Issues in Explainable AI: Blackboxes, Recommendations, and Levels of Explanation

Monday, September 30th

Rune Nyrup (LCFI, Cambridge): Explanatory Pragmatism as a Philosophy for the Science of Explainable Artificial Intelligence

A common objection to AI systems is that it can be difficult to adequately explain their decision-making to humans. Many forms of AI, especially those based on advanced machine learning techniques, are accused of being “opaque”, “black boxes”, “uninterpretable” or “incomprehensible”. In response, a new sub-field is currently emerging within AI research, aiming to create methods for making ‘interpretable’ or ‘explainable’ AI, sometimes abbreviated XAI.

Early work in this field tended to rely on researchers’ intuitive sense of whether a given model or system was more intelligible. Recently, however, a number of researchers have grown dissatisfied with this approach and started calling for more ‘rigorous’ or ‘scientific’ approaches to XAI. Two main such approaches are currently being pursued: (1) Empirical approaches, which seeks to devise experimental methods for measuring whether a system is explainable, e.g. by (a) measuring how users explanations; or (b) testing their performance on some relevant task. (2) Theoretical approaches which draw on some existing account of explanation from psychology or philosophy.

Both approaches represent plausible steps forward, but also face limitations in their current form. Regarding (1a), there is evidence that people sometimes overestimate how much understanding they get from an explanation; (1b) gets around this problem by focusing on behavioural measures but faces the question of which tasks are most relevant to determine understanding. Regarding (2), given the fact of explanatory pluralism, i.e. that there are many different explanatory models, there is unlikely to be a single account of explanation which can form the basis for XAI. This suggest a more contextual approach, but the field is currently lacking a principled approach to choosing which model of explanation to implement in a given application of AI.

I propose a package of philosophical views, which I call Explanatory Pragmatism, as a promising guiding framework for the field of XAI research to overcome these challenges.

Clemens Stachl (LMU Munich): More Than a Gut Feeling. The Importance of Interpretable Machine Learning for Psychological Science.

The accelerating digitization of our society, changes the way social scientists do research. Specifically, digital-footprint data from online social media and behavioral data from high-frequency mobile sensing applications provide completely new opportunities for the investigation of research questions, on a large scale. Additionally, more and more researchers in psychology and other social sciences adopt the usage of machine learning algorithms that enable the prediction of psychological traits and outcomes, using these new types of data. In addition to the predictive performance of models, researchers are interested to understand how models make predictions. However, due to the high-dimensional and nonlinear complexity of some models, their interpretation is inherently difficult. Model agnostic interpretable machine learning techniques provide some help to better understand the inner-workings of black-box models up to some degree. However, the full understanding of complex models remains challenging if not impossible. In this talk I will use some examples from our groups work to illustrate the importance of interpretable machine learning techniques. Additionally, I will discuss the importance of these methods for the validation of machine learning models and will talk about the prevention of model-based feedback loops and biases. Finally, I will hypothesize how interpretable machine learning will help to progress psychological science in theory and application.

Marija Slavkovik (University of Bergen): Moral Decisions and How to Explain Them

With the prevalence of Artificial Intelligence (AI) technologies, rises the concern that their adoption, use and abuse can erode the ethical values in our society. At least three different approaches are put forward to ensure that the AI technology is ethical: Fairness, Accountability, and Transparency (FAT), Explainable AI (XAI), and Machine Ethics. However there is still no methodology to access which AI technology should be considered ethic-sensitive or ethic-critical and which approach should be applied to handle it. The talk gives an overview of what we talk about when we talk about AI and ethics, the state-of-the art and the major challenges.

Tuesday, October 1st

Hans-Johann Glock (University of Zurich): Artificial General Intelligence: Explainable? Predictable? Desirable?

My paper will consider what kind of explainability is compatible with the ideal of General Artificial Intelligence. Would a system satisfying the conditions of AGI have to be explicable in the same way as a rational human subject? In that case, both its outputs and its computational processes would allow of being understood in terms of reasons. On the basis of distinctions from contemporary debates about reasons I shall consider whether this is possible in principle, paying special attention to so-called 'motivating reasons'. Furthermore, as the case of humans shows, that a system can be understood in terms of motivating reasons guarantees at most a limited degree of predictability. The presentation ends on ruminations about whether artificial systems that are predictable only to a limited degree are desirable.

Farzad Nozarian (DFKI, Saarbrücken): Make your Autonomous Driving Model More Interpretable by Letting It Say "I Don't Know!"

