KI - Künstliche Intelligenz (2022) 36:181–184
https://doi.org/10.1007/s13218-022-00776-4

INTERVIEW

Interview: AI Expert Prof. Müller on XAI

Or How Far do We have to Go in Order to Get There?

Johannes Fähndrich[1] · Roman Povalej[2] · Heiko Rittelmeier[3] · Silvio Berner[4]

Published online: 24 November 2022

1 Interview

Prof. Dr. Klaus-Robert Müller is rated one of the top cited computer scientists in Germany and number 47 internationally by Reseach.com.[1] We interviewed Prof. Dr. Müller because of his perspective on AI and its explainability. With his approaches to analyzing Deep Neural Networks, he might be one of a few leading scientists in the world who are researching the questions: "Can we trust results out of an AI?" This question is of particular interest to the community of digital forensics. On grounds of that, many of the modern challenges are about handling big amounts of data. We believe that AI could be the integral part of future investigations with digital evidence.

Originally, he studied physics in Karlsruhe, where he graduated in 1992 with a Ph.D. in theoretical computer science. Starting with his post doc, he founded the Data Analysis community in Berlin. In 2003, he became a full professor at University of Potsdam, in 2006 he became chair of the machine learning department at TU Berlin. He is an active researcher and has been granted many national and international science awards. As a scientist, he serves rsp. has served in the editorial boards of Computational Statistics, IEEE Transactions on Biomedical Engineering, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Information Theory, Journal of Machine Learning Research and in program and organization committees of various international conferences. In 2019, 2020, 2021 he became ISI Highly Cited Researcher. His research interest is in the field of machine learning, deep learning and data analysis covering a wide range of theory and numerous

✉ Johannes Fähndrich
johannesfaehndrich@hfpol-bw.de

1 Hochschule für Polizei Baden-Württemberg, Sturmbühlstr.250, 78054 Villingen-Schwenningen, Germany

2 Police Academy of Lower Saxony, Gimter Str. 10, 34346 Hann–Münden, Lower Saxony, Germany

3 Central Office for Information Technology in the Security Sector (ZITiS), Zamdorfer Straße 88, 81677 Munich, Bavaria, Germany

4 University of Applied Police Sciences Saxony, Friedensstraße 120, 02929 Rothenburg/OL, Saxony, Germany

[1] https://research.com/scientists-rankings/computer-science/de last visited: 26.11.2021.

scientific (Physics, Chemistry, and Neuroscience) and industrial applications.

With a h-index of 136 [2] his research is well cited and with broad contributions to the community of Machine Learning and data Analysis. His research areas include statistical learning theory for neural networks, support vector machines and ensemble learning techniques. He contributed to the field of signal processing, working on time-series analysis, statistical denoising methods and blind source separation. His present application interests are expanded to the analysis of biomedical data, most recently to brain computer interfacing, genomic data analysis, computational chemistry and atomistic simulations.

Interview with Prof. Dr. Klaus-Robert Müller:

KI Journal: What does explainable AI stand for?

Prof. Dr. Müller: It stands for the attempt to shed light into the inner workings of AI algorithms. See, e.g., [1, 2].

KI Journal: What are the differences between explainable AI and good old AI?

Müller: There is no relation between them. XAI (explainable AI) can be applied for explaining any AI algorithm.

KI Journal: When is a result interpretable?

Müller: A result is interpretable if a user can gain a better understanding about the AI method and its application to a certain problem. Feature selection is a method leading to more interpretability for the full data set, as it tells which features may hold the key (in the light of the AI method) for the decision-making. More complex and more recent is XAI for every single data point. Here, typically a heatmap depicts what variables are important for the decision-making (in the light of the AI method) for a prediction of a single new data point. Interpretability can allow understanding the shortcomings of AI models (see Clever Hans effect) or can allow gaining novel insights in domains like physics.

[2] https://scholar.google.com/citations?user=jplQac8AAAAJ&hl=en&oi=ao last visited: 27.11.2021.

KI Journal: What makes ML Models transparent?

Müller: XAI algorithms systematically decompose the decision-making process at different abstraction levels.

KI Journal: What is the difference to formal verification?

Müller: Formal verification is a concept from theoretical computer science aiming to e.g., prove theorems. XAI algorithms aim to analyze the learned non-linear function of an AI method, to make a decision-making process transparent. Some XAI algorithms, for example, Layer-wise-relevance propagation come with formal proofs.

KI Journal: Are there different kind (degrees) of explainable?

