Issues in XAI #5: Understanding Black Boxes — Interdisciplinary Perspectives

TU Dortmund — September 5–7, 2022

Schedule

Monday, September 5

11.00 – 11.10: Welcome & Introduction
11.10 – 12.20: Andrés Páez (philosophy): "Idealization and Non-Factive Understanding in Machine Learning" (chair: Chiara Balestra)

— 15 minutes coffee break —

12.35 – 13.45: Nicole Krämer (psychology): "Understanding versus trust: What do users need when interacting with AI systems?" (chair: Andrés Páez)

— 80 minutes lunch break —

15.05 – 16.15: Sara Mann (philosophy): "Understanding via Exemplification in XAI" (chair: Nicole Krämer)

— 15 minutes coffee break —

16.30 – 17.40: Chiara Balestra (computer science): "Coalitional Game Theory to explain the structure of the data" (chair: Sara Mann)

Tuesday, September 6

09.30 – 10.40: Emanuele Ratti (philosophy): "An Integrative and coherentist approach to XAI in scientific research: models 'for' and systems of practice" (chair: Mieke Boon)

— 15 minutes coffee break —

10.55 – 12.05: Tim Hunsicker (psychology): "Unexplainable Accuracy or Explainable Faultiness? Investigating the Accuracy-Transparency Trade-Off of AI-Based Systems" (chair: Emanuele Ratti)

— 5 minutes coffee break —

12.10 – 13.20: Florian J. Boge (philosophy): "Functional Concept Proxies and the Actually Smart Hans Problem: What's Special About Deep Neural Networks in Science" (chair: Tim Hunsicker)

— 80 minutes lunch break —

14.40 – 15.50: Anne Lauber-Rönsberg (law): "A Legal Perspective on Explainable AI: Why, How Much and to Whom?" (chair: Florian J. Boge)

— 15 minutes coffee break —

16.05 – 17.15: Sabine Ammon (philosophy): "When is a Black-Box a Problem? Epistemic Coercion and Epistemic Sovereignty in Explainable Artificial Intelligence (XAI)" (chair: Anne Lauber-Rönsberg)

— 5 minutes coffee break —

17.20 – 18.30: Mieke Boon (philosophy): "Where to Locate the Explainability of Explainable Machine Learning?" (chair: Sabine Ammon)

Wednesday, September 7

10.00 – 11.10: Philipp Cimiano (computer science): "Counterfactual explanations for image classification tasks with an application in medical decision making" (chair: Juan M. Durán)

— 15 minutes coffee break —

11.25 – 12.35: Nadine Schlicker (psychology): "Men Are From Mars, Machines Are From Venus — Explaining Audio Processing of Deep Neural Networks" (chair: Philipp Cimiano)

— 15 minutes coffee break —

12.50 – 14.00: Juan M. Durán (philosophy & computer science): "Thinking Outside of the (Black) Box: Computational Reliabilism and Epistemic Trust" (chair: Nadine Schlicker)

Book of Abstracts

Monday, September 5

Andrés Páez (University of the Andes)

"Idealization and Non-Factive Understanding in Machine Learning"

Proxy or surrogate models are used in explainable artificial intelligence (XAI) to provide some degree of understanding of opaque machine learning systems. In this talk I explore the nature of these models. In particular, I inquire whether they are akin to the idealizations used in science to understand complex phenomena. I argue that although proxy models differ in significant ways from idealizations, they cannot be understood in factive or quasi-factive terms either. The peculiar nature of proxy models, and the epistemic role they play, provide an argument in support of three different but interconnected theses. First, I argue that proxy models show that (i) non-factive understanding is a legitimate and unavoidable form of understanding. This view is closely tied to (ii) the idea that many models are epistemic tools that transcend their representational nature. Proxy models present a vivid example of that pragmatic thesis. Finally, I argue that (iii) proxy models provide objectual understanding of the target system, and that the understanding they provide cannot be reduced to a functional kind of understanding based only on inputs and outputs.

Nicole Krämer (University of Duisburg-Essen)

"Understanding versus trust: What do users need when interacting with AI systems?"

In future, numerous intelligent systems will help humans to take decisions in their private as well as occupational lives. As an important prerequisite for the acceptance of the systems, explainability and understandability are hailed – in the sense that the user is able to understand the system's opaque functioning. However, first studies show that users either do not want to "understand" too much and/or that the system's functioning is difficult to grasp since most users do not have knowledge on computational processes they could build on. Based on literature from the field of science communication, it is therefore discussed whether an alternative approach to yield acceptance can be to instill "epistemic trust".

