Value Sensitive Design based Approach for Trustworthiness including Ethics in Artificial Intelligence

Nuria Quintano[1[0000-0002-3030-0413]], Carmen Alonso-Montes[1[0000-0001-8476-0071]], Alexander Eguía[1[0009-0001-8468-5179]], Irina Marsh[2,3[0000-0002-1287-3699]], Szymon Buś[4[0000-0003-2412-0739]], Agnieszka Sprońska[4[0000-0001-8351-9577]], Michel Barreteau[5[0000-0001-8000-3401]]

[1] Tecnalia, Basque Research and Technology Alliance (BRTA), Derio, Spain
{nuria.quintano,carmen.alonso,alexander.eguia}@tecnalia.com
[2] CBRNE Ltd., London, UK
irina.marsh@cbrneltd.com
[3] Scoala Nationala de Studii Politice si Administrative (SNSPA) (National University of Political Studies and Public Administration), Romania
irina.marsh@comunicare.ro
[4] Lukasiewicz Research Network - Industrial Research Institute for Automation and Measurements PIAP, Warsaw, Poland
{szymon.bus,agnieszka.spronska}@piap.lukasiewicz.gov.pl
[5] Thales cortAIx Labs, France
michel.barreteau@thalesgroup.com

Abstract. The lack of trust in artificial intelligence (AI) slows its adoption across industries and society, usually due to limited understanding of how AI works and its advantages, and of how its trustworthiness including its ethical impact are approached during development. Designing and validating a reliable AI system is complex, requiring inputs from various stakeholders throughout the development process. Ensuring that their concerns are systematically addressed remains a challenge. In this paper, a Value Sensitive Design (VSD) based approach to AI trustworthiness including ethics is presented. The rationale of VSD is that since technologies should primarily serve societal needs, the social impact and associated ethical problems should be anticipated by involving stakeholders at the earliest possible stage [1]. This paper presents a values reference model (VRM) and an implementation method (IM) for Trustworthiness Including Ethics (TIE) defined based on the VSD rationale mentioned. Also, the piloting of the VRM-TIE and of part of the IM-TIE in a use case involving collaborative robots is presented, addressing both technical and human-related characteristics of the AI systems. The preliminary findings in the work suggest that: (1) AI Trustworthiness is a broad concept that must be analysed in detail, considering the industrial context and other priorities, (2) Achieving a shared understanding among stakeholders is essential but difficult, highlighting the need for a structured dialogue, (3) Adopting this VSD based approach, composed by VRM-TIE and IM-TIE, may have a significant impact on the stakeholders' minds.

Keywords: Artificial Intelligence, Trustworthy AI, Ethical AI, Value Sensitive Design.

1 Introduction

"The integration of AI-based solutions into products and services has elicited growing concerns regarding their potential impact on fundamental rights and safety risks posed to users. Notably, apprehensions have been raised regarding the potential infringement on key rights such as non-discrimination, freedom of expression, human dignity, protection of personal data, and privacy" [2]. Ensuring societal trust in AI, Data and Robotics is part of the 2025-2027 mission of European industries. There are many misconceptions and much misinformation about AI, Data and Robotics in societal debates, and the technology is not fully accepted by society in all application areas. This will slow uptake, particularly where mistrust is unfounded, and may also damage markets where genuine risks are not properly addressed [3].

"Trustworthiness must be built up from the design phase of the systems, and this means building responsibly from the top down, bottom up, and throughout the AI lifecycle" [4]. However, current design practices often lack structured processes that make ethical and trustworthy principles traceable and actionable across the AI lifecycle. Trustworthiness by design for AI, concepting and designing human-centric AI systems that embed the technical foundations of trustworthy AI across industrial applications, is still a challenge. It is established as a medium-term objective as part of the Strategic Research, Innovation, and Deployment Agenda 2025-2027 published by the AI Data Robotics Association for Europe [5]. Designing and validating a trustworthy AI system is complex, requiring inputs from various stakeholders throughout the AI lifecycle, from conception to retirement.

