From Intention to Action: The Fair AI Toolbox

Robbie Meyer
Systems Design Engineering, University of Waterloo
Alexander Wong
Systems Design Engineering, University of Waterloo
Email: {robbie.meyer, alexander.wong}@uwaterloo.ca

Abstract

As artificial intelligence (AI) systems become more widely used, awareness of the fairness issues associated with such systems is increasing. Real world systems have been found to discriminate against marginalized groups and governments are starting to see fairness concerns as a significant risk associated with the use of AI. AI practitioners have many tools to address these concerns, but it is difficult to determine which tools are most appropriate. To assist AI practitioners, we provide a solution-oriented systemic overview of fairness in AI. We divide the AI system into five areas of concern and identify the AI fairness tools associated with each area. We find there are numerous tools for identifying and addressing fairness concerns that can be practically implemented in AI systems.

1 Introduction

The fairness concerns associated with the use of artificial intelligence (AI) systems are becoming well known. Fairness concerns are making headlines as large corporations such as Amazon [1] and Twitter [2] find biases in their AI systems. Governments and regulators are also starting to take notice. The White House specifically addressed the problem of fairness in AI systems in their "Blueprint for an AI Bill of Rights" [3]. Unfair systems can discriminate against marginalized groups, break down trust and sully reputations. To avoid these issues, AI practitioners must not only consider the overall performance of the system, but also the fairness of the system.

To help ensure fairness in AI systems, many researchers have developed solutions to identify and mitigate system biases. Solutions have been developed to automatically generate training samples that challenge the biases of image classification models [4], discourage model biases during training through regularization [5] and more. There are numerous solutions for the problem of fairness in AI, each with their own requirements, advantages and disadvantages.

The application of these solutions is not straightforward. It is difficult to identify the potential fairness challenges an AI system might face and the actions that should be taken as a response to the challenges. Existing reviews and analyses of fairness in AI systems focus on the problem of fairness in AI systems instead of the actions that can be taken to improve fairness. While these papers may help practitioners understand the causes and impacts of fairness in their AI system, they do not help practitioners determine how to address their specific fairness concerns.

In this work, we provide a solution-oriented, systemic overview of the current AI fairness landscape. We present a schema of an AI system, shown in Figure 1, that segments the system into five different areas of concern. For each area of concern, we detail the tools that an AI practitioner can use to mitigate fairness concerns. By identifying, organizing and analyzing current solutions to various AI bias problems we seek to assist AI practitioners in identifying solutions to their fairness concerns.