Nowadays, many deep learning researchers and practitioners try to model the true distribution behind the training dataset using deep neural network architectures to provide more accurate predictions for the given task. However, the focus is usually on predictions that do not consider the "uncertainty" of the neural network model. It is a well-known fact that neural networks are not robust enough to provide reliable predictions for out of distribution samples and come up with overconfident decisions for samples far from the training distribution. Therefore using neural network's raw predictions without letting the user know where and when the model is not certain enough might have an adverse effect on many safety-critical tasks like Autonomous Driving where every single prediction in different layers of the architecture plays an important role in the final decision. Thus, to increase the safety, trust, and interpretability of predictions in complex deep neural network pipelines, measuring the uncertainty of deep neural networks is a key factor. In this presentation, I introduce a safe and interpretable Autonomous Driving policy which provides human-understandable representations as intermediate results by which the user can trust more on the final decisions of the neural network. In addition, by employing Bayesian Deep Learning methods in our architecture, the driving policy lets the users know when it does not know and might make incorrect decisions that lead to infractions or accidents in the future.

Sonja Ötting (University of Bielefeld): Just? Talk about it! Fair Decision Procedures and What to Expect from AI at the Workplace

At the workplace, employees encounter decisions in plenty. Reactions to these decisions are as diverse: satisfaction, joy, anger and sadness, commitment to the workplace, good or bad work performance, extra hours or negative behavior are just a few. One thing that importantly influences reactions to decisions is their perceived justice, especially justice of the decision procedures. With AI arising as new decision agent at the workplace we asked and tried to answer two questions: Do we need just AI? And how does AI influence reactions to just or unjust decisions?

Dan Brooks (University of Cincinatti): Levels, Hierarchical Ordering, and Explanation

Reference to “levels” is rampant and for the most part unchecked in the scientific and philosophical literature. Part of this situation stems from the lack of consensus regarding which qualifying notion adequately expresses what is at stake when invoking the idea of levels: Instead, we are mostly left to our own devices when considering which “levels of x” labels to apply in our work. In this talk I consider several common level labels found in philosophy of biology and cognitive science, and find these wanting. In particular, “levels of analysis” seems to require no levels talk at all, while “levels of explanation” leaves the levels concept underdefined. I then suggest that a robust source for levels language is available in the levels of organization tradition, and that other level labels are sufficiently derivative from this source. I turn after this to considering the distinction (if there is one) between hierarchical ordering and levels language, and likewise find the notion of “hierarchy” difficult to reconcile with established organizational levels work: Specifically, “hierarchy” seems to work better with artificial systems, while “levels” seems more fitted to natural, living systems. I conclude with some upshots for understanding scientific explanation.

Wednesday, October 2nd

Geoff Keeling (LCFI, Cambridge): Treatment Recommendation, Informed Consent and Understanding (co-authored with Rune Nyrup)

Treatment recommender algorithms use natural language processing to read and interpret patient records and then recommend medical interventions based on given and inferred features of the patient's medical history. According to the standard view in medical ethics, ideal medical decision-making requires the patient and doctor to reach a joint decision about which intervention is best based on a shared understanding of the relevant features of the patient's case. We aim to clarify the tension between our best practices of shared decision-making and algorithmic treatment recommendation. We then draw on recent work in the philosophy of science on the nature of understanding to provide a practical account of what doctors ought to disclose to patients about algorithmic treatment recommendations in order to facilitate patient understanding.

Cornelius König and Nadine Schlicker (Saarland University): XAI and Psychology - Issues and Experimental Investigations

Humans typically want to understand what is going on because it helps them gaining a feeling of control. This human tendency is challenged by the opacity of systems relying on artificial intelligence, which has resulted in the emergence of the field of eXplainable Artificial Intelligence (XAI). We argue that psychology needs to play a more important role in XAI for at least three reasons. (a) Psychology can help understanding the multiple facets of reactions to experiencing unexplainable systems. (b) Psychology can help understand individual differences because it is unlikely that "one XAI solution fits all." (c) Psychologists' expertise in running experiments should be beneficial for the XAI field because experiments help understanding mental process that people often have little or no direct introspective access to and because such empirical research can be the basis for recommendations for practice. As an example, we present the results of a recently conducted experiment. This study examined how automation affects the perception of justice and trustworthiness from the perspective of the recipient of the decision. Furthermore, we tested the effects of different explanations respectively the absence of an explanation. We used a fully randomized 2 (agent: automated vs. human) x 3 (explanation: equality-explanation vs. equity-explanation vs. no explanation) between-subjects design. Participants were recruited from the healthcare sector (N = 209) and responded to an online study in which they put themselves in the position of a professional nurse whose vacation request is denied. Results revealed that (a) interpersonal justice was perceived as more important for the human agent, (b) procedural

justice was found to be estimated differently between human and automated agents, and (c) the absence of explanation did only decrease informational justice perception towards a human agent, but not towards an automated agent. These results indicate a tension between the expectations of potential users and the demands for explainable systems.

Georg Borges (Saarland University): New Technologies and the Law – A Complicated Relationship?

In this presentation, some aspects of the complex relationship between law and technology will be discussed. Whereas law functions as a limitation to the development and use of new technologies, it also has an important role as an enabler to new technological developments.