Integrating Explainability-by-Design for Transparent and Efficient AI in Manufacturing

Andrea Capaccioli [1], Carlos Agostinho [2,3], Veronica Antonello[4], Fenareti Lampathaki[5] and Michele Sesana [4]

[1] *Deep Blue srl, Via Daniele Manin, 53 00185, Rome, Italy*
[2] *Center of Technology and Systems (UNINOVA-CTS) and LASI, Campus FCT, 2829-516 Caparica, Portugal*
[3] *Knowledgebiz Consulting, Rua Fernando Pessoa, N 4ª, 2805-998 Almada, Portugal*
[4] *TXT e-solutions SpA, Via Milano, 150, 20093 Cologno Monzese (MI), Italy*
[5] *Suite5 Data Intelligence Solutions, Alexandreias 2, Bridge Tower 3013, Limassol, Cyprus*

Abstract
Artificial Intelligence (AI) has rapidly transformed industries but concerns about opacity have led to a focus on Explainable AI (XAI). The X-by-Design paradigm, developed within the XMANAI project, embeds explainability into AI design. It prioritizes transparency, integrating explainability mechanisms from the outset to enhance understanding and trust. This paper outlines the core principles of X-by-Design and its implementation framework across Data, Model, and Results Explainability. Case studies from four industrial demonstrators demonstrate its efficacy in addressing specific business needs and empowering users with actionable insights. Overall, the X-by-Design paradigm signifies a crucial shift towards transparent and accountable AI systems, ensuring ethical deployment and user trust.

Keywords
Explainable Artificial Intelligence, X-by-Design, Transparency, AI in manufacturing

1 Introduction

In the past years, artificial intelligence (AI) has swiftly reshaped numerous industrial sectors, fundamentally altering our interaction with technology and the way we handle data [1]. However, concerns about the transparency and interpretability of AI systems have emerged alongside these advancements, leading to a growing emphasis on Explainable AI (XAI) [2], [3], [4].

AI systems are often seen as opaque, with their internal logic inaccessible or incomprehensible to humans. Models can consist of millions of features connected in a complex web of dependent behaviours. Conveying this internal state and dependencies in a humanly comprehensible way is extremely challenging. Transparency in AI systems has been identified as quintessential, but the black box nature of AI systems makes the definition of transparency requirements challenging [5]. With the rise of AI and its capacity for automated decision-making, there's a growing demand for transparency in understanding how these systems reach conclusions [6]. However, achieving transparency is challenging due to the complex nature of AI technology, particularly with advanced machine learning (ML) methods, such systems become virtually impossible to trace, even for experts. Thus, a trade-off must be made between accuracy and explainability or interpretability, as advanced systems that are more accurate in their predictions are becoming less interpretable [7].

XAI is the ability of AI systems to provide clear and understandable explanations for their actions and decisions. Its central objective is to make the behaviour of these systems understandable to

Proceedings Acronym: I-ESA 2024 12th International Conference on Interoperability for Enterprise Systems and Applications, April 10–12th, 2024, Crete, Greece
EMAIL: andrea.capaccioli@dblue.it (A. Capaccioli); ca@uninova.pt (C. Agostinho); veronica.antonello@txtgroup.com (V. Antonello); fenareti@suite5.eu (F. Lampathaki); michele.sesana@txtgroup.com (M. Sesana)
ORCID: 0000-0001-7281-2049 (A. Capaccioli); 0000-0002-2884-776X (C. Agostinho); 0000-0002-3131-4622 (F. Lampathaki)

humans by elucidating the underlying mechanisms of their decision-making processes. However, many efforts to improve explainability often lead to explanations that are tailored to the AI researchers themselves, rather than effectively addressing the needs of the intended users. This places the responsibility for defining a satisfactory explanation for complex decision models in the hands of AI experts who have a detailed understanding of these models [8]. Ideally, XAI should include the ability to explain the system's competencies and understandings to all types of users, explain its past actions, ongoing processes and upcoming steps, and disclose the relevant information on which its actions are based [9].

This paper presents an explainable-by-design (X-by-Design) paradigm that applies XAI to change the standard in the design of AI systems. The approach has been developed and validated in the XMANAI European project (www.ai4manufacturing.eu).