Müller: Yes, XAI algorithms systematically decompose the decision-making process at different abstraction levels. E.g., a prediction can be explained in terms of input variables or also in terms of more abstract concepts learned in the higher layers of a neural network.

KI Journal: How about methods of machine learning or Data Science, why should they be explainable?

Müller: ML or Data Science algorithms tend to make use of any kind of correlation in the training set. Such use of spurious correlations may later on drastically decrease the functioning of models in the real world. XAI helps to detect such flaws (see Clever Hans effect) [3].

KI Journal: Which methods of AI are explainable by design?

Müller: There is no need for algorithms that are explainable by design. My opinion is, that it is sufficient to have a post-hoc explanation.

KI Journal: What is the difference between Symbolic and connectionist approaches regarding explainability?

Müller: There is a well-known difference between symbolic and connectionist approaches, loosely speaking, both implement some nonlinear function classes that can typically be translated into a neural network, which automatically renders them explainable.

KI Journal: Which AI methods are the least explainable? or can not be explained at all?

Müller: All AI methods that implement, with a grain of salt, smooth nonlinear function classes can be made explainable.

KI Journal: Is explainability application specific?

Müller: Indeed, it should be. Users require explanation on different abstraction levels (doctors, laymen, students), so there is an interesting and rather unexplored HCI side to XAI.

KI Journal: How do you explain the current hype concerning XAI?

Müller: It is clearly not a hype. Users want their methods to be transparent, safe, fair, and trustworthy. This requires opening the black box, and XAI assumed this role and provided methods for this need. Note that this has been available only for a short time (I wrote my first XAI paper in 2010).

KI Journal: Which is your preferred tool for creating XAI? and why?

Müller: We like LRP[3] as it comes with a proof.

KI Journal: In which domain do you see XAI to make the biggest impact?

Müller: Users want their methods to be transparent, safe, fair, and trustworthy. Also, XAI can be used to debug methods (remove Clever Hans etc.) and thus improve them.

KI Journal: Which risks do you see in the application of methods out of XAI research?

Müller: I see the chance for methods to become better, more transparent, safe, fair and trustworthy. Also, to arrive at novel insights from learned models (as we demonstrated, e.g., in pathology [5] and quantum chemistry [6]).

KI Journal: What needs to be done with methods of AI, so they can become trustworthy?

Müller: Use XAI and of course other tools to improve their quality and understanding.

KI Journal: What changes in the approach to AI should happen before you would trust AI Systems in law enforcement?

Müller: In short: Personally, I find AI Systems in law enforcement a bit dubious, as all AI systems are stochastic in their nature of decision-making. 65

KI Journal: When results are understandable (white-box) does it mean why the result has been created has to be explainable?

Müller: If understandable means that we understand what the neural network is doing to achieve its decision-making in detail, then indeed that is an essential step.

KI Journal: Is there a need to adapt the law to new AI stakeholders?

Müller: This is already happening. Standardizing committees on the topics AI for Health and AI for Telecommunication have been created by WHO[4] and ITU.[5] These standards will be the

[3] LRP: Layer-wise Relevance Propagation an introduction can be found in [2, 4].

[4] See https://www.itu.int/en/ITU-T/focusgroups/ml5g/Pages/default.aspx.

[5] See https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx.

basis for future AI systems in use for critical applications. Other application domains will follow naturally.

KI Journal: Where do you believe we are at in five years?

Müller: Happily growing. I feel the wonderful thing in this community is that we witness an international effort of young researchers (and more seasoned ones like me). It is a very lively and creative community with lots of interesting research contributions still to be done. For example, just recently, provable higher order explanation methods have emerged. We will see more and more use of XAI in the sciences to make sure that the results obtained are watertight, and moreover to use XAI for reaching insights and novel hypotheses that were not available before.

References

1. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR (2019) Explainable AI: interpreting, explaining and visualizing deep learning, vol 11700. Springer Nature
2. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR (2021) Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE 109(3):247–278
3. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR (2019) Unmasking clever hans predictors and assessing what machines really learn. Nat Commun 10(1):1–8
4. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10(7):e0130140
5. Binder A, Bockmayr M, Hägele M, Wienert S, Heim D, Hellweg K, Ishii M, Stenzinger A, Hocke A, Denkert C et al (2021) Morphological and molecular breast cancer profiling through explainable machine learning. Nat Mach Intell 3:355–366
6. Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A (2017) Quantum-chemical insights from deep tensor neural networks. Nat Commun 8(1):1–8