Explainable AI and Regulatory Compliance under the AI Act

Davide Marchi

Topic 7
Team A

November 2024

Contents

1 AI Explainability in Industry 5.0

1.1 Overview of AI Explainability

Artificial Intelligence (AI) forms an integral component of Industry 5.0[1], which seeks to augment production efficiency, optimize decision-making, and facilitate the integration of human-centric technologies. Explainability in AI pertains to the capacity of AI models to provide transparent, interpretable, and comprehensible rationales for their decisions and predictions. This characteristic is indispensable for fostering stakeholder trust, ensuring transparency, and complying with stringent regulatory frameworks. Within the industrial landscape, where sophisticated systems interface to govern production processes, explainability becomes a fundamental requirement to mitigate risks, streamline operations, and empower operators with actionable insights.

In the present project, AI explainability is pivotal in ensuring that the decision-making mechanisms of machine learning models are rendered transparent, interpretable, and compliant with regulatory directives such as the European Artificial Intelligence Act (EU AI Act)[2] and various ethical AI standards. The explainability requirements are designed to ensure that all stakeholders, including operators, decision-makers, and external auditors, possess a comprehensive understanding of the AI's operational logic, particularly in high-risk and mission-critical industrial environments.

1.2 Scope of AI Explainability in the Project

This project aims to develop an AI-driven platform that integrates multiple data streams, automates Key Performance Indicator (KPI) analysis, and supports real-time decision-making through generative AI methodologies. To achieve these objectives, three core machine learning components require specific considerations for explainability:

1. **Productivity Forecasting Model**: This model is responsible for forecasting machine productivity to identify and preempt potential production bottlenecks. Explainability is essential to ensure that operators can understand the underlying factors driving these predictions, thereby enabling them to undertake informed, proactive interventions based on sound, data-driven insights.

2. **Energy Cost Outlier Detection Model**: This model aims to identify anomalous patterns in energy expenditures, with the goal of recognizing abnormal spikes. Providing transparent explanations for these classifications will enable to discern the underlying reasons for outlier detection, ensuring that corrective actions are both judicious and justifiable in their application.

3. **Retrieval-Augmented Generation (RAG) System for Personalized Dashboards**: Utilizing a large language model (LLM), this system

facilitates the generation of personalized dashboards and tailored recommendations based on user-specific inputs and data retrieval. Explainability in this context is imperative to instill confidence in users regarding the validity of the generated reports, to elucidate why specific KPIs or visualizations were prioritized, and to enhance the reliability of these dashboards for strategic decision-making.

1.3 Importance of Explainability for Compliance and Trust

Explainability is instrumental in ensuring adherence to regulatory frameworks such as the EU AI Act, which stipulates requirements for transparency and accountability. Additionally, the General Data Protection Regulation (GDPR)[3] mandates that individuals have the right to obtain ”meaningful information about the logic involved” in any automated decision-making processes that significantly affect them. Ensuring a high level of explainability across all AI models not only satisfies these regulatory requirements but also reinforces ethical AI practices by mitigating biases, fostering fairness, and safeguarding the responsible deployment of AI technologies.

Moreover, explainability plays a critical role in enhancing user trust and adoption of AI systems by demystifying the decision-making processes. In the context of manufacturing, operators and decision-makers must trust that AI-generated recommendations are precise and reflective of the true state of production environments. The absence of such trust may result in AI solutions being underutilized or misapplied, potentially leading to operational inefficiencies or non-compliance with regulatory mandates.

1.4 Objectives of Explainability Requirements

The primary objectives of the AI explainability requirements for this project are as follows:

- **Transparency**: Ensure that every prediction and recommendation generated by the AI models is traceable to its logical basis and the contributing input factors.
- **User Comprehension**: Make the decision-making processes intelligible to both technical users and non-technical stakeholders, ensuring broad accessibility of model rationale.
- **Compliance**: Achieve conformity with legal and regulatory standards, such as those established by the EU AI Act and GDPR, particularly concerning the elucidation of automated decision-making processes.
- **Operational Decision Support**: Deliver actionable insights that are comprehensible to operators, thereby facilitating timely and effective interventions—especially for mitigating productivity bottlenecks or addressing anomalous energy cost spikes.