Sara Mann (TU Dortmund)

"Understanding via Exemplification in XAI"

Artificial intelligent (AI) systems have proven to be efficient tools in numerous contexts, including high-stakes scenarios such as autonomous driving or medical diagnosis. Many of these application contexts involve image classification. Since many AI systems used for image classification are considered to be opaque, research in explainable artificial intelligence (XAI) develops approaches which aim at rendering their inner workings understandable.

I show that Elgin's work on exemplification offers a useful framework in this context. An effective example provides epistemic access to contextually relevant facts by exemplifying features it shares with its target. The outputs of most XAI methods aiming at explaining image classification can be seen as providing us with examples of the image class in question. Accordingly, we can evaluate whether and to what degree those examples are *effective*, in the sense that they help to understand why certain images are classified as belonging to a particular class by the AI system.

Based on these insights, I suggest to draw a conceptual distinction between samples, which are any images instantiating the exemplified feature(s), and exemplars, which are visualizations intentionally designed to emphasize only those features we want to exemplify in a given context. I argue that current XAI methods usually provide us with samples. In those rare cases where exemplar-like visualizations are provided, these are mostly ill-suited to bring about understanding. Since exemplification works best with exemplars, I suggest to lay more research emphasis on XAI approaches generating exemplars specifically tailored to convey understanding in a given context.

Chiara Balestra (TU Dortmund)

"Coalitional Game Theory to explain the structure of the data"

Explainable machine learning connects well to the dimensionality reduction of data, where a reduction of the data size can improve the explainability of the selected features and provide additional insights into the structure of the original features. Coalitional game theory and Shapley values have often been argued to be explainable methods to assign fair importance scores to features in the black-box models' explanation context. However, not all real-world data are labeled, and when labels are unavailable, it is often costly to obtain them. Unsupervised feature selection aims to reduce the number of features, often using feature importance scores to quantify the relevancy of single features to the task at hand. These scores can be based only on variables' distribution

and their interactions' quantification. We address the redundancy-elimination issue by introducing a synergy between coalition game theory and information theory and use a quantification of correlations among features to compute feature importance scores. The introduced scores will represent the contribution of single features in explaining the dataset's structure and include a notion of redundancy awareness, making them a tool to achieve redundancy-free feature selection. Finally, the deriving features' selection lowers the redundancy rate while maximizing the information contained in the data.

References Balestra, C., Huber, F., Mayr, A., Müller, E. (2022). Unsupervised Features Ranking via Coalitional Game Theory for Categorical Data. In: Wrembel, R., Gamper, J., Kotsis, G., Tjoa, A.M., Khalil, I. (eds) Big Data Analytics and Knowledge Discovery. DaWaK 2022. Lecture Notes in Computer Science, vol 13428. Springer, Cham. https://doi.org/10.1007/978-3-031-12670-3_9

Tuesday, September 6

Emanuele Ratti (Johannes Kepler University Linz)

"An Integrative and coherentist approach to XAI in scientific research: models 'for' and systems of practice"

In the past few years, there has been an explosion of concerned literature about the opacity of data science tools. The problem with opacity, it is said, is that it makes the epistemic warrants and the moral accountability of AI tools problematic. If we cannot understand how and why a tool has arrived at certain conclusions, how do we know if this tool is reliable and/or trustworthy? Recently, a field called Explainable AI (XAI) has advanced various solutions to 'open' the black-box of opaque algorithmic systems. Finding the right way to 'explain' AI models (e.g. data science models) or the processes leading to them, it is said, is what can ensure the epistemic and moral accountability of AI. But despite the richness of XAI proposals, it has been noticed that this emerging field suffers from several problems. First, it is not clear what the ultimate goals of XAI tools are, whether they are about trustworthiness or reliability, which are both equally problematic goals. Second, it is not clear what XAI tools are supposed to explain: are the explanations about data-generating processes, or about the models themselves? Third, there are many ways of thinking about explanations, and it is not clear how to evaluate which one is the best given a certain context.

