xxAI - Beyond Explainable Artificial Intelligence

Andreas Holzinger[1,2,3](✉), Randy Goebel[3], Ruth Fong[4], Taesup Moon[5], Klaus-Robert Müller[6,7,8,10], and Wojciech Samek[9,10]

[1] Human-Centered AI Lab, University of Natural Resources and Life Sciences, Vienna, Austria
andreas.holzinger@human-centered.ai
[2] Medical University Graz, Graz, Austria
[3] xAI Lab, Alberta Machine Intelligence Institute, Edmonton, Canada
[4] Princeton University, Princeton, USA
[5] Seoul National University, Seoul, Korea
[6] Department of Artificial Intelligence, Korea University, Seoul, Korea
[7] Max Planck Institute for Informatics, Saarbrücken, Germany
[8] Machine Learning Group, Technical University of Berlin, Berlin, Germany
[9] Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[10] BIFOLD – Berlin Institute for the Foundations of Data and Learning, Berlin, Germany

Abstract. The success of statistical machine learning from big data, especially of deep learning, has made artificial intelligence (AI) very popular. Unfortunately, especially with the most successful methods, the results are very difficult to comprehend by human experts. The application of AI in areas that impact human life (e.g., agriculture, climate, forestry, health, etc.) has therefore led to an demand for trust, which can be fostered if the methods can be interpreted and thus explained to humans. The research field of explainable artificial intelligence (XAI) provides the necessary foundations and methods. Historically, XAI has focused on the development of methods to explain the decisions and internal mechanisms of complex AI systems, with much initial research concentrating on explaining how convolutional neural networks produce image classification predictions by producing visualizations which highlight what input patterns are most influential in activating hidden units, or are most responsible for a model's decision. In this volume, we summarize research that outlines and takes next steps towards a broader vision for explainable AI in moving beyond explaining classifiers via such methods, to include explaining other kinds of models (e.g., unsupervised and reinforcement learning models) via a diverse array of XAI techniques (e.g., question-and-answering systems, structured explanations). In addition, we also intend to move beyond simply providing model explanations to directly improving the transparency, efficiency and generalization ability of models. We hope this volume presents not only exciting research developments in explainable AI but also a guide for what next areas to focus on within this fascinating and highly relevant research field as we

A. Holzinger et al. (Eds.): xxAI 2020, LNAI 13200, pp. 3–10, 2022.
https://doi.org/10.1007/978-3-031-04083-2_1

enter the second decade of the deep learning revolution. This volume is an outcome of the ICML 2020 workshop on “XXAI: Extending Explainable AI Beyond Deep Models and Classifiers.”

Keywords: Artificial intelligence · Explainable AI · Machine learning · Explainability

1 Introduction and Motivation for Explainable AI

In the past decade, deep learning has re-invigorated the machine learning research by demonstrating its power in learning from vast amounts of data in order to solve complex tasks - making AI extremely popular [5], often even beyond human level performance [24]. However, its power is also its peril: deep learning models are composed of millions of parameters; their high complexity [17] makes such “black-box” models challenging for humans to understand [20]. As such “black-box” approaches are increasingly applied to high-impact, high-risk domains, such as medical AI or autonomous driving, the impact of its failures also increases (e.g., medical misdiagnoses, vehicle crashes, etc.).

Consequently, there is an increasing demand for a diverse toolbox of methods that help AI researchers and practitioners design and understand complex AI models. Such tools could provide explanations for model decisions, suggest corrections for failures, and ensure that protected features, such as race and gender, are not misinforming or biasing model decisions. The field of explainable AI (XAI) [32] focuses on the development of such tools and is crucial to the safe, responsible, ethical and accountable deployment of AI technology in our wider world. Based on the increased application of AI in practically all domains which affects human life (e.g., agriculture, climate, forestry, health, sustainable living, etc.), there is also a need to address new scenarios in the future, e.g., explaining unsupervised and intensified learning and creating explanations that are optimally structured for human decision makers with respect to their individual previous knowledge. While explainable AI is essentially concerned with implementing transparency and tractability of black-box statistical ML methods, there is an urgent need in the future to go beyond explainable AI, e.g., to extend explainable AI to include causality and to measure the quality of explanations [12]. A good example is the medical domain where there is a need to ask “what-if” questions (counterfactuals) to gain insight into the underlying independent explanatory factors of a result [14]. In such domains, and for certain tasks, a human-in-the-loop can be beneficial, because such a human expert can sometimes augment the AI with tacit knowledge, i.e. contribute to an AI with human experience, conceptual understanding, context awareness, and causal reasoning. Humans are very good at multi-modal thinking and can integrate new insights into their conceptual knowledge space shaped by experience. Humans are also robust, can generalize from a few examples, and are able to understand context from even a small amount of data. Formalized, this human knowledge can be used to build structural causal models of human decision making, and the features can be traced back to train AI - helping to make current AI even more successful beyond the current state of the art.