2 Overview of the AI Act and Risk Levels

2.1 The European Union AI Act

The EU AI Act is a comprehensive regulatory framework designed to ensure the safe and ethical use of artificial intelligence across the European Union. It classifies AI systems into different risk categories based on their potential impact on fundamental rights, safety, and overall societal well-being. These categories range from unacceptable risk to minimal risk, each with specific compliance obligations.

2.2 Risk Levels Defined by the AI Act

The AI Act classifies AI systems into the following risk levels:

1. **Unacceptable Risk**: AI systems that manipulate human behavior, use real-time biometric identification in public spaces, or are used for social scoring fall under this category and are banned unless exemptions apply. These systems are considered too dangerous due to their potential to harm fundamental rights.

2. **High Risk**: AI systems that pose significant threats to health, safety, or fundamental rights are categorized as high risk. Examples include AI used in healthcare, recruitment, law enforcement, or critical infrastructure. Such systems must meet stringent requirements for transparency, accuracy, safety, and undergo continuous evaluation.

3. **General-Purpose AI**: Added to the AI Act in 2023, this category covers foundation models such as large language models (e.g., ChatGPT[4]). Depending on their impact, these models are subject to transparency obligations, especially when their capabilities could present systemic risks.

4. **Limited Risk**: AI systems that are not high risk but still require transparency measures fall into this category. These include applications like generating or manipulating content (e.g., deepfakes). These systems must inform users that they are interacting with an AI system and provide clear information about its capabilities.

5. **Minimal Risk**: Most AI applications, such as those used for video games or spam filters, are considered minimal risk. These systems are not regulated under the Act but are encouraged to follow a voluntary code of conduct for responsible use.

2.3 Risk Level Assessment for Project Models

In our project, the AI models developed fall into limited risk or minimal risk categories under the AI Act:

- **The Retrieval-Augmented Generation (RAG) System** for generating personalized dashboards is categorized as **limited risk**, as we are using a pretrained LLM rather than training a new one. It generates content based on user inputs, and while it offers valuable insights, it must maintain transparency to ensure user trust.

- **The Energy Cost Outlier Detection Model** and the **Productivity Forecasting Model** are classified as **minimal risk**. These models provide insights and notifications rather than making autonomous decisions, meaning they pose a minimal risk to fundamental rights or safety.

3 AI Act Compliance Requirements

3.1 The importance of the EU AI Act in our setting

With the foundational understanding of AI risk levels established in the previous chapter, we now turn to the specific compliance requirements relevant to our project. In the context of **Industry 5.0**, AI systems must navigate regulatory demands with the added complexity of managing real-time, high-stakes production environments. Compliance with the EU AI Act requires that AI models employed in industrial settings, whether they are optimizing productivity, detecting anomalies, or generating decision-support dashboards, adhere to stringent standards of transparency and accountability. This chapter focuses on the specific compliance requirements that pertain to the models in our project, ensuring that each is fully aligned with EU regulatory expectations.

3.2 Forecasting Model Compliance Requirements

The **Productivity Forecasting Model** plays a crucial role in predicting machine output and identifying potential production bottlenecks before they materialize. Given its role in a production environment, this model can have significant implications for operational efficiency and safety. To comply with the **EU AI Act**, even if considerable a minimal risk model, it must meet the following requirements:

- **Transparency and Interpretability**: The Productivity Forecasting Model must provide clear insights into the factors driving productivity predictions. To ensure compliance, it must offer interpretable justifications for each forecast, detailing the impact of the most relevant features that contributed to the prediction.

- **Human Oversight**: Operators must have sufficient insight into the predictions to supervise and have the final word on whether to act and how depending on the forecast. This implies that human operators must understand when and why the model predicts a bottleneck, ensuring they can intervene effectively.

- **Accuracy and Robustness**: The model must demonstrate accuracy in its predictions across various operational contexts. Compliance also requires frequent testing to assess robustness under changing conditions, ensuring that the model's recommendations are both reliable and actionable.

- **Documentation and Traceability**: The development, training, and decision logic of the model must be well-documented. This documentation is crucial for external auditing and for ensuring transparency, enabling stakeholders to understand how decisions are being made.

