TABLE OF CONTENTS

Association for Computing Machinery

Advancing Computing as a Science & Profession

About *SIGKDD Explorations*

Explorations is published twice yearly, in June/July and December/January each year. After the first two volumes, frequency may increase to quarterly. The newsletter is distributed in hardcopy form to all members of the ACM SIGKDD. It is also sent to ACM's network of libraries. Additionally, issues are published on the web and are free to the general public (http://www.acm.org/sigkdd/explorations/).

Our goal is to make *SIGKDD Explorations* an informative, rapid means of publication and a dynamic forum for communication with the Knowledge Discovery and Data Mining community. SIGKDD membership is growing at a very fast pace, and with KDD being a multi-disciplinary field, we hope that *Explorations* will facilitate its fusion and enhance the sense of community. Submissions will be reviewed by the editor and/or associate and guest editors as appropriate. We are particularly interested in short research and survey articles on various aspects of data mining and KDD. *Explorations* is also a forum for publishing position papers, controversial positions, challenges to the community, product reviews, book reviews, news items and other items of interest to the field. Please see:

http://www.acm.org/sigkdd/explorations/instructions.htm

Advertiser Information:

Explorations accepts advertisements related to data mining and KDD, including company, book, vendor, and service advertisements. For rates and instructions on submitting an ad, please see:

http://www.acm.org/sigkdd/explorations/instructions.htm#advertise

Notice to Contributing Authors to SIG Newsletters:

By submitting your article for distribution in this Special Interest Group publication, you hereby grant to ACM the following non-exclusive, perpetual, worldwide rights:

- to publish in print on condition of acceptance by the editor
- to digitize and post your article in the electronic version of this publication
- to include the article in the ACM Digital Library and in any Digital Library related services
- to allow users to make a personal copy of the article for noncommercial, educational or research purposes.

However, as a contributing author, you retain copyright to your article and ACM will refer requests for republication directly to you.

Introduction to The Special Section on Bias and Fairness in AI

Toon Calders, Eirini Ntoutsi, Mykola Pechenizkiy, Bodo Rosenhahn, Salvatore Ruggieri
U. Antwerpen, Freie Universität Berlin, TU Eindhoven, Leibniz University Hannover, Università di Pisa
toon.calders@uantwerpen.be, eirini.ntoutsi@fu-berlin.de, m.pechenizkiy@tue.nl,
rosenhahn@tnt.uni-hannover.de, salvatore.ruggieri@unipi.it

ABSTRACT

Fairness in Artificial Intelligence rightfully receives a lot of attention these days. Many life-impacting decisions are being partially automated, including health-care resource planning decisions, insurance and credit risk predictions, recidivism predictions, etc. Much of work appearing on this topic within the Data Mining, Machine Learning and Artificial Intelligence community is focused on technological aspects. Nevertheless, fairness is much wider than this as it lies at the intersection of philosophy, ethics, legislation, and practical perspectives. Therefore, to fill this gap and bring together scholars of these disciplines working on fairness, the first workshop on Bias and Fairness in AI was held online on September 18, 2020 at the ECML-PKDD 2020 conference. This special section includes six articles presenting different perspectives on bias and fairness from different angles.

Keywords

Bias, fairness, discrimination, fairness-aware machine learning, responsible artificial intelligence

1. INTRODUCTION

Artificial Intelligence (AI) techniques based on big data and algorithmic processing are increasingly used to guide decisions in important societal spheres, including hiring decisions, university admissions, loan granting, and crime prediction. However, there are growing concerns with regard to the epistemic and normative quality of AI evaluations and predictions. In particular, there is strong evidence that algorithms may sometimes amplify rather than eliminate existing bias and discrimination, and thereby have negative effects on social cohesion and on democratic institutions.

Despite the increased amount of work in this area in the last few years, we still lack a comprehensive understanding of how pertinent concepts of bias or discrimination should be interpreted in the context of AI and which socio-technical options to combat bias and discrimination are both realistically possible and normatively justified. The main objective of the workshop on Bias and Fairness in AI held online[1] on September 18, 2020 at the ECML-PKDD 2020 conference is a contribution to the understanding of "How can standards of unbiased attitudes and non-discriminatory practices be met in (big) data analysis, AI and algorithm-based decision-making?".

We introduce topics in Bias and Fairness in AI and describe how they were covered in the program of the workshop in Section 2 and provide a brief overview of the contributed articles to this special section in Section 3.

2. TOPICS IN AI BIAS AND FAIRNESS

Research on fairness in machine learning and data mining took off in 2008-2010 with some of the first works on discrimination discovery in databases [1] and learning classification models with (non-discrimination) independency constraints [2, 3]. These papers were followed by an exponential explosion of papers in major AI conferences, and an emergence of new cross-disciplinary workshops and conferences such as most notably FAccT[2] and AIES[3]. A recent snapshot of the frontiers of fairness in machine learning research can be found in [4].

Much of the research on fairness in machine learning can be framed in an optimization context [5], where the goal is to maintain good predictive performance while satisfying a number of group-level or individual fairness constraints. This combination can be achieved via modeling and removing representation bias and/or labeling bias in the training data, via fairness-aware representation learning [6, 7], model induction, model selection, regularization, or post-processing of specific [8] or any [9] trained models or model outputs.

In parallel, temporal dynamics of fairness in algorithmic decision making [10] and its long-term impact [11] has been studied to address feedback loops that may amplify discrimination.

Next to algorithmic approaches, also progress has been made with respect to theoretical analysis to better understand the possibility or impossibility of fairness with its different often conflicting notions [12].

Another recent avenue of fairness-aware machine learning research includes causality. The notion of counterfactual fairness and approaches of counterfactual inference have been proposed to make predictions fair across different subpopulations. Considering classification as an optimization problem with fairness constraints entailed by competing causal explanations, Russell et al. [13] demonstrated that it is possible to be approximately fair with respect to multiple pos-

[1] https://sites.google.com/view/bias-2020/programme

[2] https://facctconference.org/

[3] https://www.aies-conference.com/

sible causal models at once, thus mitigating the bottleneck of exact causal specification.

The BIAS2020 workshop solicited contributions on bias and fairness in all areas of AI (supervised and unsupervised learning, reinforcement learning, information retrieval and recommender systems, human-computer interaction, constraint solving, complex systems and networks, etc.) and encouraging interdisciplinary studies including law, philosophy and social sciences. 21 full paper submissions were received of which 7 were selected to the workshop program after peer-review. The program also featured four invited talks and a concluding panel discussion. Revised and extended contributions were invited for this special section.

3. CONTRIBUTED ARTICLES

The special section includes six contributed articles spanning a variety of topics: philosophical viewpoints on discrimination [14], applicability of different ML fairness notions [15], a new measure for viewpoint fairness in ranking applications [16], gender perception in online platforms [17], fair classification via ethical adversaries [18], and why not only serendipity but also equity should be considered to mitigate historical discrimination effects [19].

Two Kinds of Discrimination in AI-Based Penal Decision-Making. Hubner in [14] presents a viewpoint on discrimination in algorithmic decision making from the standpoint of practical philosophy and ethics of science. In his work, he distinguishes two kinds of discrimination that need to be addressed in AI-based penal decision-making: the problem of inevitable trade-offs between incompatibility of statistical fairness measures as became widely known due to the COMPAS study and analyzed theoretically in [20], and the problem referred to as the so-called *discursive fairness* that applies when *humans make decisions based on empirical evidence*. Hubner discusses the fundamental differences in approaching requirements of non-discriminatory action within the penal sector for each of these two kinds of discrimination. Whereas in the case of statistical fairness, the focus is on measuring dependency between race and (correctly and/or wrongly) predicted recidivism, in case of discursive fairness, it is necessary to analyze what types of information must be provided when justifying a court's decisions based on a machine learning model's predictions. This leads to seeking answers to the core question: *What reasons must a judge as a human decision maker provide for* ***her*** *each and every decision to grant or deny parole.*

On the Applicability of Machine Learning Fairness Notions. While many notions of fairness were introduced and many machine learning approaches and techniques have been development that can help to optimize for those notions, we also know that it is impossible to optimizing several of the competing notions of fairness at the same time [12]. Hence, a natural practitioner's question is which notion of fairness should be used. Makhlouf et al. [15] introduce a survey of fairness notions that should help find an answer to the question "which notion of fairness is most suited to a given real-world scenario and why?". The authors identify a set of fairness-related characteristics of real-world scenarios and analyze the relevance of corresponding fairness notions to these characteristics. Their findings are summarized in a decision diagram that may help different research communities, practitioners and policy makers to understand and navigate the space of fairness notions studied in fairness-aware machine learning.

Blind Spots in AI: the Role of Serendipity and Equity in Algorithm-Based Decision-Making. Van Leeuwen et al. [19] argue that designing an algorithm-based decision-making system focusing solely on serendipity might not be enough to avoid historical discrimination and therefore they suggest to also include equity in the development process. To this end, they propose a design rationale that incorporates the principles of serendipity (diversifiability) and equity (intersectionality, reflexivity and power balance) for the development of such systems.

Gendering algorithms in social media. Fosch et al. [17] investigate the impact of algorithmic bias on inadvertent privacy violations and the reinforcement of social prejudices of gender and sexuality. In particular, they conducted an online survey to understand whether and how Twitter inferred the gender of users. They found that gender-related stereotypes persist both online and offline, and platforms often appear to fail to understand that gender is not binary (male/female). Beyond Twitter's binary understanding of gender and the inevitability of the gender inference as part of Twitter's personalization trade-off, they also found that the misgendering rate is much higher for gay men (32%) and straight women (16%) as compared to straight males (8%). Their results call for attention to gender in gender classifiers to avoid amplification of existing biases that affect especially marginalized communities.

Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning. It is now common in machine learning research to address non-discrimination by introducing independency constraints into the predictive modeling process. One generic approach to do this was presented in [5]. Delobelle et al. [18] continue on this track and introduce the idea of using adversarial training for improving fairness of classification. The authors introduce a framework that makes use of two models. One model is optimized for preventing the correct guessing of the values of protected attributes, while staying as accurate as possible. The other adversary model leverages evasion attacks to generate new examples that will be misclassified and provides them to the training of the first model. The experimental evaluation of this framework on common benchmarks like the COMPAS datasets demonstrates promising results for achieving group level fairness including demographic parity and equality of opportunity.

Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics. Fairness-awareness is being considered in a variety of applications of autonomous decision making by machine learning based scoring mechanisms. Considering biases and fairness in recommender systems and web search, a graph-based algorithm that post-processes generated recommendations for improving aggregate diversity was proposed in [21]. The paper of Draws et al. [16] included in the special section, highlights the importance of researching how to measure and assess *viewpoint diversity* in real search result rankings. Depending on how the items are ranked in search results, more homogeneous or more diverse items or viewpoints will be exposed to the user. The authors show that assessing the viewpoint diversity might not be as straightforward as it may seem, considering and experimenting with a few ranking fairness metrics in a controlled simulation study.

We hope you will enjoy reading the papers on bias and fairness in AI in this special section and find them an inspiration for formulating and addressing many of the open challenges in this socio-technical problem space, advancing the current state of the art further and further.

Acknowledgements

Our special thanks go to the invited speakers, all authors who submitted to and presented their work at the BIAS2020 workshop, to the program committee members and ad hoc reviewers, and to all the participants. The workshop was part of the FAccT network `https://facctconference.org/network/`. The work of E. Ntoutsi and S. Ruggieri was partially supported by the European Community H2020 project *NoBIAS* (`nobias.eu` G.A. 860630).

4. REFERENCES

[1] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08*, page 560–568, New York, NY, USA, 2008. ACM.

[2] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, ICDMW '09*, page 13–18, USA, 2009. IEEE Computer Society.

[3] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, 21(2):277–292, 2010.

[4] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, April 2020.

[5] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

[6] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 325–333. PMLR, 2013.

[7] Tongxin Hu, Vasileios Iosifidis, Wentong Liao, Hang Zhang, Michael Ying Yang, Eirini Ntoutsi, and Bodo Rosenhahn. FairNN - conjoint learning of fair representations for fair decisions. In *Proceedings of the 23rd International Conference on Discovery Science, DS 2020*, volume 12323 of *LNCS*, pages 581–595. Springer, 2020.

