Explainability in the Artificial Intelligence Act

Radosław Pałosz, Michał Araszkiewicz, Grzegorz J. Nalepa

Jagiellonian University in Kraków

Supported by the XPM Project (Explainable Predictive Maintenance) under CHIST-ERA 2019 Call (Project No. 857925, NCN 2020/02/Y/ST6/00070)

NATIONAL SCIENCE CENTRE
POLAND

XPM

Artificial Intelligence for Europe COM(2018) 237

- „Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals”

- „Artificial intelligence (AI) is already part of our lives – it is not science fiction”

- „The stakes could not be higher. **The way we approach AI will define the world we live in**”

Artificial Intelligence for Europe COM(2018) 237

- Aims of the European Initiative:
  - Boost the EU's technological and industrial capacity and AI uptake across the economy
  - Prepare for socio-economic changes
  - **Ensure an appropriate ethical and legal framework, based on the Union's values and in line with the Charter of Fundamental Rights of the EU.** This includes forthcoming guidance on existing product liability rules, a detailed analysis of emerging challenges, and cooperation with stakeholders, through a European AI Alliance, for the development of AI ethics guidelines.

Legal Framework according to „Artificial Intelligence for Europe”

- Product-liability regulations;
- Standards in data management and security set by GDPR;
- Digital Single Market Act – management of non-personal data flow;
- Building trust through **explainability**, by ensuring the humans could understand the functioning of the legal systems;
- Product liability regulations;
- Consumer empowerment – ability to control data and to contact a human in connection with AI system operations

Ethical Guidelines by AI HLEG

- In April 2019 High-Level Expert Group on Artificial Intelligence issued Ethical Guidelines for Trustworthy AI
- Lawful, ethical and robust
- Seven key requirements for trustworthy AI:
  - **(1) human agency and oversight**, (2) technical robustness and safety, (3) privacy and data governance, **(4) transparency,** (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and **(7) accountability.**

Trust-Building Through Explainability

Human Agency and oversight

Transparency

Accountability

Human Agency and Oversight

- Protection of fundamental rights that can be both supported and endangered by the AI – like privacy or right to education

- „Users should be able to make **informed autonomous decisions** regarding AI systems. They should be given the knowledge and tools to **comprehend and interact** with AI systems to a **satisfactory** degree and, where possible, be enabled to reasonably self-assess or challenge the system. (…) Key to this is the **right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them**".
- Human oversight – HITL, HOTL or HIC

Transparency

- Grounded in *principle of explicability* – transparency of (i) data, (ii) system and (iii) the business models
- **Traceability –** documenting data sets and the processes that yield the AI system's decision in order to understand causes of the output in order to properly indentifying malfunctions;
- **Explainability** – explaining technical processes as well as reasons for human decisions connected with the operation of the system; trade-offs between explainability and accuracy; explanations adapted to the stakeholders
- **Communication –** AI systems should not represent themselves as humans for users and the latter should be able to communicate directly with human.

Accountability

- Responsibility for AI systems operations – before, during and after their deployment and use
- **Auditability** – requirement of periodical assessment of the AI systems;
- **Minimasation and reporting of negative impacts** – including protection of the reporting parties (whistleblowers, NGOs, etc.)
- **Trade-offs** – when they arise, they should be explicitly addressed. If no ethical solutions are possible, the system should be modified accordingly;
- **Redress –** especially for vulnerable parties

White Paper on AI - COM(2020) 65 final

- „A European approach to excellence and trust„
- „Simply put, AI is a collection of technologies that combine data, algorithms and computing power”.
- „The Commission is committed to enabling scientific breakthrough, to preserving the EU’s technological leadership and to ensuring that new technologies are at the service of all Europeans – **improving their lives while respecting their rights**”.
- „If the EU fails to provide an EU-wide approach, there is a real risk of fragmentation in the internal market, which would undermine the objectives of trust, legal certainty and market uptake”

White Paper – endangered fundamental rights

- freedom of expression,
- freedom of assembly,
- human dignity,
- non-discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation, as applicable in certain domains,
- protection of personal data and private life,
- right to an effective judicial remedy and a fair trial,
- consumer protection.

White Paper - transparency

- The Communicate addresses the transparency, but does not mention explainability;
- Transparency should serve primarily ensuring the product-liability standards and trust-building for innovation;
- Focus on providing information about system’s operation – on its different layers;
- Communication about interacting with the AI system;
- Requirement of human oversight.

