Toward a cost-effective fairness-aware ML lifecycle

Bussi Elisa [1], Basso Andrea [1], Maran Elena[2], Chiavarino Claudia[1], and Severini Simone [2]

[1] *Istituto Universitario Salesiano Torino Rebaudengo, Piazza Conti di Rebaudengo, 22, 10155 Turin, Italy*
[2] *Modulos AG, Technoparkstr. 1, 8005 Zurich, Switzerland*

Abstract

Machine learning technology has a profoundly transformative impact on innovation, presenting challenges to existing regulatory frameworks. Trustworthiness principles play a pivotal role in both the EU AI Act and other relevant regulations, such as the GDPR. In a broader context, trustworthiness that will be an essential requirement for AI systems necessitates the integration of fairness considerations throughout the entire ML lifecycle. This integration is a complex yet crucial endeavor to ensure the responsible development and deployment of AI systems. The challenges associated with incorporating fairness into machine learning models arise from not only the lack of standardized fairness constraints but also the hesitancy among practitioners to adopt existing fairness measures and to the cost associated with the inclusion of fairness in the process. Overcoming these challenges is essential for building AI systems that are genuinely trustworthy, compliant with regulations, and ultimately beneficial to society.

Keywords

Machine Learning, Fairness, Machine Learning Lifecycle, Algorithmic Bias, Ethical AI

1. Introduction

Machine Learning has become one of the most promising disruptive innovations of the last decade: it's been estimated that AI will reach a worldwide market value of 1,500 billion USD by 2030 [5]. ML deployment can be found in core sectors such as finance [32], healthcare [14] and human resources [31], and so is set to impact people's lives rapidly and deeply. The rise of machine learning has led to increasing concerns about the vulnerability of these systems to amplify existing social biases [34, 13], resulting in unfair treatment of entire populations [8]. Unfortunately, most of the models available in the industry today are not designed to account for fairness [20]. Numerous examples exist of discriminatory systems that discriminate based on protected attributes like race [10], gender [7], or a combination of both [24], as well as any of the proxy features correlated with protected attributes. The need to address these issues has given rise to a new research field called algorithmic fairness [25], focused on mitigating bias and ensuring fair decision-making systems.

While fairness in AI has long been recognized as an important concept [19], its significance in machine learning has grown exponentially due to the integration of trustworthiness principles in modern legislation such as GDPR [12] and the upcoming EU AI Act [11]. However, incorporating fairness into the development process of machine learning is complex. Although numerous fairness metrics and tools have been developed to provide solutions, applying them in real-world contexts is often facing difficulties. These challenges arise from a lack of adaptability to institutional realities [33], as well as computationally expensive deployment [8]. Addressing these complexities is crucial in achieving true fairness in AI systems and meeting modern legal requirements.

When it comes to development, ML practitioners often face challenges in making fair decisions [22], identifying potential risks in their specific context and domain area [18], and integrating the

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29-31, 2023, Pisa, Italy
EMAIL: 8340@studenti.ius.to (A. 1); andrea.basso@ius.to (A. 2); elena.maran@gmail.com (A. 3); claudia.chiavarino@ius.to (A. 4); severini.simone@gmail.com (A. 5)

available toolkits into their existing processes, as they find their functioning difficult to comprehend and limited in their ML lifecycle coverage [22]. Additionally, integrating these tools into organizational structures can prove to be challenging, with constraints that compromise the motivation to consider fairness in their work [23]. However, maintaining the current status quo in terms of fairness will soon be untenable. With the enactment of AI-related legislation [11, 12], compliance with fairness regulations will become a top priority. Moreover, incorporating fairness can be a strategic business practice, offering numerous benefits such as reduced risk of failure, reinforced brand reputation, and enhanced user trust [34]. Embracing fairness in AI systems, therefore, not only fulfills legal obligations but also fosters long-term organizational success. The cost of embracing fairness needs to be considered as a key element to make fairness-aware ML lifecycles adopted at large.

