ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Trust in Artificial Intelligence: Beyond Interpretability

Tassadit Bouadi[1], Benoît Frénay[2], Luis Galárraga[3],
Pierre Geurts[4], Barbara Hammer[5], Gilles Perrouin[2]

1 - University of Rennes I - Campus de Beaulieu - IRISA Lab, 35042 Rennes
2 - University of Namur - NaDI - Faculty of Computer Science
Rue Grangagnage, 21, 5000 Namur - Belgium
3 - Inria, Rennes Bretagne Atlantique - IRISA Lab, 35042 Rennes
4 - University of Liège - Montefiore Institute - Liège, Belgium
5 - Bielefeld University - CITEC - 33594 Bielefeld, Germany

Abstract. As artificial intelligence (AI) systems become increasingly integrated into everyday life, the need for trustworthiness in these systems has emerged as a critical challenge. This tutorial paper addresses the complexity of building trust in AI systems by exploring recent advances in explainable AI (XAI) and related areas that go beyond mere interpretability. After reviewing recent trends in XAI, we discuss how to control AI systems, align them with societal concerns, and address the robustness, reproducibility, and evaluation concerns inherent in these systems. This review highlights the multifaceted nature of the mechanisms for building trust in AI, and we hope it will pave the way for further research in this area.

1 Introduction

As machine learning and deep learning have become widespread, it has become apparent that trust is a major challenge faced in artificial intelligence (AI). Indeed, experts, users or even ordinary citizens interact on a regular basis with AI systems that recommend content, suggest decisions, control devices, etc. As these interactions affect people's lives, AI systems must be trustworthy. Accordingly, in the EU, the high-level expert group on artificial intelligence[1] has issued the Ethics Guidelines for Trustworthy AI that identify three components for trustworthy AI: it should be lawful, ethical and robust. Seven requirements are thus proposed: (i) human agency and oversight, (ii) technical robustness and safety, (iii) privacy and data governance, (iv) transparency, (v) diversity, non-discrimination and fairness, (vi) societal and environmental well-being and (vii) accountability. Following these principles, this tutorial paper aims to show how the AI literature goes beyond interpretability and offers more complex mechanisms to build trust with AI systems. Section 2 reviews recent trends in XAI (req. iv), Section 3 shows how to control AI systems (req. i), Section 4 discusses how to align them with societal concerns (req. v), Section 5 tackles the robustness and reproducibility of AI algorithms (req. ii), Section 6 reflects on their evaluation (req. vii) and Section 7 concludes with perspectives.

[1] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

2 Trends in XAI

eXplainable AI (XAI) is a sub-field of AI concerned with making AI systems transparent, said otherwise, capable of delivering justifications of the reasons behind a verdict or recommendation. Explainability is particularly appealing for black-box AI models that resort to latent data representations and transformations such as those conducted by neural networks and other machine learning (ML) methods. Moreover, it can be a legal requirement in some circumstances, and an increasing body of evidence suggests its positive impact on user confidence and trust in AI [1]. Research in XAI has gained traction in the last 6 years – in part thanks to the publication of GDPR [2, 3][2] – and its realm transcends the domain of AI for it intersects not only with other subfields of computer science such as human-computer interaction (HCI) or algorithmic fairness but also with disciplines such as psychology and cognitive sciences.

The first efforts to bring explainability to black-boxes models originated from the ML community and focused on post-hoc *global explanations*, mostly on supervised ML models [2]. Later efforts focused then on more precise *local explanations* that shed light on the logic of an AI model when confronted with a particular input, often called the *target instance*. Popular explanation methods such as LIME [4], SHAP [5] or LORE [6] work in a post-hoc manner. However, an important body of literature has put effort into developing neural-based approaches that can "predict" and explain at the same time [7]. Those include the extensive list of methods in neurosymbolic reasoning [8]. While explainability on supervised ML dominates the landscape, some works have also studied the problem of bringing explainability to unsupervised or self-supervised ML tasks such as clustering, representation learning or dimensionality reduction [9, 10, 11].

