Towards a process for Trustworthy AI systems development

Carlos Mario Braga[1]

[1]Alarcos Research Group – Univ. of Castilla-La Mancha – Ciudad Real – Spain

carlosmario.braga1@alu.uclm.es

Abstract. *This work aims to define a comprehensive Software Development Life Cycle (SDLC) tailored for Trustworthy Artificial Intelligence (AI) systems. These systems are understood as full software solutions that incorporate AI components—such as machine learning models or intelligent agents—whose behavior must align with ethical principles, legal requirements, and technical robustness. The proposed SDLC is grounded on a multidimensional taxonomy of trustworthiness, covering aspects such as lawfulness, non-maleficence, beneficence, autonomy, justice, explicability, and technology. By integrating these principles into all development phases, the SDLC supports the design of AI systems that are not only effective and innovative but also aligned with human values, regulations, and societal expectations. The methodology follows a Design Science Research approach, ensuring the model's relevance, feasibility, and adaptability to evolving technological and regulatory contexts.*

Keywords— Artificial Intelligence, Trustworthy, Process Model

1. Introduction

Norbert Wiener established the foundations of computer ethics in the late 1940s and early 1950s, particularly through his book The Human Use of Human Beings [Wiener 1950], where he analyzed the impact of technologies on human values.

The rise of computational power in the 20th century enabled diverse AI applications, raising ethical concerns—as noted by O'Neil [O'neil 2017]. This led to the publication of numerous ethical guidelines. Algorithm Watch has compiled over 160 such documents.[1] Several works have attempted to synthesize these guidelines into global frameworks for ethical AI [Floridi et al. 2018, Jobin et al. 2019, Thiebes et al. 2021]. However, critiques from legal and technological perspectives persist. Legally, ethics and law are often conflated, which creates a false sense of compliance known as "ethics washing" [Bietti 2020, Steinhoff 2023]. Technologically, ethical principles often lack mechanisms for practical integration into AI development [Whittlestone et al. 2019, Morley et al. 2020, Stix 2021].

To address this gap, a taxonomy that classifies trust-related concerns from ethical, legal, and technical domains is needed. This taxonomy supports the construction of a Software Development Life Cycle (SDLC) for trustworthy AI systems, understood here as complete software systems incorporating AI components whose behavior must align with legal requirements, ethical values, and technical robustness.

As a foundation, this research adopts CRISP-DM, a widely used, non-proprietary model that structures data projects into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [Shearer 2000]. Despite evolving challenges, CRISP-DM is still considered the de facto standard in data-driven development [Schröer et al. 2021], especially in domains like Data Science [Saltz 2021, Saltz et al. 2022,

[1]https://inventory.algorithmwatch.org/

Abbasi et al. 2016]. However, its usage is declining [Mohan et al. 2011, Piatetsky 2014], and improvements to process models have shown benefits in project outcomes [Mariscal et al. 2010, Saltz et al. 2016].

Despite its widespread adoption, CRISP-DM has shown limitations in addressing modern AI development needs. Its evolution towards data science trajectories and the need for modernization have been discussed in recent works [Martínez-Plumed et al. 2019].

This research proposes a CRISP-DM refinement that integrates trustworthiness dimensions throughout the SDLC. The goal is to support the ethical, lawful, and robust development of AI-enabled systems, particularly relevant now given the increasing impact and autonomy of AI technologies, including Generative AI.

2. Research Questions

This research aims to answer the following questions:

- RQ1: Can we classify and organize the core elements of Trustworthy AI, considering ethical, legal, and technological perspectives in a structured and comprehensive way?
- RQ2: How can these elements be effectively operationalized into a SDLC tailored for the development of full AI-based systems, using an established model as foundation?
- RQ3: What requirements must the proposed SDLC meet to ensure compliance with international standards, long-term applicability, and the ability to address emerging challenges in AI development?

RQ1 aims to define a taxonomy that unifies ethical, legal, and technical perspectives of trustworthiness, RQ2 focuses on translating this taxonomy into actionable tasks within a refined SDLC, and RQ3 evaluates the model's compliance with legal, ethical, and technical standards, and its adaptability to emerging regulations.

3. State of Art

According to [Barletta et al. 2023], the main issues in developing ethical and trustworthy AI systems are the diversity of perspectives, the lack of standardized development practices, and the absence of practical frameworks covering the full SDLC while being usable by all stakeholders.