[8] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *Proceedings of the 10th IEEE International Conference on Data Mining, ICDM 2010*, pages 869–874. IEEE Computer Society, 2010.

[9] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems, NIPS 2016*, pages 3315–3323, 2016.

[10] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*, page 525–534, New York, NY, USA, 2020. ACM.

[11] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 3156–3164, 2018.

[12] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143, March 2021.

[13] Chris Russell, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems, NIPS 2017*, pages 6414–6423, 2017.

[14] Dietmar Hübner. Two kinds of discrimination in AI-based penal decision-making. *SIGKDD Explorations*, 23(1), 2021.

[15] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of machine learning fairness notions. *SIGKDD Explorations*, 23(1), 2021.

[16] Tim Draws, Nava Tintarev, and Ujwal Gadiraju. Assessing viewpoint diversity in search results using ranking fairness metrics. *SIGKDD Explorations*, 23(1), 2021.

[17] Eduard Fosch-Villaronga, Adam Poulsen, Roger A. Søraa, and Bart Custers. Gendering algorithms in social media. *SIGKDD Explorations*, 23(1), 2021.

[18] Pieter Delobelle, Paul Temple, and Bettina Berendt. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *SIGKDD Explorations*, 23(1), 2021.

[19] Cora van Leeuwen, Annelien Smets, and An Jacobs. Blind spots in ai: the role of serendipity and equity in algorithm-based decision-making. *SIGKDD Explorations*, 23(1), 2021.

[20] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science*, page 43:1–43:23, 2017.

[21] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Fairmatch: A graph-based approach for improving aggregate diversity in recommender systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '20*, page 154–162, New York, NY, USA, 2020. ACM.

Two Kinds of Discrimination in AI-Based Penal Decision-Making

Dietmar Hübner
Institute of Philosophy, Leibniz University Hannover
Im Moore 21, D-30167 Hannover, Germany
dietmar.huebner@philos.uni-hannover.de

ABSTRACT

The famous COMPAS case has demonstrated the difficulties in identifying and combatting bias and discrimination in AI-based penal decision-making. In this paper, I distinguish two kinds of discrimination that need to be addressed in this context. The first is related to the well-known problem of inevitable trade-offs between incompatible accounts of *statistical fairness*, while the second refers to the specific standards of *discursive fairness* that apply when basing human decisions on empirical evidence. I will sketch the essential requirements of non-discriminatory action within the penal sector for each dimension. Concerning the former, we must consider the relevant *causes* of perceived correlations between race and recidivism in order to assess the moral adequacy of alternative standards of statistical fairness, whereas regarding the latter, we must analyze the specific *reasons* owed in penal trials in order to establish what types of information must be provided when justifying court decisions through AI evidence. Both positions are defended against alternative views which try to circumvent discussions of statistical fairness or which tend to downplay the demands of discursive fairness, respectively.

Keywords

AI-based decision-making, crime prediction, bias, discrimination, justice, fairness, statistical fairness, discursive fairness, COMPAS.

1. INTRODUCTION

This paper addresses two types of potential discrimination in algorithm-based decision-making. Both problems are well-known and have achieved widespread attention within the general public and the academic disciplines, though it may be surprising that I discuss both of them under the title "discrimination". However, I hope the following discussion will elucidate why this overarching perspective is conceptually justified and ethically insightful. In particular, it may clarify central questions within both dimensions and reveal important connections between the two issues.

Concerns of possible bias and discrimination in computer algorithms pertain to a multitude of areas, ranging from everyday applications such as image recognition software, search engines or chat bots to specialized systems used in university admission procedures, hiring decisions or loan granting. Notwithstanding this ubiquity of the topic, it is plausible to assume that normative analyses of and technical solutions to bias and discrimination in AI-based decision-making must ultimately be tailored to the concrete fields of social interaction in which these applications take place. In this paper, I will focus on the forensic sector, more specifically on AI-based crime prediction in penal court decisions. My primary example of use will be the famous COMPAS case.

COMPAS ("Correctional Offender Management Profiling for Alternative Sanctions") is a software package for crime prediction which was originally developed and marketed by the private firm Northpointe, meanwhile succeeded by Equivant (www.equivant.com). COMPAS aims to assess a defendant's probability of reoffending, thus supporting judges in their decisions concerning whether a culprit should be detained before trial or might be released on bail, whether she should go to prison or might be eligible for probation, or whether she should stay in jail or might be a candidate for parole [13, 31]. COMPAS is based on a questionnaire, collecting data such as the current charges against the defendant and her criminal history, but also socio-economic factors including education levels, employment status, family background, social environment etc. Drawing from 137 items, COMPAS generates a score predicting the risk of reoffending, ranging from 1 to 10 [9].

In 2016, a public controversy arose when Angwin and colleagues claimed that COMPAS risk scores were discriminating against black persons, pointing to apparent problems of *statistical fairness* in the algorithm's predictions [3, 28]. Beforehand, several authors had already expressed *procedural worries* about using AI evidence in court hearings [8, 10]. In the following sections, I will turn to both aspects respectively, highlighting the basic tenets of how each notion of fairness is to be assessed. Finally, I will connect these analyses into an integrated view.

2. STATISTICAL FAIRNESS

2.1 Numerical Features of COMPAS

Angwin et al.'s statistical concerns about COMPAS can be most readily retraced by arranging retrospective data on algorithmic predictions and real outcomes in an "error matrix" (or "confusion matrix"). In such a matrix, rows contain the numbers of persons predicted to exhibit a certain trait (here: predicted reoffending or predicted not reoffending), while columns display the numbers of persons indeed falling into the corresponding groups (here: in fact reoffending or in fact not reoffending). The four resulting fields of the matrix contain true positives (TP: predicted reoffending and in fact reoffending), false positives (FP: predicted reoffending, but in fact not reoffending), true negatives (TN: predicted not reoffending and in fact not reoffending), and false negatives (FN: predicted not reoffending, but in fact reoffending). Based on these numbers, several parameters can be calculated in order to evaluate the statistical performance of the algorithm.

Issues of statistical fairness can be addressed by calculating these parameters separately for different groups. In particular, discrepancies in parameters between groups distinguished by salient features such as race or gender may be taken as indications of potential discrimination. The following error matrix for COMPAS (Table 1), compiled from Larson et al., shows algorithmic predictions

and real outcomes in Broward County, Florida (2013/14), contrasting data on white (w) and black (b) defendants [28]. By analyzing these basic data, especially by comparing statistical parameters for whites and blacks, one can determine whether COMPAS satisfies different fairness conceptions [5, 6, 19, 27, 30, 34].

	In fact reoffending	In fact not reoffending
Predicted reoffending	TP w = 505; b = 1,369	FP w = 349; b = 805
Predicted not reoffending	FN w = 461; b = 532	TN w = 1,139; b = 990

Table 1: Error matrix for COMPAS, numbers from [28][1]

First, it must be stressed that COMPAS does not make *explicit use* of a *race variable* in order to generate its predictions. Race is not among the 137 items on the questionnaire, and nothing suggests that COMPAS reconstructs its value from proxy variables and then utilizes it as an additional input to its calculations. So COMPAS does comply with the standard of "fairness through unawareness", not referring to a variable that would carry the label "protected" in fields pertaining to possible racial discrimination. However, it is widely agreed that this is a minimum requirement which, in general, does not exhaust all statistical fairness issues.

Second, *predicted rates (PRs)*, i.e. relative numbers of persons predicted to reoffend, can be calculated (PR = {TP+FP}/total). Here we find a stark difference between both groups ($PR_w = 35\%$, $PR_b = 59\%$), demonstrating that COMPAS does not correspond to the standard of "statistical parity" ("demographic parity", "equal acceptance rate"). However, one might consider it adequate to compare the PRs to the *true rates p* ("base rates", "prevalences"), i.e. the relative numbers of individuals who do in fact reoffend (p = {TP+FN}/total). Although not in perfect agreement, these display at least a similar tendency ($p_w = 39\%$, $p_b = 51\%$). Against this background, it may appear incongruous to complain about unequal predicted rates PR. Rather, it might be suggested, COMPAS simply tracks social reality, as displayed in the true rates p. I will comment on this issue in the following sections, particularly on the problems of calling the prevalences p "true rates". But for the time being, it should be noted that also Angwin et al., in their critique of COMPAS, do not focus on its lack of statistical parity.

Instead, what Angwin et al. are predominantly concerned about is the *false positive rates (FPRs)*, i.e. the relative numbers of persons who, although they ultimately do not reoffend (i.e. being either a false positive or a true negative), have erroneously been predicted to reoffend (ending up as a false positive) (FPR = FP/{FP+TN}). This indicator is much higher for blacks than for whites ($FPR_w = 23\%$, $FPR_b = 45\%$), i.e. COMPAS does not fulfil "predictive equality". Maybe not surprisingly, the reverse is true for the *false negative rates (FNRs)*, i.e. the relative numbers of individuals who, although they ultimately do reoffend (i.e. being either a true positive or a false negative), have erroneously been predicted not to reoffend (ending up as a false negative) (FNR = FN/{TP+FN}). This figure is much higher for whites than for blacks ($FNR_w = 48\%$, $FNR_b = 28\%$), i.e. COMPAS does not satisfy "equal opportunity". So in both regards, we do not have "error rate balance", and taken together we do not have "equalized odds", which would require both error rates to be equal. Put simply, COMPAS seems to be too strict for blacks and too lax for whites.

However, Northpointe replied to these concerns by stating that this difference should not be misread as racial bias against black defendants [11]. In particular, they argued that the appropriate metric for judging fairness is rather the *positive predictive values (PPVs)*. This parameter measures the relative numbers of persons who, after having been predicted to reoffend (i.e. being either a true positive or a false positive), do in fact reoffend (ending up as a true positive) (PPV = TP/{TP+FP}). This value, although not fully identical, is in reasonable agreement for both groups ($PPV_w = 59\%$, $PPV_b = 63\%$). So COMPAS does establish approximate "predictive parity" (essentially equivalent to "calibration"). In this respect, COMPAS does not seem to discriminate against black people.

It appears like a natural requirement that all the above parameters, error rates as well as predictive values, be roughly equal for whites and blacks in order to avoid potential discrimination. But unfortunately, this comprehensible demand, except for degenerate cases like zero errors, is mathematically impossible to meet, where the prevalences p differ for both groups. There are several "impossibility theorems" demonstrating this unfavorable constellation [16, 25]. Maybe the most easily accessible proof is by Chouldechova [7]. She bases her argument on the equation FPR = [p/{1–p}] · [{1–PPV}/PPV] · [1–FNR]. From this formula, it can easily be seen that, if there is a difference in prevalence p for the two groups, the groups must also differ in at least one of the three quality parameters, FPR, FNR or PPV, unless these have values zero or one.[2]

2.2 A Stalemate

Against this background, some scholars have started to turn away from the debate on statistical fairness, preferring other approaches to issues of algorithmic discrimination [6, 18, 19, 24]. This reaction is comprehensible, and appears to be backed by several considerations.

First, the impossibility theorems mentioned above demonstrate that we cannot have all that we might want in terms of statistical fairness. And facing this irresolvable conflict of alternative conceptions, it is not obvious which fairness measure to prefer.

[1] Larson et al. only analyzed pretrial-detainment decisions, not probation or parole decisions, as COMPAS was predominantly used for the former in Broward County. They classified as "predicted reoffending" individuals receiving a risk score of 8–10, and as "predicted not reoffending" those with a risk score of 1–4, corresponding to Northpointe's classification of these individuals as "high risk" or "low risk", respectively. They defined "in fact reoffending" or "in fact not reoffending" with regard to whether the same person was again arrested within two years after her scoring, as COMPAS itself is supposed to predict a new offence within two years. I will comment on the problem of identifying reoffending with rearrests in due course.

