Implementing ethical principles in AI: an initial discussion

Mykhailo Danilevskyi, Fernando Perez-Tellez, Davide Buscaldi

AI and Ethics
https://doi.org/10.1007/s43681-025-00710-y

REVIEW

Implementing ethical principles in AI: an initial discussion

Mykhailo Danilevskyi[1] · Fernando Perez-Tellez[1] · Davide Buscaldi[2]

Received: 30 October 2024 / Accepted: 14 March 2025

Abstract
In recent years, there has been a lot of discussion around ethics in IT and AI. Many researchers and organisations have proposed guidelines to address privacy, fairness, and explainability challenges for creating trustworthy AI. In this paper, we discuss ethical principles in the context of AI and their significance in developing trustworthy AI solutions. We consider the problem of the categorisation of ethical principles in IT. We concentrate our discussion on privacy, fairness, and explainability. These principles, we believe, meaningfully contribute to the trust of AI systems. We overview the available privacy regulations in the EU and US. We also look at how to achieve compliance with them, including private data detection, data anonymisation techniques and toolkits. From a practical perspective, we analyse fairness and bias problems. We discuss the issue of fairness assessment and metrics. To improve the fairness of AI solutions, an enormous number of techniques have been developed. We also focus on fairness improvement techniques and a few popular toolkits in which these techniques are implemented. Explainability is another ethical principle discussed. It is one of many socially important properties, as it ensures understanding of AI system decision-making and transparency in inspection. Ensuring explainability is important for high-risk applications in healthcare, finance and criminal justice. Finally, we outline approaches that help in the level of explainability. With this review and analysis, we contribute to the knowledge of available techniques and toolkits that can be used by AI practitioners as an initial step in implementing ethical principles into AI solutions.

Keywords Ethical AI · Privacy · Fairness · Explainability

1 Introduction

In recent years, ethics in IT and, in particular AI, received a lot of attention. Ethics, as a system of principles, guides people in distinguishing between right and wrong. Ethics in AI and IT guides how to build programming solutions that people can trust on. [1] recognizes ethics as a building block of trustworthy AI and trustworthiness as a prerequisite for people and societies when developing, deploying and using AI systems.

In this paper, we concentrate our discussion on privacy, fairness, and explainability. These principles, we believe, meaningfully contribute to the trust in AI systems. Privacy is one of the fundamental rights. Its protection makes people feel safe and encourages data sharing among institutions for research and development of AI solutions. Due to long-standing historical biases and a lack of data across a wide range of demographic groups, fairness is one of the most difficult moral ideals to implement in AI systems. Explainability principle is often identified as a requirement of an AI system. This principle ensures understanding of AI system decision-making and transparency to inspection—to name one of many socially important properties [2].

From a practical perspective, ethical principles require understanding at the level of people who design and develop artificial intelligence solutions. Moreover, developers need

Mykhailo Danilevskyi, Fernando Perez-Tellez and Davide Buscaldi have contributed equally to this work.

✉ Mykhailo Danilevskyi
d22126578@mytudublin.ie

Fernando Perez-Tellez
fernando.pereztellez@tudublin.ie

Davide Buscaldi
davide.buscaldi@lipn.univ-paris13.fr

[1] School of Enterprise Computing and Digital Transformation, Technological University Dublin, Blessington Rd, Dublin D24 FKT9, Ireland

[2] Laboratoire d'Informatique de Paris Nord, Université Sorbonne Paris Nord, 99, Avenue Jean-Baptiste Clément, 93430 Paris, Villetaneuse, France

Published online: 15 April 2025

to be aware of techniques and programming tools that can help implementing ethical principles in software products. In this paper, we review and analyse some of these tools and techniques for the presented ethical principles.

The paper consists of 7 sections. In Section 2, we describe the methodology of the literature search. In Section 3, we discuss the existence of guidelines, relevance of ethical principles, their importance for various fields and the problem of categorising ethical principles. In Sections 4, 5, 6 we consider privacy, fairness and explainability principles. And finally, in Section 7 the conclusion is presented.

2 Methodology

This paper presents a narrative review of research on ethical principles in AI, focusing on privacy, fairness, explainability and their practical implementation. It aims to provide an overview of ethical principles in AI, guidelines and regulations, available techniques and toolkits that can assist AI practitioners in building trustworthy AI systems.

A structured literature search was conducted using the following keywords, including but not limited to: AI ethics, trustworthy AI, ethical AI frameworks, AI guidelines and regulations, privacy in AI, data privacy regulations, bias and fairness of AI models, bias mitigation techniques, explainable AI, AI model explainability techniques.