In this talk, I start from the assumption that these concerns are well-motivated, and that XAI is a promising field in need of a clearer goal. By limiting myself to the

context of scientific research, I propose that XAI, despite the name, does not have an explanatory purpose; rather, I formulate a new conceptualization of XAI tools that I call 'coherentist'. The notion of 'coherence' is taken from Hasok Chang's work on science as a system of practices (SoP). A SoP is a network of epistemic activities, scientific objects, and agents; these components have to stay in a relation of coherence (defined in various ways) in order to ensure the optimal functioning of the overall SoP of a given scientific project. Through Chang's lens, AI tools should not be seen as isolated entities which fully determine scientific decisions. Rather, AI tools are just one component of a dense network constituting a given SoP. In this context, the role of XAI is not to explain what AI tools do: the role of XAI is to facilitate the integration of AI tools into a given scientific project, and to make sure that AI tools themselves are in a relation of 'coherence' with the other components of a given SoP. Through a case study of biomedical data science, I will delineate (1) the idea of SoP, (2) the different ways in which 'coherence' acts as a 'glue' among different components of a given SoP, and (3) the special coherentist role that XAI plays in integrating AI tools in scientific practice.

Tim Hunsicker (Saarland University)

"Unexplainable Accuracy or Explainable Faultiness? Investigating the Accuracy-Transparency Trade-Off of AI-Based Systems"

The choice between different algorithmic approaches underlying AI-based systems is accompanied by the fact that the approaches yielding the most accurate outputs are often the least transparent and the most transparent ones are the least accurate. Therefore, when choosing between approaches in everyday use contexts, there is an accuracy-transparency trade-off. In this talk I will shed light on the decision-making and weighing process considering this trade-off from a psychological perspective.

In a between-participants online study (N = 383), we investigated whether framing (framing prediction performance of these systems as accuracy rate vs. error rate), accountability (decision-makers were informed that they need to explain their decision vs. that they do not need to) and the use context (medicine, personnel selection, finance, law) affect choosing between different versions of systems underlying the accuracy-transparency trade-off. Moreover, we examined whether the experimental manipulations and the system choice affected trustworthiness and trust perceptions. I will discuss the effects of framing, accountability and the use context on the system choice. Furthermore, I will present the results regarding the relationship between system choice and trust as well as the effects of different framing on trust in the chosen

system.

Florian J. Boge (University of Wuppertal)

"Functional Concept Proxies and the Actually Smart Hans Problem: What's Special About Deep Neural Networks in Science"

From a certain vantage point, Deep Neural Networks (DNNs) are nothing but parametrized functions f_θ(x) of some data vector x, and their 'learning' is nothing but an iterative, algorithmic fitting of the parameters θ to data. Hence, what could be special about DNNs as a scientific tool or model? Following a number of recent approaches, I argue that DNNs are capable of developing what I call concept proxies (FCPs), and that this makes them interestingly different from traditional multivariate methods in statistics. I will illustrate the salient differences by considering the possibility of what I call 'Actually Smart Hans predictors', i.e., DNNs that robustly succeed because they learn to rely on features connected to the data that are not transparent to human researchers.

Anne Lauber-Rönsberg (TU Dresden)

"A Legal Perspective on Explainable AI: Why, How Much and to Whom?"

Explainability is seen as a critical element in developing and building trustworthy AI. The requirement of transparency with respect to opaque decision-making systems has a long tradition in the legal system. The presentation will examine the legal instruments for AI transparency that already exist or are under discussion, and discuss their purpose, scope, and for whom explainability should be achieved.

Sabine Ammon (TU Berlin)

"When is a Black-Box a Problem? Epistemic Coercion and Epistemic Sovereignty in Explainable Artificial Intelligence (XAI)"

In my presentation, I am going to explore the black-box problem of AI technologies from the perspective of procedural epistemology, embedded in key concepts from philosophy of technology. I will argue that for many AI applications, once they are based on an appropriately designed milieu of reflection, the black box is less of a problem than thought. Disciplines like engineering and medicine, which always face an abundant complexity, have developed strategies to deal with (partial) ignorance, and

manage to arrive at robust knowledge, nevertheless. Key is to address an epistemic sovereign user; an attitude which many AI applications lack.

To frame algorithmic knowledge production, I suggest applying the concept of the epistemic tool (Boon & Knuuttila 2009) to XAI technologies. This allows to investigate two primary epistemic processes of XAI technologies, namely the making of the tool, and the application of the tool.

In the process of the making (design, implementation, testing), the AI technology becomes a knowledge repository. I will argue that the quality of the generation of this knowledge is essential for understandability and explainability. What is needed is an explicit communication of presuppositions, weaknesses, and limitations if the tool. This applies to the quality of the input data, underlying heuristics for the choice of the algorithm, strategies for the architecture of the tool, planned knowledge transfer, hypotheses-building, as well as the overarching epistemic model.