[2] COMPAS, in fact, has significant differences in two parameters. Both FPRs and FNRs considerably deviate for whites and blacks. Theoretically, an algorithm could achieve equality of two parameters between the groups. But at least one of the three needs to compensate for the given difference in prevalence p.

To be sure, there is some reason to share Angwin et al.'s view that the differing FPRs for whites and blacks are especially disturbing. COMPAS is applied within the penal system, where false positives appear to be particularly troubling. It may be tempting to back this normative intuition with reference to the basic standard *in dubio pro reo*. However, this classical legal tenet concerns the *ascription of past offences* to a defendant: it states that, if you are not reasonably sure that some person has committed a crime, she should rather not be prosecuted. COMPAS, by contrast, is applied in decisions concerning bail, probation or parole, where the *prediction of future offences* of the defendant is at stake: in these contexts, we are fairly certain that a person has committed some offence, but we consider waiving incarceration, given an optimistic prediction that she will not perpetrate again. Consequently, *in dubio pro reo* does not properly apply here. In particular, false positives in COMPAS do not, as is sometimes suggested, amount to "incarcerating innocent people". If that was the case, even an FPR of 23% for whites would be outrageously high, not just an FPR of 45% for blacks. Yet, it may be reasonable to hold that, although the *in dubio* principle itself is not to the point, some more basic imperative standing behind it will still apply, namely the idea that it is legally paramount to avoid unnecessary punishment. Even when defendants are highly suspect or actually convicted of some past offence, imprisonment without demonstrated need ought to be avoided where possible, given its devastating impact on individuals and their families. Consequently, once we agree that the past offence in question is of a minor kind or has been atoned for to a sufficient degree, so that good confidence in the future compliance of a defendant would justify her release, failure to grant bail, probation or parole would constitute a major wrong within a liberal state, ultimately conflicting with the rule of law. Following this line of thought, focusing on FPRs, rather than on FNRs or PPVs, would appear paramount to penal justice. Admittedly, though, it may be less obvious why the FPRs need to be equal for different groups. We should possibly *minimize* them, but it is not clear yet why we should *equalize* them.

Similar remarks hold with regard to the FNRs. Contrary to the arguments sketched above, one may insist that, in discussions of bail, probation or parole, false negatives should constitute our primary focus. In these contexts, judges are presented with highly suspect or actually convicted individuals whose incarceration would be basically justified. Under these circumstances, decisions to waive imprisonment must, first and foremost, avoid possible dangers to the general public due to potentially non-compliant, dangerous, recidivating individuals. This is why predictions of future offences are involved in these decisions. The defendants did commit an offence, or are highly suspect of having done so, and the question of whether imprisonment could be waived must focus more on the danger of future recidivism in case of false negatives than on the danger of unnecessary imprisonment in the form of false positives. Note that this argument would not, as might first appear, establish that there was no race-related problem in COMPAS: to be sure, it would shift the focus away from the disproportionate numbers of false positives in black defendants that Angwin et al. concentrate on. But instead, it would have to turn to the enlarged numbers of false negatives in white defendants: stressing the need to protect the public, of all things an FNR of 48% for whites would seem to be unbearably high, not so much an FNR of 28% for blacks. Again, however, this line of reasoning may not really bear on issues of statistical fairness. It would probably require the *minimization* of FNRs, but it may not straightforwardly suggest the importance of their *equalization*.

It is also understandable that Northpointe underlined the importance of PPVs. Given that the algorithm predicted that a person would reoffend, the PPV indicates the probability that the person will indeed do so. So in a way, the PPV announces the reliability of the algorithmic prediction, the quality of the provided service. Thus, it is not surprising that computer scientists are inclined to focus on this parameter, and that common processes of algorithm optimization tend to increase its value. In addition, the information conveyed by the PPV seems to be in better correspondence to the epistemic situation of a decision-maker than the FPR or FNR. She is not presented with a not reoffending or reoffending individual and has to make up her mind whether the algorithm might misclassify that person (FPR or FNR), but she is presented with an algorithmic prediction and has to make up her mind whether this assessment will turn out to be true (PPV). Finally, it makes sense to assume that the PPV should not only be *maximized*, but also *equalized* across groups. For, if this is the case, the decision-maker (i.e. the judge) may restrict her considerations to the prediction that she is given (i.e. the risk score), without having to pay additional attention to the defendant's group affiliation when interpreting this information. If the PPV is equal for whites and blacks, a given risk score has the same meaning for both groups. The prediction has a consistent reliability, no matter whether the person concerned is white or black.

In short, there seems to be a real stalemate between these different fairness measures. And it may appear hopeless to find a decisive argument in favor of one of them.

Second, focusing on measures of statistical fairness runs the danger of absurd solutions, ending up with an AI that simply rearranges numbers in the error matrix in the way desired, but without any substantial sense [12, 18]. For instance, an algorithm could achieve statistical parity by predicting proportionate fractions from two groups to reoffend while selecting the individuals from both groups randomly, or it could equalize false positive rates by attributing high risk scores to actually harmless individuals.

Third, largely analogous debates on different statistical fairness standards and their mutual mathematical incompatibility took place back in the 1960s and 1970s, in discussions on potential bias and discrimination in assessment tests [21]. These discussions produced no decisive results, undermining hopes that we might do now better with the parallel problems in algorithmic predictions.

Given these findings, it is not surprising that some people have become weary of discussions on statistical fairness. At the same time, something important still seems to show up in the numbers which is worth addressing. In the following sections, I will make some remarks on these issues and suggest how they might be tackled. In particular, I contend that there is no general solution stating which fairness measures should dominate in any AI-based decision-scenario, be it university admissions or loan granting, but that we need to turn to a concrete scenario, like crime prediction in the forensic sector, in order to approach these problems.

2.3 The Core Question

To adequately grasp the issue of statistical fairness in AI-based penal decision-making, one core question needs to be addressed: *What is the cause of the correlation between race and recidivism that we appear to observe both in empirical data and in AI predictions?* It is only answers to this core question, I propose, that can guide us in balancing various statistical fairness demands. Two main answers to this question seem to suggest themselves.

(1) The first answer would be: "A major cause of the correlation is the *past treatment* of black people in the US. In the US history, we witness an extensive thread of *massive discrimination* against black persons, including slavery, political exclusion, segregation, and social marginalization. This practice has clearly led to a significant *socio-economic deprivation* of the black population. And higher crime incidence, or enlarged recidivism rates, as they show up both in empirical data and in AI predictions, must be regarded, to a large extent, as another *downstream effect* of this targeted maltreatment."

This attitude would assume that there is a *true correlation* between race and recidivism in social reality, i.e. that there is in fact a higher rate of reoffending in black defendants. But it would underline that this correlation is an obvious effect of *past wrongs* done to that population. In this paper, I will not try to enter into a political debate whether this perspective is adequate. Recent reports on systemic racism in the US police may suggest that higher crime rates in the black population are, to a considerable degree, a myth [4].[3]

For my current purpose, however, I want to explore what reactions this diagnosis would entail. And I think what suggests itself would be the idea that some kind of "affirmative action" might be applicable here [17]. Affirmative action comprises political measures meant to counter the disproportionate prevalence of salient groups in certain areas of public life. Such programs have been mainly justified in two different ways.

Firstly, and predominantly, affirmative action is grounded on the aim to promote diversity, plurality, integration or participation, e.g. in classrooms, universities, workplaces and offices. The driving idea behind this conception is recognition of the fact that these social units themselves *benefit* from the presence of different experiences and world views, and that society at large *needs* e.g. black attorneys or female managers in order to retain social cohesion and provide role models. However, it seems dubious how this line of justification might apply to the case at hand. It would appear strange to argue that we need racial diversity or racial integration in person imprisoned or in persons being released.

Secondly, however, affirmative action can be justified through the definite purpose to counter social correlations based on acknowledged wrongs. The main idea behind this conception is that we should ignore or override certain criteria that we usually apply in our assessments if it should turn out that they are tainted by past discrimination, in order to prevent these *past wrongs* from further infecting our *current decisions*. For instance, when we find that test results correlate with race or gender, and when we know that these correlations obtain because blacks or women have been subjected to preceding discrimination in their development and education, we should suppress or overrule these indicators, at least to some extent, and accept those applicants, in spite of their poorer performances. We should compensate them, not in the cheap sense of giving them some arbitrary advantages in order to balance their former harms, but in the conscientious sense of not letting their past disadvantages determine their future fates.

Following this line of thought, affirmative action advocates the targeted departure from common decision criteria in order to prevent past discrimination from influencing people's future lives. Within the context of COMPAS, this would mean that predicted rates for recidivism, when underlying judgements on bail, probation or parole, should be taken at values deviating from the true rates, correcting them for their problematic background in past injustice. More precisely, the *predicted rates* should be taken as equal, or at least more equal. Consequently, *statistical parity* would be the fairness measure to adhere to, at least to some extent [6, 12, 14, 15].

There may be some debate concerning whether this perspective is persuasive. For instance, the compensatory approach to affirmative action, as opposed to the diversity logic, presupposes that the concrete individual, and not just her social group, has been personally affected by the past wrongs in order to justify her favorable treatment, which might be hard to argue for in a given case of penal justice [17]. Moreover, it may be doubted that abstaining from punishment can really count as compensating for disadvantages, comparable to offering someone a university place considering her deficient education. In addition, our reason for ignoring poor test results may ultimately be backed by our confidence that the person thus favored might eventually succeed at our university, hoping that her hitherto underdeveloped talents will be awakened through high quality teaching, whereas in ignoring high risk scores we would have to acknowledge that we do in fact underrate her probability of reoffending, as her personality structure is likely to fail again in her unimproved circumstances.

Notwithstanding these caveats, affirmative action is basically applicable to any social system. And the demand to eradicate the social influence of past wrongs has at least some argumentative weight in the penal context. At any rate, it should be noted that the concept is not meant to apply to extremely dangerous criminals expected to commit further violent felonies. Its use is restricted to persons who, given their past and present record, are realistically eligible for bail, probation or parole.

(2) A different answer would be: "A major cause of the correlation is the *current treatment* of black people in the US. In the US criminal system, we find a systematic policy of *racial targeting* of black people, consisting in more intensive surveillance, more frequent arrests, and more severe sentences. This skewed practice leads to *false data* in the training sets from which COMPAS has learned, and these exaggerated trends are now being reproduced in the algorithm's predictions. In particular, the higher 'prevalence', the differing 'base rate' or disproportionate 'true rate' that seems to show up in retrospective assessments, is actually, to a large extent, a *social artefact* and not 'true' at all."

One major problem that this position will emphasize is the fact that, while COMPAS is generally supposed to predict future *offences*, as only the probability of impending offences can have any legitimate impact on court decisions concerning bail, probation or parole, COMPAS is actually designed to predict future *arrests*, as a closer reading of the official Practitioner's Guides reveals,

simply because it has been mainly trained on data sets of past arrests [9, 13, 31]. This implicit equation of (re-)offending with (re-)arrests in the application of COMPAS is plainly *wrong*, as not every arrest is based on a verified offence, and it is clearly *biased*, as in the US blacks are much more likely than whites to be subject to unfounded arrests without having committed an actual offence [4]. Using COMPAS, however, will feed this bias back into the system and perpetuate it. Based on false data (disproportionate arrest rates), it will make distorted predictions (concerning future offences), thus producing enlarged imprisonment rates, thus suggesting exaggerated crime rates, thus encouraging more racial

[3] I will come back to this skepticism below. Essentially, it converges with the alternative answer to the core question.

targeting, thus generating more false data, thus making more distorted predictions, and so on [22, 27].[4]

However, if it is false data that underlie our decisions, differing false positive rates are particularly hard to accept. In any case, unnecessary punishment is a major problem for penal justice, but if it is based on false data, it becomes clearly untenable.

In this light, Angwin et al.'s focus on the false positive rates is most comprehensible. As stated above, we may generally debate whether false positives, false negatives or positive predictive values are of paramount importance in penal justice. But when we learn that *unnecessary imprisonments* stem from constant misinformation, *false positives* must become our major concern.

In addition, against this background it makes sense not just to demand the minimization of false positive rates, but also their equalization across groups. When data are distorted to the detriment of one group, resulting differing error rates become a real issue. When *deviating miscarriages* are based on fake differences, we must avoid *differing mistakes* in harming people.