3.3 Outlier Detection Model Compliance Requirements

The **Energy Cost Outlier Detection Model** is responsible for detecting unusual spikes or anomalies in energy consumption, thereby helping companies reduce costs and optimize energy use. Given that this model only provides insights and does not directly make decisions, it is considered to be of minimal risk under the **EU AI Act**. Since the output from this model informs operational adjustments that could lead to financial repercussions, its compliance requirements include:

- **Explanation of Anomalies**: To align with the transparency requirements of the EU AI Act, the model must clearly explain why certain spikes are classified as outliers. It must provide comprehensible rationale regarding which variables contributed to the detection of an anomaly.

- **User-Friendly Interpretability**: The model must present its findings in a manner that is accessible to both technical and non-technical stakeholders, such as energy managers. Compliance involves ensuring that the outlier explanations are interpretable without extensive technical knowledge, thereby allowing stakeholders to validate the anomalies and take appropriate actions.

- **Bias and Fairness**: The model must undergo regular assessments for potential biases, ensuring that it does not systematically misinterpret normal variations as outliers based on incorrect assumptions or flawed data patterns. Bias reduction is critical in maintaining fair and justifiable outputs.

3.4 RAG System Compliance Requirements

The **RAG System for Personalized Dashboards** utilizes a large language model (LLM) to generate customized dashboards and recommendations based on user inputs and retrieved data. Its purpose is to provide tailored insights that assist operators and decision-makers in understanding and optimizing production processes. Given its scope and nature, this model is considered limited risk under the EU AI Act and must adhere to the following compliance requirements:

- **Transparency of Generated Content**: The RAG system must provide transparency into how information was retrieved and why specific elements were chosen for the generated dashboards. This includes explaining which data sources were utilized, what parameters influenced the retrieval, and how final recommendations were made. Users must be able to trace the logic behind the dashboard's layout and KPI prioritizations.

- **Human Oversight and Explainability**: Decision-makers using the RAG system must have adequate oversight capabilities to verify the system's generated dashboards. The system should provide tools to help users understand the rationale behind the generation process, particularly for critical recommendations that may significantly affect production outcomes.

- **Mitigation of Bias**: Given that the RAG system utilizes a generative model, there is an inherent risk of biases stemming from the underlying training data. The model must implement bias detection and mitigation measures to ensure the generated content is impartial, especially in cases where biased information could lead to unfair or suboptimal operational decisions.

- **Compliance with User Rights**: The system must ensure compliance with user rights under GDPR and the EU AI Act, especially when using personal or sensitive data for generating personalized insights, even if our current requirements do not mandate it. It must provide users with access to information about the logic involved in content generation, allowing them to understand and, if necessary, contest the outcomes.

- **Clarity on AI Interaction**: Even though users will be aware that they are interacting with an AI system and not a real person, it is important that the model explicitly clarifies this in its responses. Such transparency is also a requirement for limited risk AI under the EU AI Act, helping to set accurate expectations and improve user trust in the system.

4 Available Techniques and Tools

4.1 Ensuring Explainability in AI Models

In this chapter, we present the techniques and tools necessary to ensure our AI models comply with the transparency and accountability requirements outlined in the previous chapter. Each model developed for this project is analyzed individually, with a focus on specific explainability needs and best practices achieved through state-of-the-art tools and methodologies.

4.2 Techniques and Tools for the Forecasting Model

The **Productivity Forecasting Model** plays a crucial role in predicting machine output and identifying potential bottlenecks before they occur. The model must also display a **confidence score** along with its predictions to provide stakeholders with a clear understanding of the reliability of each forecast. To ensure compliance with the EU AI Act, we need to apply explainability techniques that make the predictions comprehensible to stakeholders. Below are the specific techniques and tools used to achieve this:

- **SHAP (Shapley Additive Explanations)**: SHAP[5] is a powerful model-agnostic technique that assigns each feature a contribution value towards a prediction. By applying SHAP, we can provide both global and local interpretability to explain how different factors impact the productivity forecast.

  - **Tool**: We recommend using the SHAP library[6] in Python (compatible with PyTorch[7]), which provides robust visualization capabilities for understanding feature importance across the model's predictions.

- **Partial Dependence Plots (PDPs)**: PDPs help visualize the relationship between one or two features and the predicted outcome, allowing stakeholders to understand the influence of particular features on the model's behavior.

  - **Tool**: The scikit-learn[8] library in Python can be used to generate PDPs[9], offering easy integration with the rest of the machine learning workflow.