2 X-by-Design Paradigm

Transparency and explainability are essential quality requirements in machine learning, influencing user needs, cultural values, laws, and corporate standards. However, often in ML such property is not available, and to have it, models would need to be retrained which is a labour- and computation-intensive process [4]. X-by-Design therefore refers to the AI models and systems design process, where explainability is enabled and embedded from the start, thus improving transparency as a key non-functional requirement in ML and autonomous systems. The design of technologies significantly influences their subsequent use and effects on society, emphasizing the importance of governing the design process to ensure responsible implementation of information technologies and protect citizens from unintended negative consequences [10]. An explainable design requires extracting qualifiers from data to describe meaningful features and establish causal relationships between inputs and outputs [11].

The X-by-Design approach emphasizes integrating interpretability mechanisms directly into the architecture of AI models during the developmental phase. By prioritizing explainability from the outset, AI designers, engineers, and data scientists can construct models that inherently offer transparent insights into their decision-making processes, thereby mitigating the need for complex and often unreliable post hoc explanations. "By-design" concepts like privacy-by-design and security-by-design are already commonly employed, and **X-by-Design represents the next paradigm shift**.

2.1 X-by-Design Approach in XMANAI: Interlinked Perspectives of Data, Model and Result Explainability

To transition towards an X-by-Design paradigm, XMANAI considers explainability under three interlinked and collaborative perspectives:

- **Data Explainability** that focuses on a concrete understanding of data in terms of semantics and structure (data types) per feature in order to gain insights into the input data, achieved through mapping to a common data model. Interactive data exploration and visualization that allow viewing data distribution/profiling charts or summary statistics (e.g. number of missing values, min/max values) can be leveraged to monitor potential data drifting or bias issues.
- **Model Explainability** that concerns understanding the different AI (ML/DL) models towards global interpretability (answering how does a model work for all our predictions) and local interpretability (answering how a model is generating a specific prediction, given specific data points). Different techniques can be employed depending on the model family and type, e.g. black-box or opaque models, in order to attempt to shed light on the model design and training phase. Indicative categories include: (a) Explanations by simplification (or surrogate models) referring to the techniques that approximate an opaque model using a simpler one, which is easier to interpret; (b) Feature relevance explanations which attempt to explain a model's decision by quantifying the influence of each input variable, serving as a first step towards gaining insights into the model's reasoning; (c) Explanations through directly interpretable models (since transparent models like decision trees, linear and logistic regression are directly interpretable). Typical techniques associated to the above categories

are: Local interpretable model-agnostic explanations (LIME), SHAP (SHapley Additive exPlanation), Anchors (High-Precision Model-Agnostic Explanations), Partial Dependence Plot (PDP), etc.

- **Results Explainability** that promotes shared understanding of results and translating them into concrete actions in an appropriate style/format. At this step, post-hoc explanations over the results are created (after the model is trained) and may include: (a) Visualizations that typically act as the primer for communicating results to the involved stakeholders in order to inform them about the decision boundary or how features interact with each other; (b) Text explanations that convey in natural language how to take action and can be automatically generated (through natural language generation techniques); (c) Explanations by example that extract representative instances from the training dataset to demonstrate how the model operates in a similar manner as humans may approach explanations by providing specific examples to describe a broader outcome/process; (d) Counterfactual Explanations that aim to find the model's decision boundaries with respect to specific input values.

Additionally, in order to produce AI models and pipelines that are explainable by design, XMANAI has also delivered appropriate methods and techniques to address a number of complementary challenges that currently constitute significant data scientists' pains. Our "XAI Models Security" component performs a risk and vulnerability assessment over different Explainable ML/DL models to offer immunity and robustness. The aim is to timely anticipate and safeguard against unintended bias and adversarial attacks that may try to manipulate algorithms. Such adversarial attacks may refer to poisoning (attempting to maliciously manipulate the training dataset) and evasion attacks, in general. The XMANAI "XAI Models Performance" component sets the baseline thresholds for technical metrics like accuracy and scalability, along with evaluating explainability metrics for audience satisfaction. The XMANAI "XAI Assets Sharing" is another key component that supports interaction among different types of stakeholders who want to share their assets and collaboratively work to build a solution for a specific manufacturing case/problem that reaches consensus and is broadly accepted. The full spectrum of project solutions is detailed in [12].

