# Fairness: A Formal-Methods Perspective

Aws Albarghouthi

University of Wisconsin–Madison, Madison, WI

**Abstract.** Sensitive decisions of large-scale societal impact are increasingly being delegated to opaque software—a trend that is unlikely to slow down in the near future. The issue of fairness and bias of decision-making algorithms has thus become a multifaceted, interdisciplinary concern, attracting the attention of computer scientists, law scholars, policy makers, journalists, and many others. In this expository paper, I will outline some of the research questions we have been studying about fairness through the lens of formal methods.

## 1 Introduction

Whether it is at the industrial or governmental level, we are witnessing widespread automation of processes that have the potential to adversely impact individuals or groups. Consider, for instance, automatically assigned credit scores, automated filtering of applicant resumes, predictive policing, or algorithmic pricing. All of these automated processes have the potential to adversely affect the individuals or groups—typically minorities—who are the subjects of the algorithmic decisions. These are not mere academic concerns. Indeed, accounts of automated discrimination are being constantly reported in a range of areas.

Prompted by the importance of addressing bias in automated decision-making, researchers across many disciplines have started studying this problem. Notably, in computer science, multiple formal definitions of fairness have been proposed, and studying their merits, shortcomings, and contradictions is an active area of study (see, e.g., [4, 6, 7, 9]). Relatedly, numerous techniques—mostly in statistical machine learning—have been proposed to enforce some of those fairness definitions (see, e.g., [6, 13, 12, 7]).

In this paper, I will view the fairness problem through the lens of formal methods, and outline a number of research questions we are currently studying. A central technical theme that I wish to highlight is the reduction of probabilistic reasoning to logical reasoning, an approach that allows us to harness the power of established logical techniques—e.g., SMT solvers—for probabilistic reasoning.

## 2 Fairness Through the Lens of Formal Methods

**Fairness as a Program Property.** Suppose we have a program $P$ that, given an input $x$ representing some individual's information, decides whether to invite them for a job interview. Our view of the problem is broad: $P$ could be a machine-learned classifier, it could be an SQL query filtering out applicants from a database, a Python script, etc.

How do we ensure that $P$ is *fair*? One class of definitions, called *individual fairness*, specifies that $P$ should return a *similar* decision for similar individuals—for some definition of similarity. This is a notion of program *robustness*, ensuring that $P$ is unaffected by input perturbations. Another class of definitions, called *group fairness*, specify that the selection rate for applicants from a minority group is comparable to the selection rate of the rest of the applicants—the majority.

In our work on FairSquare [1], we showed that we can characterize a range of such fairness definitions as formal specifications of programs, and proposed automated verification algorithms to check whether a program satisfies the specification, under a given population. Specifically, we characterize the population as a probability distribution $D$. Then, $P$ can be viewed as a distribution transformer, and we can ask questions like, what is the probability that $P$ hires an applicant conditioned on them being a minority? By answering such questions, we can check a property like group fairness, e.g., following Feldman et al. [6],

$$\frac{\Pr[P(x) = \mathit{true} \mid x \text{ is a minority}]}{\Pr[P(x) = \mathit{true} \mid x \text{ is a majority}]} \geqslant 1 - \varepsilon$$

which prescribes that the selection rate from the minority group is at least $1 - \varepsilon$ that of the majority group, for some small $\varepsilon$.

At the algorithmic level, we reduce the problem of computing those probabilities to *weighted volume computation*, where our goal is to compute the volume of the region defined by an SMT formula in linear arithmetic. We demonstrated that we can solve this quantitative problem via iterative calls to an SMT solver, resulting in a sound and complete approach [1, 11].

**Fairifying Unfair Programs.** What should we do when we detect an unfair program? We extended FairSquare with what we like to call a *fairification* algorithm [2], which takes an unfair program $P$ and transforms it into a fair program $P'$, for a provided fairness definition.

Specifically, we cast the problem as follows: Find a fair program $P'$ such that

$$\Pr[P(x) \neq P'(x)] \text{ is minimized}$$

The idea is that $P$ is not trying to be egregiously unfair. What we therefore want to do is nudge it a little bit to make sure that we cover its blind spots that are causing its bias.

At the algorithmic level, we demonstrate how to solve fairification via iteratively solving constraint-based synthesis problems, whose solutions are candidate programs $P'$. Our approach is inspired by classic results in *probably approximately correct* (PAC) learning, adapted to an SMT-based program synthesis setting.

