Fairness in AI and its relation to social well-being

WHITE PAPER

Trustworthy artificial intelligence

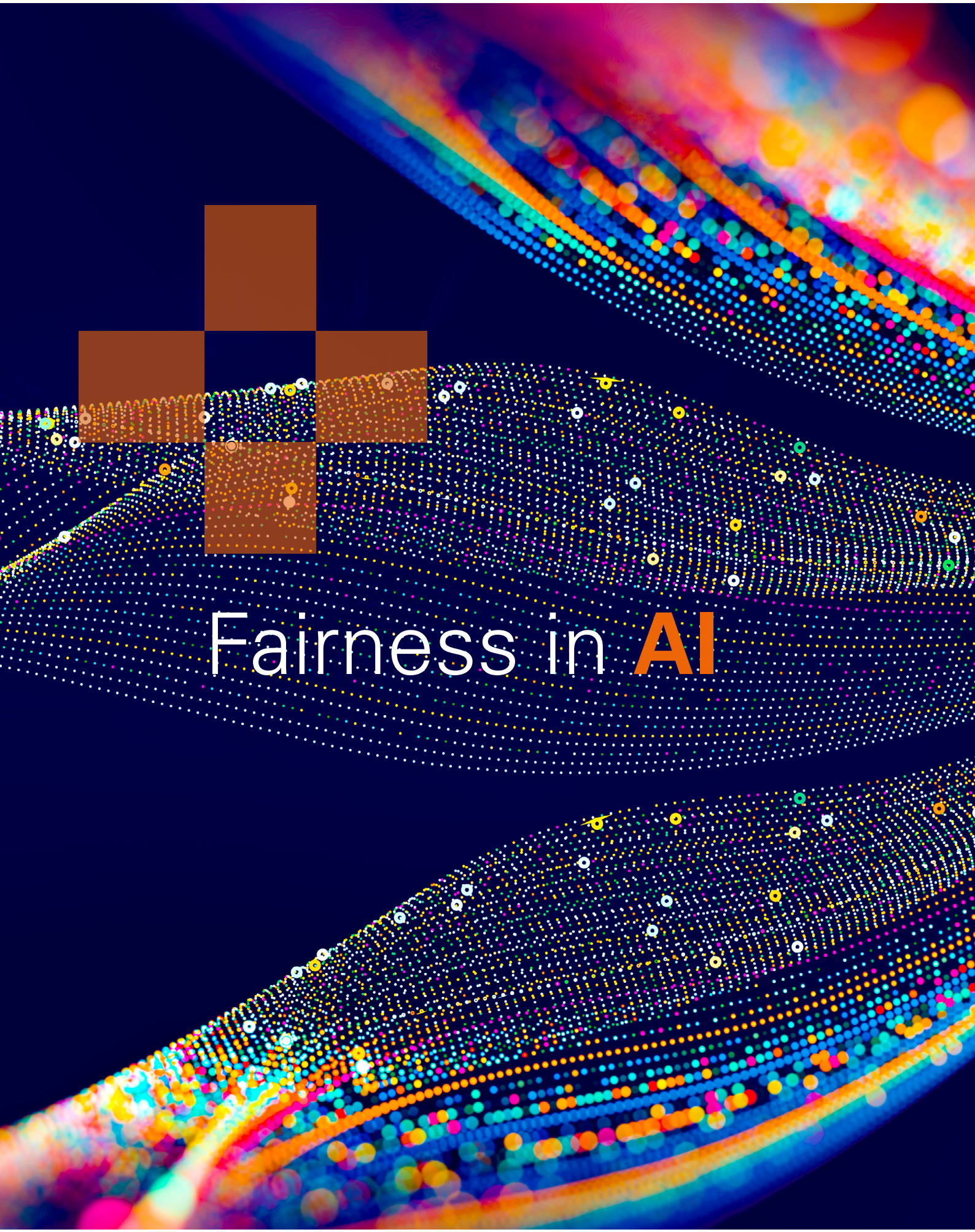
27 September 2023
Simone Kopeinik*, Sebastian Scher*, Tomislav Nad** and Dominik Kowald*