2.2 X-by-Design Service Framework for Replicability

The explained approach evolved in discussions with the project Industrial Advisory Board, charting a commercial path for enabling XAI in manufacturing. The paradigm transitioned into a consultancy and technical services framework that was adapted from the strategy followed with the project's pilot companies. It is structured around two axes of services:

1. *X-LEARN services* that focus on studying the problem of the manufacturing companies, identifying the best methodologies or models to integrate explainability in the design phase. *X-LEARN* can be split in 2 sub-applications: (a) *X-LEARN-basics*, analysing state-of-the-art research to provide consultancy training on explainability concepts and identify XAI models for enhancing technological transparency. This involves investigating manufacturing scenarios like Production Optimization or Demand Forecasting; (b) *X-LEARN-Applications,* that goes a step further, utilizing lessons learned, evaluating similar needs and devising implementation plans to drive AI-driven digital transformation. The goal consists in using use cases of similar applications with other subjects and sharing the obtained results to strengthen user trust.
2. *X-APPLY services* that focus on the most effective path to integrate digital solutions based on explainability in industrial contexts by providing IT services. The final goal is to favour XAI adoption in both new or already existing productive systems both from a consultancy and technological implementation level. The work mostly concerns the design and consecutive application of XMANAI digital solutions and analytics.

3 X-by-Design in Action

The implementation of the X-by-Design approach in the XMANAI industrial pilot cases followed a user centric process mimicking the *X-LEARN and X-APPLY* services, with the aim of defining

Explainability requirements to guide both the technical implementation of Data, Model and Results Explainability and to define the future human-XAI interactions within the 4 project industrial demonstrators. The process aimed at strengthening design and implementation of explainability with X-by-Design.

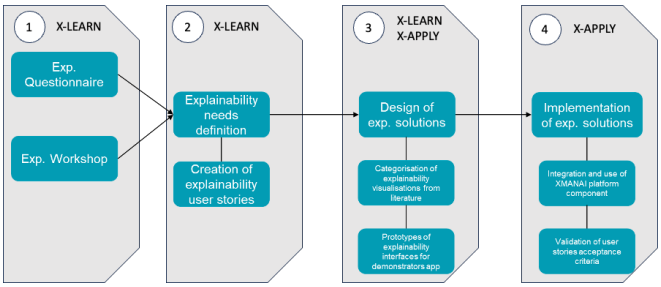

Figure 1: XAI design process

As illustrated in Figure 1, the X-by-Design process considered 4 main steps: 1) data gathering, where workshops and detailed questionnaires are utilized to understand the explainability needs of industrial demonstrators; 2) definition of explainability requirements from the collected needs as user stories like specified by [13]. This approach on formalising the why and how the user interacts with the system, provides context to developers and consultants working on the system specification. 3) Design the explainability solutions in terms of data, model, and results following the approach previously described; 4) Implementation of the designed solutions using the XMANAI platform components and the specific explainable interfaces.

3.1 Industrial Application in XMANAI

The XMANAI project has four industrial demonstrators where the X-by-Design process has been applied, here below are discussed the main results achieved in the demonstrators:

Demonstrator 1 – FORD: The initial phase focused on understanding the specific Explainable AI (XAI) requirements of end-users, especially operators on the engine production line. Close collaboration with these stakeholders was pivotal in grasping their needs and preferences regarding AI model interpretability. Through this engagement, their expectations and concerns were captured, guiding the subsequent selection of models and tools to align with their identified needs. The next step involved evaluating various methods to meet these XAI needs, considering the unique requirements and constraints of the manufacturing context. Selected methods were carefully assessed for their ability to provide interpretable insights into AI model decision-making. This rigorous selection process ensured that the chosen methods effectively addressed the identified XAI needs. By following these steps, the AI and explainability aspects of the demonstrator were tailored to meet the end-user requirements of the plant. This collaborative and user-centric approach guarantees that the selected methods offer meaningful and actionable explanations. Ultimately, this process facilitates the development of XAI models that empower plant personnel by enhancing understanding and trust in AI-driven decisions, leading to more effective decision-making and optimized manufacturing processes.