2. The need of a new Machine Learning Lifecycle

The development of a Machine Learning (ML) lifecycle involves a complex sociotechnical process that consists of several phases, each with its own set of stakeholders. While there is no standard or agreed-upon process for ML production and release, most of the ML lifecycles deployed today include similar phases [9, 35], although some valuable phases, such as the inclusion of fairness metrics in the pipeline, are underrepresented. Such metrics, while are measurable and mathematically defined, are of limited efficacy [29]. They do not consider the individual choices made by stakeholders throughout the project, which can influence the outcome of the model and hinder its long-term efficacy [30]. The ML development lifecycle in fact involves a series of decisions that, apart from algorithmic bias management, can lead to unintended consequences [21]. Understanding how each step influences the decision-making process of pipeline workers may help address the most harmful downstream consequences more directly and meaningfully [30]. For instance, incomplete documentation detailing the choices made in the lifecycle [16] can lead to biased decision-making and the loss of accountability for the choices made [34].

In order to prove our point effectively, let us reflect on a real use case scenario [26] where a fairness-aware intervention encompassing the whole ML lifecycle can seriously benefit a business practice. One of the sectors in which ML deployment is particularly looked upon is loan eligibility [3], whose problem statement perfectly fits that of a prediction task. As decisions about granting a loan to a customer is traditionally based on variables like credit history or income, these can be turned into features to train a model to predict from a customer profile whether it will return or not a loan if granted one. Three main outcomes are possible in relation to a credit decision: a creditworthy applicant receives the loan and can repay it; a creditworthy applicant is refused the loan; an applicant receives the loan and defaults after its disbursement.

This use case has been chosen because the use of ML in this sector without the rightful fairness concerns being made has already produced episodes of discrimination [2], which stem from a specific kind of harm: allocative harm [1]. The social and economic cost of a mistake in this sector can be detrimental for the consumer, and sometimes for the company itself: assuming the AI system operates in a country with an official credit score system aimed at determining the likelihood of an applicant meeting her financial obligations, tending to a significant credit score reduction would affect applicants who receive the loan but subsequently default (false positives) whilst a credit score reduction of a lesser extent would apply to those applicants who are denied access to the loan but would have repaid it (false negatives). The latter group is also likely to see a long-term deterioration of their credit score as having been denied access to financial resources may result in a long-term impact on their ability to meet future financial obligations. Basically, what makes the problem critical is that harm can go both ways: on one hand, using a traditional ML lifecycle (using historical data, maximizing fairness *etc*.) to build a loan eligibility system poses the risk of producing more false negatives; if the people affected were to eventually resort to legal means, revealing that the decision made is only explainable with protected features discrimination (like gender or race) and nothing more, the reputation of the business owner would be at risk, and customers would become more distrustful of the AI. On the other hand, though, some selection has to be made in order not to default the company itself; plus, there could be a situation where ML prediction may actually prove

correct despite being casually related to a protected variable (*e.g.* a woman not being granted a loan has nothing to do with her gender), but where the company is still called for an explanation of the decision process.

In conclusion, the problem to be solved here is how to find the most optimal trade-off between profitability, fairness, reputational concerns, but also current work practices and stakeholder engagement. For these reasons, we posit that the old ML lifecycle as it is today not cut for meeting said requirements: a new, different approach is needed, one that aims at satisfying the priorities of all involved stakeholders, including those ultimately impacted by the decision (the customers), thus leading to better societal outcomes whilst ensuring transparency on the impact of the decisions on the profitability of the business.

3. Next Generation ML Lifecycle: recent approaches

Recently, some approaches were proposed to tackle the problem of ML lifecycle.

The first one is called **Data-Centric AI** [6, 27], an approach that provides a granular understanding of how sources of noise, error, and bias in data impact the model performance both in terms of fairness and accuracy. By creating an intrinsic tie between data and model results, and operating a data-model feedback loop, Data-Centric AI ensures that a long-lasting and sustainable fairness strategy is achieved. This methodology brings several stakeholders in the process, creates accountability amongst data owners, and allows one to understand specific problems in single data samples which are impacting the model's fairness. Considering the consumer lending problem, the initial step is to agree on a fairness measure that is suitable for the application under consideration, based on the type of harm that is ensued (in this case, allocative), and the one which best reflects the above-anticipated harms from the AI system under scrutiny is the equalized odds measure [15]. Once identified the appropriate fairness measure, the Data-Centric AI approach entails the identification of the individual contribution of each observation in the dataset to the quantitative performance of the model with respect to this measure, and the results of this procedure will allow practitioners to correctly select the best strategy for improving overall fairness (e.g. improving the dataset features distributions).

