ercim-news.ercim.eu Number 134 July 2023

ERCIM NEWS

Special theme:

eXplainable AI

Also in this issue

Research and Innovation:

Fibre-Optic Sensing for Road-Traffic Monitoring in Remote Areas

Introduction to the Special Theme

Explainable AI

by Manjunatha Veerappa (Fraunhofer IOSB) and Salvo Rinzivillo (CNR-ISTI)

Artificial Intelligence (AI) has witnessed remarkable advancements in recent years, transforming various domains and enabling groundbreaking capabilities. However, the increasing complexity of AI models, such as convolutional neural networks (CNNs) and deep learning architectures, has raised concerns regarding their interpretability and explainability. As AI systems become integral to critical decision-making processes, it becomes essential to understand and trust the reasoning behind their outcomes. This need has given rise to the field of explainable AI (XAI), which focuses on developing methods and frameworks to enhance the interpretability and transparency of AI models, bridging the gap between accuracy and explainability.

Explainable AI Methodology

The lack of transparency in AI models can hinder their effectiveness and introduce potential vulnerabilities. XAI aims to address this challenge by incorporating interpretability techniques into AI models, allowing security analysts and stakeholders to understand the reasoning behind AI-driven decisions. Héder discusses the history and the evolution of the concept of explainability and its relationships with the legal context in Europe (page 9).

Amelio et al. discuss approaches to interpret and compress convolutional neural networks (CNNs), enhancing their interpretability and efficiency (page 10). Zalewska et al. introduce the BrightBox technology, which provides a surrogate model for interpreting the decisions of black-box classification or regression algorithms (page 12). Spinnato et al. present the LASTS framework, which aims to provide interpretability in black- box time series classifiers (page 14).

Explainable AI in Health Care

The healthcare industry has witnessed the integration of AI systems for various purposes, such as medical imaging analysis, disease diagnosis and personalised treatment. However, the lack of interpretability in AI-based decision-making raises concerns regarding trust and accountability. The authors Bruno et al. highlight the need for addressing the black-box problem by developing an ad-hoc built classifier for lung ultrasound images (page 16). The importance of governing and assessing ethical AI systems in healthcare is emphasised by Briguglio et al. (page 18). Rodis et al. introduce the concept of multimodal explainable AI (MXAI) and its relevance to complex medical applications (page 20). Lädermann et al. discuss the use of machine learning methods in detecting surgical outcomes, aiming to improve patient selection for surgery (page 22). Zervou et al. focus on the application of AI and generative models in precision medicine to streamline the drug discovery process (page 23). These approaches aim to enhance trust, improve patient safety, and provide actionable insights to healthcare professionals.

Explainable AI in Industry

AI plays a crucial role in enhancing productivity and efficiency in industrial applications. However, the lack of explainability in AI models hampers their adoption in critical industrial use cases. Brajovic and Huber focus on integrating AI-specific safety aspects into the automotive development process, particularly addressing the challenges associated with AI application in standard software development (page 24). Folino et al. introduce a ticket-classification framework that integrates deep ensemble methods and AI-based interpretation techniques to support customer support activities (page 27). Jalali et al. propose a counterfactual explanation approach for time-series predictions in industrial use cases, enabling interpretable insights into AI models' decisions (page 28).

Explanations for Chatbots

Generative language models are attracting a lot of attention even in non-technical populations. In many cases, the generated text may not return a faithful representation of truth. Thus, the necessity emerges to provide additional evidence of the elements that are included in the text. Mountantonakis and Tzitzikas present GPT•LODS, a prototype that validates ChatGPT responses using resource description frameworks (RDF) knowledge graphs (page 29). Prasatzakis et al. propose an easy-to-understand and flexible chatbot architecture based on the "event calculus" for high-level reasoning (page 31).