Recent contributions to Trustworthy AI often focus on assessment and verification tools, such as third-party evaluations or self-assessment checklists [Zhou et al. 2020, Zhang et al. 2021]. Examples include Z-Inspection®, based on the European Commission's framework, and Wasabi [Singh et al. 2023], which supports structured evaluation of trust dimensions. While extensive work exists on audits, verifications, and checklists [Sharma et al. 2023, Xiaomei et al. 2023, Radclyffe et al. 2023], these efforts mostly address what should be done, not how to operationalize trustworthiness in development processes.

Some proposals aim to embed trust principles into the AI lifecycle [Wu et al. 2020, Xue et al. 2022], or improve human-AI interaction across development stages [Wickramasinghe et al. 2020]. However, most of these are partial or focused on specific phases. For example, GreatAI [Schmelczer et al. 2023] targets deployment, TrustOps [Kwak 2022] and AI TRiSM [Habbal et al. 2024] center on risk management, while others address individual dimensions such as fairness or explainability [Zhang et al. 2023, Schmid et al. 2023, Sekiguchi et al. 2020].

In parallel, traditional SDLCs such as Waterfall, Spiral, or Agile are still used in AI projects [Agarwal et al. 2017, Mandal et al. 2013], but they lack explicit mechanisms for integrat-

ing legal, ethical, or trust-related aspects [Morley et al. 2020]. Specific models for data science, such as ASUM-DM or Microsoft's Team Data Science Process (TDSP), address technical lifecycle management but do not systematically incorporate trustworthiness dimensions like fairness, accountability, or legal compliance [Saltz et al. 2016, Saltz 2021].

Therefore, despite the variety of existing tools and partial frameworks, there is still a lack of a comprehensive, practical, and evolvable SDLC specifically tailored for Trustworthy AI. As pointed out in prior works [Mandal et al. 2013, Agarwal et al. 2017], adoption of new SDLC strategies remains low due to complexity, compatibility issues, and lack of validation. This research addresses that gap by proposing a refinement of the mature CRISP-DM model, integrating trustworthiness dimensions in a structured and adaptable way, suitable for current and future AI systems under evolving regulatory frameworks.

4. Methodology

This research adopts a Design Science Research (DSR) methodology [Hevner et al. 2010], structured in six iterative phases. This approach ensures that the artifact—the proposed SDLC for Trustworthy AI—is both theoretically grounded and practically validated.

- Problem Identification: The study begins by identifying a critical gap: the absence of comprehensive, practical SDLCs that integrate ethical, legal, and technical dimensions of trust in AI system development.
- Objective Definition: The goal is to design a process model that enables the structured development of full AI-based systems (not only models), ensuring compliance, explicability, and robustness from the outset.
- Design and Creation: The artifact was developed in two steps. First, a taxonomy of trustworthiness was created from a systematic literature review. Second, this taxonomy was operationalized through a refinement of the CRISP-DM model, embedding trust dimensions into each phase and task of the SDLC.
- Evaluation and Validation: The refined SDLC will be assessed via expert reviews, alignment with international standards (e.g., EU AI Act, ISO/IEC, IEEE), and application to a pilot AI project. A comparative analysis (with and without the SDLC) will be conducted to assess practical utility, completeness, and usability.
- Iteration and Improvement: Based on evaluation feedback and real-world testing, the SDLC will be iteratively refined to improve coverage and applicability.
- Practical Use: Finally, the artifact will be deployed in a real development context. A guideline and set of practical resources (techniques, checklists, templates) will be developed to support adoption by AI teams.

5. Proposed Solution and Results

The proposed solution is structured in two key contributions. First, we developed a comprehensive Trustworthy AI Taxonomy that unifies, classifies, and organizes core concepts related to trustworthiness in AI systems. This taxonomy was constructed after conducting a systematic literature review (SLR), in which 67 selected papers were analyzed to extract the ethical, legal, and technological foundations of Trustworthy AI. The taxonomy includes seven top-level principles—*Lawfulness*, *Beneficence*, *Non-Maleficence*, *Autonomy*, *Justice*, *Explicability*, and *Technology*, as shown in Fig. 1 —each decomposed into second-level concepts to ensure consistent interpretation across disciplines. It is conceived as a sociotechnical system, supporting the integrated management of trust-related aspects throughout the AI lifecycle.