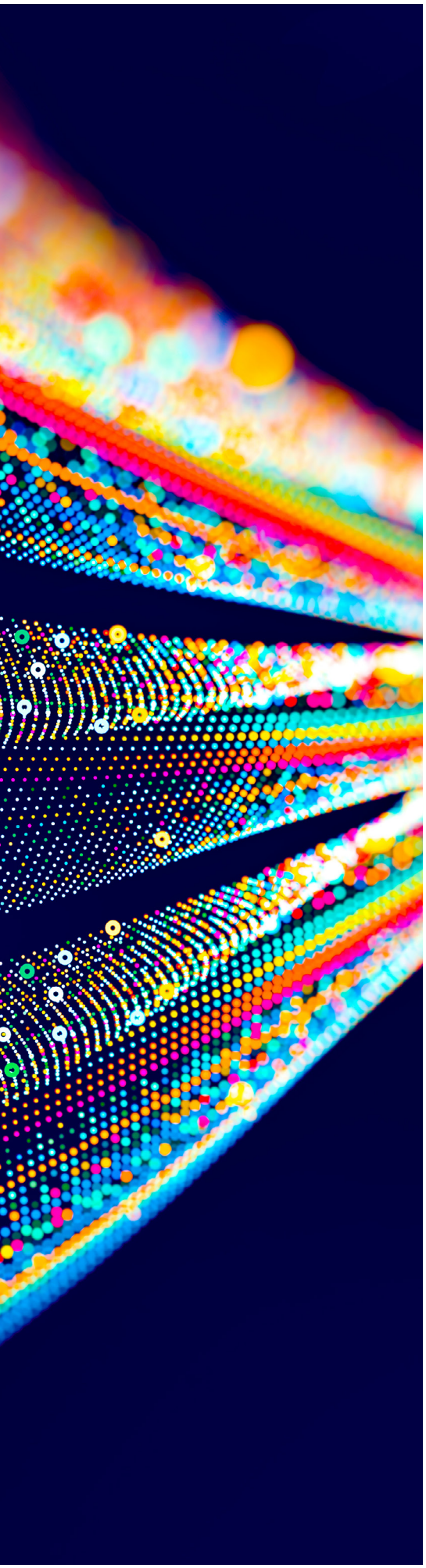
* Know-Center GmbH
** SGS Digital Trusts Services GmbH

Partners of SGS

KNOW Center

SGS


Fairness in AI



Contents:

1 Introduction

Artificial intelligence- (AI) based decision support systems are widely employed in private companies, governmental institutions and other organizations. With the rising application of AI products in various fields and domains, the influence and impact of AI results increases and becomes a matter of public interest. Depending on the application field, this might impact the well-being of individuals and society as a whole. This is especially problematic in sensitive domains like criminal justice, employment, education or health, where it can lead to serious consequences such as being denied medical treatment or an educational scholarship.

The AI incident database [1] is a collection of AI incidents listing over 2,636 incident reports so far (status as of May 2023), where many are related to fairness issues (e.g. #19, #74, #92, …, #375, #416). The most named companies are Facebook with 48 incidents, Tesla with 36, and Google with 28. Examples of notable incidents: Google ads showed high-paid jobs to women and minority groups less often [2], Apple Card gave "females lower credit limits than equally qualified males" and commercial face recognition systems turned out to work very poorly for black-skinned women [3].

Even in domains that are at first glance categorized as non-critical, unfair bias might influence individuals, society and businesses in a substantial manner. From a business perspective, there are different levels of risks and consequences that need to be considered (Fancher et al.2021): missing out on opportunities, reputational damage and regulatory and compliance problems. An example of missing out on opportunities would be a recommender system whose suggestions only benefit a dominant user group. While the individuals of the dominant group would be satisfied with the system, the AI provider would probably lose all other groups from using their product. Reputational damage can occur in cases where sensitive societal problems are hit, for example, a face recognition software that was trained on data not representing the diversity of the real population and thus works well only for e.g. white people. This can raise public debate, resulting in a potentially strong negative reputation for the company. Finally, in applications that are regulated by anti-discrimination law, such as the job market, unfair algorithms – such as a job recommender system that discriminates based on gender, age or race – can lead to legal problems, including fines and/or penalties.