Correspondingly, this second answer to the core question suggests that our major concern should in fact be to equalize *false positive rates*. In technical terms, our algorithm should strive to satisfy *predictive equality*, rather than one of the other fairness measures [20, 35].

There is a little problem with this conclusion, as it apparently presupposes the FPRs to be objectively true when calling for their equalization. If these numbers are themselves infected by false data, equal or minimal or even zero FPRs will be no real comfort as they will still perpetuate the current discrimination in the system [6]. And in fact, we must suspect that FPRs, as reported by Angwin et al., are still distorted, because they follow COMPAS in counting rearrests as reoffences. So not only the "true rate" is not "true", as the second answer stresses, but the FPRs are not true either, although the second argument seeks to equalize them.

However, this inconsistency does not undermine the argument in a fatal way. Admittedly, the call for equal FPRs should ultimately not apply to Angwin et al.'s own figures, but to ideal numbers, counting as true positives or false negatives not simply all rearrested persons, but only individuals who do indeed reoffend. But this caveat does not contradict the basic idea that false positives must be the major concern against the background of biased training data. And if the FPR in blacks is too high even for the distorted numbers, at least that obvious mistake should be reduced, all the more as we have to suspect that their true FPR is bigger still, containing all the rearrested persons who did not reoffend.

2.4 Division of Labor

I will not try to explore how adequate the two answers to the core question are, or decide which argument is more convincing. It seems reasonable to assume that both asserted ways of influence contribute to the situation, and that both suggested remedies can be supported: There may be some true correlation between race and recidivism, based on past discrimination, which can encourage the affirmative action logic and thus make us want to move towards more equal predicted rates. There may also be false data underlying the algorithmic predictions, based on racial targeting, which should make unnecessary imprisonment our major concern and hence call for more equal false positive rates.

Even if both lines of reasoning apply, though, it is helpful to highlight their divergent focuses and disentangle their logical structures. Not least, this differentiation may be important in deciding which corrections should be performed by which player, suggesting a division of labor: Carrying out compensatory adjustments to predicted rates in the spirit of affirmative action might ultimately be the business of human users at the end of the decision-making process, and so best be realized by the judges: this conforms to widespread intuitions that it is up to society, and not to the algorithms, to take charge of correcting the long-term effects of discriminatory practices that shape our communities [19]. Balancing out error rates due to false data, by contrast, should rather be regarded as part of the algorithmic service provided, and so be taken care of by the programmers: correcting for problematic input should take place before the predictive output is presented.

At the same time, this ideal disentanglement may have its realistic limits. We must be prepared to meet deeper intertwinements between the two lines of argument, at all levels from diagnoses to principles and remedies.

Factual assessment. In an indirect sense, past discrimination may as well contribute to false data: ultimately, it is these historical practices that have brought about present stereotyping, prejudice and harshness on the side of police agencies. Conversely, current racial targeting may to some extent contribute to true correlations [19]: in fact, by generating opposition, resignation or role-acceptance within the black population it may reinforce problematic behavioral patterns.

Normative demands. In a certain sense, statistical parity, i.e. the aim of having more equal predicted rates, may also be seen as an approximate correction of skewing effects due to racial targeting [15]: in any case, deviation from observed correlations due to affirmative action is easier to accept when it is clear that these allegedly "true rates" are not "true" at all. Conversely, predictive equality, i.e. the aim of having more equal false positive rates, may be regarded a minimum requirement of not adding further wrongs to past discrimination: after all, the unnecessary punishment of black people appears like an unpleasant continuation of malpractices such as slavery, exclusion, segregation or marginalization.

Technical implementation. When applying common techniques to "debias" algorithms against the background of differing base rates, we might expect to produce both corrections simultaneously. Especially when not performing some selective rearrangement of entries in the error matrix, but following a more reasoned approach (e.g., by looking for variables strongly correlating with race and eliminating these variables from the training set), the ultimate effect will be to assimilate groups, i.e. more equal predicted rates and more equal error rates. This will usually have its costs, because eliminating information from the data will generally impair the accuracy of the algorithm (e.g., the error rates will be more equal, but they will go up). But this is the kind of price you always pay for affirmative action, and if the information is dubious anyway it may not be a high price.

So there is some irony in all of this. Both positions sketched above follow distinct paths of problem assessment and suggested solutions, and also imply a division of labor between judges and programmers in fighting statistical discrimination in AI-based

[4] A defense attorney presented with a high COMPAS risk score of her black client might callously reply, with only a minor admixture of outright sarcasm: "Of course, my client has a high probability of being rearrested – she is black!"

decision-making. But then again, both approaches appear to be deeply merged. Causes and their statistical effects, aims and their moral justification, and techniques and their mathematical impacts seem to be ultimately entangled. We may think that unequal false positive rates are the most urgent problem for a just penal system, particularly as they partly stem from false data due to racial targeting. However, in fighting this phenomenon (preferably at the algorithmic level), we will probably also produce more equal predicted rates, maybe ignoring some real differences between the groups. Now this is what affirmative action (most reasonably applied at the human level) is always about, overriding true correlations that originate in past discrimination. But anyway, past discrimination also contributes to false data and encourages equalizing false positive rates, and likewise, racial targeting enforces true correlations and suggests equalizing predicted rates.

2.5 Alternative Approaches

In Section 2.2, I noted that some scholars have proposed dropping issues of statistical fairness altogether. The alternative approaches to algorithmic discrimination that they instead pursue mostly refer to conceptions of causal reasoning [6, 19, 24], often employing standards of counterfactual fairness [18, 27]. It is beyond the scope of this paper to comment on these strategies in detail. However, it seems likely that their arguments will ultimately depend on considerations largely parallel to those sketched above.

Modern accounts of causal reasoning predominantly refer to Bayesian networks, representing and quantifying deterministic or probabilistic influences between relevant variables of specific systems [32]. In the present context, these conceptions would amount to unearthing the causal paths between (i) the sensitive feature "race", (ii) other attributes collected through the items on the COMPAS questionnaire such as education levels, family background etc., (iii) the risk score arrived at by the algorithm, and (iv) the real outcome of recidivating or not recidivating [6, 19, 24]. On this basis, one could check for discriminatory causal paths within the Bayesian network. In particular, causal paths between the variables "race" and "score" that are deemed illegitimate could be marked out as indicating wrongful discrimination by the algorithm. Obviously, *direct paths* between "race" and "score" would be unacceptable in this sense, as any immediate influence of the protected variable on the algorithmic prediction would amount to straightforward discrimination. However, *indirect paths*, mediated through other variables such as employment status or social environment, could be more controversial, in particular when these mediating variables do impact on recidivating behavior. So which of these paths are to be classified as "discriminatory", and which are to be accepted as legitimate?[5] A rather *extreme position* would regard all indirect paths between "race" and "score" as illegitimate. The effect would be that COMPAS were to be rated as thoroughly discriminatory, because its scores largely depend on such mediating variables. But this attitude seems to amount to an all-too sweeping exculpation of defendants, leaving no notion of personal accountability for any predictive traits that may be statistically correlated with someone's race. A more *moderate position* would consider some of these indirect paths to be tolerable, others less so. The effect of this would be that COMPAS might be in need of some corrections, but not of complete abandonment. Such a view would probably try to distinguish between variables that should be regarded as lying within a defendant's liability and those that she ought not to be accountable for. But how can this line be drawn, without getting deeply entangled in notoriously difficult metaphysical issues of free will and moral responsibility? A promising approach is to concentrate on the distinctly normative dimension of this question, asking which of the correlating variables are linked to race due to plainly unfair practices. More precisely, we need to know whether *past discrimination* has produced stable paths between "race" and "score", and we need to know whether *present discrimination* has produced false values of influences between these variables. This implies that, instead of just putting together a Bayesian network of *causal paths*, we will have to analyze the empirical causes of the *network itself*, in order to assess whether it contains evidence of algorithmic discrimination. But then, causal reasoning on algorithmic discrimination eventually carries us back to exactly those questions concerning affirmative action and unfounded data that our above discussion on statistical fairness has already marked out as essential.

Contemporary arguments on counterfactual fairness may be understood as special variants of causal analysis, arrived at by giving causes a counterfactual interpretation [29]. Within the current debate, these approaches amount to asking whether COMPAS would confer a different risk score to some person if she belonged to a different race [18, 27]. This approach has some intuitive appeal as a guide to discrimination issues. Apparently, an algorithm should be regarded as wrongfully discriminating against a black person if it gave that individual a lower risk score when switching her race variable from "black" to "white". To be sure, the envisaged alteration of her race should not come along with all sorts of *additional changes* within her personality or behavior, as such extra variations might certainly justify corresponding adjustments of algorithmic predictions. Rather, the change must be restricted to a shift in *race alone*, if a difference in score is to be indicative of discrimination. But how precisely is this fictitious state of some person "merely" having a different race to be envisaged?[6] On a rather *naïve interpretation*, it would mean that just the variable "race" changes its value, while all the items contained on the COMPAS questionnaire remain constant [18]. In that case, of course, COMPAS would output the same risk score as before, as it does not use the feature "race" at all, but only the items on the questionnaire. However, to conclude from this fact that there is no problem of discrimination involved appears simplistic, essentially reducing the concept of non-discrimination to plain "fairness through unawareness". On a more *realistic interpretation*, it would mean that the variable "race" changes, and along with it many other items on the COMPAS questionnaire which are socially correlated with race, including education levels, family background etc. [27]. In that case, COMPAS might certainly change its risk score. But it is not obvious that such a change would necessarily indicate discrimination, if we admit that these variables might correlate with criminal behavior. How then should we conceive of an imaginary change in "race" which would imply

[5] Within causal approaches to algorithmic fairness, this central question is framed as the distinction between "unfair" or "fair" causal paths, or between "unresolved" or "resolved" causal influences, leading from "race" to "score" [6, 24].

[6] Assuming a specific (social, non-biological, constructivist, non-reductionist) understanding of "race", some authors claim that the counterfactual notion of some person having (merely) a different race is incomprehensible in the first place and useless for debates on discrimination [23, 26].

discrimination if accompanied by a change in "score"? Again, this can be taken as an inherently normative question, and an adequately designed answer will be largely parallel to the above one. All those changes in other variables that can be traced back to *past or present discrimination* should be in the counterfactual picture of the person belonging to a different race. For if an algorithm's predictions changed with those variables, it would *perpetuate these discriminating effects* of historical wrongs or false data, and so be in need of debiasing corrections. Consequently, counterfactual reasoning has again brought us back to exactly those issues that already appeared pivotal for discussions of statistical fairness.

In short, whether (indirect) causal paths leading from "race" to risk scores are to be regarded as instances of bias, or which (imagined) counterfactual scenarios where "race" is switched and risk scores shift too should count as indicators of discrimination largely depends on the causes that establish these relationships in the first place. And adequate countermeasures against the workings of these correlations must ultimately be based on assessments of past or present discrimination that bring them about. At the same time, one may doubt whether it is necessary, and actually possible, to conduct these assessments down to the levels of single causal paths or even individual persons, as approaches of causal reasoning and counterfactual fairness tend to suggest. General policies that need to be established in fighting algorithmic discrimination in the penal sector may well be allowed to, and may ultimately have to, restrict themselves to a critical awareness of the sociological impact of discriminatory practices on correlations between "race" and scores, without spelling out their psychological mechanisms through concrete traits in specific persons.

3. DISCURSIVE FAIRNESS

3.1 Another Problem in COMPAS

It must be emphasized that nothing in the above discussion of statistical fairness is unique to AI predictions. Human predictions, whether in the forensic sector or in other social spheres, are affected by the same problems, i.e. the basic plurality of fairness measures and the general impossibility of their simultaneous fulfilment. But COMPAS opens up another problem, which we might phrase as a problem of discursive fairness. This problem, in contrast to statistical fairness, is specific to AI predictions, or more precisely to human decisions based on AI predictions.