- **Integrated Gradients**: For models built using neural networks, Integrated Gradients[10] can help assess the importance of input features relative to the prediction. This method is particularly useful for ensuring the model's decision-making process aligns with expert knowledge in the production domain.

  - **Tool**: The Captum[11] library from PyTorch is an excellent tool to implement Integrated Gradients and visualize feature attributions in the productivity forecasting model.

4.3 Techniques and Tools for Outlier Detection

The **Energy Cost Outlier Detection Model** is designed to identify anomalies in energy consumption, enabling companies to make more efficient operational adjustments. This model must also display a **confidence score** alongside its predictions to inform stakeholders of the reliability of identified outliers. The following techniques and tools are recommended to explain the model's decisions and ensure compliance:

- **LIME (Local Interpretable Model-agnostic Explanations)**: LIME[12] is effective for generating localized explanations of why certain energy consumption data points were classified as outliers. It helps stakeholders understand specific predictions by perturbing the input and observing the resulting changes in output.
  - **Tool**: The LIME package[13] in Python can be used to create local explanations, and it integrates seamlessly with models developed in PyTorch.
- **Feature Importance Analysis**: This technique identifies the features that most influence the detection of anomalies. By providing insight into variables such as machine type or historical consumption patterns that contribute to outlier detection, stakeholders can validate the model's reliability
  - **Tool**: The Yellowbrick library[14] in Python is suitable for generating feature importance visualizations and is compatible with the typical machine learning frameworks used.
- **Isolation Forests**: In some cases, using techniques like Isolation Forests can help identify anomalous patterns and provide an intuitive explanation of how outliers are isolated within the data structure.
  - **Tool**: scikit-learn can be used to implement Isolation Forests[15], which can then be combined with visualization tools to further enhance interpretability.

4.4 Techniques and Tools for the RAG System

The **Retrieval-Augmented Generation (RAG) System** for personalized dashboards uses a large language model (LLM) to generate insights and recommendations based on user-specific data. Given its content generation capabilities, ensuring transparency and user-rights compliance is paramount. It is also important that the system provides information on the sources of retrieved data, explicitly showing where the data was taken from to build trust and clarify the origin of the information used. The following techniques and tools are recommended for this purpose:

- **Attention Mechanisms**: Attention mechanisms help identify which parts of the input are most influential in the model's generation process, providing a layer of interpretability for LLMs. This is particularly important in explaining why specific data points were highlighted in the dashboards.
  - **Tool**: PyTorch offers built-in capabilities for implementing and visualizing attention mechanisms, making it suitable for adding transparency to the RAG model.

- **Counterfactual Explanations**: Counterfactual explanations involve modifying input data slightly to observe how it affects the outcome. This helps users understand what changes would have led to different recommendations, enhancing trust in the system's outputs.
  - **Tool**: The AI Explainability 360 toolkit[16] can be used to generate counterfactual explanations for models. This tool helps visualize the alternative scenarios, providing end-users with actionable insights.
- **SHAP for LLMs**: Even for generative models, SHAP can be adapted to provide insights into feature importance during the retrieval and generation stages. This helps explain why certain KPIs or recommendations were prioritized over others.
  - **Tool**: The SHAP library can be used in combination with other tools to generate explanations, providing stakeholders with a clear understanding of model decisions.

5 Conclusions: Compliance and XAI Approaches

5.1 Summary of Compliance Requirements

Ensuring compliance with the EU AI Act is critical for deploying AI models in Industry 5.0, where real-time decision-making and safety are key. This report has outlined the risk levels under the EU AI Act and the compliance needs of the models in this project: Productivity Forecasting, Energy Cost Outlier Detection, and the RAG System. These requirements focus on transparency, accountability, fairness, robustness, and human oversight to foster user-friendly, explainable, and compliant AI systems.

5.2 Summary of XAI Techniques and Suggested Tools

To meet compliance requirements, various Explainable AI (XAI) techniques and tools are suggested to make models interpretable for stakeholders:

- **Productivity Forecasting Model**: Techniques include SHAP, Partial Dependence Plots (PDPs), and Integrated Gradients. Tools like SHAP Library, scikit-learn, and Captum are recommended for transparency into predictions, offering both local and global explanations.
- **Energy Cost Outlier Detection Model**: LIME, Feature Importance Analysis, and Isolation Forests are recommended for explaining flagged anomalies. Tools like LIME Package, Yellowbrick, and scikit-learn help generate clear visualizations and explanations.
- **RAG System**: Techniques like Attention Mechanisms, Counterfactual Explanations, and SHAP help explain generated insights. Tools such as

PyTorch, AI Explainability 360, and SHAP Library provide transparency regarding data sources and feature influences.