Searches were performed on multiple academic databases, including ScienceDirect, IEEE Xplore, Springer Link, ACM Digital Library, Google Scholar, and arXiv.

The selection of studies was based on their relevance to AI ethics principles, particularly privacy, fairness, and explainability. Studies that discussed AI governance frameworks, ethical guidelines, and regulations were included, as well as those that presented tools, techniques, or methodologies for implementing AI ethics principles. Only peer-reviewed journal articles, conference papers, and reputable industry reports were considered.

3 Ethical principles in AI

The value of AI and the harm it can cause to people are recognised by society. As a result, a large number of guidelines for trustworthy AI have been developed [3–5]. By 2019, 84 documents containing ethical principles or guidelines were identified [4]. It is also observed that multiple definitions of each ethical principle exist, for example, fairness has more than a hundred metrics and definitions [6, 7]. As presented in [4] "a global convergence emerging around five ethical principles—transparency, justice and fairness, non-maleficence, responsibility and privacy". However, the research made by [8] outlines numerous gaps in the AI governance frameworks, including the lack of details on how to achieve a fair society, respect diversity, and address the potential for AI to drive inequalities and bias in the labour market.

In 2021, the European Commission proposed the Artificial Intelligence Act [9]. AI Act would be the first comprehensive law to regulate the development and use of AI systems. It differentiates AI systems depending on the application risk—unacceptable, high and limited-risk applications. Unacceptable risk applications, including social scoring and biometric identification systems, will be banned. High-risk applications, for instance, in toys, aircrafts, cars, will be assessed during development and before becoming open for usage. Limited risk applications, such as media content editors, have to comply only with minimal transparency requirements and the user can decide whether to use them or not. The presented categorisation helps to imply specific rules for each group to reduce potential harm to people and increase the level of compliance with ethical principles.

The importance of ethical principles varies. According to the survey [5], the most important principle in the ICT industry is responsibility, followed by protection of data privacy, transparency, robustness, minimised bias, and "an AI should have a purpose" meaning that the highest purpose of AI is to support people, not replace them.

The World Health Organization recognises the great potential of AI and to reap the benefits of AI, ethical issues must be addressed [10]. In the medical field, the full effectiveness of AI solutions is limited by the inability of algorithms to explain results to human experts. The authors of [11] recognise explainability, transparency, accountability and trust as very important ethical principles in the healthcare domain.

Diversity of views on ethical principles as well as on their categorisation certainly exists. The High-Level Expert Group on AI that was set up by European Commission in June 2018 formulated the Guidelines to promote Trustworthy AI [1]. They outlined four ethical principles, including respect of human autonomy, prevention of harm, fairness and explicability. Preservation of privacy is not on their list of ethical principles. They identified privacy as a technical requirement among seven others: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being and accountability.

The authors of [12] created a graph of relationships between different aspects of AI trustworthiness. They define privacy, fairness and accountability as ethical requirements, and explainability as a technical requirement. From a practical perspective, the ethical principles presented as

requirements to an AI system are more natural for developers and technicians. Therefore, we need a methodology that translates ethical principles into technical requirements and actions that need to be taken during AI solution development lifecycle.

4 Privacy

Privacy is one of the fundamental human rights included in the Universal Declaration of Human Rights [13]. GDPR protects the right to know about the use of personal data, the right to update personal information, the right to restrict the use of data, and others. Nevertheless, it does not forbid the free movement of data [14]. American regulations are more specific. HIPAA regulates privacy in the healthcare domain, FERPA in education and CCPA protects consumers from unwanted data usage. Identification and anonymisation of personal information are two key actions required for mitigating privacy concerns. The authors in [15] distinguish two groups of identification methods—rule-based and machine learning (ML) methods, including deep learning (DL). "The final solution to the problem can also be a combination of several approaches used together for different parts of the task, or one approach can be the input to another approach" [15]. For example, the authors in [16] experimented with a neural network model that feeds output vectors of long short-term memory (LSTM) network into a conditional random fields (CRF) layer for NER tasks. The choice of approaches to identifying private information depends on the structure of the text dataset, the length of the text instances, and the need to understand the context. The shorter and more structured text means rule-based methods may be suitable, otherwise—less structured and long text—require ML or DL methods.

Data anonymisation involves modification of sensitive data including personally identifiable information. The level and type of data anonymisation should be chosen in consultation with legal experts. As a "gold standard" of anonymisation, differential privacy is considered within legal circles [17]. "Differential privacy is a system for sharing information about a dataset by describing patterns about groups in the dataset without disclosing information about specific individuals" [18, 19]. The recent advances in large language models that are able to understand context can make the data anonymisation task less challenging. In the experiment, the authors in [20] used a chain of GPT-3 for NER task and ChatGPT for pseudonymisation. As a result, it was shown that LLMs can accurately pseudonymise sensitive information while preserving the syntactic and semantic integrity of the original text.