The second one is called **Z-Inspection** [35]. The Z-Inspection process is a versatile tool that can be used to evaluate and audit AI systems before they go into production. Its primary purpose is to raise awareness among relevant stakeholders about the potential ethical, social, technical, and legal risks associated with implementing an AI system. Z-Inspection is inspired by the seven requirements outlined in the "Framework for Trustworthy AI" [17]. Z-Inspection brings together two distinct approaches into a single process. The first is a holistic approach that aims to consider the entire sociotechnical system. The second is an analytical approach that considers each part of the problem domain in greater detail. The outcome is a multi-perspective view that is capable of assessing, discussing, and resolving the tensions that arise during the assessment process through a set of recommendations.

To illustrate how it operates, let's consider the example of customer lending. The first phase of the Z-Inspection process involves forming an interdisciplinary team of investigators, including engineers, ethicists, case owners, and company practitioners, to define the boundaries of the assessment. In the second phase, each team identifies all possible ethical and legal issues and maps them to the trustworthy AI ethical values and requirements, such as protected features and discrimination danger. Finally, in the third phase, the team addresses ethical tensions and solves them whenever possible, such as recommending a specific fairness measure in favor of other ones [*cfr*. 28]. One of the main advantages of the Z-Inspection process is that it considers fairness as an integral part of the assessment process from the outset. It also allows a multidisciplinary team of professionals to collaborate and discuss together, leading to a more comprehensive and effective approach to addressing ethical issues.

4. Toward a cost-effective fairness aware ML lifecycle

The contributions made toward understanding and standardizing the next generation of ML lifecycles are incredibly valuable. However, it is crucial to consider the cost factor in the development of these lifecycles. To ensure that an ML lifecycle is fairness conscious, it must incorporate fairness and other ethical considerations from problem formulation to deployment. Therefore, optimizing each step of the lifecycle is crucial.

To address these issues, we propose the redesign of a novel ML Lifecycle that places fairness at its core and factors in also the cost/benefit for every step of the process. Our approach will draw inspiration from the FATE paradigm (*fairness, accountability, transparency, explainability* [*cfr*. 28]) and incorporate Human-Centered Design and Ergonomics techniques. An initial hypothesis of the different phases is the following:

Problem Formulation: In the problem formulation stage, the objective is to define the business problem and identify the relevant stakeholders. This includes identifying the key performance indicators (KPIs) and the decision-making criteria that will be used to evaluate the model's performance. At this stage, it is also essential to identify the fairness concerns and metrics that will be used to assess fairness.

Data Collection: The data collection stage is crucial for fairness-aware machine learning. The goal is to collect data that is diverse and representative of the population being studied [13]. This includes taking steps to ensure that the data collection process is unbiased and does not discriminate against any group [4]. At this stage, it is also important to identify any potential sources of bias in the data.

Data Pre-processing: In the data pre-processing stage, the goal is to clean and transform the data so that it is suitable for machine learning. This includes removing outliers, filling in missing data, and transforming the data into a format that is suitable for modeling. At this stage, it is also essential to check for any biases that may have been introduced during data collection [4].

Model Training: In the model training stage, the goal is to build a model that is fair and accurate. This involves choosing an appropriate algorithm, tuning hyperparameters, and evaluating the model's performance on training and validation data. It is also essential to check for any biases that may have been introduced during model training.

Model Evaluation: In the model evaluation stage, the goal is to evaluate the model's performance on test data. This includes measuring accuracy and fairness using appropriate metrics [*cfr*. 27]. It is also essential to check for any biases that may have been introduced during model evaluation.

Model Deployment: In the model deployment stage, the goal is to deploy the model into a production environment. This involves testing the model in real-world scenarios and monitoring its performance over time. It is also essential to continually evaluate the model for fairness and accuracy.

Model Maintenance: In the model maintenance stage, the goal is to ensure that the model remains fair and accurate over time. This includes monitoring the model's performance, updating the model when new data becomes available, and re-evaluating the model for fairness and accuracy.

At each step of the machine learning lifecycle, a cost-benefit analysis is performed to optimize fairness and cost simultaneously. The cost-benefit analysis considers the trade-offs between fairness, accuracy, and cost and identifies the optimal balance between these factors.