Despite its prominent position within the AI landscape, the notion of explainability, and the related concept of *interpretability* [12], also depend on a handful of user-related aspects such as the user's background, the stakes behind the interaction with the AI system, and the purpose of the explanation process itself. This realisation has motivated the human-centred AI research community – at the crossroads of the XAI and HCI communities – to study the impact of explanations on some cognitive aspects such as understanding, trust or perceived fairness [13, 14, 15]. The overarching goal of such studies is to derive design lessons for the development of efficient and trustworthy AI systems. This builds upon the assumption that complacency effects aside, explainability increases the trustworthiness of AI systems – an assumption that is confirmed by a large body of literature [1]. On the other hand, the diversity of use cases for XAI has made it impossible to develop a one-size-fits-all explainability technique. Instead, current research in XAI has shifted towards defining or improving different explanation paradigms (feature attribution, rules, abductive, counterfactual, adversarial explanations, etc.) applied to specific models or tasks. Other approaches have focused on endowing explanations with desirable properties such as diversity, realism (for counterfactual explanations) or robustness [16, 17, 18, 19], or sim-

[2] `https://gdpr-info.eu/`

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

ply studying the theoretical and functional (e.g., adherence, stability) properties guaranteed by a particular explanation approach [20, 21].

Finally, the emergence of large language models, smart assistants, and generative AI have triggered new research questions on how to explain the inner workings of such complex models. They pose challenges for XAI researchers not only because of the nature of their underlying techniques but also due to their multi-modal nature, which will require them to extend the notions of explanations. But XAI novel research trends are also concerned with how to exploit the human-like language capabilities of such models to build comprehensible explanations, like the ones a human pedagogue could offer [22].

3 Constraints

One of the major issues with ML and deep learning models is inherent in how they are learned. Indeed, one typically collects information about a task in the form of a dataset, i.e., a collection of pairs of instances (patients, images, etc.) and targets (pathology, class, etc.). Then, the dataset supervises the optimisation of the ML or deep learning model through an objective function (typically, a measure of discrepancy between expected targets and model predictions). Yet, albeit this approach is efficient in obtaining good predictors, it provides no guarantee that it satisfies constraints that arise from the domain itself (e.g., physical constraints) or societal imperatives (e.g., legal or ethical considerations). In many cases, this may be a sufficient reason to not trust and reject models: credit denial based on illegal grounds, predictions that violate physical laws, etc.

Trust in models can be improved if users get some control over them. This can be achieved through different mechanisms, including constraint enforcement, interactivity and alignment with societal imperatives (see Section 4). Constraints can be either model-specific or model-agnostic. For example, decision trees (DTs) can be constrained at three levels [23, 24, 25]. First, one can control the structure of a DT by restricting its number of nodes, depth, number of leaf nodes, etc. Second, constraints can also be expressed w.r.t. attributes that can be associated with costs (of using them in a decision), ordering (of the attributes in decision paths), incompatibilities (some attributes cannot be used in the same decision path), etc. For example, in a medical application, a DT would be expected to start with cheap, harmless tests (e.g., medical examination or blood work) and then move to expensive, harmful tests (e.g., X-ray or invasive procedures). Third, instance-level constraints make sure that some instances are classified (or not) in the same leaf or that the prediction is correct for some critical instances. Another well-known example of model-specific constraints is the L_1 regularisation [26] of linear models to balance their complexity and interpretability.

Model-agnostic constraints are expressed in terms of the prediction properties. For example, the prediction can be constrained to be monotonic w.r.t. some key feature, like in credit rating [27]. Monotonicity can be enforced even in complex neural networks by simply adding a regularisation term, like in Monteiro *et*

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

al. [28]. Otherwise, the model often does not satisfy such constraints, or doing so requires a large amount of data. More complex domain constraints can also be expressed, for example in physics-informed ML, where physical laws are used both to guide and control learning algorithms. Karniadakis *et al.* [29] describe an example where a neural network is learned with two aggregated losses: a classical loss based on data measurements and a "physical" loss that penalises the neural network if its output does not conform to a partial differential equation.