In addition to existing legal regulations for non-discrimination in general, there is rapid development of regulation of AI in particular. The AI ACT is a legal proposal of the European Commission that guides the understanding of requirements towards lawful AI in the European market. This includes the formulation of fairness, diversity and non-discrimination in AI.

[1] *https://incidentdatabase.ai/*
[2] *https://incidentdatabase.ai/cite/19/*
[3] *https://incidentdatabase.ai/cite/92*
[4] *http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline*

Definition of terms

Bias is a term used in many contexts. In the context of fairness in AI it refers to outcomes that are of disproportionate advantage or disadvantage for a specific group of individuals, e.g." Systematic discrimination combined with an unfair outcome is considered to result in bias." (Bird et al. 2019). Accordingly, **fairness** is the absence of discriminatory bias.

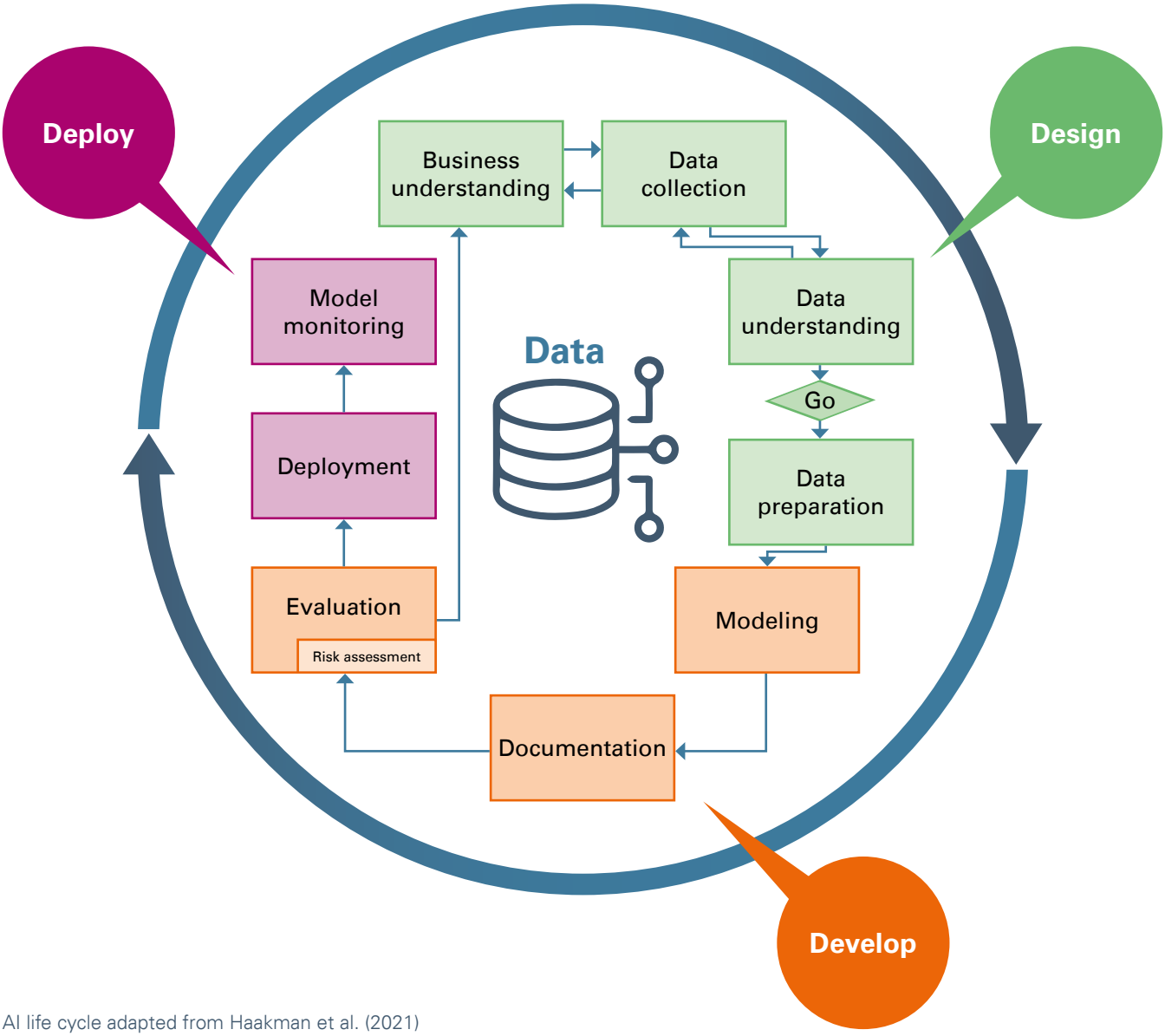
People's perception of fairness is generally highly context-dependent. It depends on a variety of factors such as socio-political views, personal preferences and the particular use case. This is also true for AI systems. **Algorithmic fairness** describes the absence of bias in AI decisions that would favor or disadvantage a person or group in a way that would be considered unfair in the context of the application (Ntoutsi et al., 2020).