For the identification and anonymisation of sensitive data, multiple toolkits have been developed. For instance, Microsoft Presidio[1] can be used for private data identification and anonymisation in text and images. It leverages regular expressions, rule-based logic, NER and can detect names, credit card numbers, emails, etc. Anonymisation stage is performed by operators such as redact, hash, mask, replace, encrypt. Another tool or even platform named Kodex[2] can anonymise almost any kind of structured data in batch and streaming modes. For data anonymisation, it leverages cryptography, aggregation, randomisation, probabilistic techniques, and differential privacy. Another algorithm is Format Preserving Encryption (FPE) [21]. It is applicable for transforming sensitive data in a specified format with a fixed length, such as credit card numbers or phone numbers into identically formatted data.

The choice of data protection methods can be challenging. On the one hand, the data must be protected; on the other hand, it must not lose its value for downstream tasks. In this case, there is a trade-off between data security and data quality that must be balanced to preserve the value of the AI system.

5 Fairness

The requirement for AI systems to be fair in their decisions is more important than ever. In the literature, no universal definition of fairness exists. Additionally, it is common for AI systems aim to meet multiple fairness constraints, for instance, group and individual fairness.

The terms fairness and bias are often used interchangeably. A bias measurement or fairness evaluation determines whether individuals with various protected characteristics, such as gender, age, and ethnicity, are treated fairly and are not subjected to discrimination. The evaluation can be performed on a dataset prepared for training the ML model, as well as on a dataset that includes decisions made by the trained ML model. Among the metrics that can be used for the assessment are demographic parity [22], equalized odds [23], equal opportunity [23] and others.

Depending on the outcome of the fairness assessment, particularly if bias exceeds acceptable limits, mitigation may be necessary. In total, there are more than 100 unique bias mitigation methods [6]. These methods are grouped into three categories depending on the stage of their application in the AI solution development lifecycle: pre-processing, in-processing and post-processing [6, 7]. According to the survey [6], the most used are in-processing methods, followed

[1] https://microsoft.github.io/presidio/.

[2] https://heykodex.com/.

by less commonly used pre-processing and post-processing methods. Pre-processing techniques aim to modify or weight data instances to reduce bias in the dataset. For instance, Re-weighting technique [24] adds additional column with weights, Disparate Impact Remover [25] modifies numeric attributes, whereas Learning Fair Representation [26] and Optimized pre-processing [27] create new data representations. Adversarial Debiasing [28] and Prejudice Remover [29] methods are examples of in-processing approaches that apply constraints to the loss function and control fairness with regularisation parameters, respectively. Post-processing techniques, including Calibrated eq. odds post-processing [30] and Reject option classifications [31], adjust model output to satisfy applied fairness constraints. As a response to the fairness problem, toolkits that include fairness metrics and mitigation techniques have been developed. Two of the most popular open-source toolkits, AI Fairness 360 (AIF360) [32] and Fairlearn [33], were compared among practitioners [34]. AIF360 was created by IBM Research. It contains 71 fairness metrics and 9 bias mitigation techniques. Among the metrics implemented in this framework are popular metrics such as statistical parity difference [18], disparate impact [25], and equalized odds [23]. Bias mitigation algorithms are implemented for the pre-processing, in-processing, and post-processing stages of the ML pipelines. Furthermore, the toolkit includes a facility for metric explanations to help end users understand the meaning of bias detection results. This feature is also suitable for building bias audit reports.

Fairlearn toolkit is grounded in the understanding that fairness is a sociotechnical challenge [33]. For fairness assessment, the toolkit contains all popular metrics, including demographic parity, equalized odds, equal opportunity. The toolkit contains one pre-processing algorithm, Correlation Remover, which removes correlation with sensitive features. Among the in-processing mitigation techniques, two approaches are implemented: redaction approach [22] and adversarial mitigation approach [28]. A post-processing technique, proposed in the paper "equality of opportunity in supervised learning" [23], is implemented in the toolkit and identifies a separate threshold for sensitive groups.

Another toolkit, called Fairkit-learn [35], is based on scikit-learn and AIF360. It has the same interface as scikit-learn, and utilises all the metrics and bias mitigation strategies from AIF360. A major contribution of the toolkit is the ability to search for an optimum set of models that balance fairness and performance. The toolkit performs a grid search over multiple models, hyperparameters, applies different bias mitigation techniques available in AIF360, and selects Pareto optimal set of models that satisfy one or multiple fairness metrics.