Let us imagine that there were no statistical fairness problems in COMPAS, i.e. no true or spurious correlations between race and recidivism, no differing error rates for blacks and whites, and perfect matches of positive predictive values. Even if this were the case, we still might question the use of COMPAS predictions in court, pointing to issues which are commonly framed as "black box problem", "lack of transparency", or "right to explanation" [8, 10, 22]. But it is possible to reconstruct these issues once more under the heading of discrimination. This will help to make more explicit what the black box problem amounts to, what kind of transparency is required, and what rights are at stake, in the given penal context.

3.2 Definition of Discrimination

To see this it will be useful to start off from the following working definition of (wrongful) discrimination, which aims to capture the essential factual and normative dimensions of the concept [1]: (Wrongful) discrimination consists in *differentiating* between persons, particularly *disadvantaging* certain persons belonging to *salient groups*, for no *relevant reason*, notably *just because* of their belonging to a *salient group*. Focusing on "salient groups" brings in a *historical dimension*. More accurately, it is the history of a given society which determines whether some social group has been exposed to widespread disadvantaging, so that the corresponding feature marks out a vulnerable subpopulation, defined e.g. by race, gender, ethnicity, or religion. Clarifying what amounts to a "relevant reason" opens up a *contextual dimension*. More precisely, the question whether something is a relevant reason or not will depend on the social system and the corresponding decision processes envisaged, e.g. university admissions, loan granting, school tests, or job hiring.

Admittedly, within the given context of crime prediction for penal decision-making in court hearings there may be different accounts of what a "relevant reason" for denying a defendant bail, probation or parole must ultimately amount to. The alternative imperatives of focusing either on the avoidance of unnecessary imprisonment or on the protection of the general public leave this issue largely open (see Section 2.2). We may have been able to make some statements concerning statistical fairness, pointing out which imbalances between groups appear to be particularly troubling, given their past or present causes. But this does not give us concrete guidance on which levels of disposition to recidivism might justify waiving punishment and which might not.

Fortunately, however, for our present purpose we need not make any definite statements on these matters. What is important for the current discussion is merely that, whatever precise standard of penal justice we may subscribe to, if we decide to deny a defendant bail, probation or parole, we need to justify this decision by providing reasons for it. This is a basic demand of discursive fairness. For if we cannot provide reasons, we will have an instance of wrongful discrimination, differentiating between persons for no relevant reason (see definition above).

Note that according to this perspective, the focus is no longer on AI predictions and their statistical qualities, which may need to be adjusted, either by a judge or by a programmer. Rather, the focus is on human decision-making based on AI predictions, and on the specific justificatory demands that social decisions concerning other persons' fates entail. In our context, it is the specific discursive setting of a penal trial that will determine what may count as a relevant reason. As stated above, we will not need to establish substantial sets of reasons that are valid in this regard, but we can restrict ourselves to narrowing down the formal types of reasons that might serve in such justifications.

3.3 The Core Question

In order to address this problem of discursive fairness in AI-based penal decision-making, we need to answer another core question: *What reasons must a judge provide for her decision when denying a defendant bail, probation or parole?* Such a decision amounts to differential treatment of the defendant, compared to other defendants who were granted these advantages, and so the judge must give relevant reasons for this differential treatment, in order not to generate a clear instance of straightforward discrimination. However, when basing her decision on COMPAS risk scores, a couple of answers that the judge might want to produce are clearly insufficient for this purpose.

First attempt: "I did not decide, but COMPAS did!" – This answer is plainly wrong: the judge signed the judgment, and this very procedure of signing the judgment is what making a judicial decision consists in. Besides, if the answer was true, it would be bad news for the judge: judges are paid to make these decisions,

and if she could demonstrate that she did not make the decision, she would have to return her salary.

Second attempt: "I did decide, and my reason was the COMPAS risk score!" – This answer is beside the point: we did not ask for the subjective reason that may have prompted the judge's decision, i.e. a psychological explanation of her action. What we require as an answer is something entirely different: we ask for the objective reason that may account for her decision's adequacy, i.e. a legal justification of her action.

Third attempt: "But COMPAS is very reliable. Accordingly, its risk score should be taken as an objective reason!" – As a matter of fact, COMPAS's reliability is not all too impressive: a positive predictive value of around 60%, implying that only 6 out of 10 of COMPAS's predictions turn out to be true, is not that good.[7] But even if it were, or if we concluded that better estimates for future human criminal behavior are not available, either because of epistemic limits to such foreknowledge or because of ontological indeterminacies in human behavior, the answer would still be misplaced: COMPAS's position in the judicial process is comparable to that of an expert witness, and so positive evidence needs to be provided for its current recommendations, beyond just pointing to its past performances and general reliability.

Fourth attempt: "I know important details of COMPAS's structure, training, working, mechanism, including its problems, such as the diverging FPRs and the mathematical incompatibility of different fairness measures. Against this background, I do have an objective reason to take its predictions into account!" – Indeed, this is not true: the basic structure and subsequent training of COMPAS is a commercial secret of Northpointe, not revealed to the judges or the public.[8] But even if these facts about COMPAS were available, comparing its position to that of an expert witness once again demonstrates why the answer is not satisfying: we do not want details about an expert witness's brain structure, school training, mental processes, or reasoning styles either, but rather, we want a reason why some person is not eligible for bail, probation or parole.

3.4 Basic Requirements of Fair Trials

The four responses sketched in the preceding section are all flawed. But they bring us closer to what would in fact be required for something to count as a relevant reason in a court setting, when basing a denial of bail, probation or parole on a prediction of future criminal behavior. Certainly, some explanation of this prediction is necessary for a justification of the decision. But how much and what kind of explanation is demanded?

This problem must be addressed with regard to the discursive standards of a fair trial. Against this background, there seem to be two questions that the judge has to answer: First, which *feature of the defendant* makes her suggest that the defendant might reoffend (and thus is not eligible for bail, probation or parole)? Second, what *psychosocial regularity or causal mechanism* is assumed in this prediction (and thus in her decision)?

These questions must be answered within the discursive setting of a criminal proceeding. This is because the defense must be able to challenge the decision, and this can happen in two essential ways: The defense may either provide evidence that the defendant *does not have* the feature in question (by calling a witness, by submitting relevant documents, etc.). Or the defense may provide evidence that the regularity or the mechanism presumed *does not hold* (by hearing an expert, by pointing to recent research, etc.).

As a consequence, the following suggests itself as a first approximation to the above problem: a relevant reason required to prevent wrongful discrimination in a court decision on bail, probation or parole must specify *(i) decisive features of the defendant* presumed to make future offences from his side sufficiently probable, and *(ii) empirical regularities or scientific mechanisms* assumed to support this predictive verdict. This is a formal requirement which must be met independently of substantial debates on the factual reliability of both pieces of evidence, or the normative impact that they should have: We may argue about whether a defendant has the feature in question, or whether it is indeed predictive of criminal behavior. We may debate whether these facts should suffice to foreclose bail, probation or parole. But this discussion can only proceed, within a fair trial, when the two pieces of information are provided as reasons for the decision.

So this is the kind of "transparency" that is required for algorithmic predictions in penal settings. Its content is specified with regard to the discursive setting of a fair trial. COMPAS does not fulfil this criterion, because of its "black box" character, and so it violates a defendant's "right to explanation", in a very clear sense. In particular, the problem is not the lacking transparency of COMPAS's basic construction, owed to its commercial background, but the lacking transparency of COMPAS's concrete predictions, due to their unspecified references.

Note that this is a qualitative difference to human experts, such as psychologist consultants. To be sure, human experts, when appearing in court, may make risk predictions based on specialist theories that are not fully intelligible to judges or defense attorneys. However, they are still able to, and they will be asked to, state clearly the features of the case they consider paramount for their assessments and the regularities of behavior they assume in their prognoses. This is not just to let them demonstrate the epistemic quality of their predictions, but rather to enable others to challenge their predictions through targeted counter-evidence.

3.5 Diverging Opinions

The above argument is contrary to *State vs. Loomis*, a famous judgment of the *Supreme Court of Wisconsin* (July 13, 2016) in which it was decided that the due process rights of (white) defendant Eric Loomis were not infringed by the use of COMPAS risk scores in the trial against him [33]. In particular, his right to be sentenced based upon accurate information, including his opportunity to assess this accuracy and challenge its validity, was

[7] Indeed, much simpler algorithms, referring to considerably fewer items (age, sex, and number of past convictions) and working in a completely transparent way (through a decision tree), seem to perform as well as COMPAS in terms of reliability [2]. This fact might provide courts with additional doubts concerning whether it is worth carrying the costs of this commercial product and accepting its inherent lack of transparency.

[8] This fact, along with the problem of mislabeling (re-)arrests as (re-)offences (see Section 2.3) and the lack of transparency in results (see Section 3.4), seems to constitute a major knock-out argument against the use of COMPAS in its present form [22]. Public institutions in general, and penal courts in particular, should not accept this policy and rather demand that private firms, if they want to sell their services to public authorities, fully disclose the structures and workings of their products.

not considered to have been compromised in his trial. In its justification, the *Supreme Court of Wisconsin* argued that COMPAS used individual information on Loomis himself (collected from his criminal file and personal interviews) and proved statistically reliable in published validation studies (notwithstanding certain limitations and race correlation issues). Both pieces of evidence could be checked by Loomis so that his rights to due process were not infringed.

However, both levels of information are far too unspecific in order to grant the defense adequate opportunity to mount a legal challenge. The Loomis side needs to know, first, *which of the 137 items* were used in his case (and to what extent), and, second, *which empirical regularity* was assumed to obtain in the prediction (and how it was supported). It is only when provided with this information that the defense can launch a targeted challenge of the impending decision. So contrary to the opinion of the *Supreme Court of Wisconsin*, the use of COMPAS risk scores constitutes a clear violation of procedural justice.

4. CONCLUSION

Both accounts, statistical fairness and discursive fairness, allow for no general answer concerning what definition of fairness we should apply, or what standards of fairness we should adhere to. Adequate answers can only be approached with close regard to the historical facts we face and the concrete systems we are talking about. In discussing *statistical fairness* we need to look into the *empirical causes* for perceived correlations between race and recidivism, in order to establish which aims are paramount in debiasing AI predictions, i.e. which fairness measures are most relevant. In discussing *discursive fairness* we need to specify the *relevant reasons* in judicial proceedings on bail, probation or parole, which are determined by the justifications that must be provided for court decisions based on AI predictions in a fair trial.

Finally, both dimensions must be brought into close contact, as it is exactly the combination of the two fairness dimensions which may help us to avoid plainly insufficient accounts. As mentioned above, focusing exclusively on statistical fairness may end up in devising algorithms that satisfy fairness standards by issuing absurd predictions (equalizing predicted rates by way of randomization, equalizing false positives rates through deliberate misclassification of harmless persons). Taking into account discursive fairness may safeguard against these obvious malpractices (requiring justification of penal decisions precludes tossing coins, or detaining innocuous people). So being forced to mark out decisive features of persons and assumed regularities in predictions in the name of discursive fairness may prevent misguided versions of statistical fairness.

The considerations in this paper have closely referred to the specifics of discrimination against black people in the US, and the basic tenets of fair trials. The concrete statements arrived at are not directly transferrable to other settings and systems, such as gender discrimination in job hiring. However, similar observations might hold for these alternative applications as well. Monitoring empirical causes for observed correlations and defining relevant reasons for justifiable decisions may prove to be of paramount importance in many fields of AI-based decision-making.

5. ACKNOWLEDGMENTS

This work was carried out as part of the project *Bias and Discrimination in Big Data and Algorithmic Processing – BIAS* (www.bias-project.org), funded by Volkswagen Foundation. Many thanks to Markus Ahlers, Philippe van Basshuysen, Uljana Feest, Mathias Frisch, Caroline Gentgen, Christian Heinze, Jan Horstmann, Tina Krügel, Wolfgang Nejdl, Eirini Ntoutsi, Bodo Rosenhahn, Arjun Roy and Jannik Zeiser for numerous discussions, and to Lucie White for careful checking of the final manuscript.