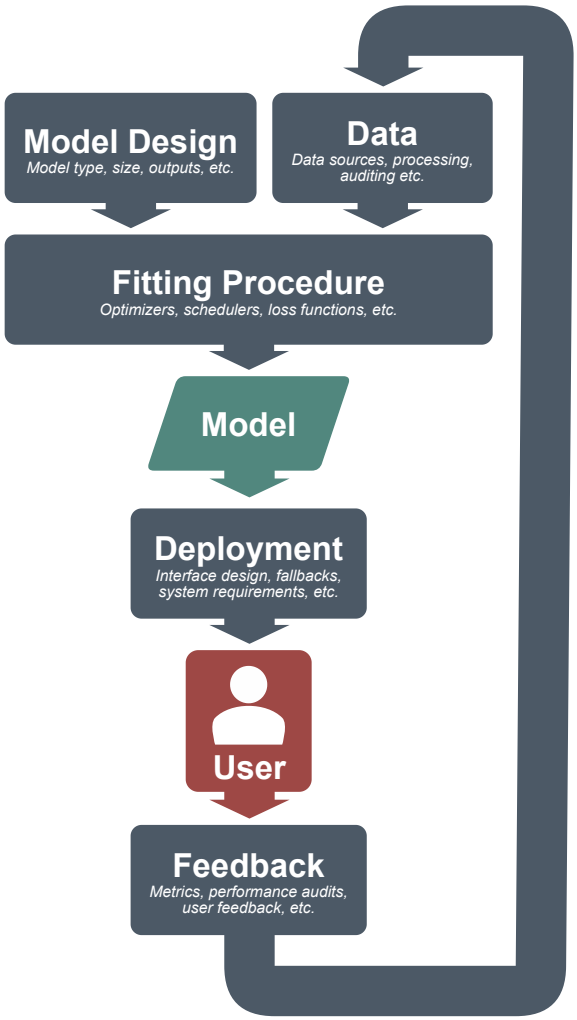

Fig. 1: Our schema of an AI system. Each area of concern, represented by a grey box, is associated with design decisions that can affect the fairness of the system.

2 Defining Fairness

The term fairness is not well defined. In the broadest sense, one might say that an AI system has a fairness problem if the system's performance varies over an attribute which is not intrinsic to the system's task. For instance, a speech recognition system may be seen to be unfair if its performance varied between different accents as a subject's accent. If the accent of a subject was changed, we would not expect the system to behave differently and therefore the varying performance between accents can be seen as an instance of unfairness.

While this definition is intuitive, it is not particularly useful. A useful definition of fairness would be specific and measurable. Unfortunately, there is no single ideal fairness metric that is appropriate for all projects.

To illustrate this, we can consider fairness for binary classification. For binary classification, three definitions of fairness are demographic parity, equalized odds and equal opportunity [6]. As described in Table 1, all three definitions appear similar yet have slightly different conditions. Demographic parity is the notion that the prediction should be independent of protected attributes. While demographic parity is straightforward, it breaks down when the label distribution differs between attribute groups. Hardt et al. [6] proposed the fairness definitions of equalized odds and equal opportunity that are able to account for this problem. Equalized odds requires the prediction to be independent of protected attributes conditional on the label [6]. This means that the true positive and false negative rates are equal between attribute groups. Equal opportunity relaxes the equalized odds condition to only consider the case when the label is true [6]. In other words, the equal opportunity definition is only concerned with samples that should receive a positive prediction.

Different tasks with different goals will require different definitions. Ultimately, fairness is a social concept. There is no single value which fully encompasses the notion of fairness. AI practitioners should consider both the nature of their system as well as the social context surrounding their system when determining which fairness metrics would be most relevant.

Table 1: Three Definitions of Fairness for Binary Classification

Name	Definition	Explanation
Demographic Parity	$Pr\{\hat{Y}=1\|A=0\}=Pr\{\hat{Y}=1\|A=1\}$	The prediction, $\hat{Y}$, is independent of protected attribute A
Equalized Odds	$Pr\{\hat{Y}=1\|A=0,Y=y\}=Pr\{\hat{Y}=1\|A=1,Y=y\} \quad y\in\{0,1\}$	The prediction, $\hat{Y}$, is independent of protected attribute A conditional on the label, Y
Equal Opportunity	$Pr\{\hat{Y}=1\|A=0,Y=1\}=Pr\{\hat{Y}=1\|A=1,Y=1\}$	The prediction, $\hat{Y}$, is independent of protected attribute A when $Y=1$

3 Tools For Fair AI

AI systems are complex. The design of an AI system involves numerous decisions involving data sources, model architectures, loss functions and more. Consequently, there are numerous approaches for improving the fairness of an AI system. It is a daunting task to identify the design decisions that will produce a system that is both fair and performant.

In this section, we provide an overview of the design space of an AI system with respect to designing for fairness. By presenting the variety of solutions that have been proposed for improving the fairness of AI systems, we hope to help AI practitioners understand the tools that are available to them for designing fair systems.

To organize the overview, we associate each presented tool with an area of concern. Each area of concern represents a different part of the system, forming our schema for an AI system as shown in Figure 1. In this schema, 'data' represents all information used by the system to make decisions, 'model design' represents the solution space for the AI model, 'fitting procedure' represents the procedure by which model is identified from the solution space, 'deployment' represents how the model is used by the user, and 'feedback' represents how the behaviour of model is monitored.