One core aspect that makes the topic of bias and fairness in AI challenging is the fact that bias can be introduced in every phase of the AI life cycle (Baeza-Yates, 2018). In the following paragraph, we describe how bias might enter a system.

In the **design phase**, human-made decisions may introduce bias in neglecting the interest of specific stakeholder groups, for instance in requirements engineering and task definition, by focusing the data collection on one-sided sample groups (e.g. image data is primarily available for white people) or historically primed datasets (e.g. there is a strong gender bias in labor market data).

During the **development phase**, it is essential which model is selected, how weights are set and what features to focus on. The evaluation can reveal potential misconceptions, but since it is mainly done in-house, it can also verify them through a biased design.

When the AI system is running **(deployment phase)**, the bias may be introduced due to a specific ordering or visual presentation of the results, decisions made around how the system is monitored and through interaction data that enter the system and contribute to the learning of an evolving system.



AI life cycle adapted from Haakman et al. (2021)

2 Methods to address fairness issues in AI

"Unfair" algorithmic bias is a replication of human bias and can be introduced in all phases of the AI-life cycle i.e. design, development and deployment. In the mitigation of bias, we distinguish between pre-, in- and post-processing methods, which is again in accordance with the AI life cycle. Pre-processing techniques are methods that are applied in the design phase, e.g. during requirements engineering, data preparation and data selection. In-processing methods are mitigation strategies that are implemented as part of the algorithm design and development. Finally, post-processing methods are applied to mitigate unfair effects in the result set by the algorithms, e.g. by reranking them (Baeza-Yates, 2018).

Metrics and methods to evaluate fairness

Fairness of AI can, in principle, be evaluated quantitatively. The challenge, however, is to find the right quantitative metric. Many different definitions of algorithmic fairness exist and, accordingly, many different quantitative metrics. Verma and Rubin (2018) give an overview of the 20 most prominent definitions. The highest-level separation between different fairness definitions is between individual fairness on the one hand and group fairness on the other hand. Individual fairness is hard to define in mathematical terms, therefore most metrics are related to group fairness. Here one can differentiate between three principal approaches, namely fairness in acceptance rates, fairness in error rates and fairness in outcome frequency (Barocas et al. 2019). In most settings, these definitions contradict each other – thus, it is usually not possible for an AI model to be fair in all three aspects. For a given application at hand, the appropriate metrics have to be selected. These metrics are then evaluated on the output of the AI model, and one can evaluate whether the model is fair with respect to the chosen definition of fairness.

Methods to ensure fairness in AI

A wide range of methods are available to make AI models fairer (Bellamy et al. 2019, Barocas et al. 2019). Since, as we saw above, there is not a single definition of fairness that is always appropriate, and thus, making AI models fair actually means making them fair with respect to a certain definition of fairness. This is often referred to as "bias mitigation techniques". These techniques can be grouped into three categories, depending on which stage of the AI life cycle they intervene:

- Data pre-processing
- Adaption of the training algorithm (in-processing)
- Post-processing of results

Each approach has its own pros and cons. Data pre-processing as well as post-processing of results works for all training algorithms, in contrast to the adaption of training algorithms (in-processing), which only works for specific algorithms/models. All approaches have in common that they potentially negatively affect the accuracy of the models, but this is less an issue with methods that adapt the training algorithm (in-processing) than with the other two approaches. Finally, pre-processing techniques can solely ensure fairness in acceptance rates and not for other fairness definitions, and post-processing techniques typically are computationally expensive.