The purpose of the study conducted is to define, implement and pilot a Values Reference Model (VRM) and an Implementation Method (IM) for Trustworthiness Including Ethics (TIE) as an end-to-end approach for AI solutions. The aim is to answer the following research questions (RQ): (RQ1) Is trustworthiness and ethics requirements implementation being assessed from technical and human viewpoints?; (RQ2) Is trustworthiness and ethics requirements implementation being assessed by different means?; (RQ3) Are the different stakeholders prioritising the same trustworthiness and ethics requirements?; (RQ4) What is the end-user feedback regarding trustworthiness including ethics?; (RQ5) How are the ethics and legal safeguards being implemented?. The insights obtained at the time of writing this paper allow to answer research questions RQ1, RQ2 and RQ3, since the implementation and piloting of the IM-TIE is work in progress. Answering RQ4 and RQ5 requires the analysis of the results of future work (see section 3.1 for more details).

This paper is organised into the following sections: (1) Introduction, (2) Values reference model and implementation method for trustworthiness including ethics in AI – this section presents the VRM-TIE and the IM-TIE defined, (3) Results – this section presents the findings of the study done in a collaborative robots context and (4) Conclusions – this section presents general conclusions, identified limitations and future work.

2 Values reference model and implementation method for trustworthiness including ethics in AI

The development of a trustworthy and ethical AI requires both (1) a values reference model (VRM) and (2) an implementation method (IM) to integrate trustworthiness and ethical values during the AI lifecycle. In this section both works are presented and explained. This dual contribution – the VRM and the IM – lays the foundation for a structured, traceable, and stakeholder-inclusive integration of trustworthiness including ethical concerns into AI design. The concept "trustworthiness including ethics" (TIE) is used to reflect an integrated approach that encompasses both technical trustworthiness aspects (such as robustness and explainability) and ethical concerns (such as human agency, fairness, and privacy).

2.1 Values Reference Model for Trustworthiness including Ethics (VRM-TIE)

Trustworthiness and ethics are broad concepts, which meaning varies among stakeholders due to different perspectives and priorities within the industrial context. The definition of a values reference model is required to standardise a common understanding and the information gathering among the stakeholders. This section presents the VRM-TIE defined for studying the trustworthiness and ethical requirements of AI solutions. Although multiple sources of literature were examined, the principal sources used to define the VRM-TIE are: Ethically Aligned Design (EAD) [6] and Assessment List for Trustworthy Artificial Intelligence (ALTAI) [7]. From this analysis, a set of nine values combining those suggested by EAD and ALTAI, and each one split into sub-values, are identified, selected and included in the VRM-TIE (see Fig. 1). Please, refer to [6] and [7] for the values' definitions.

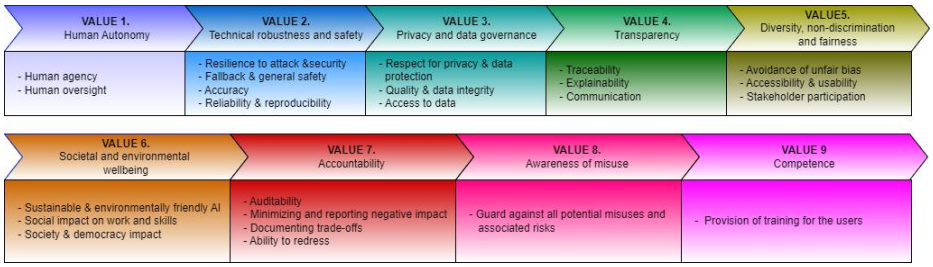

Fig. 1. VRM-TIE: values and sub-values

2.2 Implementation Method for Trustworthiness Including Ethics (IM-TIE)

The IM-TIE is composed by two phases (see Fig. 2): Phase 1. Trustworthiness including ethics by design and Phase 2. Ethics and legal impact assessment.

