Trimming the Thorns of AI Fairness Research

Jared Sylvester
Independent
j@jsylvest.com

Edward Raff
Booz Allen Hamilton
Raff_Edward@bah.com

Abstract

Machine learning practitioners are often ambivalent about the ethical aspects of their products. We believe anything that gets us from that current state to one in which our systems are achieving some degree of fairness is an improvement that should be welcomed. This is true even when that progress does not get us 100% of the way to the goal of "complete" fairness or perfect alignment with our personal belief on which measure of fairness is used. Systems being built with some measure of fairness implemented would still put us in a better position than the status quo by increasing the number of systems caring about the problem enough to invest effort toward its remediation. Impediments to getting fairness and ethical concerns applied in real applications, whether they are abstruse philosophical debates or technical overhead such as the introduction of ever more hyper-parameters, should be avoided. In this paper we further elaborate on our argument for this viewpoint and its importance.

1 Introduction

General questions regarding the fairness of machine learning models have increased in their frequency and study in recent years. Such questions can quickly enter philosophical domains and subjective world views [1], but are crucial as machine learning becomes integrated in the fabric of society. The attention and critical thought is well deserved as we see applications emerge which dramatically and grievously impact people's lives and families, such as predictive policing [2] and sentencing [3, 4].

Despite this, we argue that a significant portion of the machine learning community are missing important questions regarding how to maximize the amount of fair machine learning deployed in the world. In particular, there are practical considerations for applied fairness with respect to current fairness that are being ignored. Stated simply if we want to increase fairness of real world machine learning systems, we should not delay solutions over concerns of optimal fairness when there currently exists no fairness at all.

As such we must ask: how do we maximize the number of people implementing/deploying fair/ethical machine learning solutions? We posit that the answer to such a question is to minimize the amount of *mental* and *computational* work that must be done to gain fairness. This applies to any practitioner with varying degrees of education or training in ethics and machine learning. If the incremental cost to deployment is too significant, we argue the concern of fairness will often be dropped in the name of expediency and financial cost. To be explicit: we do not think that fairness considerations *ought* to be dropped if their implementation is perceived as being onerous, but that unfortunately they *are* dropped as a "practical matter."

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

Under this general belief, we have identified three areas where we feel the community could increase social good by instead tempering its emphasis on some optimal notion of fairness. These three areas relate to the debate around what ideal of fairness should be used (mental cost), over-reliance on trolley car hypotheticals (mental cost), and the nature of the algorithms people are developing (computational cost). (While there is ongoing debate about whether AI solutions are appropriate at all for some high-impact fields such as criminal justice, it seems indisputable that there exist some domains in which *(i)* AI will be used and *(ii)* AI should be made as fair as possible. It is these domains we are discussing in this paper. Furthermore, in this paper we are concerning ourselves with fairness interventions via technical means only. We are aware that many ethical considerations in AI may be affected by social means such as altering the cultural landscape the AI system operates in, or involving stakeholders with different experiences and viewpoints. We consider these issues outside our this paper's scope and our expertise.)

In section 2 we will argue for changing how we discuss and interact with parties about fairness & ethics concerns. Because of the nature of answering "what is fair" is fraught with complexity that can lead to heated debate, we believe better outcomes can be achieved sooner by avoiding characterizing alternative approaches as "wrong" but instead as opportunities for discussion. In section 3 we will discuss the nature of the importance endowed to the equation of ethics. It is undeniably important, but can obscure practical considerations of what can actually be achieved. We will highlight some of the debate around self-driving cars as an example where unconstrained concerns have potentially moved past what could ever exist. To ensure that such valid concerns are addressed we must remember to design with respect to what a system could reasonably do. In section 4 we will discuss the need to perform more research on making the questions and mitigations to fairness, bias, and ethics issues more accessible to implementers and decision makers. The research to date has largely focused on technical aspects, and will not have a practical impact until we bridge the gap to those who are in a position to impact change in their organizations. Finally we will conclude in section 5.

