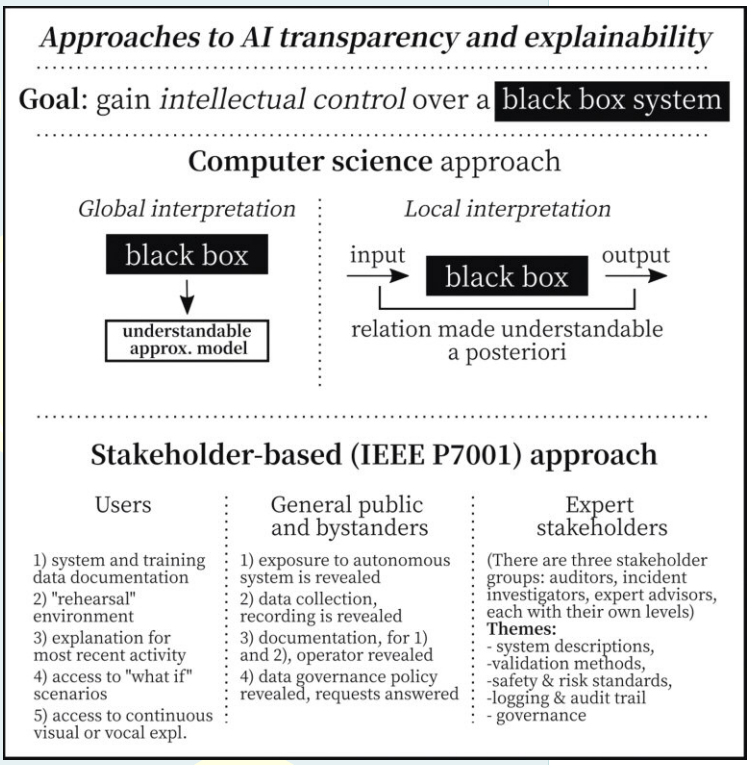
empowering stakeholders to understand and trust AI systems. The advancements in XAI contribute to improved diagnostic accuracy, enhanced customer support experiences, ethical AI governance, theoretical developments in model compression and surrogate modelling, interpretability in tree-growth models, 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.

There are two main strategies to interpret, that is, understand these models: first, the entire model may be interpreted, in which case the resulting explanation is called "global" – continuing the tradition of poor choice of terminology in AI (alternatives could have been "comprehensive", "broad", etc.).

This can be achieved by a surrogate system, which helps by faithfully representing the original model while allowing for simplification, and uses elements that humans easily understand. If such a surrogate is successfully made, the entire model is made transparent. Moreover, we can predict its behaviour to imagined inputs before it happens, providing us with intellectual control. Other global explanations visualise the model or map out concepts used by a model. We can only speculate regarding the etymology, but most probably, this approach is called "global" because it is the apparent linguistic opposite of "local". This word takes us to the second interpretation strategy, local interpretation. The usage of "local" is much better justified by the concept of local fidelity – it means that an explanation is made for one particular output of a system, but in a way that it may be used for similar inputs, where similarity is measured as the distance in a mathematical space. Therefore, we are talking here about true spatial locality.

This epistemic approach to transparency is inevitably relative to the knowledge of the particular persons trying to achieve intellectual oversight. This fact is best engaged by the IEEE P7001 standard draft [3], which is expected to become a harmonized EU standard as a part of the EU AI Act; legislation that makes transparency (and therefore explainability) central.

This approved draft uses a stakeholder-based approach and divides humans into "users", the "general public" or "bystanders" (non-users who may still be affected), and "experts". The last group is further divided into certification agencies and auditors, incident investigators and expert advisers in litigation. This draft is very helpful, as it clarifies that the mathematical method under the XAI umbrella term is for the experts, while user transparency may be created by layperson explanations, like for clustering the term "other users who listened to this also liked the following". Agencies are catered for yet another, more administrative modus of transparency, tuned for accountability.

As transparency is widely believed to be essential to build trust, the methods to achieve it are here to stay, and therefore explainable AI has a long future.

References:

[1] T. Winograd, F. Flores and F. F.Flores, Understanding Computers and Cognition: A New Foundation for Design, Intellect Books, 1986.
[2] M. Héder, "The epistemic opacity of autonomous systems and the ethical consequences," AI & Society, 1–9, 2020.
[3] A. F. T. Winfield, et al., "IEEE P7001: A proposed standard on transparency," Frontiers in Robotics and AI, vol. 8, 665729, 2021.

Please contact:
Mihály Héder, SZTAKI, Hungary
mihaly.heder@sztaki.hu

A Multilayer Network-Based Approach for Interpreting and Compressing Convolutional Neural Networks

by Alessia Amelio, Gianluca Bonifazi, Domenico Ursino and Luca Virgili (Polytechnic University of Marche)

We propose an approach to map a convolutional neural network (CNN) into a multilayer network. It allows the interpretability of the internal structure of deep learning architectures. Then, we use this representation to compress the CNN.

Researchers have recently become more aware of the necessity to scale back the size and complexity of deep neural networks. As a result, a number of techniques are being suggested to shrink the size of current networks without significantly impacting their performance. Exploring the many layers and components of a deep learning model is crucial in order to achieve this goal. In fact, one could pinpoint the most important components, the most relevant patterns and features, the information flow and so on. We want to make a contribution in this setting by proposing a new way of interpreting and exploring a CNN through a multilayer network representation of it, which is then used for compressing it [1].

We operate under the assumption that deep learning networks may be represented, analysed, explored and otherwise greatly supported by complex networks, particularly multilayer ones. Accordingly, we first introduce a method to transform deep learning networks into multilayer ones and then exploit the latter to explore and manipulate the former. Our study focuses on the CNN, which is a specific type of deep learning network widely adopted in different fields, especially computer vision; however, it can easily be extended to other kinds of deep learning networks. The multilayer network is a particular graph-based data structure composed of different layers. Each layer represents a graph with a specific type of connection among the nodes. Multilayer networks are a type of complex networks sophisticated enough to represent all the main components of a CNN. In fact, all the typical elements of a CNN (i.e. nodes, connections, filters, weights, etc.) can be represented through the basic components of a multilayer network (i.e. nodes, arcs, weights and layers). Once the representation of the CNN by the multilayer network has been obtained, the latter is adopted to explore and manipulate the former. To prove its potential, we use this representation to provide a method for removing unnecessary convolutional layers from a CNN. This method looks for layers in the CNN that can be pruned without significantly affecting the CNN performance and, if it finds any, it goes ahead and removes those layers, returning a new CNN [1].

More specifically, mapping the CNN into a multilayer network is performed in different steps (see Figure 1). In the first step, the CNN is trained from a database of images labelled with