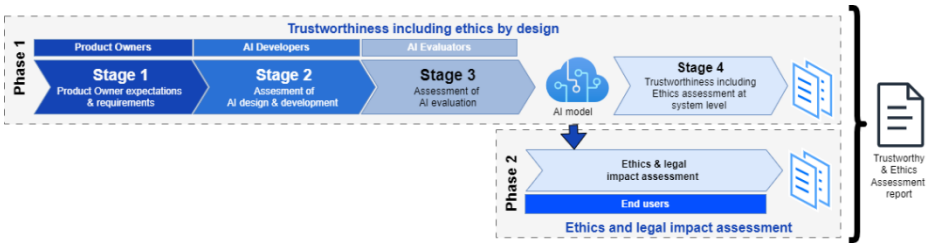

Fig. 2. IM-TIE: phases and stages

Considering the VRM-TIE and its values presented in section 2.1, two value domains are distinguished:

- **Technical values**, which are internal to the system and relate to its functionality and performance. These include robustness, explainability, traceability, and safety. During Phase 1 these values are first assessed by means of AI algorithm-level toolchains (e.g., robustness testing, confidence calibration). Then, they are embedded into the AI system and assessed again at system level.
- **Ethical values**, which are outward-facing and pertain to the system's interaction with users and its impact on society. These include human agency, fairness, privacy, and prevention of misuse. These are addressed through design decisions and governance features such as oversight interfaces, auditability, and fallback mechanisms, and are later examined during Phase 2.

One of the main benefits of this dual approach is to be able to look at AI-based solutions from two points of view, making it possible to collect different types of information, obtained by different means, from different stakeholders (product owners, AI developers, AI evaluators and end-users) and throughout the AI lifecycle. Also, this implementation method does not only ensure compliance with current and evolving regulatory expectations, but also reinforces traceability between design intentions and deployment consequences, in line with the industry challenges for responsible AI development.

Phase 1. Trustworthiness including ethics by design.

Phase 1 is split into four stages (see Fig 2). Stage 1 (product owner expectations and requirements) is initiated with the establishment of an inventory of requirements associated with each of the VRM-TIE sub-values. This requirements inventory (RI) is used to support product owners in the identification of relevant trustworthiness and ethical values and sub-values for their AI-based solution, since they are asked to determine the relevance and the priority of each requirement. In stages 2 and 3, the aim is to gain insight on how trustworthiness and ethical requirements identified by the product owners during stage 1, are covered during the design development of the AI solution, and finally during its evaluation. With this purpose, a checkpoint document is used to gather the information from AI developers (stage 2) and AI evaluators (stage 3). Finally, the information obtained as an output of the four stages is analysed and reviewed

considering all the perspectives, providing a global trustworthiness and ethical assessment of the AI-based solution. The aim of this phase is to provide traceability among relevance and priority, mitigating trustworthiness, ethical and legal misalignment among different stakeholders.

Phase 2. Ethics and legal impact assessment.

The second phase of the IM-TIE addresses the verification of ethical and legal risks associated with the AI system, complementing the values captured during the AI system engineering lifecycle. While Phase 1 embedded values within the system architecture, Phase 2 evaluates their real-world implications, focusing on system accountability, legal exposure, and societal alignment. This phase adopts a formal methodology based on the AI Impact Assessment framework described in ISO/IEC 42005 [ISO/IEC 42005: AI System Impact Assessment, International Organization for Standardization, 2024], adapted for research contexts. It is structured around a detailed questionnaire covering technical, legal, social, and ethical dimensions of impact. The methodology also incorporates regulatory references, including the GDPR, AI Act, and related cybersecurity legislation, synthesised into a Legal Impact Summary Chart.

3 Results

The presented VRM-TIE and IM-TIE have been applied in an industrial context involving collaborative robotics (cobots). This context consists in a human-shared workshop, where robotic manipulators pick components from containers and place them on an assembly table for human operators, who then assemble the parts and return them for transport by a mobile robot to the warehouse. The primary goal is to enhance the reliability and efficiency of robotic support using AI functionalities, contributing to improved safety through human detection and tracking, as well as anomaly detection (unsafe situations, safety gestures). Also, it facilitates smoother human-robot coordination during tasks such as part handovers, ultimately boosting overall operational efficiency.