2 The Unfair Criticism for the Wrong Fairness

One of the largest impediments stopping the adoption of fairness by non-expert practitioners is answering "*what is fairness?*" There is a rich history of philosophical debate around this very question from which we can build upon as a community [1]. At the same time, a philosophical conclusion has not been reached after hundreds of years — so we arguably should not expect there to be one universal definition that the machine learning community will ever rally around. If people can not agree on a single definition of "fair" when defining it with natural language, why would we expect a single definition to be found when we move into the more rigorous, less ambiguous language of mathematics and algorithms? Throughout this paper ourselves we will talk about "fairness" and "social good" as if they are quantities that can be objectively improved, but they are not. They exist as intricate social constructs that vary from each person based on life experiences, exposure, history, and belief systems. For the sake of brevity in exposition, we use this oversimplifying verbiage precisely because their discussion is too complex to do justice in the constraints of this page.

The complex nature of specifying a robust definition is part of what makes fairness a moving target: the definition of "fair" changes over time as societal views change and new issues are surfaced, unlike technical definitions we are used to working with.[1] Ordinary Least Squares has always been and will always be the "optimal" linear unbiased estimator.[2] No future social views will change the optimality of the OLS algorithm. In contrast, a racially-dependent algorithm for mortgage underwriting which we view as being patently unfair would have not only been widely considered fair in prior years, but potentially *mandatory* [5]. As a result of the

[1]The royal "we" of persons trained in STEM oriented programs that did not necessarily prepare them to ask or answer moralistic, social, or philosophical questions.

[2]Subject to the usual conditions of homoscedastic, non-autocorrelated errors, and using "unbiased" in the mathematical as opposed to the ethical sense.

mutability of our notions of fairness, there may in fact be no one, final, universal definition of "fair" to be found. Even if all current parties agreed to adopt a particular definition, that would not guarantee any longevity to that consensus. As such, we must address how we as a community interact when confronting disagreements on these important issues.

Just as there are many differing notions of fairness in our society at large, many differing definitions and metrics for fairness and discrimination in predictive machine learning have been developed [6, 7, 8]. Unfortunately, these definitions have shown to be at some level incompatible with each other, making it impossible to maximize fairness as measured along one of these dimensions without sacrificing fairness measured according to another [8]. These incompatibilities have been evinced primarily in regards to binary prediction problems in Machine Learning; seeking to reconcile nascent definitions in sub-areas like recommender systems [9, 10], regression [11, 12], and clustering [13] or those that may have been defined in neighboring fields like economics [14] will only introduce further contradictions.

Given these competing definitions of fairness, it is important that we as a community avoid being overly critical on what *specific* definition of fairness is selected for an individual project or system. For those applications where no measures of fairness are currently considered, we should even go further and applaud and encourage the selection of *any* reasonable and beneficial fairness criteria, even if it is not the one we would have personally preferred.

Selecting any fairness criteria is not necessarily satisficing, but we argue it is an improvement from the status quo by developing buy-in to the need for fairness, ethics, and bias to be considered. The implementers of the system, by selecting any measure, are now invested in the fairness of their product and thus may become more open to improving the fairness as a type of feature. Even if another measure is superior in some context, having a less-ideal metric implemented opens the door for revisiting and adjusting the fairness portion of the system at another point in time.[3]

Encouraging this could prove to be an advantageous path of least resistance. Not only does it allow for a transitional nature, but can yield positive network effects within larger organizations. For example, Team A gets to add a monthly update report that fairness was added as a feature, which could get other mangers or engineers thinking about fairness for their project. This may not happen if members of the team fear being censured for choosing the "wrong" metric of fairness, or for implementing a system which increases fairness without completely maximizing it on some measure.