European AI Alliance

- Platform for wide-spread discussion about developing particular elements of AI strategy
- Forum
- Blog
- Documents
- Library
- Events

A European Approach to Artificial Intelligence

- EU as a global hub for innovative and trustworthy AI
- Safety based on transparency achieved i.e. with explainability
- Explanation of technical processes (both human and non-human) within a system, adjusted to the capabilities of a person demanding it

Artificial Intelligence Act - overview

- Regulatory flagship for further normative developments
- According to the Memorandum preceeding the project, AI Act aims to provide an environment for preserving "the EU's technological leadership" while at the same time mitigate risks connected with the use of AI
- Proceedings since April 2021, now during the works in Parliamentary Commitees

AI in AI Act

- According to the art. 3(1) of AIA, an artificial intelligence system is:
  - a software "developed with one or more of the techniques and approaches listed in Annex I”, namely:
    - Machine learning approaches;
    - Logic- and knowledge-based approaches;
    - Statistical approaches
  - can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.

Is the Artificial Ingelligence **just any software?**

High-risk AI systems

- Focus on regulating **high-risk AI systems**. According to the art. 6.1 and art. 6.2, these are:
  - systems being safety components of products to which EU product-safety legislation applies or the products themselves or those that undergo a third-party conformity assessment. Additionally,
  - those listed in Annex III enumerates further categories of high-risk systems, that influence spheres of fundamental human rights and law enforcement.
- The group of high-risk systems is very diverse, containing systems operating on different principles, being used both by public and private subjects and serving purposes from product safety control to biometric identification of individuals.

Obligations on high-risk AI systems providers

- preparing a risk management system that "shall be established, implemented, documented and maintained" (art. 9.1).
    - monitoring possible risks related to the system operation
    - eliminating, reducing and mitigating them where possible
- Providers should also introduce „appropriate data governance”
    - data sets used by the system should be "relevant, representative, free of errors and complete (…), have the appropriate statistical properties" (art. 10.3)
- Relevant technical documentation, "to demonstrate that the high-risk AI system complies with the requirements" (art. 11.1).
- Systems should also automatically record their operation to "**ensure a level of traceability** (…) that is appropriate to the intended purpose of the system" (art. 12.2).
- Design allowing oversight by natural persons to minimize the risks to health, safety, or fundamental rights (art. 14.1 and 14.2).
- „Appropriate level of accuracy, robustness and cybersecurity" (art. 15).

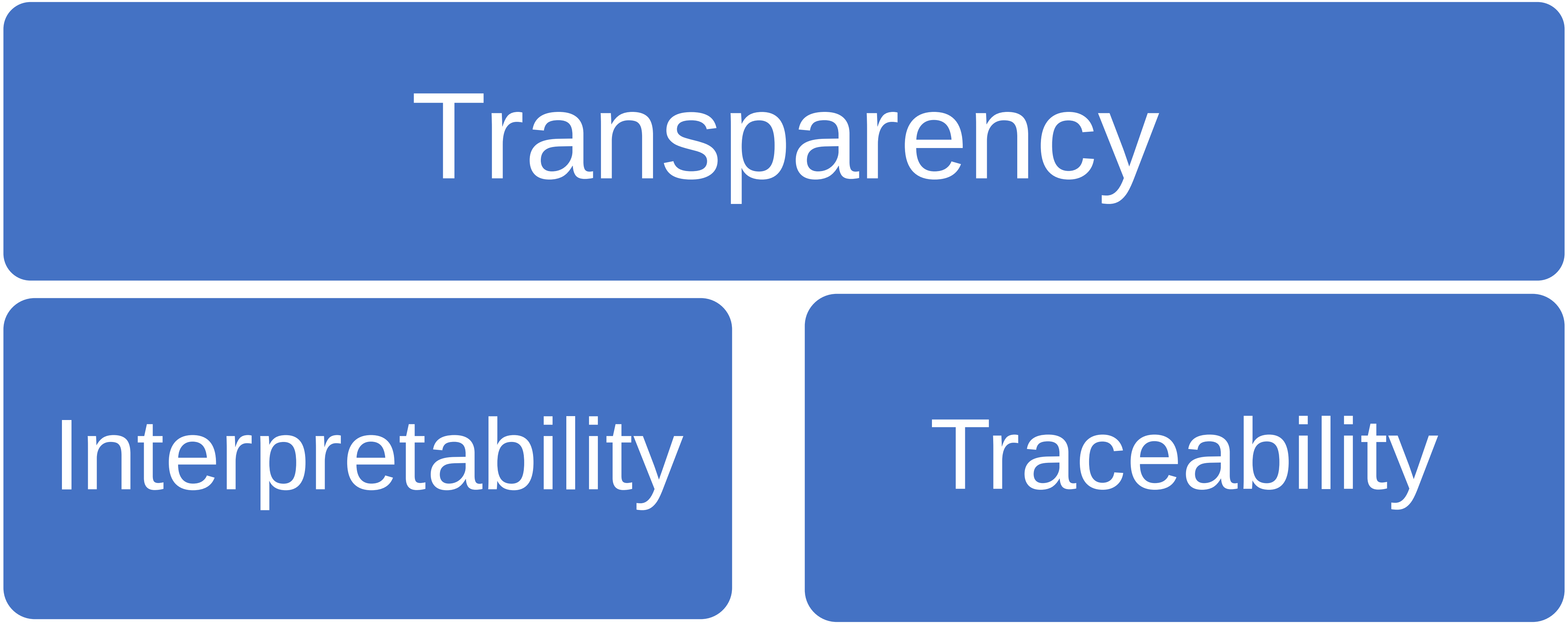
Does AIA support the strategy of innovation and trasparency?