There is one key solution that does not fit into this schema: not using an AI system. AI systems are not applicable for every task. If the danger imposed by the fairness concerns is too significant or if the actions required to mitigate the concerns are infeasible, it is likely wiser to consider alternate types of systems. In these situations, systems designed around transparent rules or human decision makers may be more appropriate.

3.1 Data

Data is the foundation for any AI system and no conversation on fairness in AI systems is complete without a discussion of data. While researchers often consider their datasets as fixed components, AI practitioners do have influence over data sources, collection processes and processing. A survey of practitioners found that most of those surveyed considered data collection to be the most important place to intervene to improve system [7].

Data can be used for tasks such as training, testing and auditing. For all of these tasks to be effective, the data needs to represent the target population. Biases in the data can induce unfairness in system outputs, even when the model is adjusted for fairness [8]. Using sound sampling methodologies is essential for this goal. Data collection strategies that are geographically centralized or only consider certain types of users may not produce datasets that effectively represent the target population. Simply using a popular available dataset may not be sufficient. Representation issues that affect model performance have been identified in many popular datasets [9, 10]. If the data does not fully represent the population it may be possible to adjust some metrics if the true composition of the population is known [8].

While the data collection process is important, assuming one's data collection strategy is effective is foolhardy. Developers should audit and monitor their datasets to verify quality. Auditing is most effective when high quality metadata is available. If possible, metadata such as collection time and data demographics should be collected. If it is not feasible to collect this information for all data, it may still be feasible to use a model-driven approach to generate estimations of annotations [10] or to draw conclusions from annotated subsets of the data.

It is important to note that representing the attribute distributions of the population is not a sufficient condition for a dataset to be fair. Different subsets of data may be more complex and consequently require more data for the system. For instance, a pose estimation system may display higher performance with a simple rigid object than it would with a complex flexible object. To ensure fairness, we may want to collect more samples of the complex flexible object. Instead of ensuring that there is an equal number of samples for each subgroup, practitioners should ensure that there are enough samples for each subgroup.

Data can also be processed using data augmentation techniques to rectify any deficiencies. Simple data augmentation approaches such as random affine transformations can improve model performance, but the developer should take care to ensure that the transformations are appropriate as inappropriate transformations can degrade performance. For example, due to the inherent consistency of the data, in one instance standard data augmentations actually degraded self-driving car performance [11]. If the augmentation approach is only appropriate for a subset of samples, data augmentation should cause fairness issues.

More sophisticated data augmentation approaches can directly address fairness issues in the dataset. For example, BiaSwap generates new training images by merging automatically identified 'bias-guiding' and 'bias-contrary' samples [4].

3.2 Model Design

It can be easy to overlook the impact of the model design on fairness. Many state of the art model designs are highly general and make few assumptions that would affect the system fairness. This is not to say that the model design does not impact fairness. However, we do not see significant research relating specific model design decisions such as model type and model architecture decisions to fairness. Instead, the developer should be concerned with high-level decisions as well as the ways in which the model design affects other decisions in the system.

One high-level design decision that does directly affect fairness is the model capacity. A model with a small capacity might exhibit high bias and may not capture the patterns of subgroups with less representation in the data. Low capacity models that underfit the data have been shown to exhibit underestimation bias in which low frequency events are not predicted enough [12]. In contrast, a system with a large capacity may overfit the training data, decreasing the systems ability to generalize. Any subgroups with inadequate representation in the training data would therefore see degraded performance with a model that is too large.

The model design also influences the type of data that can be used in the system. For instance, some models may require meaningful hand-crafted features while others can work with raw unstructured data using learned features. The use of handcrafted features allows the developer to directly control the information used by the system whereas learned features may capture information that should not be used in decision making. For example, an image classifier may consider skin tone as a feature even in situations where racial features may be inappropriate. However, the use of raw unstructured data may make it easier to construct large, representative datasets. Similarly, a semi-supervised model design may be able to improve fairness when used in place of a supervised model by incorporating additional unlabeled data [13].