In addition, a machine learning system of suboptimal fairness may still be more fair than the non-quantitative, human system it augments or replaces. In fairness, as in other aspects of assessing engineering systems, designers should be encouraged to always ask "compared to what?" — i.e. not "is this system fair with respect to an ideal?" but "is this system fair compared to the system it supplants?" An only-partially fair quantitative system may be preferred because it can be measured, logged and inspected in a way that no human-driven, qualitative decision making can be [15, 16]. This greater legibility can lead to greater transparency.

An important component of this success is respectful discourse between groups on disagreements about what is or is not fair, and openness about how one is measuring fairness in a given system. If these do not exist, disagreements may devolve to stronger accusations and acrimony. Our concern in this is not to dismiss the valid and important concerns in underrepresented or impacted populations, but a practical matter on the nature of achieving desired outcomes that require convincing others of a new viewpoint. By the very nature of having multiple choices to select from (i.e., criteria for fairness and potential interventions) without full understanding, people can become more invested in the option they have chosen simply by making the choice [17]. If the party with sub-optimal fairness has a sincere belief that their current course of action is correct, it is possible that accusatory attempts to alter their opinions will have the opposite effect [18, 19]. Instead, engaging in the

[3]This is essentially Ousterhout's Dictum that "the best performance improvement is the transition from the nonworking state to the working state," but applied to fairness: better to have some fairness mechanisms implemented, affording the opportunity for future iteration and improvement, than to have implemented no fairness mechanism at all.

conversation in a nature that reaffirms the value of what has been done (even if we disagree), can be a more effective method to changing beliefs [20].

Johndrow and Lum [21] highlighted an example of this with the maligned COMPAS system for predicting criminal recidivism. Angwin et al. [22] from ProPublica published an article about bias in the COMPAS system. In response, the company which developed COMPAS, Northpointe, released a report showing the metrics by which their system was fair [23]. (These were, perhaps predictably, not the metrics Angwin et al. used to evaluate COMPAS, nor were they compatible with them.) Clearly the issue under consideration is of critical importance, to a degree such that the debate about what the best measure of fairness is and how to make the system more fair as a whole should be mandatory and continuous. But the nature of how this debate has unfolded in this particular instance has lead to considerable negative publicity when it appears that Nortpointe made an earnest good faith effort to address the issue before it became newsworthy. The issue appears be not a manichaean struggle of fair-vs-unfair, but which of two competing and somewhat incompatible definitions of fairness should be prioritized.

As a community we must avoid exchanges like COMPAS to avoid scaring off future leaders and decision makers from the issue of fairness (and machine learning more broadly). Put simply, COMPAS sets a precedent for social risk via negative publicity even when attempting to imbue machine learning systems with fairness. Even if one were to go well beyond what Northpointe did, there is still a risk of censure from critiques simply because they may adopt a different definition of fairness. This risk may prevent adoption, and thus lower the total fairness within the world.

If we instead accept that there is no single supreme definition of fairness, the situation can be improved. When we accept that others may not have considered certain factors in selecting their fairness measure, or may have reached their conclusion under different but equally valid philosophical beliefs, the conversation about fairness can be lifted to a more civil and less accusatory tone. In doing so the social risk can be transformed into social reward, as feedback will no longer be perceived as an attack that must be defended — but as genuine interest from the larger community. In this way we can make the matter of fairness more accessible, and increase the success rate of intervention when an insufficient consideration or effort to address a fairness concern is present.

3 Should Autonomous Vehicles Brake for the Trolley Car?

The trolley car problem [24, 25] has been the subject of much debate recently, coinciding with the increased interest in both fairness and autonomous vehicles. While many variants exist, the general trolley car problem is as follows: if a vehicle continues on it's current path, it will kill five people in its way; if some action is taken it can instead strike and kill only a single person. The specifics of the dilemma change (if the vehicle continues driving it will hit a child; if it swerves it will kill its passenger; etc.), but at its core it is a contrived situation with a pair of differing, but both negative, outcomes.