In the process of the application, XAI technology and human user enter in a joint process of knowledge production. Here, the abilities of the person dealing with the technical system and the affordances of the technical system (Norman 1988) come together and form a milieu of thinking and reasoning, aiming at an epistemic ascent in the sense of a cognitive achievement (Ammon 2017, 2019). A successful process brings together the affordances of the system, namely explainability and understandability, and the abilities of the human user (such as prior knowledge, education, competencies, skills) into a reflective equilibrium (comp. Goodman 1954).

To achieve an explanation, the user needs to know whether the output (in the form of a knowledge claim) is plausible and robust. I argue that knowledge about presuppositions, weaknesses, and limitations of the AI tool is needed to formulate good reasons for a justification. This is achieved not only by aligning the design to specific cognitive constraints and affordances, but also by an appropriate training of the user. To achieve understanding, the user needs to embed the output (in the form of a knowledge claim) in her or his background knowledge. This cognitive operation is successful if it leads to an epistemic ascend.

I claim that for the design of responsible XAI technologies, explainability and understandability need to be implemented in such a way that the cognitive constraints and affordances of the system match the abilities of the user. The resulting reflective equilibrium needs to enable satisficing (Simon 1996) cognition and to secure and to respect the epistemic sovereignty of the human user instead of enforcing epistemic coercion.

Mieke Boon (University of Twente)

"Where to Locate the Explainability of Explainable Machine Learning?"

with Koray Karaca (University of Twente)

In Logical Empiricism, Carl Hempel defended a covering law account of scientific explanation: the law that applies to (i.e., covers) a specific event constitutes the explanation of the event. Conversely, the law arises from inductive reasoning, i.e., through finding a regularity in a series of similar observations or events (Hempel 1948). In this way a correlation between variables is found that represents a regularity, which is then called a law. Subsequently, this found law explains specific cases in the future. However, this account of (scientific) explanation turns out to be untenable and leads to well-known unsolvable puzzles, for example, how to distinguish between accidental regularities and real laws, and thus how can we make sure that explanations genuinely account for their target phenomena (Boon 2020a). Against this backdrop, we examine the way in which explainability is sought in the context of machine learning (ML) where model construction is based on correlations founded in training data sets by what are called learning algorithms. However, due to the problem of algorithmic opacity, we do not understand how these correlations are founded by the learning algorithms. As a result, even though ML models can provide accurate predictions, we need explanations as to why (or for what reasons) they are able to provide these predictions. What is called explainable ML (XML) is a methodology that aims at extracting adequate reasons from ML models in order to account for their predictions (Xie et al., 2020).

In this talk, we will question whether explanations obtained through XML enable understanding the reasons as to why ML models makes predictions. Our motivation is based on the analogy with Hempel's covering-law account, suggesting that just like accidental generalizations XML explanations fail to be genuine explanations of real-world phenomena, as, e.g., they may be based on spurious correlations in training data sets on which ML models are constructed. We will rely on mechanistic view of explanation (Craver and Tabery, 2016), according to which the mechanism is the explanation of the law, and the mechanism thus makes the law intelligible. In this view, the distinction between accidental regularities and real laws can be accounted for because the real law is based on a mechanism.

In analogy with the mechanistic view of explanation, we will suggest that ML explainability should not be sought within the inner structure of ML models (including learning algorithms), but rather outside it. Just like a mechanism is needed for genuine explanations according to the mechanistic view, as we will argue, developing a specific

ML application first requires a conceptual model of the system for which the MLT is being developed, and that the explanation is then localized in that conceptual model. In other words, the development of an ML starts with the construction of a conceptual model of how the system works. This construction encompasses all kinds of choices based on theoretical and empirical knowledge about which factors are important and which are not. It also includes choices for the simplification of the conceptual model (Boon, 2020b). Next, the ML model is developed by using data-sets that are (or should be) chosen on the basis of the conceptual model (such as to prevent that the ML is not fed with irrelevant data). Eventually, the conceptual model represents a form of mechanism that explains the algorithm (ML model) and its outcomes. When asked for an explanation for a certain outcome, it will then be possible to fall back on the conceptual model that formed the basis for the MLT.

In this account, the MLT is reduced to an instrument that is able to make quantitative connections in a set of variables that are presented to it (the learning set). The conceptual model ensures that (1) the set that is presented consists of variables relevant to the results of the system, and (2) that results from the MLT can be explained on the basis of the CM – and not by just referring to the algorithm that somehow speaks truth.