These methods support compliance by ensuring AI-driven outcomes are explainable and align with EU AI Act requirements.

5.3 Recommendations for Implementation

For effective implementation of these techniques, here are our recommendations:

1. **Early Integration of Explainability Tools**: Use tools like SHAP, LIME, and Attention Mechanisms from the start to ensure transparency throughout the model development lifecycle.

2. **User-Centric Interface**: Develop user-friendly interfaces to visualize feature attributions, accuracy scores, and data sources, aiding decision-makers.

3. **Regular Validation for Fairness**: Periodically assess models for biases, especially the RAG System and Outlier Detection Model. Tools like AI Explainability 360 can help evaluate fairness and reduce biases.

4. **Human Oversight and Documentation**: Ensure thorough documentation and provide training for using tools like Captum and SHAP. Clear documentation will also assist while checking for compliance.

5. **Transparency of Data Sources**: For the RAG System, be clear about data sources used to generate the dashboards to build user trust.

6. **Confidence Scores with Predictions**: Ensure that the Productivity Forecasting Model and Energy Cost Outlier Detection Model provide a confidence score alongside their predictions.

7. **Continuous Monitoring and Improvement**: Maintain ongoing monitoring with feedback loops to improve explanations based on user needs. Update explainability tools regularly to align with model evolution.

5.4 Final Thoughts

This project outlines the approach to building explainable, ethical AI models that comply with the EU AI Act. By adopting the suggested techniques and tools, our AI solutions aim to be compliant, transparent, and aligned with Industry 5.0 principles. Focusing on transparency, user rights, and explainability fosters a trustworthy AI ecosystem, setting the stage for future ethical and effective AI applications.

References

[1] European Union. *Industry 5.0 - What this approach is focused on, how it will be achieved and how it is already being implemented.* Official website of the European Union. URL: https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry-50_en.

[2] European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).* Official Journal of the European Union, 12 July 2024. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[3] European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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).* Official Journal of the European Union, L 119, 4 May 2016. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[4] OpenAI. *ChatGPT.* Accessed: 2024-11-04. 2024. URL: https://openai.com/chatgpt.

[5] Scott Lundberg. *SHAP Documentation.* Accessed: 2024-11-04. 2024. URL: https://shap.readthedocs.io/en/latest/.

[6] Scott Lundberg. *SHAP (SHapley Additive exPlanations).* Accessed: 2024-11-04. 2023. URL: https://github.com/slundberg/shap.

[7] PyTorch Team. *PyTorch: An Open Source Machine Learning Framework.* Accessed: 2024-11-07. 2023. URL: https://pytorch.org.

[8] Scikit-Learn Developers. *scikit-learn: Machine Learning in Python.* Accessed: 2024-11-04. 2023. URL: https://scikit-learn.org.

[9] Scikit-Learn Developers. *scikit-learn: Machine Learning in Python - Partial Dependence Plots.* Accessed: 2024-11-05. 2023. URL: https://scikit-learn.org/stable/auto_examples/inspection/plot_partial_dependence.html.

[10] Captum Team. *Integrated Gradients - Captum Documentation.* Accessed: 2024-11-05. 2024. URL: https://captum.ai/docs/extension/integrated_gradients.

[11] PyTorch Team. *Captum: Model Interpretability for PyTorch.* Accessed: 2024-11-05. 2023. URL: https://captum.ai.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[13] Marco Tulio Ribeiro. *LIME: Local Interpretable Model-agnostic Explanations*. Accessed: 2024-11-06. 2023. URL: `https://github.com/marcotcr/lime`.

[14] District Data Labs. *Yellowbrick: Visual Analysis and Diagnostic Tools*. Accessed: 2024-11-06. 2023. URL: `https://www.scikit-yb.org`.

[15] Scikit-Learn Developers. *scikit-learn: Machine Learning in Python - Isolation Forest*. Accessed: 2024-11-07. 2023. URL: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html`.

[16] IBM Research AI. *AI Explainability 360*. Accessed: 2024-11-08. 2023. URL: `https://aix360.res.ibm.com/`.