5. Conclusions

In this position paper, we have examined the challenges that arise in current ML lifecycles regarding fairness and introduced different strategies to address the issue. Specifically, we have explored data-centric approaches and inspection-based perspectives to ensure fairness in the lifecycle.

Data-centric approaches involve methods that prioritize the selection, preparation, and use of data to mitigate any potential biases in the ML lifecycle. Inspection-based perspectives, on the other hand, focus on evaluating the outcomes of the ML model to identify and correct any instances of unfairness.

However, in the adoption of new generation ML lifecycles, cost is a critical factor that must be taken into account. The cost of developing and implementing a new ML lifecycle can vary significantly depending on the complexity of the model, the size and quality of the data, and the availability of resources.

Thus, when designing and implementing new ML lifecycles that prioritize fairness, it is essential to consider the cost factor to ensure practical and feasible solutions. This includes assessing the cost-benefit of different approaches and optimizing individual steps in the ML lifecycle to achieve the desired outcomes while minimizing costs.

For example, a cost-effective strategy for ensuring fairness in the ML lifecycle could involve using readily available datasets and implementing guided data cleaning strategies to reduce bias. Additionally, optimizing the algorithm's performance through regular monitoring and testing can reduce the need for costly and time-consuming manual inspections.

In conclusion, while addressing fairness in ML lifecycles is crucial, it is equally important to

consider the cost factor in the adoption of new generation ML lifecycles. By optimizing the individual steps and assessing the cost-benefit of different approaches, we can ensure that fair and practical solutions are developed and implemented.

6. Acknowledgements

We thank the people at IUSTO, Modulos and Lawfultech.ai for the help provided with examples, discussions and contributions given to make this paper possible.

7. References

[1] S. Barocas, K. Crawford, A. Shapiro, and H. Wallach, The problem with bias: from allocative to representational harms in machine learning, in: 9th Annual Conference of the Special Interest Group for Computing, Information and Society, SIGCIS '17, Philadelphia, PA, USA, 2017.

[2] R. Bartlett, A. Morse, R. Stanton, and N. Wallace, Consumer-lending discrimination in the FinTech era, Journal of Financial Economics (2021), 143(1), 30-56. https://doi.org/10.1016/j.jfineco.2021.05.047

[3] S. Bhatore, L. Mohan, and Y.R. Reddy, Machine learning techniques for credit risk evaluation: a systematic literature review, JBFT 4 (2020), 111-138. https://doi.org/10.1007/s42786-020-00020-3

[4] S. Biswas, and H. Rajan, Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline, in: Proceedings of the 29th ACM ESEC/FSE Symposium on the Foundations of Software Engineering, ACM, New York, NY, 2021, 981–993. https://doi.org/10.1145/3468264.3468536

[5] Bloomberg.com, Artificial Intelligence Market USD 1,581.70 Billion By 2030, Growing At A CAGR of 38.0% - Valuates Reports, 2022. URL: https://www.bloomberg.com/press-releases/2022-06-13/artificial-intelligence-market-usd-1-581-70-billion-by-2030-growing-at-a-cagr-of-38-0-valuates-reports

[6] S. Brown, Why it's time for 'data-centric artificial intelligence', 2022. URL: https://mitsloan.mit.edu/ideas-made-to-matter/why-its-time-data-centric-artificial-intelligence

[7] Y.T. Cao, and H. Daumé, Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle, Computational Linguistics (2021), 47 (3), 615–661. https://doi.org/10.1162/coli_a_00413

[8] G. Castiglione, G. Wu, C. Srinivasa, and S. Prince, fAux: Testing Individual Fairness via Gradient Alignment, arXiv e-prints (2022). arXiv:2210.06288

[9] D. Crankshaw, A Short History of Prediction-Serving Systems, 2018. URL: https://rise.cs.berkeley.edu/blog/a-short-history-of-prediction-serving-systems/

[10] W. Douglas Heaven, Predictive policing algorithms are racist. They need to be dismantled. MIT Technology Review, 2020. URL: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

[11] Future of Life Institute, The Artificial Intelligence Act, European Union, 2021. URL: https://artificialintelligenceact.eu/

[12] Garanteprivacy.it, Regolamento UE 2016 679. Arricchito con riferimenti ai Considerando Aggiornato alle rettifiche pubblicate sulla Gazzetta Ufficiale dell'Unione europea 127 del 23 maggio 2018, 2016. URL: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/6264597