Trust-Building Through Explainability

Ongoing, iterative proces that possesses:

- ➢ assumptions concerning the typical or expected behavior in a given situation type,
- ➢ normative criteria serving as tools of evaluation of either party's behavior and their expectations and
- ➢ appropriate liability rules becoming effective in case of breach of trust

Explainability in AIA



Transparency

Interpretability

Traceability

Main Flaws of the Regulation

- Too much focus on risk management
- Constraints on development
- Neutrality towards different types of AI systems
- Treating AI as a source of problems, not a tool
- Different branches of AI use treated the same

Two contexts – art. 13

- According to **art. 13**
  - AI systems should be "**sufficiently** transparent to enable users to interpret the system's output and use it appropriately, primarily to achieve compliance with risk control measures described earlier.
  - Instructions containing comprehensible information about:
    - the intended purpose of the system
    - its accuracy
    - Robustness
    - cybersecurity,
    - data on the appropriate use of the system and how to interpret its output.

Two contexts – art. 52

- Article 52.1. uses the term "transparency" with regards to informing natural persons that they interact with an AI system (not only high-risk ones). In art. 52.2-52.3. there are some specific obligations for the design of emotion recognition systems or biometric categorization systems, as well as the system used for the creation of so-called "deep fakes".

Place of explainability

- Transparency used in AI Act, in the meaning adopted in art. 13 resembles notion of *interpreatbility*
- Explainability should be understood differently – as providing means to answer *why* the system acts in particular way
- Does transparency include explainability?

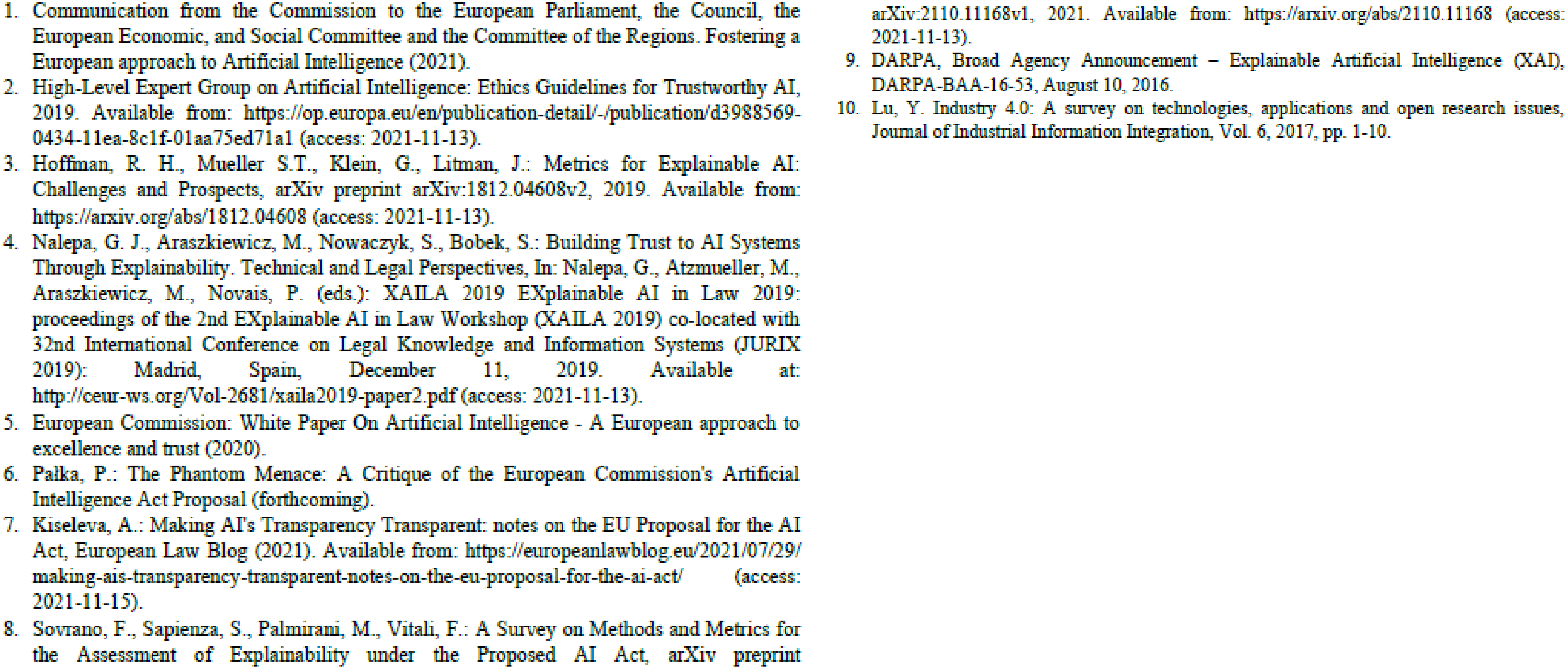
Explainability?