In such sensitive and safety-critical application domains, there will be an increasing need for trustworthy AI solutions in the future [13]. Trusted AI requires both robustness and *explainability* and should be balanced with human values, ethical principles [25], and legal requirements [36], to ensure privacy, security, and safety for each individual person. The international XAI community is making great contributions to this end.

2 Explainable AI: Past and Present

In tandem with impressive advances in AI research, there have been numerous methods introduced in the past decade that aim to explain the decisions and inner workings of deep neural networks. Many such methods can be described along the following two axes: (1) whether an XAI method produces *local* or *global* explanations, that is, whether its explanations explain individual model decisions or instead characterize whole components of a model (e.g., a neuron, layer, entire network); and (2) whether an XAI method is *post-hoc* or *ante-hoc*, that is, whether it explains a deep neural network after it has been trained using standard training procedures or it introduces a novel network architecture that produces an explanation as part of its decision. For a brief overview on XAI methods please refer to [15]. Of the research that focuses on explaining specific predictions, the most active area of research has been on the problem of feature attribution [31], which aims to identify what parts of an input are responsible for a model's output decision. For computer vision models such as object classification networks, such work typically produce *heatmaps* that highlight which regions of an input image most influence a model's prediction [3,8,28,33–35,38,41].

Similarly, *feature visualization* methods have been the most popular research stream within explainable techniques that provide global explanations. Such techniques typically explain hidden units or activation tensors by showing either real or generated images that most activate the given unit [4,27,35,38,40] or set of units [10,18,42] or are most similar to the given tensor [21].

In the past decade, most explainable AI research has focused on the development of post-hoc explanatory methods like feature attribution and visualization.

That said, more recently, there have been several methods that introduce novel, *interpretable-by-design* models that were intentionally designed to produce an explanation, for example as a decision tree [26], via graph neural networks [29], by comparing to prototypical examples [7], by constraining neurons to correspond to interpretable attributes [19,22], or by summing up evidence from multiple image patches [6].

As researchers have continued to develop explainable AI methods, some work has also focused on the development of disciplined evaluation benchmarks for explainable AI and have highlighted some shortcomings of popular methods and the need for such metrics [1–3,9,11,16,23,28,30,37,39].

In tandem with the increased research in explainable AI, there have been a number of research outputs [32] and gatherings (e.g., tutorials, workshops, and conferences) that have focused on this research area, which have included some of the following:

- NeurIPS workshop on "Interpreting, Explaining and Visualizing Deep Learning – Now what?" (2017)
- ICLR workshop on "Debugging Machine Learning Models" (2019)
- ICCV workshop on "Workshop on Interpretating and Explaining Visual AI Models" (2019)
- CVPR tutorial on "Interpretable Machine Learning for Computer Vision" (2018–ongoing)
- ACM Conference on Fairness, Accountability, and Transparency (FAccT) (2018–ongoing)
- CD-MAKE conference with Workshop on xAI (2017–ongoing)

Through these community discussions, some have recognized that there were still many under-explored yet important areas within explainable AI.

Beyond Explainability. To that end, we organized the ICML 2020 workshop "XXAI: Extending Explainable AI Beyond Deep Models and Classifiers," which focused on the following topics:

1. *Explaining beyond neural network classifiers* and explaining other kinds of models such as random forests and models trained via unsupervised or reinforcement learning.
2. *Explaining beyond heatmaps* and using other forms of explanation such as structured explanations, question-and-answer and/or dialog systems, and human-in-the-loop paradigms.
3. *Explaining beyond explaining* and developing other research to improve the transparency of AI models, such as model development and model verification techniques.

This workshop fostered many productive discussions, and this book is a follow-up to our gathering and contains some of the work presented at the workshop along with a few other relevant chapters.