Among other tools, it is worth mentioning Google What-If[3] and Aequitas [36]. Google What-If is good for data analysts due to its excellent graphical interface for data exploration. Aequitas can generate tables, charts, PDF reports with metric values across analysed groups that are interpretable by technical users and policymakers.

Bias mitigation techniques may have a negative effect on AI model accuracy. We face a trade-off issue with fairness protection that is comparable to that with privacy. Additionally, pre-processing techniques also change data that might not be acceptable for high-risk applications. The choice of fairness metrics and bias mitigation techniques, as well as their effect on AI system decisions, should be thoroughly analysed by domain experts and technicians.

6 Explainability

Explainability is an active area of research due to its importance in the development of AI solutions for healthcare, finance, criminal justice and education. It is also one of the requirements for building trustworthy AI according to the Ethics Guidelines for Trustworthy AI [1], though, it is named as a principle of explicability. Explainability helps to understand how an AI model makes decisions [37]. The level of explainability limits AI application in the medical domain due to the inability to "explain" its results to human experts [11]. However, models with high explainability level have been recognised as being disadvantageous in terms of performance [12].

The problem of explainability of ML and DL models arises due to the internal structure of such models. Models, by design, can be divided into models explainable by design, and black-box models. The first group includes ML models such as linear regression, trees, KNN. These models require manual feature development; therefore, a data scientist explores the influence of each feature on the data analysis stage and has intuitive understanding of feature importance. The other group of models, mainly DL models, do not require manual feature generation and generally have higher accuracy, but explainability is their disadvantage. These models include various kinds of neural networks—CNN, RNN, and LSTM. To explain the outcomes of such models so-called post-hoc methods are used. The idea behind them is to perform experiments with inputs and further analysis of intermediate state and outcome. The approaches for post-hoc model explanations include model approximation with a more explainable model, analysis of feature importance using tools such as LIME [38] and SHAP [39], feature introspection, example or rule-based

[3] https://pair-code.github.io/what-if-tool/.

explanation. A comprehensive guide of explainability methods is presented by [40].

The problem of explainability has led to the research and development of new ML algorithms that are nonlinear, highly accurate and directly interpretable, such as explainable neural networks. Other examples of interpretable models are Explainable boosting machines (EBMs, also known as GA2M), Monotonically constrained gradient boosting machines, Skope-rules, Supersparse linear integer models (SLIMs), RuleFit [19].

7 Conclusion

This paper examines the problem of compliance with ethical principles in the field of IT and AI. We review the current state of regulations and guidelines, as well as the categorisation and relevance of ethical principles. The problem of the declarative nature of the guidelines and the lack of their influence on decision-making when developing an AI system is outlined. We discuss privacy, fairness, and explainability principles, as well as a limited number of existing techniques and toolkits for ensuring compliance with ethical principles. To comply with privacy principle, AI creators must address four major challenges—ensure compliance with applicable regulations, detect and, if needed, transform private information and preserve data quality. To ensure an AI system complies with fairness principle, it is important to identify fairness, determine ways to measure it for a specific use-case and consider applying bias mitigation techniques. AIF360 and Fairlearn are among widely used toolkits for fairness assessment and bias mitigation. Explainability principle is identified as very important for the medical domain. Explainable models as well as post-hoc approaches, including LIME and SHAP, are used to improve explainability. Ensuring compliance with privacy, fairness and explainability has a negative influence on AI decision accuracy; therefore, metrics and techniques should be chosen considering the final efficiency of AI system. This work contributes to the knowledge of AI ethics through a summary and analysis of the methods and toolkits available to practitioners for ensuring implementation of ethical principles in AI solutions.

Author Contributions The initial concept and structure of this manuscript were proposed by F.PT. M.D. took primary responsibility for drafting the manuscript. F.PT. and D.B. contributed to shaping the paper's structure and provided ongoing feedback. All authors reviewed the manuscript.

Funding Open Access funding provided by the IReL Consortium
No funding was received to assist with the preparation of this manuscript.

Data availibility No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

Clinical trial number Not applicable.

