Spring School & Workshop on

Ethics of AI
(Un-)Explainability

Philosophical-conceptual questions meet applications

Universität Münster

19 and 20 March 2024,
Institute for Theoretical Physics,
Lecture room 404

Program & Abstracts

Organizers:

Dr. Paul Näger, Department of Philosophy
Dr. Katrin Schmietendorf, CeNoS and InterKI

Center for Nonlinear Science

INTER KI WWU

ZfW Zentrum für Wissenschaftstheorie

Program

Tuesday, 19th March

Spring School

09.00 – 09.10	Opening
09.10 – 10.30	Prof. Dr. Benjamin Risse (Institute for Geoinformatics, UM): *Unintuitive? Yes. — Intelligent? No! From poorly chosen scientific terminology to superfluous AI questions*
	— coffee break —
11.00 – 12.20	Dr. Stefan Roski (ZfW & University of Hamburg): *Explanations and Explainability*
	— lunch break —
14.00 – 15.20	Dr. Paul Näger (Department of Philosophy, UM): *Basics of AI Ethics*
	— coffee break —
16.00 – 17.30	Colloquium lecture:
	Prof. Dr. Gitta Kutyniok (LMU, München): *A Mathematical Perspective on Legal Requirements of the EU AI Act: From the Right to Explain to Neuromorphic Computing*

Wednesday, 20th March

Workshop

09.00 – 09.10	Opening
09.10 – 10.10	Carlos Zednik, PhD (TU Eindhoven): *Disentangling XAI Concepts: Explanation, Interpretation, and Justification*
	— coffee break —
10.40 – 11.40	Prof. Dr. Eva Schmidt (TU Dortmund): *Reasons of AI System*
11.40 – 12.40	Dr. Astrid Schomäcker (University of Bayreuth): *That's not fair! Explainability as a means to increase algorithmic fairness*
	— lunch break —
14.10 – 15.10	Prof. Dr. Kristian Kersting (TU Darmstadt): *Where there is much light, the shadow is deep. XAI and Large Language Models*
	— coffee break —
15.40 – 16.40	Prof. Dr. Florian Boge (TU Dortmund): *Put it to the Test: Getting Serious about Explanations in Explainable Artificial Intelligence*
16.40 – 17.40	Dr. Thomas Grote (University of Tübingen): *The Double-Standard Problem in Medical ML Solved: Why Explainability Matters*

Abstracts

Prof. Dr. Benjamin Risse:
Unintuitive? Yes. — Intelligent? No! From poorly chosen scientific terminology to superfluous AI questions

Artificial Intelligence (AI) is one of the most discussed scientific technologies of our times. More and more areas of life will be affected by AI so that there is a growing number of questions regarding the capabilites, reliability and trustworthiness of this technology. For example, is it possible to explain the result of AI computations? And what are the legal implications of these supposedly intelligent systems?

In my presentation I will give an easy-to-follow introduction how AI systems derive their results and I will argue that, despite being unintuitive, no intelligence can be found along the potentially complex chain of computations. Based on this absence I hypothesise that many current discussions are potentially rooted in, or at least influenced by a poorly chosen AI terminology such as 'artificial reasoning', 'decision systems', 'intelligent algorithms', 'learning' and 'explainable AI'. This hypothesis will be contextualized by several state-of-the-art examples in order to enable a critical yet more pragmatic perspective on some of the recent breakthroughs such as large language models (e.g. ChatGPT).

My overall goal to rectify some of the over-hyped expectations of AI while providing a general technical foundation which will hopefully help to address socio-ethical questions regarding the usage of AI.

Dr. Stefan Roski:
Explanation & Explainability

Artificial neural networks are often characterized as black boxes. Allegedly, we cannot explain why an input of a given network produces a specific output. Nor can we explain exactly how the individual weights of a network come about. But what do we actually mean when we say that we are unable to explain? To answer this question, we will explore some of the basic of current theories of explanation from the philosophy of science. Against this background, we will discuss different interpretations of the thesis that artificial neural networks and their outputs are not explainable. We will see that the plausibility of these theses depends heavily on how we explicate the connection between explanation and understanding.

Dr. Paul Näger:
Basics of AI Ethics

Ethical concerns about AI applications worry computer scientists, policy makers and the public alike. Surprisingly, however, many discussions do not involve too much ethical expertise – as a consequence, judgements are often made on an intuitive basis by the persons involved. As with all intuitive judgments, this procedure bears the risk of not meeting established epistemic standards.

In contrast, the philosophical discipline of ethics has a long tradition of thinking rationally about what is right and wrong in a given situation. This talk is supposed to provide an introduction for non-philosophers to such professional applied ethics in the realm of AI ethics. It introduces to some of the basic concepts, principles and insights that are relevant in the field: It treats the status of moral claims (especially their objectivity), the basic principles of normative reasoning, the role of explanations and justifications in ethics, the fundamental values that

are at stake in AI ethics, the question for an appropriate ethical theory, and how to balance ethical values in cases of conflict.