Demonstrator 2 – WHIRLPOOL: This demonstration involved key stakeholders such as Central Demand Planners, Direct-to-Consumer (D2C) marketing and sales specialists, and Data Scientists, each with specific needs for explainability. The Central Demand Planner requires insights into forecasting processes, root causes affecting accuracy, visualization of forecasting plots, trend analysis, customer behaviors, and marketing strategy impacts. Conversely, the D2C Marketing & Sales specialist focuses on understanding demand evolution, customer purchasing patterns, and marketing strategy effects. To address these needs, two categories of explainability approaches were identified: Explanation at the data level and Explanation at the instance and model level. For explanations at instance and model level, post-hoc processing of trained models was conducted using techniques such as SHAP and permutation importance. SHAP force plots provided insights at the instance level, indicating feature contributions and impact direction. Alluvial plots were utilized for enhanced user understanding based on business user feedback. What-if scenarios were also explored, enabling users to grasp the correlation between input features and model outputs. These tailored explainability methods empower business users to interpret and utilize AI models for effective decision-making, promoting a user-centric approach in AI.

Demonstrator 3 – CNH: In the X-by-Design process for CNH, substantial efforts were dedicated to understanding end users' specific needs and translating them into graphical representations for

explanations. For operators seeking algorithmic suggestions on sensors contributing to machine errors, a selection of plots (force plots, bar plots, cohort plots, and heatmaps) was chosen to explain algorithm results within the mobile nature of the demonstrator app. These charts were adapted in format and visual style to suit the target audience, considering the plant operators' limited familiarity with statistical charts and data science. A simplified force plot assists operators in swiftly identifying primary failure components, enabling targeted actions or exclusion of less likely contributors. Additionally, a more advanced correlation matrix visualization serves white-collar workers, offering insights into correlations between anomalies and causing sensors. Anomalies are categorized, and a relationship score is assigned to each pair, aiding maintenance engineers in identifying patterns for proactive maintenance planning. Overall, the proposed visualization provides a user-friendly yet powerful tool for maintenance operators and engineers to interpret and act upon algorithmic results effectively.

Demonstrator 4 – UNIMETRIK: The UNIMETRIK demonstrator targets industrial users, particularly metrologists and process engineers, seeking to understand how parameters like Lateral Density, Direction Density, and Exposure Time influence product measurement scanning processes and quality. Employing explainability techniques, the demonstrator enhanced metrologists' understanding of parameter impacts on measurement accuracy and quality. Visualizations and explanations at both data and model levels offer essential insights, optimizing measurement scanning processes and overall system efficacy. Data-level visualizations, including line plots and box plots, provide a holistic understanding of parameter interactions, aiding decision-making. Model-level explanations like SHAP summary plots and global surrogate models offer insights into feature contributions and model output. During the X-by-Design process, technical partners created descriptive visualizations based on user discussions, ensuring user-friendliness. Examples include joint plots illustrating measurement errors and correlation heatmaps showcasing feature-target relationships. These measures optimize measurement plans and point cloud quality, enhancing accuracy and efficiency in industrial metrology. Overall, explainability fosters user comprehension, actionable insights, and trust in AI systems, facilitating their successful integration into metrology workflows.

3.2 UXAI Tool

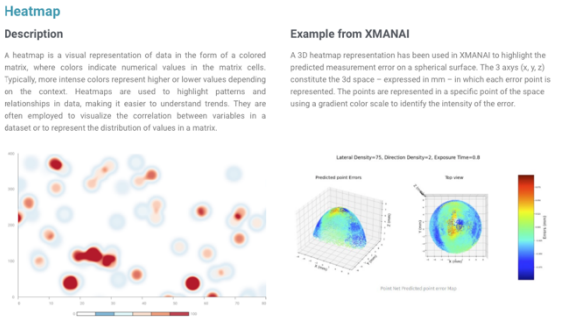

Figure 2: UXAI tool (example on selecting the best-fit visualization to identify anomalies)

UXAI is a valuable tool for users navigating Results Explainability within the X-by-Design framework, assisting in selecting optimal visualizations (Figure 2). Accessible via the project website (ai4manufacturing.eu/uxai), UXAI draws from the experience of the XMANAI pilots, offering design support for companies and designers developing XAI solutions in manufacturing. By highlighting visualization options, UXAI facilitates user-friendly XAI implementations, enhancing the human-centered approach of the X-by-Design process. From the experiences of the XMANAI pilots, five clusters of XAI goals in manufacturing have been identified (Anomalies, Forecast, Pattern and Trends, Planning, Scenarios). For each cluster specific questions have been defined, related to possible XAI requirements that a company wants to achieve with the implementation of XAI. According to the work done in the project, to each question are associated XAI data visualisations suggestions. The user of the tool can then navigate among the different cluster and see the example of possible visualisation for the defined questions. The tool is easily expandible in the future to integrate future insights and development.