6. REFERENCES

[1] Altman, A. Discrimination. The Stanford Encyclopedia of Philosophy (Winter 2020 Edition), ed. E.N. Zalta. https://plato.stanford.edu/archives/win2020/entries/discrimination.

[2] Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. Learning Certifiably Optimal Rule Lists for Categorical Data. Journal of Machine Learning Research 18, 234 (2018), 1–78.

[3] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 23, 2016. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] Balko, R. There's overwhelming evidence that the criminal justice system is racist. Here's the proof. The Washington Post, June 10, 2020, www.washingtonpost.com/graphics/2020/opinions/systemic-racism-police-evidence-criminal-justice-system/#Policing.

[5] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. University of Pennsylvania, Department of Criminology, Working Paper No. 2017-1.0, May 25, 2017. https://crim.sas.upenn.edu/sites/default/files/2017-1.0-Berk_FairnessCrimJustRisk.pdf.

[6] Chiappa, S., and Isaac, W.S. A Causal Bayesian Networks Viewpoint on Fairness. arXiv:1907.06430v1, July 15, 2019.

[7] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:1703.00056v1, February 28, 2017.

[8] Citron, D.K. Technological Due Process. Washington University Law Review 85, 6 (2008), 1249–1313.

[9] COMPAS Risk Assessment. www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html.

[10] Crawford, K., and Schultz, J. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. Boston College Law Review 55, 1 (2014), 93–128.

[11] Dieterich, W., Mendoza, C., and Brennan, T. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe, July 8, 2016. www.documentcloud.org/ documents/2998391-ProPublica-Commentary-Final-070616.html.

[12] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness Through Awareness. arXiv:1104.3913v2, November 29, 2011.

[13] Equivant. Practitioner's Guide to COMPAS Core. April 4, 2019. www.equivant.com/practitioners-guide-to-compas-core.

[14] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. arXiv:1412.3756v3, July 16, 2015.

[15] Fish, B., Kun, J., and Lelkes, Á.D. A Confidence-Based Approach for Balancing Fairness and Accuracy. arXiv:1601.05764v1, January 21, 2016.

[16] Friedler, S.A., Scheidegger, C., and Venkatasubramanian, S. On the (im)possibility of fairness. arXiv:1609.07236v1, September 23, 2016.

[17] Fullinwider, R. Affirmative Action. The Stanford Encyclopedia of Philosophy (Summer 2018 Edition), ed. E.N. Zalta. https://plato.stanford.edu/archives/sum2018/entries/affirmative-action.

[18] Galhotra, S., Brun, Y., and Meliou, A. Fairness Testing: Testing Software for Discrimination. Proceedings of 11th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, Paderborn, Germany, September 4–8, 2017 (ESEC/FSE '17), 498−510. https://doi.org/10.1145/3106237.3106277.

[19] Glymour, B., and Herington, J. Measuring the Biases that Matter. The Ethical and Casual Foundations for Measures of Fairness in Algorithms. Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta (GA), USA, January 29–31, 2019 (FAT* '19), 269–278. https://doi.org/10.1145/3287560.3287573.

[20] Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. arXiv:1610.02413v1, October 7, 2016.

[21] Hutchinson, B., and Mitchell, M. 50 Years of Test (Un)fairness: Lessons for Machine Learning. Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta (GA), USA, January 29–31, 2019 (FAT* '19), 49–58. https://doi.org/10.1145/3287560.3287600.

[22] Joh, E.E. Feeding the Machine: Policing, Crime Data, & Algorithms. William & Mary Bill of Rights Journal 26, 2 (2017), 287–302.

[23] Kasirzadeh, A., and Smart, A. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21). ACM, New York, NY, USA, 228–236. DOI: https://doi.org/10.1145/3442188.3445886.

[24] Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding Discrimination through Causal Reasoning. arXiv:1706.02744v2, January 21, 2018.

[25] Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:1609.05807v2, November 17, 2016.

[26] Kohler-Hausmann, I. Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination. Northwestern University Law Review 113, 5 (2019), 1163–1227.

[27] Kusner, M., Loftus, J., Russell, C., and Silva, R.: Counterfactual Fairness. arXiv:1703.06856v3, March 8, 2018.

[28] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica, May 23, 2016. www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[29] Lewis, D. Causation. The Journal of Philosophy 70, 17 (1973), 556−567.

[30] Makhlouf, K., Zhioua, S., and Palamidessi, C. On the Applicability of Machine Learning Fairness Notions. SIGKDD Explorations 23(1), ACM, 2021.

[31] Northpointe. Practitioners Guide to COMPAS. August 17, 2012. www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf.

[32] Pearl, J. Causality. Models, Reasoning, and Inference. 2nd ed., New York: Cambridge University Press 2009.

[33] State v. Loomis. 881 N.W.2d 749 (2016) 749 (Wis. 2016).

[34] Verma, S., and Rubin, J. Fairness Definitions Explained. Proceedings of the International Workshop on Software Fairness, Gothenburg, Sweden, May 29, 2018 (FairWare '18), 1–7. https://doi.org/10.1145/3194770.3194776.

[35] Zafar, M.B., Valera, I., Gomez Rodriguez, M., and Gummadi, K.P. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. arXiv:1610.08452v2, May 8, 2017.

About the author:

Dietmar Hübner is Professor of Practical Philosophy, particularly Ethics of Science at Leibniz University Hannover. His main research is in general ethics, applied ethics, and political philosophy. He holds a diploma (University of Bonn) and an M.Phil. (University of Cambridge) in physics. He earned his Ph.D. in philosophy with a dissertation on decision theory and philosophy of history, and completed his habilitation in philosophy with a book on metaphorical accounts in distributive justice (both University of Bonn). Dietmar Hübner is principal investigator in the interdisciplinary research project *Bias and Discrimination in Big Data and Algorithmic Processing – BIAS* (www.bias-project.org), funded by Volkswagen Foundation.

On the Applicability of Machine Learning Fairness Notions

Karima Makhlouf
Université du Québec à Montréal
Montréal, Canada
makhlouf.karima@courrier.uqam.ca

Sami Zhioua
Higher Colleges of Technology
Dubai, UAE
szhioua@hct.ac.ae

Catuscia Palamidessi
INRIA, École Polytechnique, IPP
Paris, France
catuscia@lix.polytechnique.fr

ABSTRACT

Machine Learning (ML) based predictive systems are increasingly used to support decisions with a critical impact on individuals' lives such as college admission, job hiring, child custody, criminal risk assessment, etc. As a result, fairness emerged as an important requirement to guarantee that ML predictive systems do not discriminate against specific individuals or entire sub-populations, in particular, minorities. Given the inherent subjectivity of viewing the concept of fairness, several notions of fairness have been introduced in the literature. This paper is a survey of fairness notions that, unlike other surveys in the literature, addresses the question of "which notion of fairness is most suited to a given real-world scenario and why?". Our attempt to answer this question consists in (1) identifying the set of fairness-related characteristics of the real-world scenario at hand, (2) analyzing the behavior of each fairness notion, and then (3) fitting these two elements to recommend the most suitable fairness notion in every specific setup. The results are summarized in a decision diagram that can be used by practitioners and policy makers to navigate the relatively large catalogue of ML fairness notions.

1. INTRODUCTION

ML-based decision-making (MLDM)[1] is beneficial as it allows to take into consideration orders of magnitude more factors than humans do and hence outputting decisions that are more informed and less subjective. However, in its quest to maximize efficiency, ML algorithms can systemize discrimination against a specific group of population, typically, minorities. As an example, consider the automated candidates selection system of St. George Hospital Medical School [32; 36]. The aim of the system was to help screening for the most promising candidates for medical studies. The automated system was built using records of manual screenings from previous years. During those manual screening years, applications with grammatical mistakes and misspellings were rejected by human evaluators as they indicate a poor level of english. As non-native english speakers are more likely to send applications with grammatical and misspelling mistakes than native english speakers do, the automated screening system built on that historical data ended up correlating race, birthplace, and address with a lower likelihood of acceptance. Later, while the overall english level of non-native speakers improved, the race and ethnicity bias persisted in the system to the extent that an excellent candidate may be rejected simply for her birthplace or address.

In the context of automated decision-making, a consensual definition of fairness can be formulated as: "*absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits*" [34]. Mathematically, however, there is no consensual definition of fairness. Very often, research papers focus on a specific real-world scenario of automated decision system and propose a fairness definition tailored to that scenario and its specificities. Consequently, several fairness notions have been introduced in the literature. These notions are the subject of several survey papers [2; 6; 19; 34; 35; 41; 44].

The very reason of having different flavors of fairness notions is how suitable each one of them is for specific real-world scenarios. But none of the existing surveys addressed this aspect. Discussion about the suitability (and sometimes the applicability) of the fairness notions is very limited and scattered through several papers [3; 10; 29; 35; 43]. In this survey paper we show that each ML-based automated decision system can be different based on a set of criteria such as: whether the ground-truth exists, difference in base-rates between sub-groups, the cost of misclassification, the existence of a government regulation that needs to be enforced, etc. We then revisit exhaustively the list of fairness notions and discuss the suitability and applicability of each one of them based on the list of criteria.

The results of this survey are finally summarized in a decision diagram that hopefully can help researchers, practitioners, and policy makers to identify the subtleties of the ML-based automated decision system at hand and to choose the most appropriate fairness notion to use, or at least rule out notions that can lead to wrong fairness/discrimination result.

2. REAL-WORLD SCENARIOS WITH CRITICAL FAIRNESS REQUIREMENTS

As the paper is focusing on the applicability of fairness notions, we provide here a list of notable real-world MLDMs where fairness is critical. In each of these scenarios, failure to address the fairness requirement will lead to unacceptably biased decisions against individuals and/or sub-populations. These scenarios will be used to provide concrete examples of situations where certain fairness notions are more suitable

[1]We focus on automated decision-making system supported by ML algorithms. In the rest of the paper we refer to such systems as MLDM.

than others.

Job hiring: MLDMs in hiring are increasingly used by employers to automatically screen candidates for job openings.Typically, the input data used by the MLDM include: affiliation, education level, job experience, IQ score, age, gender, address, etc. The MLDM outputs a decision and/or a score indicating how promising the application is for the job opening. A biased MLDM leads to rejecting a candidate because of a trait that she cannot control (gender, race, sexual orientation, etc.). This can be damaging for the employer as excellent candidates might be missed.

Granting loans: Since decades, statistical and MLDM systems are used to assess loan applications and determine which of them are approved and with which repayment plan and annual percentage rate (APR). The assessment proceeds by predicting the risk that the applicant will default on her repayment plan. Loan Granting MLDMs currently in use include: FICO, Equifax, Lenddo, Experian, TransUnion, etc. The common input data used for loan granting include: credit history, purpose of the loan, loan amount requested, employment status, income, marital status, gender, age, address, housing status and credit score. An unfair loan granting MLDM will either deny a deserving applicant a requested loan, or give her an exorbitant APR, which on the long run will create a vicious cycle as the candidate will be very likely to default on her payments.

College admission: Given the large number of admission applications, several colleges are now resorting to MLDMs to reduce processing time and cut costs[2]. Typically, the candidates' features used include: the institutions previously attended, SAT scores, extra-curricular activities, GPAs, test scores, interview score, etc. The predicted outcome can be a simple decision (admit/reject) or a score indicating the candidate's potential performance in the requested field of study [18]. Unfair college admission MLDMs may discriminate against a certain ethnic group (e.g. African-American [38]) which could lead, in the long term, to economic inequalities and corrupting the role of higher education in society as a whole.

Criminal risk assessment: There is an increasing adoption of MLDMs that predict risk scores based on historical data with the objective to guide human judges in their decisions. The most common use case is to predict whether a defendant will re-offend (or recidivate). Predicting risk and recidivism requires input information such as: number of arrests, type of crime, address, employment status, marital status, income, age, housing status, etc. Unfair risk assessment MLDMs, as revealed by the highly publicized 2016 proPublica article [1], may result in biased treatment of individuals based solely on their race. In extreme cases, it may lead to wrongful imprisonments for innocent people, contributing to the cycle of violation and crime.