Fairness of AI can, in principle, be evaluated quantitatively. The challenge, however, is to find the right quantitative metric.

Open issues and challenges

Driven by the scientific community, huge achievements have been made regarding understanding, measuring and mitigating different kinds of bias. The proposed methodologies promise to hold in specific contexts but might not generalize to other AI systems, as both the composition of AI systems and the perception/definition of fairness can vary greatly with the application setting. The proper selection and application of methods to ensure fairness still demands a high level of expertise in legal, ethical and technical perspectives. This causes a gap between the general availability of scientific methods and their application in more complex practical settings where challenges must be addressed in a flexible and continuous manner.

Aspects without a current solution are, amongst others:

- Assuring fairness when combining multiple AI components, e.g. when reusing AI tools/algorithms developed independently of a specific application, with limited access to source code, as well as data that is audited for one use case but might incline bias intolerable in another algorithm/context
- Assuring fairness in varying cultural/legal contexts (Srivastana et al., 2019). What is perceived as fair or unfair varies between different cultural and legal contexts, and it is thus unclear how one can assure the fairness of an AI application that is intended to be used in multiple contexts
- How to assess fairness in evolving (learning) AI systems. Some AI systems are constantly updated (in some cases with every single use). Monitoring of fairness in production is, in principle, possible (e.g. Vasudevan & Kenthapadi, 2020). However, in a changing system, it is much more challenging to define in which situations fairness criteria are met or not because the performance of the algorithm may change over time (e.g. Lazer et al., 2014)
- How to apply/assess existing regulations, standards and ethical constraints in practice (Constanza-Chock et al., 2022). For example, there is no standard to determine the adequate trade-off between different fairness metrics nor between fairness and accuracy

Summary and outlook

AI applications can generate unfair outcomes that discriminate against groups or individuals. Research has provided a wide range of fairness definitions and accompanying quantitative metrics. These metrics can be used to measure fairness both in training data as well as in predictions of AI models. Which definition(s) of fairness is appropriate for an application needs to be decided for each application separately, taking into account the technical, cultural and legal settings. There are methods to enforce certain fairness requirements in an AI model (i.e. mitigation techniques), which can be implemented during the design, development or deployment of an AI system. Some international standards that discuss fairness and AI have been published, but several more are in development and soon to be presented. While independent auditing of fairness has been done before, there are no well-established auditing standards yet. From the research side, more work is needed on the robustness of evaluation metrics (e.g. the trade-off between individual and group fairness) and on the generalizability of different AI applications (e.g. ranking problems as in the case of recommender systems). From the standardization perspective, work on guidelines and tools is needed to allow the useful evaluation of AI fairness for a broader audience (e.g. policymakers or third-party users of AI libraries). Finally, from the policy-making side, regulations for the accountability of distributed AI systems and the bias-aware labeling of single AI components are still pending.

There are methods to enforce certain **fairness requirements in an AI model** (i.e. bias mitigation techniques), which can be implemented during the design, development or deployment of an AI system.

Acknowledgement

Know-Center is a leading European research center for big data, artificial intelligence (AI) and data-driven business models. Know-Center is a COMET Centre within COMET – Competence Centers for Excellent Technologies. This program is funded by the Austrian Federal Ministries for Climate Policy, Environment, Energy, Mobility, Innovation and Technology (BMK) and for Labor and Economy (BMAW), represented by Österreichische Forschungsförderungsgesellschaft mbH (FFG), Steirische Wirtschaftsförderungsgesellschaft mbH (SFG) and the Province of Styria, Wirtschaftsagentur Vienna and Standortagentur Tyrol GmbH.

References

Balayn, A., Lofi, C., & Houben, G.-J. (2021). Managing bias and unfairness in data for decision support: A survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. The VLDB Journal, 30(5), 739–768. https://doi.org/10.1007/s00778-021-00671-8

Barocas S., Hardt M., & Arvind, N. (2019). Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org. http://www.fairmlbook.org

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4-1.