Interaction is another way to constrain ML algorithms. Interactive ML [30, 31] has been an active field of research in dimensionality reduction for information visualisation [32], but also in classification or regression to align models with users' expectations and to make them easier to understand.

4 Ethical, Fairness and Acceptability Issues

A fundamental aspect of trust in AI systems is their adherence to strong ethical requirements. The inclusion of advanced ML algorithms has put *fairness* [33] at the centre of the attention of AI researchers, but also the general public. Indeed, unfair systems can now impact everyone, from being attributed higher chances of crime recidivism because of skin colour [34, 35], or misclassified as "primates" in generated video caption featuring black people[3] [36] to being denied credit because of gender[4]. Such tip-of-the-icebergs abound, are not always easy to detect nor to prevent and have detrimental impact not only on the users, but also on the reputation of the companies owning and deploying these systems.

This problem is not new and the community offers various mitigation strategies, as mapped by recent surveys [37, 38]. The main goal of such techniques is to ensure that a given AI-enabled system gives the same predictions for inputs that differ only on sensitive attributes (gender, age, etc.) that we want to protect from unfair decisions. Equal treatment can occur for individual inputs or groups. Mehrabi *et al.* classify fairness mitigation approaches in three categories [37]. Pre-processing techniques transform the data [39] so that the underlying discrimination is not presented to the learning algorithm. In-processing techniques affect the learning algorithm by itself, by ensuring it cannot predict the value of the sensitive attribute while predicting the main attribute [40]. Finally, post-processing methods learn an auditing model to identify the inputs where the model under study makes more mistakes. Based on such an auditor, one can derive weights to fix predictions [41]. Bellamy *et al.* integrates other post-processing strategies [42]. It is also possible to combine categories by leveraging adversarial ML [43]. In addition, it is also possible to assess the sensitivity to fairness perturbations once the model is deployed in production [44].

Despite the large variety of existing methods, some problems remain. There is no unique definition of fairness as noted by Mehrabi *et. al* [37]. It leads to a plethora of fairness metrics, sometimes conflicting with each other [45]. Verma and Rubin demonstrated that a given classifier may appear fair or unfair depend-

[3] https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html

[4] https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

ing on the used notion of fairness [46]. This also challenges the interpretability of fairness-aware models. The EU guidelines for trustworthy AI mention one should seek an "adequate" definition of fairness. The evaluation of such adequacy remains an open problem [38].

5 Robustness and Reproducibility

Generally speaking, robustness and reproducibility are concerned with the desire that ML models and their explanations do not change significantly if experiments or evaluations are repeated under possibly slightly changing conditions. It is well known that deep architectures, and more generally ML models for high dimensional data, are brittle to small changes in inputs [47]. This effect transfers to explanations of models in different aspects [17, 48, 49].

Explanation schemes can be vulnerable to small variations of the scenario due to their design choices; e.g., gradient-based schemes such as saliency maps suffer from non-differentiable activation functions such as ReLu, hence non-continuous gradients. This renders alternative reference points such as those used in DeepLift preferable [50]. Further, explanations can be adversarially attacked [51], enabling the possibility to manipulate, fool, or fairwash explanations, i.e., explanations get useless or misleading on purpose. As for defences of models against attacks, various defences against adversarial attacks on explanations have been proposed such as robustness to data shift [52]; yet these cannot faithfully prevent all possible attacks. A third challenge is given by the fact that faithful explanations depend on the faithfulness of the underlying model to the modelled process, hence perturbations and noise which cause challenges to process inference do also affect the robustness of explanations [48].

One fundamental problem underlying this brittleness of XAI technologies lies in the fact that many modern explanation techniques are based on correlations rather than causal effects, as their purpose is to increase the accessibility of statistical black box models to humans. Such models are usually based on families of functions which possess the universal approximation capability, the underlying model parameters are neither meaningful nor identifiable, and the uniqueness of a function which fits the given data satisfactorily is not necessarily given. Given such fuzzy starting conditions for XAI techniques, it is not surprising that the latter do not necessarily provide the (unique) explanation a human would expect in typical model situations [53].