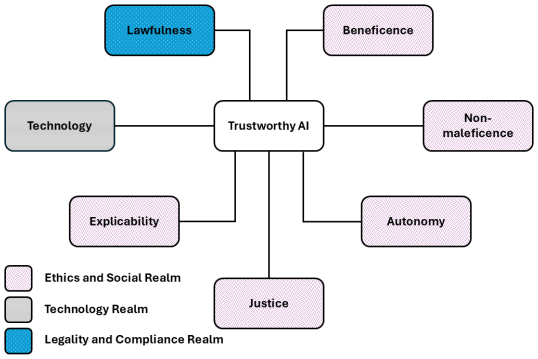

Figure 1. First Level of Trustworthy AI Taxonomy.

The taxonomy includes seven key dimensions: *Lawfulness*, concerning compliance with laws and regulations; *Beneficence*, focused on promoting human and environmental well-being; *Non-maleficence*, ensuring safety, privacy, and harm prevention; *Autonomy*, supporting human oversight and self-determination; *Justice*, targeting fairness and bias mitigation; *Explicability*, ensuring transparency and interpretability; and *Technology*, addressing infrastructure, algorithms, and security to operationalize trust principles.

Each of these dimensions has been further decomposed into a second level of concepts, derived from literature and aligned with the principles of a sociotechnical system. These subcomponents provide additional granularity and semantic clarity, ensuring consistent interpretation and application across different domains and contexts. The complete structure of this second level is illustrated in Figure 2, which presents the detailed Trustworthy AI taxonomy and the classification of its elements into ethical, legal, and technological realms.

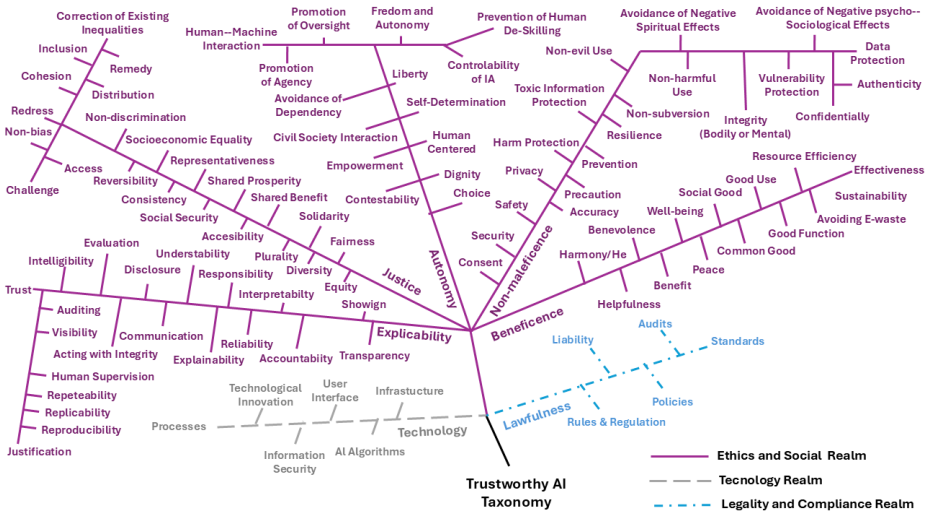

Figure 2. Detailed Trustworthy AI Taxonomy.

The second contribution is the development of CRISP-TAI, a modernized and specialized refinement of CRISP-DM tailored for the development of trustworthy AI systems. To construct this model, a second SLR was conducted to analyze 35 existing CRISP-DM refinements, applying a comparison framework based on eight evaluation metrics (including alignment with trust principles, risk management, and tool compatibility). From this process, 18 high-quality refinements were selected, leading to the extraction of 130 design considerations, which were consolidated

into 100 structured changes to the original CRISP-DM model.

These changes include structural modifications, new tasks and deliverables, and the introduction of a new phase (Continuous Deployment and Monitoring) that addresses post-deployment responsibilities such as fairness evaluation, explainability for end users, continuous risk mitigation, and regulatory compliance.

The structure of CRISP-TAI including its phases and tasks is summarized in Figure 3. This figure provides a high-level visual overview of how trustworthiness principles are embedded across the entire process lifecycle, highlighting the integration of a new iterative phase for continuous deployment and monitoring.