Prof. Dr. Gitta Kutyniok:
A Mathematical Perspective on Legal Requirements of the EU AI Act: From the Right to Explain to Neuromorphic Computing

Artificial intelligence (AI) is currently leading to one breakthrough after the other, both in public life with, for instance, autonomous driving and speech recognition, and in the sciences in areas such as medical imaging or molecular dynamics. However, problems with reliability and the danger of abuse of AI recently led to the EU AI Act and the G7 Hiroshima AI Process.

In this lecture, we will provide an introduction into this vibrant research area. We will then focus on legal requirements such as the "Right to Explain" and "Algorithmic Transparency", and analyze those from a mathematical standpoint, including discussing suitable explainability approaches. Finally, we will also touch upon limitations of AI methods trained on digital hardware and the necessity to consider novel computing hardware.

Carlos Zednik, PhD:
Disentangling XAI Concepts: Explanation, Interpretation, and Justification

The last decade has seen an explosion of interest in so-called Explainable AI. Insofar as many state-of-the-art AI systems are considered opaque, Explainable AI aims to render these systems explainable, or transparent. Alongside the development of sophisticated mathematical and computational methods, the maturation of Explainable AI as a discipline has witnessed numerous detailed analyses of key terms such as 'explanation', 'interpretation', and 'justification', among others. These analyses are important insofar as they state the goals of the discipline, and insofar as they can guide the development of methods with which these goals might be achieved. However, disagreements still abound about the precise meaning and interconnections between these terms. In this talk, I will present an overarching conceptual framework in which these terms can be related to one another, to current technical developments in XAI and ML, and to current policy and standardization frameworks that aim to regulate and promote trustworthy AI.

Prof. Dr. Eva Schmidt:
Reasons of AI Systems

This talk connects the fields of *philosophy of action* and of *explainable artificial intelligence* (AI). We investigate whether it can ever be appropriate to explain the outputs of AI systems by appeal to practical reasons of these systems. We argue that this can indeed be fitting, and respond to objections to our claim.

Dr. Astrid Schomäcker:
That's not fair! Explainability as a means to increase algorithmic fairness

Increasingly, important decisions are being delegated to AI decision systems: Who should get a loan? Who should be hired? Who should be released from jail? Such decisions have a massive impact on the lives of individuals, which is why they need to be fair. However, many existing systems disadvantage already marginalized groups like women or POC. At the same time, given modern machine learning approaches, the decision process of such systems is often too complex even for their developers to fully understand.

Consequently, many researchers have suggested explainable AI (XAI) as a means to ensure the fairness of AI decision-making. However, while intuitively plausible, it needs to be clarified how exactly explainability can increase fairness: Are existing XAI methods useful to detect unfair decisions? Who needs to understand the systems and to what end? And what information do they really need?

Prof. Dr. Kristian Kersting:
Where there is much light, the shadow is deep. XAI and Large Language Models

Artificial intelligence (AI) systems translate texts, help to treat patients, make purchasing decisions and optimize workflows. They manage increasingly complex human activities in ever more autonomous ways. But where is the moral compass of AI? Who decides on what is „right" and „wrong"? How can we provide revise systems if they are wrong? Who designs the perfect world? Not easy questions, and not just a question of AI.

Prof. Dr. Florian Boge:
Put it to the Test: Getting Serious about Explanations in Explainable Artificial Intelligence

Joint work with Prof. Dr. Axel Mosig, Competence Area Bioinformatics, Center for Protein Diagnostics (ProDi) at the Ruhr-University Bochum

Artificial Intelligence (AI) now pervades both science and society, but many present AI systems are known to be notorious black boxes. This can become a pressing issue, especially in high-stakes contexts such as decision-making in medical practice. Here, explanations of the outputs of an AI system seem desirable for the sake of calibrating trust, or may even constitute the basis of claims to moral and legal accountability. However, in contrast to standard scientific practice, current practice in the field of eXplainable AI (XAI) falls short in an important respect: While scientific explanations are usually required to be accompanied by testable predictions (Douglas, 2009), explanations in XAI are usually only validated on existing and well-known data. In our paper, we integrate insights from the philosophy of testability with recent developments in XAI to suggest a way out of the loop. This will be done by building on the framework for 'Falsifiable eXplanations of Artificial Intelligence' (FXAI), recently proposed by Schumacher et al., as a case study, and on its applications in cancer-research.

Dr. Thomas Grote:
The Double-Standard Problem in Medical ML Solved: Why Explainability Matters

In a widely influential paper, Alex London (2019) argues that, contrary to the received view, appeals to explainability are misguided when trying to establish warranted trust in machines learning models in healthcare. Rather, the assessment of machine learning models should be consistent with existing epistemic norms in medicine. Just like drugs, we should try to establish their safety and clinical benefit via clinical trials. Call this the 'double-standard' problem in medical machine learning. In my talk, I map out various objections to the double-standard problem. In particular, I discuss differences between the kind of evidence that clinical trials can yield for drugs as opposed to machine learning models. The hope is therefore to get a more nuanced view about what guarantees are necessary to warrant trust in medical applications of machine learning.