With self-driving cars on the precipice of deployment, the trolley car problem makes intuitive sense for study. Hardware failures, sudden changes in environment (like an earthquake), or actions of bystanders & non-autonomous vehicles are all factors outside of a self-driving agent's control that could lead to a potentially fatal situation. All of this is made more pertinent due to the first, unfortunate, death at the hand of an autonomous vehicle [26].

Even before this sad death, many have been debating the trolley car problem and arguing that a solution is needed for deployment [27, 28, 29, 30]. This circles back to the problems we discussed in section 2: what measure of ethical behavior we should be using to decide who lives and who dies in the myriad of possible trolley car scenarios? Surveys reveal that people prefer that cars be willing to sacrifice the driver, but simultaneously would not personally want to own such a car [31]. That this would create a dichotomy is understandable, but it makes reaching a consensus on what should be done difficult. Further studies have looked at presenting varied

trolley car scenarios and simply asking people which way the car should swerve, and then attempting to quantify the resulting empirical ethos [32].

Despite all of these questions of research and debate, we do not see it asked: do drivers today consider the trolley car problem when they are about to enter an accident? We argue that no such consideration exists today or even could with human drivers. The small amount of time to react in any such scenario likely means people are simply relying on gut reactions and are not performing any meaningful, ethical consideration of who will and will not survive an accident. Nor do we prepare people to make these sorts of decisions: no ethical training or testing is undertaken before issuing people with drivers' licenses.

If people are not considering this problem today, why should we *require* self driving cars to do the same? It results in a moving of goal posts, requiring cars to reach super-human abilities before we let them take over a task.[4] If self driving cars can reduce the number of fatalities by 90% [33], then we reduce the incident rate of trolley car situations by 90%. In this way we are in a sense solving the trolley car problem exogenously by reducing its frequency, as the best possible scenario is the one where the trolley car problem never occurs. We argue this increases social good without having to solve such a difficult problem, and that delaying deployment until a satisfactory solution is obtained may in-fact needlessly delay improved safety for everyone.

We take a moment to emphasize that we are not arguing self driving cars should be deployed as soon as possible. Considerable and thorough safety and validation testing should be mandatory before public deployment; we can not afford to cut corners. We are arguing that certain fairness considerations that are being debated, such as the trolley car problem, have been imbued with an importance beyond the reality of their application.

Along these lines, we need to further consider what situations will lead to trolley car problems. It seems likely that one of the most common culprits is mechanical failures: brakes stop working effectively, steering or sensor systems malfunction, etc. In such a case, even if the car had an oracle that solved the trolley car problem, it is not obvious to us that it would be able to execute on that solution due to the aforementioned mechanical failure.

Going further, even if we did have a oracle that can solve the trolley car problem, we likely could not effectively use it. This is because the car itself will need to be predicting attributes of people involved, such as their ages, their risk of fatality, and a myriad of other factors that would be necessary inputs to the trolley car problem. But each of these predictions will have their own error rates, and some, like risk of fatality, may not even have any reliable predictive models developed. Realistically any trolley car solution would also require an understanding of risk and uncertainty about the situation itself. This is an issue we don't see discussed, and is contributes to why we believe a trolley car solution is an unreasonable expectation.

To delay a potential life-saving innovation is itself deadly. We are engaging in a real-life meta-trolley problem: our meta-trolley is currently running on a track that allows human drivers to kill a million people a year [34], and could be switched to an alternate track that is likely to be far less deadly. Meanwhile we stand by, arguing about the propriety of pulling the lever to allow the meta-trolley to switch to that lower lethality track.

4 Fairly Complicated Fair Algorithms

We've discussed two situations in which the emphasis on getting fairness exactly right may lead to reduced fairness in practice. Now we discuss a matter with regard to practitioners in making fairness algorithms as usable as possible. This means reducing the number of hyper parameters, and computational and cognitive costs in adding fairness to current algorithms. We feel this issue is dreadfully under-studied.