Just as understanding the decision making process of the model is also important for monitoring and rectifying fairness concerns. Explainability methods have been developed for some model types that allow the developer to use proxy models or features to understand the decision making process of the model [14]. In simple models such as simple rule-based and linear regression models, the parameters of the models can be interpreted directly. It has been argued that an interpretability approach that prioritizes understanding the actual decision making process is more trustworthy than an explainability approach that uses proxies to provide insight [15]. However, the explainability approach does not require that the

model is simple enough to be understood by a human, allowing for more complex and dynamic models.

System developers can also use multiple models in their system. This may mean developing distinct models for different subgroups or developing a simple fallback model to use when the primary model is determined to be inappropriate for a given situation. Out of distribution detection approaches could be used to trigger the use of the fallback when the input is not well represented in the training data.

To ensure that all of the model design decisions are optimal, appropriate validation procedures should be used. All the principles of good validation procedures for AI systems still apply in a fairness context. The primary difference between a validation procedure that does not consider fairness and one that does is simply the use of fairness metrics in addition to general performance metrics.

3.3 Fitting Procedure

The fitting procedure defines how the final model is derived using the data and the model design. It is the step in the system where any potential fairness issues in the data or model design become actual issues in the model. Consequently, many methods for mitigating fairness concerns target the fitting procedure, often employing modifications to the loss function or post-fitting procedures.

The loss function is a very common target for fairness boosting methods. Penalty terms, sample weighting schemes and regularization can all be used to push the fitting process towards fairer solutions. These approaches incorporate additional information such as the source data distribution [16] or to directly incentivize the reduction of fairness imbalances [5].

The use of pretraining schemes in which the models that were previously trained on unrelated datasets are used as the initial model can also impact fairness. While pretraining can assist in training more robust models [17], it also introduces the dataset and fitting procedure used to train the original model as new potential sources of biases.

Post-fitting procedures can also affect fairness. Some procedures can directly improve the fairness of a fitted model [18]. Other procedures have other aims, but still affect bias. For instance, model quantization and pruning methods which aim to compress a neural network can exacerbate fairness issues [19].

3.4 Deployment

The manner in which a model is deployed determines the manner in which users will be able to access the model. This means that while the deployment of a model may not affect its output, it can affect how the model's output will affect users. Consequently, the deployment of a model is an important consideration for fairness.

If the deployment of a model only encourages use for certain groups of users, the benefits of the system will not be distributed fairly. This may occur if the system requires expensive or uncommon tools, or if their are prior social dynamics that could affect a user's perception of the system. For instance, if certain population groups are more likely to avoid the system due to a lack of trust in the institutions providing the system, they would not benefit from the system. In this context, promoting trust through transparency and consumer-friendly policies could improve fairness.

Deployment can also affect the system's ability to adapt to situations in which use of the model is inappropriate. A simple yet reliable fallback system could take over in such situations. Automated decisions using out of distribution detection or model confidence scores, user preferences, and developer overrides could all trigger the use of a fallback system. Additionally, robust deployment pipelines could enable developers to rectify any identified fairness concerns quickly.

3.5 Feedback

It is impossible to anticipate and prevent every possible fairness issue that may arise. It is therefore essential that the system is built with feedback mechanisms that enable prompt responses to fairness issues.

Simply auditing real-world performance regularly can help developers catch when the system is not performing as expected. To catch fairness issues, the auditing should include the use of fairness metrics and investigations into the nature of the systems output instead of just focusing on overall performance metrics. Explainability tools and interpretation of model parameters can be used to verify the decision making process is being conducted in an expected manner. A fairness framework can be applied to ensure effective and thorough auditing [20–22].

System developers should also consider the role of humans in the feedback loop. It has been shown that malicious decision-makers can fool fairness audits [23]. It is therefore important that an appropriate incentive structure is created to encourage unbiased fairness audits. On the other side of the feedback loop, users can be empowered to report any suspected fairness issues they encounter. User feedback can catch issues that may be overlooked by a high-level audit.