References

1. High-Level Expert Group on AI: Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (2019)
2. Bostrom, N., Yudkowsky, E.: The ethics of artificial intelligence. In: Artificial Intelligence Safety and Security, pp. 57–69. Chapman and Hall/CRC (2018)
3. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
4. Jobin, A., Ienca, M., Vayena, E.: Artificial intelligence: the global landscape of ethics guidelines, 42 (2019)
5. Rothenberger, L., Fabian, B., Arunov, E.: Relevance of ethical guidelines for artificial intelligence: a survey and evaluation. In: Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden. ISBN 978-1-7336325-0-8 Research-in-Progress Pape. https://aisel.aisnet.org/ecis2019_rip/26June 8–14 (2019)
6. Hort, M., Chen, Z., Zhang, J., Sarro, F., Harman, M.: Bias mitigation for machine learning classifiers: a comprehensive survey arXiv:2207.07068 (2022)
7. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021).
8. Bird, E., Fox-Skelly, J., Jenner, N., Larbey, R.: The Ethics of Artificial Intelligence: Issues and Initiatives. Publications Office. https://data.europa.eu/doi/10.2861/6644 (2020). Accessed 21 08 2023
9. European Commission: Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206 (2021). Accessed 14 02 2024
10. World Health Organization: Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. World Health Organization, Geneva (2021)

11. Muller, H., Mayrhofer, M.T., Van Veen, E.-B., Holzinger, A.: The ten commandments of ethical medical AI. Computer **54**(7), 119–123 (2021). https://doi.org/10.1109/MC.2021.3074263.
12. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B.: Trustworthy AI: from principles to practices. ACM Comput. Surv. **55**(9), 1–46 (2023)
13. United Nations: Universal Declaration of Human Rights (1948)
14. European Commission: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). European Commission. https://eur-lex.europa.eu/eli/reg/2016/679/oj (2016)
15. Kuzina, V., Vusak, E., Jovic, A.: Methods for automatic sensitive data detection in large datasets: a review. In: 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), pp. 187–192. IEEE. https://doi.org/10.23919/MIPRO52101.2021.9596735. https://ieeexplore.ieee.org/document/9596735/ (2021). Accessed 08 08 2023
16. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1064–1074. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1101. http://aclweb.org/anthology/P16-1101 (2016). Accessed 14 09 2023
17. Jarmul, K.: Practical Data Privacy: Enhancing Privacy and Security in Data. O'Reilly media (2023)
18. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness arXiv:1104.3913 (2011)
19. Hall, P., Gill, N., Cox, B.: Responsible Machine Learning. O'Reilly Media, Incorporated (2020)
20. Yermilov, O., Raheja, V., Chernodub, A.: Privacy- and utility-preserving NLP with anonymized data: a case study of pseudonymization. https://doi.org/10.48550/arXiv.2306.05561arXiv:2306.05561 (2023)
21. Dworkin, M.: Recommendation for Block Cipher Modes of Operation: Methods for Format-Preserving Encryption. (2016). https://doi.org/10.6028/NIST.SP.800-38G.https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-38G.pdf Accessed 17 09 2023
22. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 60–69. PMLR. https://proceedings.mlr.press/v80/agarwal18a.html (2018)
23. Hardt, M., Price, E., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf (2016)
24. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012). https://doi.org/10.1007/s10115-011-0463-8
25. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15, pp. 259–268. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2783258.2783311
26. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 325–333. PMLR. Issue: 3. https://proceedings.mlr.press/v28/zemel13.html (2013)
27. Calmon, F.P., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R.: Optimized pre-processing for discrimination prevention. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 3995–4004. Curran Associates Inc., Red Hook, NY, USA (2017)
28. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 335–340 (2018)
29. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) Machine Learning and Knowledge Discovery in Databases, pp. 35–50. Springer (2012)
30. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf (2017)
31. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: Proceedings—IEEE International Conference on Data Mining, ICDM, pp. 924–929. (2012) https://doi.org/10.1109/ICDM.2012.45
32. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias arXiv:1810.01943 (2018)
33. Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., Madaio, M.: Fairlearn: Assessing and Improving Fairness of AI Systems. arXiv. arXiv:2303.16626 (2023)
34. Harshitaa, P.: Comparison of the usage of fairness toolkits amongst practitioners:AIF360 and Fairlearn (2022)
35. Johnson, B., Brun, Y.: Fairkit-learn: A fairness evaluation and comparison toolkit. In: 2022 IEEE/ACM 44th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), pp. 70–74 (2022). https://doi.org/10.1145/3510454.3516830
36. Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., Ghani, R.: Aequitas: A bias and fairness audit toolkit arXiv:1811.05577 (2018)
37. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI **58**, 82–115 (2020) https://doi.org/10.1016/j.inffus.2019.12.012. Place: NLD Publisher: Elsevier Science Publishers B. V
38. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, (2016), pp. 1135–1144
39. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf (2017)
40. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S.: Benchmarking and survey of explanation methods

for black box models. Data Min. Knowl. Disc. **37**(5), 1719–1778 (2023). https://doi.org/10.1007/s10618-023-00933-9