3 Book Structure

We organized this book into three parts:

1. Part 1: Current Methods and Challenges
2. Part 2: New Developments in Explainable AI
3. Part 3: An Interdisciplinary Approach to Explainable AI

Part 1 gives an overview of the current state-of-the-art of XAI methods as well as their pitfalls and challenges. In Chapter 1, Holzinger, Samek and colleagues give a general overview on popular XAI methods. In Chapter 2, Bhatt et al. point out that current explanation techniques are mainly used by the internal stakeholders who develop the learning models, not by the external end-users who actually get the service. They give nice take away messages learned from an interview study on how to deploy XAI in practice. In Chapter 3, Molnar et al. describe the general pitfalls a practitioner can encounter when employing model agnostic interpretation methods. They point out that the pitfalls exist when there are issues with model generalization, interactions between features etc., and called for a more cautious application of explanation methods. In Chapter 4, Salewski et al. introduce a new dataset that can be used for generating natural language explanations for visual reasoning tasks.

In Part 2, several novel XAI approaches are given. In Chapter 5, Kolek et al. propose a novel rate-distortion framework that combines mathematical rigor with maximal flexibility when explaining decisions of black-box models. In Chapter 6, Montavon et al. present an interesting approach, dubbed as neuralization-propagation (NEON), to explain unsupervised learning models, for which directly applying the supervised explanation techniques is not straightforward. In Chapter 7, Karimi et al. consider a causal effect in the algorithmic recourse problem and presents a framework of using structural causal models and a novel optimization formulation. The next three chapters in Part 2 mainly focus on XAI methods for problems beyond simple classification. In Chapter 8, Zhou gives a brief summary on recent work on interpreting deep generative models, like Generative Adversarial Networks (GANs), and show how human-understandable concepts can be identified and utilized for interactive image generation. In Chapter 9, Dinu et al. apply explanation methods to reinforcement learning and use the recently developed RUDDER framework in order to extract meaningful strategies that an agent has learned via reward redistribution. In Chapter 10, Bastani et al. also focus on interpretable reinforcement learning and describe recent progress on the programmatic policies that are easily verifiable and robust. The next three chapters focus on using XAI beyond simple explanation of a model's decision, e.g., pruning or improving models with the aid of explanation techniques. In Chapter 11, Singh et al. present the PDR framework that considers three aspects: devising a new XAI method, improving a given model with the XAI methods, and verifying the developed methods with real-world problems. In Chapter 12, Bargal et al. describe the recent approaches that utilize spatial and spatiotemporal visual explainability to train models that generalize better and possess more desirable characteristics. In Chapter 13, Becking et al. show how explanation techniques like Layer-wise Relevance Propagation [3] can be leveraged with information theory concepts and can lead to a better network quantization strategy. The next two chapters then exemplify how XAI methods can be applied to various kinds of science problems and extract new findings. In Chapter 14, Marcos et al. apply explanation methods to marine science and show how a landmark-based approach can generate heatmaps to monitor migration of whales in the ocean. In Chapter 15, Mamalakis et al. survey interesting recent results that applied explanation techniques to meteorology and climate science, e.g., weather prediction.

Part 3 presents more interdisciplinary application of XAI methods beyond technical domains. In Chapter 16, Hacker and Passoth provide an overview of legal obligations to explain AI and evaluate current policy proposals.

In Chapter 17, Zhou et al. provide a state-of-the-art overview on the relations between explanation and AI fairness and especially the roles of explanation on human's fairness judgement. Finally, in Chapter 18, Tsai and Carroll review logical approaches to explainable AI (XAI) and problems/challenges raised for explaining AI using genetic algorithms. They argue that XAI is more than a matter of accurate and complete explanation, and that it requires pragmatics of explanation to address the issues it seeks to address.

Most of the chapters fall under Part 2, and we are excited by the variety of XAI research presented in this volume. While by no means an exhaustive collection, we hope this book presents both quality research and vision for the current challenges, next steps, and future promise of explainable AI research.

Acknowledgements. The authors declare that there are no conflict of interests. This work does not raise any ethical issues. Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554, explainable AI, by the German Ministry for Education and Research (BMBF) through BIFOLD (refs. 01IS18025A and 01IS18037A), and by the European Union's Horizon 2020 programme (grant no. 965221), Project iToBoS.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NeurIPS (2018)
2. Adebayo, J., Muelly, M., Liccardi, I., Kim, B.: Debugging tests for model explanations. In: NeurIPS (2020)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
4. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: quantifying interpretability of deep visual representations. In: CVPR (2017)
5. Bengio, Y., Lecun, Y., Hinton, G.: Deep learning for AI. Commun. ACM **64**(7), 58–65 (2021). https://doi.org/10.1145/3448250
6. Brendel, W., Bethge, M.: Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In: ICLR (2019)
7. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. In: NeurIPS (2019)
8. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: ICCV (2019)
9. Fong, R., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: ICCV (2017)
10. Fong, R., Vedaldi, A.: Net2Vec: quantifying and explaining how concepts are encoded by filters in deep neural networks. In: Proceedings of the CVPR (2018)
11. Hoffmann, A., Fanconi, C., Rade, R., Kohler, J.: This looks like that... does it? Shortcomings of latent space prototype interpretability in deep networks. In: ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI (2021)
12. Holzinger, A., Carrington, A., Müller, H.: Measuring the quality of explanations: the System Causability Scale (SCS). KI - Künstliche Intelligenz **34**(2), 193–198 (2020). https://doi.org/10.1007/s13218-020-00636-z

