Explainable AI – Explanations in AI

Käte Hamburger Kolleg: Cultures of Research
RWTH University, Aachen
2th - 3th February 2022

Understanding Explanatory Structures in Computer-Assisted Proofs– Explanation and Understanding together Again?
Daniel Wenz (RWTH Aachen)
Wed. 2th Feb, 13:00–13:40

Abstract: In my presentation, I will apply a fairly recent discussion about the relationship between explanation and understanding from the philosophy of mathematics to the field of explainable AI. At the beginning of the last decade, there emerged a new interest in the distinction between explanation and understanding in the philosophy of mathematics. Although not always clear-cut, there is a tendency to differentiate both by referring to their respective objective or subjective natures (de Regt 2009). Roughly speaking, being explanatory is deemed as the feature of a structure like a proof or some other kind of justification. It is in this sense an explanation is objective - it is the property of an object. From this perspective, to make something explainable is to transform something that lacks this property into something that does not. Understanding on the other hand is seen as something that is characterized by changes that occur in a subject when it grasps something. It is therefore seen as something subjective. One of the most prominent approaches in this direction originates from the ability approach to understanding (Avigad 2008): According to this, the changes that occur in a subject that understands something concern primarily not a set of beliefs the subject holds but a set of abilities the subject acquires. This is not the trivial thesis that the subject acquires some abilities by understanding something, but the more radical view that understanding this very something just is acquiring those abilities. I will try to disentangle some strains of the nexus the mentioned approaches to the concepts of explanation and understanding are caught up in and highlight some of the uncovered interdependencies. I then will apply the result to some examples in explainable AI.

Theorem Proving in Deep Artificial Neural Networks
Markus Pantsar (University of Helsinki, c:o/re Aachen)
Wed. 2th Feb, 13:40–14:20

Abstract: Computer assisted theorem proving is an increasingly important part of mathematical methodology. However, the theorem proving programs show little or no intelligence: they are not able to discriminate interesting theorems and proofs from trivial ones. In order for computers to develop intelligence in theorem proving, there would need to be a radical change in how the software functions. Recently, machine learning results in solving mathematical tasks have shown promise that deep artificial neural networks can learn symbolic mathematical processing. In this paper, I analyse the possibility that such neural networks could develop genuine intelligence in theorem proving.

Peirce as a Philosopher of AI
Frederik Stjernfelt (Aarhus University, c:o/re Aachen)
Wed. 2th Feb, 14:50–15:30

Charles Peirce the pragmatist and semiotician, had strong views about the abilities of computers. This may be surprising to the extent that he died in 1914, long before Turing or Church entered the scene. Still, he played a tiny if significant role in the history of the computer – he was probably the first to propose a computer constructed from electrical circuits, in the 1880's. More importantly, machine and animal intelligence provided a constant comparative backdrop to his reflections on the characteristics of human cognition.

Machine Learning - An Emerging Experimental Culture in Science
Gabriele Gramelsberger (RWTH Aachen University & c:o/re Aachen)
Thur. 3th Feb.,10:30–11:10

Abstract: Since a few years machine learning (ML) algorithms are invading science, tremendously changing the style of research. Considering what research has experienced in the 1970s and proliferating in the 1990s about numerical experiments based on computer simulation the current situation is comparable: a new experimental culture is emerging. However, it is a computational culture using a new form of computational experiments based on artificial neural networks. The paper takes a closer look to this new experimental culture in case of material science.

Explainability in Reinforcement Learning
Lincoln V. Schreiber, Gabriel de O. Ramos (Universidade do Rio dos Sinos, UNISINOS), and Ana L. C. Bazzan (Universidade Federal do Rio Grande do Sul)
Thur. 3th Feb.,11:10–12:00
Contribution in two parts. A video and a talk by Ana L.C. Bazzan

Part I: *Explainable AI for Traffic Signal Control*
Authors and Presenter: **Lincoln V. Schreiber, Gabriel de O. Ramos and Ana L. C. Bazzan**
Abstract: Artificial intelligence systems are increasingly becoming part of our daily lives and have been shown to outperform humans in a range of complex tasks. However, most machine learning (ML) models can be seen as black boxes, making it challenging to explain why or how they came to specific conclusions. Consequently, when employing ML models, one needs to consider the model interpretability vs. performance trade-off. In fact, the widespread adoption of ML in the real world has accelerated research on Explainable Artificial Intelligence (XAI), which focuses on bringing transparency and interpretability to AI models. In this talk, we present an overview of XAI, focusing on different explainability approaches. Finally, we discuss a case study on XAI applied to traffic signal control.

