted in responsible AI considerations. Here, a link between ethical considerations, trustworthiness, and risk is done to impose the validity of handling responsible AI requirements and ethical based considerations as a RMP. This RMP should be driven within AI development, deployment, use, and decommissioning stages. This work intends to give the first view of a broad methodology under development that will introduce a well defined and structured approach. Future works will help define processes (i.e. risk assessment) and metrics to perform an e-RMP.

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, visithttp://creativecommons.org/licenses/by/4.0/.

References

1. Lockey, S., Gillespie, N., Holm, D., Someh, I. A.: "A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions," (2021)
2. Mcknight, D.H., Carter, M., Thatcher, J.B., Clay, P.F.: Trust in a specific technology: an investigation of its components and measures. ACM Trans. Manag. Inform. Syst. **2**, 1–25 (2011)
3. Bartneck, C., Lütge, C., Wagner, A., Welsh, S.: *An Introduction to Ethics in Robotics and AI*. SpringerBriefs in Ethics, Cham: Springer International Publishing, (2021)
4. ISO, "ISO 31000—Risk management," 2018
5. Dignum, V.: *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Artificial Intelligence: Foundations, Theory, and Algorithms, Cham: Springer International Publishing, (2019)
6. H.-L. E. G. on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," *European Commission*, 2019
7. Commission, E.: "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)
8. Commission, E.: "Charter of Fundamental Rights of the European Union," p. 17, (2012)
9. 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)
10. 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)
11. Flowers, J. C.: "Strong and Weak AI: Deweyan Considerations," *Papers of the 2019 Towards Conscious AI Systems Symposium (TOCAIS 2019), Stanford, CA, March 25-27, 2019.*, p. 7, (2019)
12. I. C. of Data Analitics, "Home page - ALTAI," 2020
13. Hopkin, P.: *Fundamentals of risk management: understanding, evaluating and implementing effective risk management*. New York: Kogan Page, 5th edition ed., (2018)
14. A. Limited, ed., *Managing successful projects with PRINCE2*. London Norwich: TSO, 6th edition ed., (2017)
15. ITIL, "ITIL Risk Management | The Risk Matrix & Management Process | ITIL Europe."
16. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E.: AI4People–an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach **28**, 689–707 (2018)
17. 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
18. European Parliament. Directorate General for Internal Policies of the Union., *The white paper on artificial intelligence*. LU: Publications Office, 2020
19. P. O. o. t. E. Union, "COM/2021/206 final, Proposal for a 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," Apr. 2021. Publisher: Publications Office of the European Union
20. P. O. o. t. E. Union, "Liability for artificial intelligence and other emerging digital technologies.," Nov. 2019. ISBN: 9789276129592 9789276129585 Publisher: Publications Office of the European Union
21. Porat, A., Stein, A.: "Liability for Uncertainty: Making Evidential Damage Actionable,'' in Tort Liability Under Uncertainty. Oxford University Press, Oxford (2001)
22. Brey, P., Søraker, J. H.: "Philosophy of Computing and Information Technology," in *Philosophy of Technology and Engineering Sciences*, pp. 1341–1407, Elsevier, (2009)
23. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Minds Mach. **30**, 99–120 (2020)
24. IEEE, "IEEE SA - Standards Store | IEEE 7000-2021," 2021
25. Eitel-Porter, R.: Beyond the promise: implementing ethical AI. AI Ethics **1**, 73–80 (2021)
26. T. I. f. E. A. . M. Learning, "The Institute for Ethical AI & Machine Learning."
27. Microsoft, "Home."
28. U. N. I. Global, "10 Principles for Ethical AI."
29. IEEE, "IEEE Global A/IS Ethics Initiative Newsletter."
30. Lauer, D.: You cannot have AI ethics without ethics. AI Ethics **1**, 21–25 (2021)
31. Markets, M. a.: "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," (2020)
32. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat Mach Intell **1**, 389–399 (2019)
33. Schühly, A., Becker, F., Klein, F.: *Real Time Strategy: When Strategic Foresight Meets Artificial Intelligence*. Emerald Publishing Limited, (Apr. 2020)
34. Ward, S.C., Chapman, C.B.: Risk-management perspective on the project lifecycle. Int. J. Project Manag. **13**, 145–149 (1995)
35. Beumer, C., Figge, L., Elliott, J.: The sustainability of globalisation: including the 'social robustness criterion. J. Clean. Prod. **179**, 704–715 (2018)
36. Vallance, S., Perkins, H.C., Dixon, J.E.: What is social sustainability? A clarification of concepts. Geoforum **42**, 342–348 (2011)
37. Leprince-Ringuet, D.: "AI's big problem: Lazy humans just trust the algorithms too much," (2020)
38. Soldatos, J., Kyriazis, D.: *Trusted Artificial Intelligence In Manufacturing. A review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production*. Now Publishers Inc., (2021)
39. Hois, J., Theofanou-Fuelbier, D., Junk, A. J.: "How to Achieve Explainability and Transparency in Human AI Interaction," in *HCI International 2019 - Posters* (C. Stephanidis, ed.), Communications in Computer and Information Science, (Cham), pp. 177–183, Springer International Publishing, (2019)
40. Higgins, J. P. T., James, T., Chandler, J., Cumpston, M., Li, T., Page, M. J., W. , Vivian A, and Cochrane Collaboration, *Cochrane handbook for systematic reviews of interventions*. 2020. OCLC: 1167575221
41. Chui, M., Harrysson, M., Manyka, J., Roberts, R., Chung, R., Nel, P., van Heteren, A.: "Applying AI for social good | McKinsey," (2018)
42. J. T. C. OB/7, *AS/NZS 4360:1995 Australing/New Zealand Standard; Risk Management*. (1995)
43. ISO, "IEC 31010:2019," 2019
44. ISO, "ISO Guide 73:2009," 2009

45. ISO, "ISO/IEC TR 24028:2020," 2020
46. C. of Sponsoring Organizations of the Treadway Comission, "About Us."
47. Hopkin, P., Thompson, C.: *Fundamentals of risk management: understanding, evaluating and implementing effective enterprise risk management*. London ; New York: Kogan Page, sixth edition ed., (2022)
48. I. of Risk Management, *A structured approach to enterprise risk management (ERM) and the requirements of ISO 31000*. London: The Institite of Risk Management, 2010

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