The current focus, even within the human-computer interaction (CHI) field, has advocated for more holistic understanding of a problem and a ML algorithm's application and context to design a more fair system [35]. Others have studied how users may perceive fairness criteria or their degree of understanding of different fairness

[4] Some argue that AI should only have to be as ethical as the humans whose decisions they are supplanting. Other claim that since AI may have super-human abilities, it is not unreasonable that they have super-human ethical responsibilities as well. We would contend that holding AI to a higher standard than humans may be acceptable, but holding them to a standard of *perfection* is not.

criteria [36, 37]. These studies are valuable, but all still require expertise on by implementers of a system in order to make the decisions work by bringing them to fruition. If the implementers can't successfully reify these ideas in practice, it does not matter how well those affected by an ML system may understand it. We need CHI research on how to ease the process of incorporating fairness and bias into a ML system for the decision makers, engineers, and researchers that will perform and supervise the work. In essence, it does not matter how successful any intervention we design is if practitioners at large are unable to wield them.

It is common for fairness mechanism to introduce multiple new hyper-parameters to an algorithm, in addition to the ones that existed before [38, 39, 40]. This can get particularly out of hand when multiple different parts of the model must be specified for any new problem [21]. Such solutions necessitate a more expensive parameter search, thus increasing the financial cost of developing deployable AI solutions. This reduces the incentive for companies to invest in the time to make fair models, and thus should be something we attempt to minimize. This has become more pronounced as Transfomer based models are becoming more popular due to their reported successes [41]. Such models can cost millions to train once[5] and can capture a large amount of harmful bias [42]. We are not criticizing the active work into mitigating these very new problems. Our goal is only to emphasize that an ideal solution should hopefully minimize the additional financial cost of training a Transformer in the first place. If a hypothetical solution was developed with a $25\times$ cost to train (e.g., due to additional parameters to tune) it would inflate a GPT-3 model's cost to over $100 million. This would leave the solution fiscally untenable.

Indeed, the high cost of training such state-of-the-art systems generates an ethical question in its own right, namely, will the most advanced methods be under the control of only the most financially well-endowed institutions? [43] If altering the training algorithms of such large-scale models requires the introduction of many additional hyperparamters and thereby greatly increases the training cost, we find ourselves with a similar trade-off to the one we discussed in Section 2. For example, we may be able to improve a model's demographic parity but in a way that increases the cost of the training procedure several orders of magnitude. If in so doing we limit the model's use to only those with the deepest pockets, we are faced with a conflict between two different ethical concerns.

While we have no expectation of a magic black box which will produce fair algorithms and require no work (human or computational), we do believe there is room for considerable simplification of the approaches being developed. Having one or zero hyper-parameters may not lead to a perfectly optimized balance between fairness and predictive performance, but it may lead to faster adoption and integration within organizations today, thus increasing fairness from our current baseline. Such simple, low-computation fairness algorithms may not lead to ideal results. However, it would still be good to have them in our collective toolbox as a community for certain situations, such as for use by those with limited computation resources.

In a similar vein, we would like to see research along automatically assessing measures of fairness to optimize for and provide human-readable reports about what the ramifications would be. As far as the authors are aware, these two notions have yet to receive study in the machine learning community. The automatic selection of a fairness metric could be done with respect to a maximum acceptable loss in accuracy (e.g., which measure can be maximally satisfied at a fixed cost?). Though the solution may not be optimal, it could prove better than the default state of no fairness consideration. To be explicit, we are not arguing that AI systems should decide how other AI systems should be ethical, but we should develop systems that help humans understand how they can modify the system and what the consequences of those modifications are.

A tool that can generate human-readable reports on the impacts of different fairness measures and provide some "map" of the potential options would also provide value. It better enables product developers and practitioners who are not experts to weigh the costs of various fairness methods and potentially integrate them, as well as the impacts of any measure selected in the aforementioned auto-fairness idea.