Part II: *XRL: Not as Critical as XAI ?*
Ana Bazzan

Abstract: Machine learning techniques can be classified in three categories: supervised learning, unsupervised learning, and reinforcement learning (RL). While these three are often mixed (e.g., the use of artificial neural network (ANN) for function approximation in RL or clustering methods to improve classification or to cluster RL experiences), here I focus on ANN and RL. The former is being criticized for being a "black box" and thus opaque to humans, even to experts in a given application or domain. Currently, this is motivating an agenda on explanaible AI (XAI). On the other hand, pure RL (i.e., when RL techniques are not combined with ANN) are able to create a model that is more accessible to humans. This is due to the fact that an agent that uses RL ends up learning a policy (a model) that maps states to actions. Although a policy cannot always be traced to the history of interactions that has led to it, it is nevertheless less opaque to human experts in the sense that it quantifies the value of a state-action pair.

Absolute Limits of Mathematical Modeling in AI
Jobst Landgrebe (Cognotekt, Cologne)
Thur. 03th Feb.,13:30–14:10

Abstract: Three questions are of central interest when thinking about AI:

- – What are the essential marks of human intelligence? – What is it that researchers are trying to do when they talk of achieving ‘Artificial Intelligence’ (AI)?
- To what extent can AI be achieved? The book brings together results from mathematics, physics, computer science, philosophy, linguistics, and biology. Its core argument is that an artificial intelligence with powers of a sort that would equal or exceed human intelligence – sometimes called general artificial intelligence (AGI) – is for mathematical reasons impossible. The reasons are that

1. intelligence of this sort is a capability of a complex dynamic system, and such systems cannot be modelled mathematically in a way that yields exact predictions;
2. but only what can be modelled mathematically in this way can be engineered to operate inside a computer; as we will see in the Introduction and in the final chapter, there is a great deal which AI can achieve that will be of benefit to mankind; but it does not include the work that a human intelligence can do; it does not include AI systems more powerful than humans; and it does not include AI systems which are ‘evil’ in any sense of this word.

One consequence of our argument is that much of what is discussed in the wider world concerning the potential of AI to bring about radical changes in the very nature of human beings and of the human social order is founded on an unfortunate error.

Explainable AI = transparent AI?
Andreas Kaminski (RWTH Aachen)
Thur. 3th Feb, 14:10–14:50

Abstract: A number of advanced forms of machine learning are resulting in models that are largely opaque. In response, research on explainability of AI models has emerged. Nonetheless, it is far from obvious whether explainable AI also leads to models being transparent. The talk will explore this question.

Making Sense of Intelligent Systems
From Conception Practices to Interaction Studies
Joffrey Becker (Collège de France, c:o/re Aachen)
Thur. 3th Feb.,14:50–15:30

Abstract: In a text wrote in 1988, Susan Leigh Star noted that artificial intelligence research pursues two main goals. The first is to understand intelligence by simulating biological functions, and the second is to produce intelligible objects that can easily be used by humans. It seems therefore important to grasp these two dynamics of the animation of objects (which is technical and mental) in order to better understand the relationships in which machines place us today and the role that they are intended to see us play. The aim of this contribution is to explore these two aspects jointly as they illustrate two different ways to deal with intelligence. The presentation will thus focus on both the conception of artificial life forms and the interactions we have with them. It will pay a particular attention to the analogies which
characterize intelligent systems design and to the mental processes they give rise to. Leaning on ethnographic case studies, the contribution will try to show that intelligent systems are devices of a recursive kind which convey norms, representations and various ideas about the life processes and the social relations they intend to imitate.