Societal Challenges

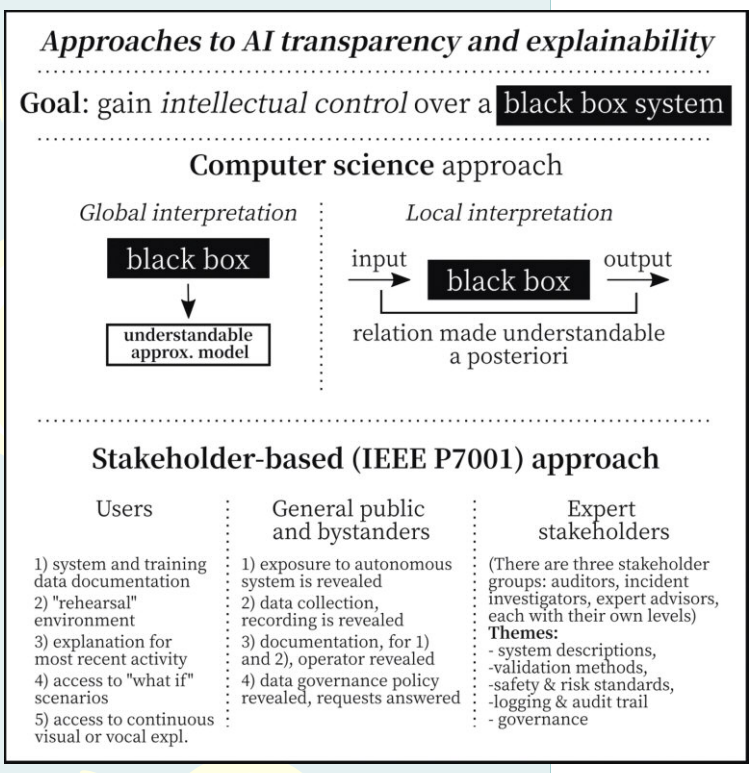
This special issue covers a range of cross-domain applications of XAI that have an impact on several societal challenges, like forest preservation, quality assessment of information and astronomy object detection. It involves developing AI models and systems that can provide transparent and interpretable explanations for their decision-making processes. Jalali and Schindler propose the integration of long short-term memories (LSTMs) with example-based explanations to enhance interpretability in tree-growth models. The aim is to identify critical features impacting outcomes, engage domain experts, address privacy protection, and select appropriate reference models to support informed decision-making in forestry and climate change mitigation (page 33). Ceolin and Qi discuss the design of AI pipelines for automated information quality assessment, which are fully transparent and customisable by end-users. By leveraging reasoning, natural language processing (NLP), and crowdsourcing components, these pipelines enhance transparency, mitigate biases, and aid in the fight against disinformation. Jaziri and Parisot apply XAI techniques to ensure the reliability and absence of bias in deep sky objects classification models used in astronomy.

In conclusion, this special issue showcases several explanation methods and the diverse applications of explainable AI (XAI) across various fields, including healthcare, industry, ethics, climate change, and generative language models. The projects showcased in this issue highlight the importance of transparency and interpretability in complex machine learning models, providing insights into decision-making processes and

empowering stakeholders to understand and trust AI systems. The advancements in XAI contribute to improved diagnostic accuracy, enhanced customer support experiences, ethical AI governance, theoretical developments in model compression and surrogate modelling, interpretability in tree-growth models, integration of AI-specific safety aspects, and combating disinformation. The papers not only provide valuable insights into XAI but also promote further research on XAI, fostering innovation and advancements in understanding AI's internal mechanisms and its impact on various industries.

Please contact:
Manjunatha Veerappa
Fraunhofer Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Germany
manjunatha.veerappa@iosb.fraunhofer.de

Salvo Rinzivillo
CNR-ISTI, Italy
rinzivillo@isti.cnr.it



Explainable AI: A Brief History of the Concept

by Mihály Héder (SZTAKI)

Understandability of computers has been a research topic from the very early days, but more systematically from the 1980s, when human-computer interaction started to take shape. In their book published in 1986, Winograd and Flores [1] extensively dealt with the issues of explanations and transparency. They set out to replace vague terms like "user-friendly", "easy-to-learn" and "self-explaining" with scientifically grounded design principles. They did this by relying on phenomenology and, especially, cognitive science. Their key message was that a system needs to reflect how the user's mental representation of the domain of use is structured. From our current vantage point, almost four decades later, we can see that this was the user-facing variation of a similar idea, but for developers – object-oriented programming, a method on the rise at the time.

The 1980s precedes the now widespread success of machine learning at creating artificial intelligence (AI). In the days of "good old-fashioned" AI, with fewer tools and fewer computational resources, success was built on data structures and logic. These constraints resulted in systems that the creators and adaptors could keep under their intellectual oversight, or at least they knew it was possible to look under the hood and see exactly what was going on.

With machine learning, the designed structures and curated rulesets were replaced by machine-generated models. But, due to the nature of computers, every detail and bit of these models can still be examined easily. This posed a challenge from the terminological point of view: why would we call something a black box (a term Rosenblatt used in the context of artificial neural networks already in 1957, but for a single neuron) if every detail can be readily known? While the word "complexity" is sometimes used – quite confusingly due to its many adjacent meanings – it is more accurate to talk about the lack of understandability or not having adequate explanations about curious behaviour. Understanding is an epistemic value to be achieved by a human investigating a system; therefore, the term "epistemic opacity" [2] was introduced. The opposite of this is then (epistemic) transparency, a feature of a system that affords human understanding and intellectual oversight.

Machine learning, especially deep learning, does not produce models and systems built on these models with this feature, therefore, they create epistemic deficit. Yet, they are here to stay because of their performance. They need to be made transparent, then.