The goal of all of these preferences for usable fair algorithms is not to directly solve fairness by any means, but to maximize social good in the near and long term. They create a path of least resistance for novices who are

[5] e.g., see analysis from `https://lambdalabs.com/blog/demystifying-gpt-3`.

concerned about fairness so that *something* can be integrated immediately. This also opens the door to future exploration and improvement of fairness as its own feature, and provides, in our opinion, a viable method for integration into the maximal number of systems. If such work continues to be unstudied, we may leave businesses and developers a daunting task: a whole world of literature, competing definitions, and philosophical questions fraught with ethical and social complexities that must be understood before even being able to start. The apparent gap itself may become the biggest deterrent to adoption, and so we wish to implore the community to build these bridges.

5 Conclusions

Our current machine learning systems are becoming more powerful and being deployed more widely each day, and yet they — and their creators — are often completely oblivious to issues of fairness. There is a broad chasm between the current state of machine learning and ideally ethical systems. It is our contention that we should welcome any efforts which narrow that gap, even if they fall short of bridging it completely.

We believe that some fairness is better than no fairness. Arguments, attitudes and techniques for perfect fairness are impeding our ability to get any improvements relative to the status quo. We should not let the perfect be the enemy of the good in our ongoing quest for a more just world.

We call on people in this discussion to realize that other researchers and practitioners are trying to make the world better and more just, even if they aren't making the exact improvement that you might prefer. We do not mean that anyone should be beyond reproach or that anyone's concerns should go unexpressed or unheard, merely that we should try whenever feasible to make critiques constructively and civilly so that we can work together toward a more fair society.

We propose that researchers and practitioners in this field should not ask "does this meet some Platonic ideal of fairness?" but rather they should be concerned with "does this increase the amount of fairness in the world?"

6 Authors' Note

Unlike the fields we typically publish in, this topic is one which is fraught with high emotion and potentially direct impact on people's lives. (Indeed, the former is because of the latter.) As such, we think it may be wise to make explicit what we are *not* saying in this paper.

- We do not advocate an "anything goes" moral relativism when it comes to different definitions of fairness, nor do we think that all definitions are equally appropriate in all conditions just because it is impossible to pick a single universal definition.
- We are not advocating any particular definition of fairness, or suggesting that any particular definitions should be used as defaults.
- We do not think that making any attempt at introducing fairness mechanisms to an AI system is sufficient, or that anyone who does so should be beyond critique. Nor should any cursory effort at fairness entitle one to praise merely for good intent.
- Our core argument is that we should not let the perfect be the enemy of the good. To the extent that an attempted fairness-increasing mechanism fails to even be "good" then we do not wish to defend it merely on the basis of its possible pure intentions. (For example, a classifier which picked labels completely at random may achieve parity, but would be an ethical disaster, and we do not recommend its adoption.)
- As stated in the introduction, we are not considering the question of whether a particular problem domain is too sensitive for AI to be used at all. Nor are we considering ethical improvements that could be made

via changes to the surrounding culture, such as increasing the diversity of ML engineers. Not all problems have technical solutions nor should all solutions be technical. Nevertheless it is only the subset of technical solutions that we address here.

- Many ethical problems with ML systems are partially a result of "garbage-in-garbage-out" issues from fundamentally biased data. There exist technical approaches to "de-bias" such data (e.g. Amini et al. [44]). That does not mean we think that such technical approaches give us reason to ignore the underlying causes that gave rise to the biased data in the first place, only that such technical approaches are superior to pretending that no problem exists at all.
- We do not think that any improvement in fairness from the status quo is sufficient as an end state, but that such improvements are useful stepping stones to even more fair systems. Criticism is an indispensable part of that continuing process of improvement, and it should be both offered and welcomed freely.
- We do not wish to downplay the harm that AI systems can cause or the suffering of those harmed. Indeed, it is because of this harm, both actual and potential, that we wish to see improvements made.
- While we hope participants in this debate will act with comity, that is only a hope. We are in no position to impose a tone on anyone, nor would we seek to.
- Finally, we wish to remark that this paper is based on our own experiences implementing machine learning systems in academic and industry settings. It is unavoidably the product of our own backgrounds, and we recognize these experiences are not universal.