13. Holzinger, A., et al.: Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Inf. Fusion **79**(3), 263–278 (2022). https://doi.org/10.1016/j.inffus.2021.10.007
14. Holzinger, A., Malle, B., Saranti, A., Pfeifer, B.: Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. Inf. Fusion **71**(7), 28–37 (2021). https://doi.org/10.1016/j.inffus.2021.01.008
15. Holzinger, A., Saranti, A., Molnar, C., Biececk, P., Samek, W.: Explainable AI methods - a brief overview. In: Holzinger, A., et al. (eds.) xxAI 2020. LNAI, vol. 13200, pp. 13–38. Springer, Cham (2022)
16. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: NeurIPS (2019)
17. Hu, X., Chu, L., Pei, J., Liu, W., Bian, J.: Model complexity of deep learning: a survey. Knowl. Inf. Syst. **63**(10), 2585–2619 (2021). https://doi.org/10.1007/s10115-021-01605-0
18. Kim, B., et al.: Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: Proceedings of the ICML (2018)
19. Koh, P.W., et al.: Concept bottleneck models. In: ICML (2020)
20. Lakkaraju, H., Arsov, N., Bastani, O.: Robust and stable black box explanations. In: Daumé, H., Singh, A. (eds.) International Conference on Machine Learning (ICML 2020), pp. 5628–5638. PMLR (2020)
21. Mahendran, A., Vedaldi, A.: Visualizing deep convolutional neural networks using natural pre-images. Int. J. Comput. Vis. **120**(3), 233–255 (2016)
22. Marcos, D., Fong, R., Lobry, S., Flamary, R., Courty, N., Tuia, D.: Contextual semantic interpretability. In: Ishikawa, H., Liu, C.-L., Pajdla, T., Shi, J. (eds.) ACCV 2020. LNCS, vol. 12625, pp. 351–368. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69538-5_22
23. Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., Weller, A.: Do concept bottleneck models learn as intended? In: ICLR Workshop on Responsible AI (2021)
24. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
25. Mueller, H., Mayrhofer, M.T., Veen, E.B.V., Holzinger, A.: The ten commandments of ethical medical AI. IEEE Comput. **54**(7), 119–123 (2021). https://doi.org/10.1109/MC.2021.3074263
26. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14933–14943 (2021)
27. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill **2**(11), e7 (2017)
28. Petsiuk, V., Das, A., Saenko, K.: Rise: randomized input sampling for explanation of black-box models. In: Proceedings of the BMVC (2018)
29. Pfeifer, B., Secic, A., Saranti, A., Holzinger, A.: GNN-subnet: disease subnetwork detection with explainable graph neural networks. bioRxiv, pp. 1–8 (2022). https://doi.org/10.1101/2022.01.12.475995
30. Poppi, S., Cornia, M., Baraldi, L., Cucchiara, R.: Revisiting the evaluation of class activation mapping for explainability: a novel metric and experimental analysis. In: CVPR Workshop on Responsible Computer Vision (2021)
31. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: a review of methods and applications. Proc. IEEE **109**(3), 247–278 (2021)

32. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6
33. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
34. Shitole, V., Li, F., Kahng, M., Tadepalli, P., Fern, A.: One explanation is not enough: structured attention graphs for image classification. In: NeurIPS (2021)
35. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: ICLR Workshop (2014)
36. Stoeger, K., Schneeberger, D., Holzinger, A.: Medical artificial intelligence: the European legal perspective. Commun. ACM **64**(11), 34–36 (2021). https://doi.org/10.1145/3458652
37. Yang, M., Kim, B.: Benchmarking attribution methods with relative feature importance (2019)
38. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
39. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 543–559. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_33
40. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene CNNs. In: Proceedings of the ICLR (2015)
41. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)
42. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Interpretable basis decomposition for visual explanation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11212, pp. 122–138. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01237-3_8