The result is a complete and operational process model named CRISP-TAI, which embeds trustworthiness into all phases of the SDLC while preserving the structure and usability of CRISP-DM. The model is supported by a full definition of phases, tasks, activities, deliverables, and traceability logs. The two main scientific contributions—(i) the Trustworthy AI taxonomy and (ii) the CRISP-TAI process model—have been submitted to peer-reviewed JCR journals and are currently under review. The venue names have been intentionally omitted to preserve the anonymity required by double-blind review processes.

These contributions lay the foundation for a practical and actionable methodology to build trustworthy AI systems, operationalizing abstract ethical, legal, and technical principles into concrete development tasks.

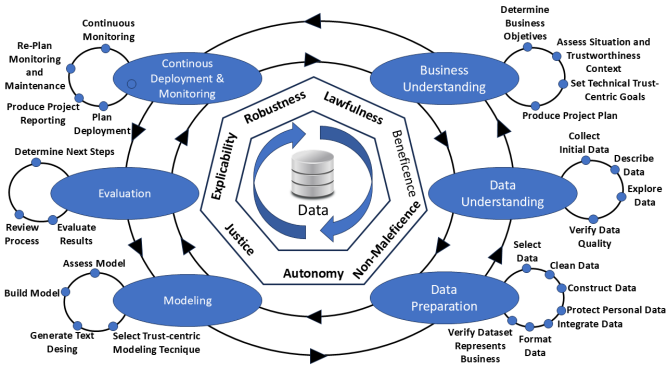

Figure 3. Phases and task of the CRISP-DM refinement for Trustworthy AI.

6. Key Distictive Aspects of Our Proposal

The key differentiating factors of our proposal compared to existing approaches are the following:

- It is based on an exhaustive analysis of the core concerns around AI trustworthiness, integrating ethical, legal, and technological dimensions into a unified sociotechnical system.
- Rather than focusing only on evaluation or audit phases, it provides practical guidance from the earliest stages of system development, embedding trust requirements into all lifecycle phases.
- It specializes and modernizes the most widely used SDLC for data projects, CRISP-DM, ensuring compatibility with industry practices while enhancing it to support Trustworthy AI.
- It introduces a new, iterative phase for continuous deployment and monitoring, supporting long-term compliance, retraining, fairness evaluation, and stakeholder engagement.
- Finally, the model was built following a structured design methodology and is supported by a detailed taxonomy, systematic refinements, and traceability of changes, allowing both transparency and future adaptability.

7. Next Steps

With CRISP-TAI fully defined, upcoming work focuses on validation, refinement, and dissemination. First, we will assess its alignment with international standards such as the EU AI Act [European Parliament and the Council 2024], IEEE P7000 series, and ISO/IEC 24027. Recent studies have translated these frameworks into actionable developer-oriented requirements, reinforcing the importance of integrating trust principles throughout the lifecycle [Baldassarre et al. 2024].

We will test CRISP-TAI in real-world AI projects to evaluate its practical applicability, completeness, and ease of use compared to conventional practices. These pilots will also help identify barriers and inform future improvements.

To foster adoption, a practical guide will be developed with checklists, templates, and traceability aids, ensuring accessibility for diverse teams and maturity levels.

Long-term sustainability will be addressed through mechanisms for updating the model as regulations, risks, and technologies evolve, ensuring its continued relevance.

Lastly, we will explore dissemination via publications and open-source formats.

8. Acknowledgments

This research would not have been possible without the support of the following projects: Di4SPDS (PCI2023145980-2), funded by MCIN/AEI/10.13039/501100011033 and by the European Union (Chist-Era Program), AETHER-UCLM (PID2020-112540RB-C42) funded by MCIN/AEI/10.13039/501100011033, ALBA (TED2021-130355B-C31), funded by MICIU/AEI /10.13039/501100011033 and the European Union NextGeneration EU/ PRTR, and MESIAS (2022-GRIN-34202) funded by FEDER.

References

Abbasi, A. et al. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the association for information systems*, 17(2):3.

Agarwal, P. et al. (2017). SDLC model selection tool and risk incorporation. *Int. J. Comput. Appl*, 172(10):6–10.

Baldassarre, M. T., Gigante, D., Kalinowski, M., and Ragone, A. (2024). Polaris: A framework to guide the development of trustworthy ai systems. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pages 200–210.

Barletta, V. S. et al. (2023). A rapid review of responsible AI frameworks: How to guide the development of ethical AI. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, pages 358–367.

Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 210–219.

European Parliament and the Council (2024). Artificial intelligence act. `https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html`. Final text approved on 13 March 2024.

Floridi, L. et al. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28:689–707.

Habbal, A. et al. (2024). Artificial intelligence trust, risk and security management (AI TRiSM): Frameworks, applications, challenges and future research directions. *Expert Systems with Applications*, 240:122442.

Hevner, A. et al. (2010). Design science research in information systems. *Design research in information systems: theory and practice*, pages 9–22.

Jobin, A. et al. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9):389–399.

Kwak, J. H. (2022). Trustops: A risk-based AI engineering process. In *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 2252–2254. IEEE.

Mandal, A. et al. (2013). Investigating and analysing the desired characteristics of software development lifecycle (SDLC) models. *International Journal of Software Engineering Research & Practices*, pages 9–15.

Mariscal, G. et al. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25(2):137–166.

Martínez-Plumed et al. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering*, 33(8):3048–3061.

Mohan, K. et al. (2011). What methodology attributes are critical for potential users? understanding the effect of human needs. In *Advanced Information Systems Engineering: 23rd International Conference, CAiSE 2011, London, UK, June 20-24, 2011. Proceedings 23*, pages 314–328. Springer.

Morley, J. et al. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4):2141–2168.

O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, New York City, U.S.

Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. *KDD News*.

Radclyffe, C. et al. (2023). The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in artificial intelligence*, 6:1020592.

Saltz, J. S. (2021). CRISP-DM for data science: strengths, weaknesses and potential next steps. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2337–2344. IEEE.

Saltz, J. S. et al. (2016). Big data team process methodologies: A literature review and the identification of key factors for a project's success. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2872–2879. IEEE.

Saltz, J. S. et al. (2022). Current approaches for executing big data science projects—a systematic literature review. *PeerJ Computer Science*, 8:e862.

Schmelczer, A. et al. (2023). Trustworthy and robust AI deployment by design: A framework to inject best practice support into AI deployment pipelines. In *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)*, pages 127–138. IEEE.

Schmid, A. et al. (2023). The importance of an ethical framework for trust calibration in AI. *IEEE Intelligent Systems*.

Schröer, C. et al. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181:526–534.

Sekiguchi, K. et al. (2020). Organic and dynamic tool for use with knowledge base of AI ethics for promoting engineers' practice of ethical AI design. *AI & Society*, 35(1):51–71.

Sharma, V. et al. (2023). Framework for evaluating ethics in AI. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, pages 307–312. IEEE.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

Singh, A. M. et al. (2023). Wasabi: A conceptual model for trustworthy artificial intelligence. *Computer*, 56(2):20–28.

Steinhoff, J. (2023). AI ethics as subordinated innovation network. *AI & Society*, pages 1–13.

Stix, C. (2021). Actionable principles for artificial intelligence policy: three pathways. *Science and Engineering Ethics*, 27(1):15.

Thiebes, S. et al. (2021). Trustworthy artificial intelligence. *Electronic Markets*, 31:447–464.

Whittlestone, J. et al. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 195–200.

Wickramasinghe, C. S. et al. (2020). Trustworthy AI development guidelines for human system interaction. In *2020 13th International Conference on Human System Interaction (HSI)*, pages 130–136. IEEE.

Wiener, N. (1950). *The human use of human beings: Cybernetics and society*. The riverside press, Cambridge, Massachusetts.

Wu, W. et al. (2020). Ethical principles and governance technology development of AI in china. *Engineering*, 6(3):302–309.

Xiaomei, S. et al. (2023). Research on trustworthiness analysis technology of artificial intelligence software. In *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*, pages 802–806. IEEE.

Xue, L. et al. (2022). Ethical governance of artificial intelligence: An integrated analytical framework. *Journal of Digital Economy*, 1(1):44–52.

Zhang, J. et al. (2023). Fairness in design: A framework for facilitating ethical artificial intelligence designs. *International Journal of Crowd Science*, 7(1):32–39.

Zhang, T. et al. (2021). Trusted artificial intelligence: technique requirements and best practices. In *2021 International Conference on Cyberworlds (CW)*, pages 303–306. IEEE.

Zhou, J. et al. (2020). A survey on ethical principles of AI and implementations. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 3010–3017. IEEE.