References

[1] R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 2018, pp. 149–159. [Online]. Available: http://proceedings.mlr.press/v81/binns18a.html

[2] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian, "Runaway Feedback Loops in Predictive Policing," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 2018, pp. 160–171. [Online]. Available: http://proceedings.mlr.press/v81/ensign18a.html

[3] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," in *FAT ML Workshop*, 2017. [Online]. Available: http://arxiv.org/abs/1703.00056

[4] C. Barabas, M. Virza, K. Dinakar, J. Ito, and J. Zittrain, "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 2018, pp. 62–76. [Online]. Available: http://proceedings.mlr.press/v81/barabas18a.html

[5] J. Kimble, "Insuring inequality: The role of the federal housing administration in the urban ghettoization of african americans," *Law & Social Inquiry*, vol. 32, no. 2, pp. 399–434, 2007.

[6] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, vol. 29, no. 05, pp. 582–638, 11 2014. [Online]. Available: http://www.journals.cambridge.org/abstract_S0269888913000039

[7] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 2 2009, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/4909197/

[8] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

[9] R. Burke, N. Sonboli, and A. Ordonez-Gauger, "Balanced Neighborhoods for Multi-sided Fairness in Recommendation," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 2018, pp. 202–214. [Online]. Available: http://proceedings.mlr.press/v81/burke18a.html

[10] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera, "All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 2018, pp. 172–186. [Online]. Available: http://proceedings.mlr.press/v81/ekstrand18b.html

[11] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling Attribute Effect in Linear Regression," in *2013 IEEE 13th International Conference on Data Mining*. IEEE, 12 2013, pp. 71–80. [Online]. Available: http://ieeexplore.ieee.org/document/6729491/

[12] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A Convex Framework for Fair Regression," in *FAT ML Workshop*, 2017. [Online]. Available: http://arxiv.org/abs/1706.02409

[13] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Fair Clustering Through Fairlets," in *FAT ML Workshop*, 2017.

[14] W. J. Baumol, "Applied Fairness Theory and Rationing Policy," *The American Economic Review*, vol. 72, no. 4, pp. 639–651, 1982. [Online]. Available: http://www.jstor.org/stable/1810007

[15] L. A. Henkel and M. Mather, "Memory attributions for choices: How beliefs shape our memories," *Journal of Memory and Language*, vol. 57, no. 2, pp. 163–176, 2007.

[16] M. Lind, M. Visentini, T. Mäntylä, and F. Del Missier, "Choice-supportive misremembering: A new taxonomy and review," *Frontiers in Psychology*, vol. 8, p. 2062, 2017.

[17] J. Brehm, "Postdecision changes in the desirability of alternatives." *Journal of abnormal psychology*, vol. 52, no. 3, pp. 384–389, 1956.

[18] D. Gal and D. D. Rucker, "When in Doubt, Shout!: Paradoxical Influences of Doubt on Proselytizing," *Psychological Science*, vol. 21, no. 11, pp. 1701–1707, nov 2010. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0956797610385953

[19] D. D. Rucker and R. E. Petty, "When Resistance Is Futile: Consequences of Failed Counterarguing for Attitude Certainty." *Journal of Personality and Social Psychology*, vol. 86, no. 2, pp. 219–235, 2004. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.86.2.219

[20] G. L. Cohen, J. Aronson, and C. M. Steele, "When Beliefs Yield to Evidence: Reducing Biased Evaluation by Affirming the Self," *Personality and Social Psychology Bulletin*, vol. 26, no. 9, pp. 1151–1164, nov 2000. [Online]. Available: http://journals.sagepub.com/doi/10.1177/01461672002611011

[21] J. E. Johndrow and K. Lum, "An algorithm for removing sensitive information: application to race-independent recidivism prediction," pp. 1–25, 2017. [Online]. Available: http://arxiv.org/abs/1703.04957

[22] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[23] W. Dieterich, C. Mendoza, and T. Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity," Northpointe, Tech. Rep., 2016. [Online]. Available: http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

[24] P. Foot, "The problem of abortion and the doctrine of double effect," *Oxford Review*, no. 5, 1967.