References Boon, M. (2020a). How Scientists Are Brought Back into Science—The Error of Empiricism. In M. Bertolaso, & F. Sterpetti (Eds.), A Critical Reflection on Automated Science: Will Science Remain Human? (Vol. 1, pp. 43-65). (Human Perspectives in Health Sciences and Technology; Vol. 1). Springer. https://doi.org/10.1007/978-3-030-25001-0_4

Boon, M. (2020b). Scientific methodology in the engineering sciences. In: D. Michelfelder, & N. Doorn (Eds.), The Routledge Handbook of the Philosophy of Engineering (pp. 80-94). Routledge Tailor & Francis Group. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315276502-8

Craver, C. F., and J. Tabery, (2016). "Mechanisms in Science", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2016/entries/sciencemechanisms/

Hempel, C.G., (1948/1965). Studies in the logic of explanation. In: Hempel C.G. (ed.) Aspects of scientific explanation. Free Press, New York, pp 245–295.

Xie, N., G. Ras, M. van Gerven, and D. Doran, (2020). Explainable deep learning: A field guide for the uninitiated, arXiv preprint. https://arxiv.org/abs/2004.14545

Wednesday, September 7

Philipp Cimiano (University of Bielefeld)

"Counterfactual explanations for image classification tasks with an application in medical decision making"

With the increase use of deep learning models in several application fields, the explanation of black box models is of high importance to increase user trust and confidence. We consider the case of using counterfactual explanations to explain classifications of images. Counterfactual approaches applied to image classification highlight the area of an image that, if changed, would yield a different classification and thus deliver a causal explanation. However, current counterfactual generation approaches do not have a spatial constraint on the area that can be changed, sometimes changing the entire image. In the generation of counterfactual explanations, minimal necessary changes are desired to have easily interpretable causal explanations.

To yield minimal changes in areas of interest, the classifier's attention can be included in the generation process. This helps the generator to focus on the important areas in the image and leave the other parts untouched. In this talk we introduce two systems that combine attention-based CycleGANs and counterfactual Image-to-Image generation. The systems are developed for the MURA wrist X-ray dataset and tested on the Horse to Zebra dataset during development. The experiments show that combining counterfactual generation with attention-based CycleGANs can yield counterfactuals that incur a minimal change, achieving smaller Kullback-Leibler Divergence values than approaches without attention mechanisms. We show applications of our method on highlighting the areas of an X-ray image that is causally responsible for a given classification.

Nadine Schlicker (University Hospital of Marburg)

"Men Are From Mars, Machines Are From Venus — Explaining Audio Processing of Deep Neural Networks"

with Markus Langer (Saarland University)

Humans and systems process information differently. This becomes especially apparent in the processing of audio information. Whereas humans process audio signals as composition of different sound waves, machines are able to decompose audio signals into single frequencies. For instance, when humans hear a dog barking, they might describe the sound as loud and unpleasant and maybe as deep or squeaky. In contrast, machines do not have this semantic, qualitative information; they process different

frequency strengths over time. It seems that humans and systems speak different languages when it comes to audio processing.

Since automated classification of audio signals provides many opportunities to support human decision-making (e.g., in medical decision making), this communication issue may lead to suboptimal human-system interaction. For example, if we aim for transparent systems that humans can trust adequately, we need to find ways to translate audio processing of systems in a way humans can understand. In other words, we need to enable communication about decision-making in audio classification between systems and humans.

In this talk I will present challenges in explaining audio classification of non-semantic sounds and discuss potential ideas to overcome them. Specifically, I will talk about audio classification in cardiac auscultation, where medical doctors use (digital) stethoscopes to classify heart sounds as pathological or normal. Although standard in every medical consultation, cardiac auscultation is a difficult task that requires intense and continuous training: heart sounds are often subtle and medical professionals need to make audio classifications in noisy environments (e.g., in the emergency room). Therefore, AI supported heart sound classification might provide valuable decision support in medical practice. Yet, in order to establish calibrated trust in respective systems, medical doctors may desire ways to trace and comprehend decision making strategies of AI assistants – a desire that may be jeopardized by the challenges associated with the differences in audio processing between humans and systems.

Juan M. Durán (TU Delft)

"Thinking Outside of the (Black) Box: Computational Reliabilism and Epistemic Trust"

A recurrent approach to justify our belief in the output of Machine Learning (ML) algorithms consists of "looking into its inner logic." Such an approach (e.g., "transparency") proposes an internalist to the algorithm perspective on justification. That is, one that requires some form of surveyance of the algorithm for the justification of its output. In this talk, I present and discuss Computational Reliabilism (CR) as the externalist alternative. CR justifies our beliefs in the output of ML by rendering the algorithm reliable. To this end, I will propose three families of reliability indicators corresponding to three layers of analysis of scientific research with ML. I close by discussing in what respects CR is superior to internalists alternatives and which are its current shortcomings.