[13] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumeé, and K. Crawford, Datasheets for Datasets, arXiv e-prints (2018). arXiv:1803.09010

[14] T. Grote, and P. Berens, On the ethics of algorithmic decision making in healthcare, Journal of Medical Ethics (2020), 46 (3), 205–211. https://doi.org/10.1136/medethics-2019-105586

[15] M. Hardt, E. Price, and N. Srebo, Equality of opportunity in supervised learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, 3323–3331. https://doi.org/10.48550/arXiv.1610.02413

[16] J.M. Hellerstein, V. Sreekanti, J.E. Gonzalez, J. Dalton, A. Dey, S. Nag, K. Ramachandran, S. Arora, A. Bhattacharyya, S. Das, M. Donsky, G. Fierro, C. She, C. Steinbach, V.R. Subramanian, & E. Sun, Ground: A Data Context Service, in: Conference on Innovative Data Systems Research, Chaminade, CA, USA, 2017.

[17] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, European Commission, 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[18] K. Holstein, J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach, Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2019, 1–16. https://doi.org/10.1145/3290605.3300830

[19] B. Hutchinson, and M. Mitchell, 50 Years of Test (Un)fairness: Lessons for Machine Learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, 49–58. https://doi.org/10.1145/3287560.3287600

[20] G. Jones, J. Hickey, P. Di Stefano, C. Dhanjal, L. Stoddart, and V. Vasileiou, Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. arXiv e-prints (2020). https://doi.org/10.48550/arXiv.2010.03986

[21] M.S.A. Lee, and J. Singh, The Landscape and Gaps in Open Source Fairness Toolkits, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021, 1–13. https://doi.org/10.1145/3411764.3445261

[22] M.S.A. Lee, and J. Singh, Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle, in: Proceedings of the 2021 Conference on AI, Ethics, and Society, Association for Computing Machinery, AAAI/ACM ’21, New York, NY, USA, 2021, 704–714. https://doi.org/10.1145/3461702.3462572

[23] M.A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, 1–14. https://doi.org/10.1145/3313831.3376445

[24] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, Model Cards for Model Reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, 220–229. https://doi.org/10.1145/3287560.3287596

[25] S. Mitchell, E. Potash, S. Barocas, A. D’Amour, and K. Lum, (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application (2021), 8, 141-163. https://doi.org/10.1146/annurev-statistics-042720-125902

[26] Modulos.ai, Exploring AI Fairness in Consumer Lending, 2022. URL: https://www.modulos.ai/resources/exploring-ai-fairness-in-consumer-lending/

[27] Modulos, Fairness in Credit Risk with Data-Centric AI, Video, 2022. URL: https://www.youtube.com/watch?v=clNJOy17YXM

[28] D. Shin, User Perceptions of Algorithmic Decisions in the Personalized AI System: Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability, Journal of Broadcasting & Electronic Media (2020), 64, 1-25. https://doi.org/10.1080/08838151.2020.1843357

[29] M. Srivastava, H. Heidari, A. and Krause, 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 and Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, 2459–2468. https://doi.org/10.1145/3292500.3330664

[30] H. Suresh, and J. Guttag, A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, in: Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '21, Association for Computing Machinery, New York, NY, USA, 2021, 1–9. https://doi.org/10.1145/3465416.3483305

[31] P. Tambe, P. Cappelli, and V. Yakubovich, Artificial Intelligence in Human Resources Management: Challenges and a Path Forward, California Management Review (2019), 61(4), 15–42. https://doi.org/10.1177/0008125619867910

[32] S. Townson, AI can make bank loans more fair, Harvard Business Review (2020). URL: https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair

[33] M. Veale, M. Van Kleek, and R. Binns, Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, 1–14. https://doi.org/10.1145/3173574.3174014

[34] A. Woodruff, S.E. Fox, S. Rousso-Schindler, and J. Warshaw, A Qualitative Exploration of Perceptions of Algorithmic Fairness, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, 1–14. https://doi.org/10.1145/3173574.3174230

[35] R. V. Zicari et al., Z-Inspection®: A Process to Assess Trustworthy AI, IEEE Transactions on Technology and Society (2021), 2(2), 83-97. doi:10.1109/TTS.2021.3066209