6 Experimental evaluation

Experimental evaluation of XAI technologies faces the challenge that there is no unique mathematical notion about what is a good explanation. Hence evaluation schemes can follow several avenues: there exist general desiderata which are intrinsic to the model or explanation scheme such as content-oriented criteria including correctness, completeness, or consistency, and representation-based aspects such as compactness [54].

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Besides intrinsic quantitative evaluation criteria, extrinsic evaluation can be based on benchmark datasets, where data with known ground truth has been generated. Thereby, the ground truth is typically available as data have been manipulated in a specific way, i.e. relevant features or causal relations are known. Examples include benchmarks for the visual domain [55], approaches tailoring the evaluation of large-scale comparison of post-hoc XAI methods [56], or unit tests for attribution methods [53], to name just a few. Thereby, evaluation suites are available in the form of XAI benchmarking tools such as OpenXAI [57] or other toolboxes as summarized in the work [58].

Ultimately, XAI technologies need to be successful in settings where humans are using or interacting with AI technologies, i.e., XAI methodologies need to be evaluated in user studies. Here the problem arises that many effects are dependent on the specific application domain and user expertise, such that systematic overarching effects of XAI technologies are hard to evaluate. Further, it is unclear whether effects are due to the specific XAI technology or they can be traced back to how explanations are expressed, such as demonstrated w.r.t. the direction of an explanation (i.e., whether the same information is phrased positively or negatively [14]). Further, several user studies have been published, targeting not only different XAI technologies, but also evaluation criteria, targeting trust, understanding, or usability [59], whereby effects of XAI technologies are partially positive but partially insignificant, or they do not transfer to effects when transferred across scenarios. Hence overarching insight about positive effects which has been validated in user studies is yet limited.

7 Conclusion

Albeit regulations such as the EU's AI Act demand human agency and oversight of high-risk AI systems as well as foundational models, the question of how far this can reliably be implemented by current XAI technologies and how its auditing can take place remains a challenge. Indeed, it has been debated how far EU's formalization falls prey to a too simplistic conceptualization of trust [60], especially in the light of the stark difference of trust and trustworthiness [1], and the richness of human explanations in social practice [61]. Accordingly, beyond recent advances in dynamic XAI schemes and possibilities of explaining effects of interaction [62, 63], there is a need for contextualized XAI schemes which can appropriately take into account human expertise and expectations as well as the situated objectives of a given scenario [64]. Further, the role of uncertainty and limitations of both, models and XAI technologies, constitutes a very important aspect which is yet only partially understood [65].

Examples in this paper show that trust needs to be achieved through different mechanisms (explicability, constraints, fairness, robustness, auditing, etc.) and offer a bright future for the field that will continue to grow and expand.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Acknowledgements

This work was supported by the *ARIAC by Digital Wallonia4.AI* project number 2010235 of the SPW Recherche / Wallonia Research. Further, we acknowledge support by MKW NRW for the project SAIL, project number NW21-059A. Gilles Perrouin is an FNRS Research Associate.

References

[1] Roel Visser, Tobias M. Peters, Ingrid Scharlau, and Barbara Hammer. Trust, distrust, and appropriate reliance in (X)AI: a survey of empirical evaluation of user trust, 2023.

[2] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black-box Models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[3] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, 2023.

[4] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the Predictions of Any Classifier. In *KDD*, 2016.

[5] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *NeurIPS*, 2017.

[6] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *CoRR*, 2018.

[7] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[8] A. Sheth, K. Roy, and M. Gaur. Neurosymbolic Artificial Intelligence (Why, What, and How). *IEEE Intelligent Systems*, 38(03):56–62, may 2023.

[9] Mohamed H Gad-Elrab, Daria Stepanova, Trung-Kien Tran, Heike Adel, and Gerhard Weikum. ExCut: Explainable Embedding-Based Clustering over Knowledge Graphs. In *International Semantic Web Conference*, 2020.