3.1 Implementation and results in the cobots industrial context

This section provides a summary of the results obtained by considering the VRM-TIE and implementing the IM-TIE, showing the relationship with the research questions (RQ) identified in the introduction. It can be said that RQ1 and RQ2 are answered in the affirmative by adopting the VRM-TIE and the IM-TIE described in section 2, since their goals can be traced to the RQs: Goal 1) Throughout the AI lifecycle, different stakeholders are consulted on the technical and ethical aspects that they consider to be a priority to focus on; Goal 2) Technical evaluations of AI-based solutions are conducted and the opinions of end-user representatives are gathered. This approach enables the collection of both qualitative and quantitative insights on the trustworthiness and ethics of the AI-based solution. Goal 1 is traced to RQ1 while goal 2 is traced to RQ2.

At the time of writing this paper, stages 1 and 2 of IM-TIE have been fully implemented in the cobots industrial context, while stages 3 and 4 as well as Phase 2 are

ongoing work. RQ4 and RQ5 cannot yet be answered based on the current results of the study. The rest of this section explains the insights obtained as an outcome of Stages 1 and 2, which can be traced to RQ3.

Table 1. Number of requirements (req) considered for each VRM-TIE value.

VRM-TIE Value	Total req	Req prioritised by product owner (Stage 1)	Req prioritised by AI developer (Stage 2)
Accountability	6	5	0
Awareness of misuse	1	0	0
Competence	3	1	0
Diversity, non-discrimination and fairness	6	1	0
Human autonomy	7	5	3
Privacy and data governance	7	6	2
Societal and environmental wellbeing	3	1	0
Technical robustness and safety	12	10	2
Transparency	7	7	4
Total	**52**	**36**	**11**

Table 1 shows the number of requirements considered during Stages 1 & 2 within the cobot use case. From the total of 52 requirements of the requirements inventory associated to the VRM-TIE, analysed during Stage 1 by product owners, 36 requirements (69,23%) are considered by product owners as key requirements for the AI-based solution to be developed. During Stage 2, AI developers have considered 11 requirements for being addressed during the implementation activities. Analysing data in Table 1, it can be appreciated that trustworthiness and ethics requirements are mainly concentrated into 4 values: *human autonomy, privacy and data governance, technical robustness and safety* and *transparency.* It is worth noting that these values are the ones with the highest percentage of requirements identified in the VRM-TIE, showing their importance in terms of accomplishment for the AI-based solutions. Both, product owners and AI developers have focused their main interest in those values. This first finding provides an insight into RQ3 “Are the different stakeholders prioritising the same trustworthiness and ethics requirements?”.,

Also, Fig. 3 shows how those values more related with the use of AI solutions and its influence on humans, are prioritised by product owners and AI developers. Focusing on transparency and accountability, which stand as pillars for building trust between technology providers and users [8], it may seem that those two values are prioritized from different viewpoints by product owners and AI developers. A relationship between accountability and transparency values is identified since auditing an AI system (part of accountability) will need first to define mechanisms for traceability and explainability (part of transparency). This finding is consistent with the relationship identified between transparency and accountability for trustworthiness assurance of IA systems in the health sector [9]. One conclusion that can be drawn is that, although a priori it seems that product owners and AI developers are focusing on different values, as the

values of accountability and transparency are closely related, it can be determined that both stakeholders are prioritizing the same functionalities for the AI system. This second finding also provides an insight into RQ3 “Are the different stakeholders prioritising the same trustworthiness and ethics requirements?”.