Bird, S., Hutchinson, B., Kenthapadi, K., Kıcıman, E., & Mitchell, M. (2019). Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned. Companion Proceedings of The 2019 World Wide Web Conference, 1297–1298. https://doi.org/10.1145/3308560.3320086

Castelnovo, A., Crupi, R., Inverardi, N., Regoli, D., & Cosentini, A. (o. J.). Investigating Bias with a Synthetic Data Generator: Empirical Evidence and Philosophical Interpretation.

Castelnovo, A., Crupi, R., Inverardi, N., Regoli, D., & Cosentini, A. (2022). Investigating Bias with a Synthetic Data Generator: Empirical Evidence and Philosophical Interpretation (arXiv:2209.05889). arXiv. http://arxiv.org/abs/2209.05889

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022, June). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1571-1583).

Corrales-Barquero, R., Marin-Raventos, G., & Barrantes, E. G. (2021). A Review of Gender Bias Mitigation in Credit Scoring Models. 2021 Ethics and Explainability for Responsible Data Science (EE-RDS), 1–10. https://doi.org/10.1109/EE-RDS53766.2021.9708589

Davoudi, A., Sajdeya, R., Ison, R., Hagen, J., Rashidi, P., Price, C. C., & Tighe, P. J. (2023). Fairness in the prediction of acute postoperative pain using machine learning models. Frontiers in Digital Health, 4, 970281. https://doi.org/10.3389/fdgth.2022.970281

Deldjoo, Y., Anelli, V. W., Zamani, H., Bellogín, A., & Di Noia, T. (2021). A flexible framework for evaluating user and item fairness in recommender systems. User Modeling and User-Adapted Interaction, 31(3), 457–511. https://doi.org/10.1007/s11257-020-09285-1

El Gayar, N., Trentin, E., Ravanelli, M., & Abbas, H. (Hrsg.). (2023). Artificial Neural Networks in Pattern Recognition: 10th IAPR TC3 Workshop, ANNPR 2022, Dubai, United Arab Emirates, November 24–26, 2022, Proceedings (Bd. 13739). Springer International Publishing. https://doi.org/10.1007/978-3-031-20650-4

Fancher D., Ammanath B., Holdowsky J., & Buckley N. (2021) AI model bias can damage trust more than you may know. But it doesn't have to. Deloitte Insights. https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/ai-model-bias.html

Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of google flu: traps in big data analysis. Science 343(6176): 1203–1205. https://doi.org/10.1126/science.1248506

Mac Namee, B., Cunningham, P., Byrne, S., & Corrigan, O. I. (2002). The problem of bias in training data in regression problems in medical decision support. Artificial Intelligence in Medicine, 24(1), 51–70. https://doi.org/10.1016/S0933-3657(01)00092-6

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining and Knowledge Discovery, 10(3). https://doi.org/10.1002/widm.1356

Pitoura, E. (2020). Social-minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias. Journal of Data and Information Quality, 12(3), 1–8. https://doi.org/10.1145/3404193

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2459-2468).

Varona, D., Lizama-Mue, Y., & Suárez, J. L. (2021). Machine learning's limitations in avoiding automation of bias. AI & SOCIETY, 36(1), 197–203. https://doi.org/10.1007/s00146-020-00996-y

Vasudevan, S., & Kenthapadi, K. (2020). Lift: A scalable framework for measuring fairness in ml applications. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 2773-2780).

Verma, S., & Rubin, J. (2018). Fairness definitions explained. Proceedings of the International Workshop on Software Fairness, 1–7. https://doi.org/10.1145/3194770.3194776

Xu, R., Cui, P., Kuang, K., Li, B., Zhou, L., Shen, Z., & Cui, W. (2020). Algorithmic Decision Making with Conditional Fairness. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2125–2135. https://doi.org/10.1145/3394486.3403263


Fairness in AI

SGS.COM/DIGITAL

CONTACT US

SGS

Emerging Technology

✉ Enquiry.Emerging-Technology@sgs.com

KNOW CENTER

Leading Research and Innovation Center for Trustworthy AI

https://know-center.at/

✉ info@know-center.at

WHEN YOU NEED TO BE SURE