[10] Adrien Bibal, Viet Minh Vu, Geraldin Nanfack, and Benoît Frénay. Explaining t-sne embeddings locally by adapting lime. In *ESANN 2020*, pages 393–398, October 2020.

[11] Adrien Bibal, Antoine Clarinval, Bruno Dumas, and Benoît Frénay. Ixvc: An interactive pipeline for explaining visual clusters in dimensionality reduction visualizations with decision trees. *Array*, 11:100080, 2021.

[12] Adrien Bibal and Benoît Frénay. Interpretability of machine learning models and representations: an introduction. In *24th european symposium on artificial neural networks, computational intelligence and machine learning*, pages 77–82. CIACO, 2016.

[13] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404, 2021.

[14] Ulrike Kuhl, André Artelt, and Barbara Hammer. For better or worse: The impact of counterfactual explanations' directionality on user behavior in xai. In Luca Longo, editor, *Explainable Artificial Intelligence*, pages 280–300, Cham, 2023. Springer Nature Switzerland.

[15] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proc. CHI*. ACM, 2021.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

[16] Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, and Alexandre Termier. Generating robust counterfactual explanations. In Danai Koutra, Claudia Plant, Manuel Gomez Rodriguez, Elena Baralis, and Francesco Bonchi, editors, *Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023, Turin, Italy, September 18-22, 2023, Proceedings, Part III*, volume 14171 of *Lecture Notes in Computer Science*, pages 394–409. Springer, 2023.

[17] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Robust counterfactual explanations in machine learning: A survey. *ArXiv*, abs/2402.01928, 2024.

[18] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 344–350, New York, NY, USA, 2020. Association for Computing Machinery.

[19] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 607–617, New York, NY, USA, 2020. Association for Computing Machinery.

[20] Gwladys Kelodjou, Laurence Rozé, Véronique Masson, Luis Galárraga, Romaric Gaudel, Maurice Tchuente, and Alexandre Termier. Shaping Up SHAP: Enhancing Stability through Layer-Wise Neighbor Selection. volume 38, pages 13094–13103, Mar. 2024.

[21] Julien Delaunay, Luis Galárraga, and Christine Largouët. When Should We Use Linear Explanations? In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 355–364, New York, NY, USA, 2022. Association for Computing Machinery.

[22] Dylan Z Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Explaining machine learning models with interactive natural language conversations using talktomodel. *Nature Machine Intelligence*, 5, 2023.

[23] Jan Struyf and Sašo Džeroski. Clustering trees with instance level constraints. In *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings 18*, pages 359–370. Springer, 2007.

[24] Siegfried Nijssen and Elisa Fromont. Optimal constraint-based decision tree induction from itemset lattices. *Data Mining and Knowledge Discovery*, 21:9–51, 2010.

[25] Géraldin Nanfack, Paul Temple, and Benoît Frénay. Constraint enforcement on decision trees: A survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–36, 2022.

[26] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–451, 2004.

[27] Chih-Chuan Chen and Sheng-Tun Li. Credit rating with a monotonicity-constrained support vector machine model. *Expert Systems with Applications*, 41(16):7235–7247, 2014.

[28] João Monteiro, Mohamed Osama Ahmed, Hoseein Hajimirsadeghi, and Greg Mori. Monotonicity regularization: Improved penalties and novel applications to disentangled representation learning and robust classification. In *Uncertainty in Artificial Intelligence*, pages 1381–1391. PMLR, 2022.

[29] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[30] Liu Jiang, Shixia Liu, and Changjian Chen. Recent research advances on interactive machine learning. *Journal of Visualization*, 22:401–417, 2019.

[31] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

[32] Dominik Sacha, Leishi Zhang, Michael Sedlmair, John A Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C North, and Daniel A Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE transactions on visualization and computer graphics*, 23(1):241–250, 2016.

[33] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[34] Lauren Kirchner Jeff Larson, Surya Mattu and Julia Angwin. How we analyzed the compas recidivism algorithm, 2016.

[35] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. PMID: 28632438.