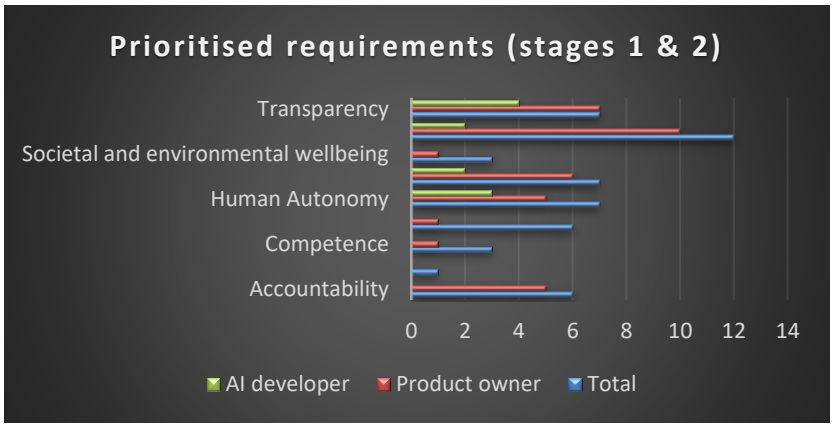

Fig. 3. Requirements per value prioritized by product owners and AI developers

4 Conclusion, limitations and future work

Currently the main limitation is related to the unavailability of definitive results for all the IM-TIE phases and stages, as explained in section 3. The future work is to collect the final feedback to assess the extent to which VRM-TIE and IM-TIE enhance existing industry practices. This feedback will be collected from the cobot context mentioned in this paper as well as from other industrial and space business contexts cases on which VRM-TIE and IM-TIE are being implemented. The analysis of the assessment performed will allow answering RQ4 and RQ5. The preliminary findings in the work suggest that AI trustworthiness including ethics is a broad concept that must be analysed taking into consideration the different stakeholder perspectives and priorities together with industrial context. The proposed VRM-TIE and IM-TIE bridge this gap, transforming abstract ethical concerns into actionable and traceable requirements that are progressively addressed across the AI lifecycle. Therefore, the approach presented in this paper provides one integrated end-to-end process addressing trustworthiness and ethical alignment across the AI lifecycle and engaging all relevant stakeholders. Preliminary results demonstrate that achieving trustworthy AI, including its ethical and societal dimensions, requires not only sound technical design and evaluation, but also the structured integration of values and stakeholder priorities, as shown for the collaborative robotics use case. The approach presented in this paper provides a traceable chain from early design values to final AI product accountability, standardising both technical and legal perspectives under this framework.

Acknowledgements. The work leading to these results has received funding from the European Union's Horizon Europe research and innovation programme within the ULTIMATE project under the Grant Agreement no. 101070162.

References

1. V. d. P. I., Investigating ethical issues in engineering design, in science and engineering ethics, 2001.
2. European Comission, "Artificial intelligence act Briefing," 2024. [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf. Last accessed 2024/09/08.
3. The AI Data Robotics Association (Adra), "Strategic orientation towards an AI, Data, Robotics roadmap 2025-2027.," 2023. [Online]. Available: https://adra-e.eu/sites/default/files/2024-02/ADRA-roadmap-May2023_PostConsultationVersion.pdf. Last accessed 2025/04/12.
4. E. Seger, Experts and AI systems, explanation and trust: A comparative investigation into the formation of epistemically justified belief in expert testimony and in the outputs of AI-enabled expert systems, Apollo - University of Cambridge Repository, 2022.
5. The AI Data Robotics Association (Adra), "Strategic Research, Innotavaion and Deployment Agenda (2025-2027)," 2024. [Online]. Available: https://adr-association.eu/sites/default/files/2024-09/Adra%20Strategic%20Research%20Jul24_v2-2_0.pdf. Last accessed 2025/04/12.
6. Ethically Aligned Design. A Vision For Prioritizing Wellbeing With Artificial and Autonomous Systems. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead v1.pdf. Last accessed 2023/09/28.
7. High-Level Expert group on Artificial Intelligence (European Commission), "The Assessment List for Trustworthy Artificial Intelligence (ALTAI)," 2020. [Online]. Available: https://altai.insight-centre.org/. Last accessed 2023/09/28.
8. Transparency and Accountability in AI Systems: Building Trust Through Openness. [Online]. Available: https://coxandpalmerlaw.com/publication/transparency-and-accountability-in-ai-systems-building-trust-through-openness/. Last accessed 2025/07/16.
9. P. Moreno-Sanchez, "A design framework for operationalizing trustworthy artificial intelligence in healthcare: requirements, tradeoffs and challenges for its clinical adoption.," 2025.