4 Conclusion

Recognizing fairness concerns within AI systems is important, but it is only the first step in the process of designing fair AI systems. Practitioners must also consider each component of their system and implement measures that promote fairness. Fortunately, practitioners have many tools at their disposal to build fair systems. These tools can help practitioners audit their datasets, discourage the formation of model biases, understand their models and more. Through careful consideration of the requirements of their system and the available tools for addressing identified concerns, AI practitioners can design systems that minimize the potential for fairness issues.

References

[1] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," *Reuters*, Oct 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[2] R. Metz, "Twitter says its image-cropping algorithm was biased, so it's ditching it | cnn business," May 2021. [Online]. Available: https://www.cnn.com/2021/05/19/tech/twitter-image-cropping-algorithm-bias/index.html

[3] T. W. House, "Blueprint for an ai bill of rights - ostp," Oct 2022. [Online]. Available: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[4] E. Kim, J. Lee, and J. Choo, "Biaswap: Removing dataset bias with bias-tailored swapping augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 992–15 001.

[5] B. Jain, M. Huber, and R. Elmasri, "Increasing fairness in predictions using bias parity score based loss function regularization," *arXiv preprint arXiv:2111.03638*, 2021.

[6] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[7] 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*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–16. [Online]. Available: https://doi.org/10.1145/3290605.3300830

[8] N. Kallus and A. Zhou, "Residual unfairness in fair machine learning from prejudiced data," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2439–2448. [Online]. Available: https://proceedings.mlr.press/v80/kallus18a.html

[9] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[10] C. Dulhanty and A. Wong, “Auditing imagenet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets,” *arXiv preprint arXiv:1905.01347*, 2019.

[11] M. Cooper, “When conventional wisdom fails: Revisiting data augmentation for self-driving cars,” Dec 2018. [Online]. Available: https://towardsdatascience.com/when-conventional-wisdom-fails-revisiting-data-augmentation-for-self-driving-cars-4831998c5509

[12] P. Cunningham and S. J. Delany, “Underestimation bias and underfitting in machine learning,” in *Trustworthy AI - Integrating Learning, Optimization and Reasoning*, F. Heintz, M. Milano, and B. O’Sullivan, Eds. Cham: Springer International Publishing, 2021, pp. 20–31.

[13] T. Zhang, T. Zhu, J. Li, M. Han, W. Zhou, and P. S. Yu, “Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 4, pp. 1763–1774, 2022.

[14] N. Burkart and M. F. Huber, “A survey on the explainability of supervised machine learning,” *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[15] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[16] A. Liu and B. Ziebart, “Robust classification under sample selection bias,” in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/d67d8ab4f4c10bf22aa353e27879133c-Paper.pdf

[17] D. Hendrycks, K. Lee, and M. Mazeika, “Using pre-training can improve model robustness and uncertainty,” in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2712–2721. [Online]. Available: https://proceedings.mlr.press/v97/hendrycks19a.html

[18] Y. Wu, D. Zeng, X. Xu, Y. Shi, and J. Hu, “Fairprune: Achieving fairness through pruning for dermatological disease diagnosis,” *arXiv preprint arXiv:2203.02110*, 2022.

[19] S. Hooker, N. Moorosi, G. Clark, S. Bengio, and E. Denton, “Characterising bias in compressed models,” *arXiv preprint arXiv:2010.03058*, 2020.

[20] S. Segal, Y. Adi, B. Pinkas, C. Baum, C. Ganesh, and J. Keshet, “Fairness in the eyes of the data: Certifying machine-learning models,” in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 926–935.

[21] K. Cachel and E. Rundensteiner, “Fins auditing framework: Group fairness for subset selections,” in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 144–155. [Online]. Available: https://doi.org/10.1145/3514094.3534160

[22] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, “Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 33–44. [Online]. Available: https://doi.org/10.1145/3351095.3372873

[23] K. Fukuchi, S. Hara, and T. Maehara, “Faking fairness via stealthily biased sampling,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, pp. 412–419, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5377