**Fairness and Privacy.** Finally, I would like to point out the connection between fairness—particularly, individual fairness—and notions of statistical privacy. *Differential privacy* [5] prescribes that minor modifications to a single record in a database do not yield large changes in the output of a query; this ensures the privacy of the individual to whom the record pertains. Differential privacy is established by randomizing the query evaluation algorithm, instilling noise in its results. The standard definition is that given query $q$, for all similar databases, $d$ and $d'$, and every possible output $o$, we have

$$\Pr[q(d) = o] \leqslant e^{\varepsilon} \cdot \Pr[q(d') = o]$$

where $\varepsilon > 0$ is a parameter. As $\varepsilon$ approaches 0, the outputs of the query $q$ on $d$ and $d'$ approach each other, increasing privacy.

Perhaps predictably, algorithms for enforcing differential privacy have been ported for ensuring fairness [4, 8]. For instance, Kearns et al. [8] employ differential-privacy mechanisms to ensure individually fair selection of individuals from a database—e.g., choosing top football players to send to the world cup.[1] Intuitively, the randomization enforced by differential privacy ensures that minor differences between players do not translate into large changes in the chosen team.

Given the difficulty of designing randomized algorithms for differential privacy (mistakes have been found in published proofs [10]), we have developed automated techniques for proving differential privacy [3]. We demonstrated that we can solve this problem via a careful reduction to solving a system of recursive constraints, defined via *constrained Horn clauses*. Specifically, we showed that a rich space of proofs, called *coupling proofs*, can be logically characterized, allowing us to eliminate probabilistic reasoning.

## 3 Conclusion

I discussed some of our progress in applying automated verification and synthesis to addressing problems in fairness of decision-making programs. I believe that the formal methods community, broadly defined, has plenty to contribute to the discourse on fairness and bias—for instance, by developing formally verified implementations of fair algorithms, runtime verification techniques for detecting unfairness, programming languages where fairness is a first-class construct, debugging techniques for detecting potential bias, and others.

**Acknowledgements.** The work described in this talk is in collaboration with a fantastic group of students and researchers: Samuel Drews, Aditya Nori, Loris D'Antoni, Justin Hsu, Calvin Smith, and David Merrell. The work described is generously supported by the National Science Foundation (NSF) grant #1704117.

## References

1. Albarghouthi, A., D'Antoni, L., Drews, S., Nori, A.V.: Fairsquare: Probabilistic verification of program fairness. Proceedings of the ACM on Programming Languages 1(OOPSLA), 80:1–80:30 (Oct 2017), http://doi.acm.org/10.1145/3133904

[1] World events at the time of writing strongly influenced my choice of example.

2. Albarghouthi, A., DAntoni, L., Drews, S.: Repairing decision-making programs under uncertainty. In: International Conference on Computer Aided Verification. pp. 181–200. Springer (2017)
3. Albarghouthi, A., Hsu, J.: Synthesizing coupling proofs of differential privacy. Proceedings of the ACM on Programming Languages 2(POPL), 58:1–58:30 (2018), http://doi.acm.org/10.1145/3158146
4. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. In: Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012. pp. 214–226 (2012)
5. Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy, vol. 9. Now Publishers, Inc. (2014)
6. 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, Sydney, NSW, Australia, August 10-13, 2015. pp. 259–268 (2015), http://doi.acm.org/10.1145/2783258.2783311
7. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. CoRR abs/1610.02413 (2016), http://arxiv.org/abs/1610.02413
8. Kearns, M., Roth, A., Wu, Z.S.: Meritocratic fairness for cross-population selection. In: International Conference on Machine Learning. pp. 1828–1836 (2017)
9. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: ITCS (2017)
10. Lyu, M., Su, D., Li, N.: Understanding the Sparse Vector Technique for differential privacy. In: Appeared at the International Conference on Very Large Data Bases (VLDB), Munich, Germany. vol. 10, pp. 637–648 (2017), http://arxiv.org/abs/1603.01699
11. Merrell, D., Albarghouthi, A., DAntoni, L.: Weighted model integration with orthogonal transformations. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 4610–4616. AAAI Press (2017)
12. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 560–568. ACM (2008)
13. Zemel, R.S., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013. pp. 325–333 (2013), http://jmlr.org/proceedings/papers/v28/zemel13.html