Child maltreatment prediction: The objective of the MLDM in child maltreatment prediction is to estimate the likelihood of substantiated maltreatment (neglect, physical abuse, sexual abuse, or emotional maltreatment) among children. The system generates risk scores, which would then trigger a targeted early intervention in order to prevent children maltreatment. The features considered in this type of MLDM include both contemporaneous and historical information for children and caregivers. An unfair MLDM may use a proxy variable to predict decisions based on the community rather than which child get harmed. For example, a major cause of unfairness in AFST is the rate of referral calls; the community calls the child abuse hotline to report non-white families at a much higher rate than it does to report white families [17].

Health care: Since decades, ML algorithms are able to process anonymized electronic health records and flag potential emergencies, to which clinicians are invited to respond promptly. Examples of features that might be used in disease (chronic conditions) prediction include vital signs, blood test, socio-demographics, education, health insurance, home ownership, age, race, address. The outcome of the MLDM is typically an estimated likelihood of getting a disease. A biased disease prediction MLDM can misclassify individuals in certain sub-populations in a disproportionally higher rate than the dominant population. For instance, diabetic patients have known differences in associated complications across ethnicities [40].

Facial analysis: Automated facial analysis systems are used to identify perpetrators from security video footages, to detect melanoma (skin cancer) from face images [16], to detect emotions [12], and to even determine individual's characteristics such as IQ, propensity towards terrorist crime, etc. based on their face images [42]. A flawed MLDM may lead to biased outcomes such as wrongfully accusing individuals from specific ethnic groups (e.g. asians, dark skin populations) for crimes (based on security video footages) at a much higher rate than the rest of the population. For instance, African-Americans have been reported to be more likely to be stopped and investigated by law enforcement due to a flawed face recognition system [22].

Others: Other MLDMs with fairness concerns include: insurance policy prediction [39], income prediction [34], teachers evaluation and promotion [7], online recommendation [25] and university ranking [33; 36].

3. FAIRNESS NOTION SELECTION CRITERIA

In order to systemize the procedure for selecting the most suitable fairness notion for a specific MLDM system, we identify a set of criteria that can be used as as roadmap. For each criterion, we check whether it holds in the problem at hand or not. Telling whether a criterion is satisfied or not does not typically require an expertise in the problem domain. We note here that in some cases, these criteria can, not only indicate if a fairness notion is suitable, but whether it is "acceptable" to use in the first place. We tried to be exhaustive when listing the decision criteria based on the existing literature. However, there are no guarantees about the completeness of this list.

Ground truth availability: A ground truth value is the true and correct *observed* outcome corresponding to given sample in the data. It should be distinguished from an *inferred* subjective outcome in historical data which is decided by a human. An example of a scenario where ground truth is available is when predicting whether an individual has a disease. The ground truth value is observed by submitting

[2]While the final acceptance decision is taken by humans, MLDMs are typically used as a first filter to "clean-up" the list from clear rejection cases.

the individual to a blood test[3] for example. An example of a scenario where ground truth is not available is predicting whether a job applicant is hired. The outcome in the training data is inferred by a human decision maker which is often a subjective decision, no matter how hard she is trying to be objective.

Base rate is the same across groups: The base rate is the proportion of positive outcome in a population (Based on Table 2, $BR = \frac{TP+FN}{TP+FP+TN+FN}$). This rate can be the same or differs across sub-populations. For example, the base rates for diabetes disease occurrence for men and women is typically the same. But, for another disease such as prostate cancer, the base rates are different between men and women[4].

(Un)reliable outcome: In scenarios where ground truth is not available, the outcome (label) in the data is typically inferred by humans. The outcome in the training data in that case can or cannot be reliable as it can encode human bias. The reliability of the outcome depends on the data collection procedure and how rigorous the data has been checked. Scenarios such as job hiring and college admission may be more prone to the unreliable outcome problem than recommender system for example. A "one-size-fit-all" MLDM model in disease prediction that does not take into consideration the ethnic group of the individual may result in unreliable outcome as well.

Presence of explaining variables: An explaining variable[5] is correlated with the sensitive attribute (e.g. race) in a legitimate way. Any discrimination that can be explained using that variable is considered legitimate and is acceptable. For instance, if all the discrepancy between male and female job hiring rate is explained by their education levels, the discrimination can be deemed legitimate and acceptable.

Emphasis on precision vs recall: Precision (the complement of target population error [13]) is defined as the fraction of positive instances among the predicted positive instances ($\frac{TP}{TP+FP}$). In other words, if the system predicts an instance as positive, how precise that prediction is. Recall (the complement of model error [13]) is defined as the fraction of the total number of positive instances that are correctly predicted positive ($\frac{TP}{TP+FN}$). In other words, how many of the positive instances the system is able to identify. There is always a tradeoff between precision and recall (increasing one will lead, very often, to decreasing the other). Depending on the scenario at hand, the fairness of the MLDM may be more sensitive to one on the expense of the other. For example, granting loans to the maximum number of deserving applicants contribute more to fairness than making sure that an applicant who has been granted a loan really deserves it. When firing employees, however, the opposite is true: fairness is more sensitive to wrongly firing an employee, rather than, firing the maximum number of under-performing employees.

Emphasis on false positive vs false negative: Fairness can be more sensitive to false positive misclassification (type I error) rather than false negative misclassification (type II error), or the opposite. For example, in criminal risk assessment scenario, it is commonly accepted that incarcerating an innocent person (false positive) is more serious than letting a guilty person escape (false negative).

Cost of misclassification: Depending on the scenario at hand, the cost of misclassification can be significant (e.g. incarcerating an individual, firing an employee, rejecting a college application, etc.) or mild and without consequential impact (e.g. useless product recommendation, misleading income prediction, offensive online translation, abusive results in online autocomplete, etc.)

Prediction threshold is fixed or floating: Decisions in MLDM are typically made based on predicted real-valued score. In the case of binary outcome, the score is turned into a binary value such as $\{0, 1\}$ by thresholding[6]. In some scenarios, it is desirable to interpret the real-value score as probability of being accepted (predicted positive). The threshold used as a cutoff point where positive decisions are demarcated from negative decisions can be fixed or floating. A fixed threshold is set carefully and tends to be valid for different datasets and use cases. For instance, in recidivism risk assessment, high risk threshold is typically fixed. A floating threshold can be selected and fine-tuned arbitrarily by practitioners to accommodate a changing context. Acceptance score in loan granting scenarios is an example of a floating threshold as it can move up or down depending on the economic context.

Likelihood of intersectionality: Intersectionality theory [11] focuses on a specific type of bias due to the combination of sensitive factors. An individual might not be discriminated based on race only or based on gender only, but she might be discriminated because of a combination of both. Black women are particularly prone to this type of discrimination.

Likelihood of masking: Masking is a form of intentional discrimination that allows decision makers with prejudicial views to mask their intentions [4]. Masking is typically achieved by exploiting how fairness notions are defined. For example, if the fairness notion requires equal number of candidates to be accepted from two ethnic groups, the MLDM can be designed to carefully select candidates from the first group (satisfying strict requirements) while selecting randomly from the second group just to "make the numbers".

The existence of regulations and standards: In some domains, laws and regulations might be imposed to avoid discrimination and bias. For instance, guidelines from the *U.S. Equal Employment Opportunity Commission* state that a difference of the probability of acceptance between two subpopulations exceeding 20% is illegal [3]. Another example might be an internal organizational policy imposing diversity among its employees.

4. FAIRNESS NOTIONS

Let V, A, and X be three random variables representing, respectively, the total set of attributes, the sensitive attributes, and the remaining attributes describing an individual such that $V = (X, A)$ and $P(V = v_i)$ represents the probability of drawing an individual with a vector of values v_i from the population. For simplicity, we focus on the case where A is a binary random variable where $A = 0$ designates the protected group, while $A = 1$ designates the non-protected group. Let Y and $\hat{Y}$ be binary random variables representing, respectively, the actual outcome and the predicted outcome where $Y = 1$ designates a positive instance, while $Y = 0$ a negative one. Typically, the predicted outcome $\hat{Y}$ is derived from a score represented by a random variable S where $P(S = s)$ is

[3]assuming the blood test is flawless.

[4]While male prostate cancer is the second most common cancer in men, female prostate cancer is rare [14].

[5]Referred also as resolving variable.

[6]The threshold is defined by the decision makers depending on the context of interest.

the probability that the score value is equal to s.

All fairness notions presented in this survey (Table 1) address the following question: "is the outcome/prediction of the MLDM fair towards individuals?". So fairness notion is defined as a mathematical condition that must involve either $\hat{Y}$ or S along with the other random variables. As such, we are not concerned by the inner-workings of the MLDM and their fairness implications. What matters is only the score/prediction value and how fair/biased is it.

A simple and straightforward approach to address fairness problem is to ignore completely any sensitive attribute while training the MLDM system. This is called *fairness through unawareness*[7]. We don't treat this approach as fairness notion since, given MLDM prediction, it does not allow to tell if the MLDM is fair or not. Besides, it suffers from the basic problem of proxies. Many attributes (e.g. home address, neighborhood, attended college) might be highly correlated to the sensitive attributes (e.g. race) and act as proxies of these attributes. Consequently, in almost all situations, removing the sensitive attribute during the training process does not address the problem of fairness.

Statistical parity[15]: is one of the most commonly accepted notions of fairness. It requires the prediction to be statistically independent of the sensitive attribute ($\hat{Y} \perp A$). In other words, the predicted acceptance rates for both protected and unprotected groups should be equal. Using the confusion matrix (Table 2), statistical parity implies that $\frac{TP+FP}{TP+FP+FN+TN}$ is equal for both groups. Statistical parity is appealing in scenarios where there is a preferred decision over the other. For example, being accepted to a job, not being arrested, being admitted to a college, etc.[8]. Statistical parity is also well adapted to contexts in which some regulations or standards are imposed. For example, a law might impose to equally hire or admit applicants from different sub-populations. The main problem of statistical parity is that it doesn't consider a potential correlation between the label Y and the sensitive attribute A. In other words, if the underlying base rates of the protected and unprotected groups are different, statistical parity will be misleading. In the ideal case ($\hat{y} = y$), this will lead to loss of utility [24]. Another issue with this notion is its "laziness"; if we hire carefully selected applicants from male group and random applicants from female group, we can still achieve statistical parity, yet leading to negative results for the female group as its performance will tend to be worse than that of male group. This practice is an example of *self-fulfilling prophecy* [15] where a decision maker may simply select random members of a protected group rather than qualified ones, and hence, intentionally building a bad track record for that group. Barocas and Selbst refer to this problem as masking [4]. Masking is possible to game several fairness notions, but it is particularly easy to carry out in the case of statistical parity.

Conditional statistical parity [10]: this notion is a variant of statistical parity obtained by controlling on a set of legitimate attributes[9]. The legitimate attributes (we refer to them as E) among X are correlated with the sensitive attribute A and give some factual information about the label at the same time leading to a *legitimate* discrimination. In other words, this notion removes the illegal discrimination, allowing the disparity in decisions to be present as long as they are explainable [10]. In practice, conditional statistical parity is suitable when there is one or several attributes that justify a possible disparate treatment between different groups in the population. More seriously, conditional statistical parity gives a decision maker a tool to game the system and realize a self-fullfilling prophecy. Therefore, it is recommended to resort to domain experts or law officers to decide what is unfair and what is tolerable to use as legitimate discrimination attribute [26].