[25] J. J. Thomson, "Killing, letting die, and the trolley problem," *The Monist*, vol. 59, no. 2, pp. 204–217, 1976.

[26] D. Lee, "Sensor firm Velodyne 'baffled' by Uber self-driving death," 2018. [Online]. Available: http://www.bbc.com/news/technology-43523286

[27] J. Achenbach, "Driverless cars are colliding with the creepy Trolley Problem," 12 2015. [Online]. Available: https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/?utm_term=.bad98a87d67e

[28] G. Corfield, "Kill animals and destroy property before hurting humans, Germany tells future self-driving cars," 8 2017. [Online]. Available: https://www.theregister.co.uk/2017/08/24/driverless_cars_ethics_laws_germany/

[29] P. Lin, "Why Ethics Matters for Autonomous Cars," in *Autonomous Driving: Technical, Legal and Social Aspects*, M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 69–85. [Online]. Available: https://doi.org/10.1007/978-3-662-48847-8_4

[30] N. J. Goodall, "Can you program ethics into a self-driving car?" *IEEE Spectrum*, vol. 53, no. 6, pp. 28–58, 6 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7473149/

[31] J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles," *Science*, vol. 352, no. 6293, pp. 1573–1576, 6 2016. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.aaf2654

[32] A. Shariff, J.-F. Bonnefon, and I. Rahwan, "Psychological roadblocks to the adoption of self-driving vehicles," *Nature Human Behaviour*, vol. 1, no. 10, pp. 694–696, 10 2017. [Online]. Available: http://www.nature.com/articles/s41562-017-0202-6

[33] M. Bertoncello and D. Wee, "Ten ways autonomous driving could redefine the automotive world," 2015. [Online]. Available: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-could-redefine-the-automotive-world

[34] World Health Organization, "Global status report on road safety," http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/, October 2015.

[35] M. K. Lee, N. Grgić-Hlača, M. C. Tschantz, R. Binns, A. Weller, M. Carney, and K. Inkpen, "Human-centered approaches to fair and responsible AI," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–8. [Online]. Available: https://doi.org/10.1145/3334480.3375158

[36] D. Saha, C. Schumann, D. Mcelfresh, J. Dickerson, M. Mazurek, and M. Tschantz, "Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. Virtual: PMLR, 2020, pp. 8377–8387. [Online]. Available: http://proceedings.mlr.press/v119/saha20c.html

[37] M. Srivastava, H. Heidari, and A. 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 & Data Mining*, ser. KDD '19. New York, NY, USA: ACM, 2019, pp. 2459–2468. [Online]. Available: http://doi.acm.org/10.1145/3292500.3330664

[38] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 3. Atlanta, Georgia, USA: PMLR, 2013, pp. 325–333. [Online]. Available: http://proceedings.mlr.press/v28/zemel13.html

[39] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The Variational Fair Autoencoder," in *International Conference on Learning Representations (ICLR)*, 2016. [Online]. Available: http://arxiv.org/abs/1511.00830

[40] H. Edwards and A. Storkey, "Censoring Representations with an Adversary," in *International Conference on Learning Representations (ICLR)*, 2016. [Online]. Available: http://arxiv.org/abs/1511.05897

[41] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *NeurIPS*, 2020. [Online]. Available: http://arxiv.org/abs/2005.14165

[42] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3356–3369. [Online]. Available: https://www.aclweb.org/anthology/2020.findings-emnlp.301

[43] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.

[44] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 289–295.