4 Challenges and Future Perspective

XAI serves as a crucial interpreter, bridging the gap between complex algorithmic processes and human understanding, rendering intricate information accessible. Identifying the benefits of XAI over conventional AI systems is crucial, especially considering the 'cost of poor decisions,' which underscores the need for explaining AI decisions as their human impacts increase [14]. The X-by-Design approach aims to promote understanding of AI systems at the data, model, and results dimensions, offering a comprehensive insight into AI logic and empowering companies to transparently harness AI benefits from the system design phase. Within the XMANAI project, the X-by-design approach enhanced explainability in demonstrator applications, prioritizing exceptional performance alongside meaningful human-AI interactions. The process involves user research to define main activities and consider contexts for application use, defining user needs and explainability requirements, and prototyping interfaces. Applying the X-by-design approach to production can benefit the manufacturing industry by streamlining processes, reducing costs and risks, and ensuring business continuity. Deploying AI-driven ML models with sensor data can develop robust AI, alerting when equipment performance degrades and needs maintenance, thus reducing downtime. The proposed approach aligns products with customer interest by understanding behaviour patterns and demands. Although the path to X-by-Design presents challenges, its potential impact on manufacturing productivity and XAI uptake is significant.

Acknowledgements

The research leading to this work has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No: 957362

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Palanivelu, V. R., and B. Vasanthi. "Role of artificial intelligence in business transformation" International journal of advanced science and technology 29.4 (2020): 392-400.
[2] Felzmann, H., Fosch-Villaronga, E., Lutz, C. et al. Towards Transparency by Design for Artificial Intelligence. Sci Eng Ethics 26, 3333–3361 (2020). https://doi.org/10.1007/s11948-020-00276-4
[3] Adadi A. and Berrada, M. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)" IEEE Access, vol. 6, pp. 52138–52160, 2018.
[4] Branco, R. et al., "Explainable AI in Manufacturing: an Analysis of Transparency and Interpretability Methods for the XMANAI Platform". In: 29th ICE IEEE/ITMC Conference (2023).
[5] Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith T., and Bohlender, D. "Explainability as a Non-Functional Requirement". In 2019 IEEE 27th International Requirements Engineering Conference (RE), Jeju, Korea (South), 2019, pp. 363-368, doi: 10.1109/RE.2019.00046
[6] Felzmann, H., Fosch-Villaronga, E., Lutz, C. et al. "Towards Transparency by Design for Artificial Intelligence". Sci Eng Ethics 26, 3333–3361 (2020). https://doi.org/10.1007/s11948-020-00276-4
[7] Lampathaki, F., Agostinho, C., Glikman, Y. and Sesana, M. "Moving from 'black box' to 'glass box' Artificial Intelligence in Manufacturing with XMANAI". In 2021 IEEE ICE/ITMC, Cardiff, United Kingdom, 2021, pp. 1-6, doi: 10.1109/ICE/ITMC52061.2021.9570236.
[8] Miller, T. (2019). "Explanation in artificial intelligence: Insights from the social sciences". Artificial intelligence, *267*, 1-38.
[9] Gunning, D. S. (2019). "XAI—Explainable artificial intelligence". Science robotics, 4(37).
[10] Kudina, O., & Verbeek, P. P. "Ethics from within: Google Glass, the Collingridge dilemma, and the mediated value of privacy". Science, Technology and Human Values, 44(2), 291–314. 2019
[11] Bäckström, T. "X by design: From privacy to explainability, trust and robustness" (2022). Available online: https://tbackstr.medium.com/x-by-design-d1f4b4cf31ef
[12] Agostinho, C. et al., "Explainability as the key ingredient for AI adoption in Industry 5.0 settings". Front. Artif. Intell., vol. 6, 2023. https://doi.org/10.3389/frai.2023.1264372

[13] Lucassen, G., Dalpiaz, F., Van Der Werf, J. M. E. M., & Brinkkemper, S. ”Forging high-quality User Stories: Towards a discipline for Agile Requirements”. In 2015 IEEE RE, 126–135. 2015.
[14] Lecue, F. “Explainable AI – The Story So Far”. 2019. Retrieved from: https://www-sop.inria.fr/members/Freddy.Lecue/presentation/Aug29-Lecue-Thales-XAI-The-Story-So-Far-Final.pdf . Accessed: 16-02-2024