- Used to describe safe and productive AI systems in earlier documents of the EU, especially HLEG on AI (2019) and White Paper on AI (2020)
- What are the relations between explainability, transparency and interpretability?
- Metric proposed by Sovrano, Sapienza, Palmirani and Vitali:
  - Risk-focused,
  - Model-agnostic,
  - Goal-aware,
  - Intelligible&accessible,
  - Compliance-oriented,
  - User-empowering

Further research

- Is ensuring explainability even necessary?
- Exploring possible practical use in industry – branch that seems to be the most suited for AIA regulation
- Could the high level of abstraction of AIA's provisions be helpful for the industry?
- How to adjust means of explainability to specific groups of stakeholders?
- To what extent constraints set by AIA would tamper innovation?

References



1. Communication from the Commission to the European Parliament, the Council, the European Economic, and Social Committee and the Committee of the Regions. Fostering a European approach to Artificial Intelligence (2021).
2. High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI, 2019. Available from: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1 (access: 2021-11-13).
3. Hoffman, R. H., Mueller S.T., Klein, G., Litman, J.: Metrics for Explainable AI: Challenges and Prospects, arXiv preprint arXiv:1812.04608v2, 2019. Available from: https://arxiv.org/abs/1812.04608 (access: 2021-11-13).
4. Nalepa, G. J., Araszkiewicz, M., Nowaczyk, S., Bobek, S.: Building Trust to AI Systems Through Explainability. Technical and Legal Perspectives, In: Nalepa, G., Atzmueller, M., Araszkiewicz, M., Novais, P. (eds.): XAILA 2019 EXplainable AI in Law 2019: proceedings of the 2nd EXplainable AI in Law Workshop (XAILA 2019) co-located with 32nd International Conference on Legal Knowledge and Information Systems (JURIX 2019): Madrid, Spain, December 11, 2019. Available at: http://ceur-ws.org/Vol-2681/xaila2019-paper2.pdf (access: 2021-11-13).
5. European Commission: White Paper On Artificial Intelligence - A European approach to excellence and trust (2020).
6. Pałka, P.: The Phantom Menace: A Critique of the European Commission's Artificial Intelligence Act Proposal (forthcoming).
7. Kiseleva, A.: Making AI's Transparency Transparent: notes on the EU Proposal for the AI Act, European Law Blog (2021). Available from: https://europeanlawblog.eu/2021/07/29/making-ais-transparency-transparent-notes-on-the-eu-proposal-for-the-ai-act/ (access: 2021-11-15).
8. Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F.: A Survey on Methods and Metrics for the Assessment of Explainability under the Proposed AI Act, arXiv preprint arXiv:2110.11168v1, 2021. Available from: https://arxiv.org/abs/2110.11168 (access: 2021-11-13).
9. DARPA, Broad Agency Announcement – Explainable Artificial Intelligence (XAI), DARPA-BAA-16-53, August 10, 2016.
10. Lu, Y. Industry 4.0: A survey on technologies, applications and open research issues, Journal of Industrial Information Integration, Vol. 6, 2017, pp. 1-10.

Thank you for attention!

radoslaw.palosz@uj.edu.pl
araszkiewicz.m@gmail.com
grzegorz.j.nalepa@uj.edu.pl

Supported by the XPM Project (Explainable Predictive Maintenance) under CHIST-ERA 2019 Call
(Project No. 857925, NCN 2020/02/Y/ST6/00070)