[36] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 429–435, New York, NY, USA, 2019. Association for Computing Machinery.

[37] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021.

[38] Ezekiel Soremekun, Mike Papadakis, Maxime Cordy, and Yves Le Traon. Software fairness: An analysis and survey. *arXiv preprint arXiv:2205.08809*, 2022.

[39] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1 – 33, 2011.

[40] Harrison Edwards and Amos Storkey. Censoring representations with an adversary, 2016.

[41] Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 247–254, New York, NY, USA, 2019. Association for Computing Machinery.

[42] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019.

[43] Pieter Delobelle, Paul Temple, Gilles Perrouin, Benoit Frénay, Patrick Heymans, and Bettina Berendt. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *SIGKDD Explor. Newsl.*, 23(1):32–41, may 2021.

[44] Camille Molinier, Paul Temple, and Gilles Perrouin. Fairpipes: Data mutation pipelines for machine learning fairness. In *AST 2024-5th ACM/IEEE International Conference on Automation of Software Test*, pages 1–11. ACM, 2024.

[45] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[46] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

[47] Jia Liu and Yaochu Jin. A comprehensive survey of robust deep learning in computer vision. *Journal of Automation and Intelligence*, 2(4):175–195, 2023.

[48] Benedikt Kantz, Clemens Staudinger, Christoph Feilmayr, Johannes Wachlmayr, Alexander Haberl, Stefan Schuster, and Franz Pernkopf. Robustness of explainable artificial intelligence in industrial process modelling, 2024.

[49] Xuanxiang Huang and Joao Marques-Silva. From robustness to explainability and back again, 2023.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

[50] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. *Gradient-Based Attribution Methods*, pages 169–191. Springer International Publishing, Cham, 2019.

[51] Hubert Baniecki and Przemyslaw Biecek. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 107:102303, July 2024.

[52] Anna P. Meyer, Dan Ley, Suraj Srinivas, and Himabindu Lakkaraju. On minimizing the impact of dataset shifts on actionable explanations. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI '23. JMLR.org, 2023.

[53] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features?, 2021.

[54] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, 55(13s), jul 2023.

[55] Yifei Zhang, Siyi Gu, James Song, Bo Pan, and Liang Zhao. Xai benchmark for visual explanation, 2023.

[56] Samuel Sithakoul, Sara Meftah, and Clément Feutry. Beexai: Benchmark to evaluate explainable ai. In Luca Longo, Sebastian Lapuschkin, and Christin Seifert, editors, *Explainable Artificial Intelligence*, pages 445–468, Cham, 2024. Springer Nature Switzerland.

[57] Chirag Agarwal, Dan Ley, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations, 2024.

[58] Phuong Quynh Le, Meike Nauta, Van Bach Nguyen, Shreyasi Pathak, Jörg Schlötterer, and Christin Seifert. Benchmarking explainable ai - a survey on available toolkits and open challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6665–6673, 8 2023. Survey Track.

[59] Yao Rong, Tobias Leemann, Thai trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: A survey of user studies for model explanations, 2023.

[60] Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32, 2024.

[61] Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath, Eyke Huellermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. Explanation as a social practice: Toward a conceptual framework for the social design of ai systems. *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–1, 2020.

[62] Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: Unified Approximation of any-order Shapley Interactions. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 11515–11551. Curran Associates, Inc., 2023.

[63] Fabian Fumagalli, Maximilian Muschalik, Eyke Hüllermeier, and Barbara Hammer. Incremental permutation feature importance (ipfi): towards online explanations on data streams. *Machine Learning*, 112(12):4863–4903, Dec 2023.

[64] Tjeerd A.J. Schoonderwoerd, Wiard Jorritsma, Mark A. Neerincx, and Karel van den Bosch. Human-centered xai: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, 154:102684, 2021.

[65] Teodor Chiaburu, Frank Haußer, and Felix Bießmann. Uncertainty in xai: Human perception and modeling approaches. *Machine Learning and Knowledge Extraction*, 6(2):1170–1192, 2024.