Equalized odds [23]: this notion considers both the predicted and the actual outcomes. Thus, the prediction is conditionally independent from the protected attribute, given the actual outcome ($\hat{Y} \perp A \mid Y$). In other words, equalized odds requires that both sub-populations to have the same $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$ (Table 2). By contrast to statistical parity, equalized odds is well-suited for scenarios where the ground truth exists such as: disease prediction or stop-and-frisk [5]. It is also suitable when the emphasis is on recall (the fraction of the total number of positive instances that are correctly predicted positive) rather than precision (making sure that a predicted positive instance is actually a positive instance). A potential problem of equalized odds is that it may not help closing the gap between the protected and unprotected groups. Because equalized odds requirement is rarely satisfied in practice, two variants can be obtained by relaxing its equation (Table 1). The first one is called **equal opportunity** [23] and is obtained by requiring only TPR equality among groups. As TPR does not take into consideration FP, equal opportunity is completely insensitive to the number of false positives. This is an important criterion when considering this fairness notion in practice. More precisely, in scenarios where a disproportionate number of false positives among groups has fairness implications, equal opportunity should not be considered. The second relaxed variant of equalized odds is called **predictive equality** [10] which requires only the FPR to be equal in both groups. Since FPR is independent from FN, predictive equality is completely insensitive to false negatives. Predictive equality is particularly suitable to measure the fairness of face recognition systems in crime investigation where security camera footages are analyzed. Fairness between ethnic groups with distinctive face features is very sensitive to the FPR. A false positive means an innocent person is being flagged as participating in a crime. If this false identification happens at a much higher rate for a specific sub-population (e.g. dark skinned ethnic group) compared to the rest of the population, it is clearly unfair for individuals belonging to that sub-population. Looking to the problem from another perspective, choosing between equal opportunity and predictive equality depends on how the outcome/label is defined. In scenarios where the positive outcome is desirable (e.g. hiring, admission), typically fairness is more sensitive to false negatives rather than false positives, and hence equal opportunity is more suitable. In scenarios where the positive outcome is undesirable for the subjects (e.g. firing, risk assessment), typically fairness is more sensitive to false positives rather than false negatives, and hence predictive equality is more suitable.

Conditional use accuracy equality [6]: with this notion,

[7]Known also as: blindness, unawareness [35], anti-classification [9], and treatment parity [31].

[8]This might not be the case in other scenarios such as disease prediction, child maltreatment, where imposing a parity of positive predictions is meaningless.

[9]Called explanatory attributes in [26].

Table 1: Classification of fairness notions.

Fairness Notion	Reference	Formulation	Classification	Type
Statistical Parity	[15]	$P(\hat{Y} \mid A=0) = P(\hat{Y} \mid A=1)$	Independence	Group Fairness
Conditional Statistical Parity	[10]	$P(\hat{Y}=1 \mid E=e, A=0) = P(\hat{Y}=1 \mid E=e, A=1) \quad \forall e$	$\hat{Y} \perp A$	
Equalized Odds	[23]	$P(\hat{Y}=1 \mid Y=y,\ A=0) = P(\hat{Y}=1 \mid Y=y,\ A=1) \quad \forall y \in \{0,1\}$	Separation	
Equal Opportunity		$P(\hat{Y}=1 \mid Y=1, A=0) = P(\hat{Y}=1 \mid Y=1, A=1)$	$\hat{Y} \perp A \mid Y$	
Predictive Equality	[10]	$P(\hat{Y}=1 \mid Y=0, A=0) = P(\hat{Y}=1 \mid Y=0, A=1)$		
Balance for Positive Class	[29]	$E[S \mid Y=1, A=0)] = E[S \mid Y=1, A=1]$		
Balance for Negative Class		$E[S \mid Y=0, A=0] = E[S \mid Y=0, A=1]$		
Conditional Use Accuracy Equality	[6]	$P(Y=y \mid \hat{Y}=y, A=0) = P(Y=y \mid \hat{Y}=y, A=1) \quad \forall y \in \{0,1\}$	Sufficiency	
Predictive Parity	[8]	$P(Y=1 \mid \hat{Y}=1, A=0) = P(Y=1 \mid \hat{Y}=1, A=1)$	$Y \perp A \mid \hat{Y}$	
Calibration		$P(Y=1 \mid S=s, A=0) = P(Y=1 \mid S=s, A=1) \quad \forall s \in [0,1]$		
Well-calibration	[29]	$P(Y=1 \mid S=s, A=0) = P(Y=1 \mid S=s, A=1) = s \quad \forall\, s \in [0,1]$		
Overall Accuracy Equality	[6]	$P(\hat{Y}=Y \mid A=0) = P(\hat{Y}=Y \mid A=1)$	Other metrics from confusion matrix	
Treatment Equality		$\frac{FN}{FP}_{(a=0)} = \frac{FN}{FP}_{(a=1)}$		
Total Fairness		-	Independence, Separation and Sufficiency	
No unresolved discrimination	[27]	-	Causality	
No proxy discrimination		$P(\hat{Y} \mid do(P_x = p)) = P(\hat{Y} \mid do(P_x = p\prime)) \quad \forall P_x \quad and \quad \forall\, p, p\prime$		
Counterfactual Fairness	[30]	$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X=x, A=a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X=x, A=a)$		Individual Fairness
Causal Discrimination	[20]	$X_{(a=0)} = X_{(a=1)} \ \wedge\ A_{(a=0)} \neq A_{(a=1)} \Rightarrow \hat{y}_{(a=0)} = \hat{y}_{(a=1)}$	Similarity Metric	
Fairness Through Awareness	[15]	$D(M(v_i), M(v_j)) \leq d(v_i, v_j)$		

Table 2: Confusion matrix

	Actual Positive $Y = 1$	Actual Negative $Y = 0$
Predicted Positive $\hat{Y} = 1$	**TP** (True Positive)	**FP** (False Positive) *Type I error*
Predicted Negative $\hat{Y} = 0$	**FN** (False Negative) *Type II error*	**TN** (True Negative)

fairness is achieved when all population groups have equal $PPV = \frac{TP}{TP+FP}$ and $NPV = \frac{TN}{FN+TN}$. In other words, the probability of subjects with positive predictive value to truly belong to the positive class and the probability of subjects with negative predictive value to truly belong to the negative class should be the same. By contrast to equalized odds, one is conditioning on the algorithm's predicted outcome not the actual outcome. In other words, the emphasis is on the precision of the MLDM system rather than its recall. **Predictive parity** [8] is a relaxation of conditional use accuracy equality requiring only equal PPV among groups. Like predictive equality, predictive parity is insensitive to false negatives. Hence in any scenario where fairness is sensitive to false negatives, predictive parity should not be used. Choosing between predictive parity and equal opportunity depends on whether the scenario at hand is more sensitive to precision or recall. For precision-sensitive scenarios, typically predictive parity is more suitable while for recall-sensitive scenarios, equal opportunity is more suitable. Precision-sensitive scenarios include disease prediction, child maltreatment risk assessment, and firing from jobs. Recall-sensitive scenarios include loan granting, recommendation systems, and hiring. Very often, precision-sensitive scenarios coincide with situations where the positive prediction ($\hat{Y} = 1$) entails a higher cost [43]. For example, a predicted child maltreatment case will result in placing the child in a foster house which will generally entail a higher cost compared to a negative prediction (low risk of child maltreatment) in which case the child stays with the family and typically no action is taken.

Balance [29] : The predicted outcome ($\hat{Y}$) is typically derived from a score (S) which is returned by the ML algorithm. All aforementioned fairness notions do not use the score to assess fairness. **Balance for positive class** focuses on the individuals who constitute positive instances and is satisfied if the average score S received by those individuals is the same for both groups. The intuition behind this notion is that a balance for the positive class should be assured, thus, a violation of this balance means that individuals belonging to the positive class in one group might receive steadily lower predicted score than individuals belonging to the positive class in the other group. **Balance of negative class** is an analogous fairness notion where the focus is on the negative class. Both variants of balance can be required simultaneously which leads to a stronger notion of balance[10]. Balance fairness notions are relevant in the criminal risk assessment scenario because a divergence in the score values of individuals from different races may indicate a difference in the type of crime that can be committed (high risk score typically means a serious crime).

Calibration [8]: To satisfy calibration, for each predicted probability score $S = s$, individuals in all groups should have the same probability to actually belong to the positive class. Interestingly, calibration is not always stronger than predictive parity [21]. Calibration is suitable to use in scenarios where the threshold is not fixed and is very likely to be tuned to accommodate a changing context. A first example is the acceptance score in loan granting applications which may change abruptly due to economic instability. A second example is the child maltreatment risk assessment where the threshold for intervention (withdrawing a child from his family) depends on the available seats in foster houses.

Well-calibration [29] is a stronger variant of calibration. It requires that (1) calibration is satisfied, (2) the score is interpreted as the probability to truly belong to the positive class, and (3) for each score $S = s$, the probability to truly belong to the positive class is equal to that particular score.

No unresolved discrimination [27]: similarly to counterfactual fairness, no unresolved discrimination is assessed using causal reasoning. Given a causal graph, no unresolved discrimination is satisfied when no directed path from the sensitive attribute A to the predictor $\hat{Y}$ are allowed, except via a resolving variable. A resolving variable is any variable in a causal graph that is influenced by the sensitive attribute in a manner that it is accepted as nondiscriminatory (this is very similar to the use of the explanatory attributes in conditional statistical parity but in a non-causal context). Compared to counterfactual fairness, no unresolved discrimination is a weaker notion. That is, a counterfactually unfair scenario may be identified as fair based on no unresolved discrimination. This can happen in case one or several variables in the causal graph are identified as resolving. The application of unresolved discrimination is completely based on the definition of the causal graph. Thus, this notion is well-suited when a reliable and trustworthy causal graph that describes best the domain at hand including all relevant relations and features (in particular, the resolving attributes) is available. Hence, it is mandatory that the choice of the resolving variables along with their causal relationships to the other attributes is in reliance on policy makers and domain professionals expertise.

No proxy discrimination [27]: a causal graph exhibits potential proxy discrimination if there exists a path from the protected attribute A to the predicted outcome $\hat{Y}$ that is blocked by a proxy variable P_x. A proxy is merely a descendant of A that is chosen to be labelled as a proxy because it is significantly correlated with A. Given a causal graph, a predictor $\hat{Y}$ exhibits no proxy discrimination if the equality of the following equation is valid for all potential proxies P_x:

$$P(\hat{Y} \mid do(P_x = p)) = P(\hat{Y} \mid do(P_x = p\prime)) \quad \forall\, p, p\prime$$

In other words, this notion implies that changing the value of P_x should not have any impact on the prediction. As with the previous two fairness notions, the applicability of proxy discrimination is based on the construction of a reliable and plausible causal graph. In particular, the main goal of this notion is to carefully investigate and analyze the relations between attributes (in particular, those related to the sensitive attributes) in order to discover all potential proxies that might result in unfair decisions.

Counterfactual fairness [30] : counterfactual is a con-

[10]No previous work reported such fairness notion.

cept from causal inference which goes beyond mere statistical correlation between variables and relies on causal relationship between them. Causal relationships are represented using causal graph where nodes represent variables (attributes) and edges represent causal relationships between variables. U represents all *exogenous* variables such that each assignment $U = u$ corresponds to a unique individual in the population or to a situation in nature [37]. Counterfactual fairness is achieved if for every individual $(U = u, X = x, A = a)$ of the entire population, the probability to be predicted as hired is the same, *had A been* a'. That is, $P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a)$ is equal to $P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)$ where $\hat{Y}_{A\leftarrow a'}(U)$ is called the counterfactual and corresponds to the predicted outcome in case the variable A is *forced*[11] to be equal to a' for an individual with exogenous variable $U = u$. The probability of the counterfactual is conditioned on $(X = x, A = a)$ which is called the evidence. In other words, counterfactual fairness requires that for every individual $(X = x, A = a)$ in the population (evidence), the probability of the outcome is the same in both the actual world $(A \leftarrow a)$ and the counterfactual world $(A \leftarrow a')$. Compared to causal discrimination where all variables are measured in the same world but on different individuals, counterfactual fairness measures variables on the same individual but in different worlds (the world of the evidence, and another hypothetical world). This notion is satisfied if the probability distribution of the predicted outcome $\hat{Y}$ is the same in the actual and counterfactual worlds, for every possible individual. A simple but important implication of counterfactual fairness formulation is that, given a causal graph, a predictor $\hat{Y}$ is counterfactually fair if it is a function of non-descendants of the sensitive variable A. Consequently, one can tell if a predictor is counterfactually fair by simply checking the causal graph[12]. Hence, the main challenge to using counterfactual fairness in practice is the construction of the causal graph which typically requires domain expertise. It is important to note that generally, data can be used to validate a proposed causal graph. That is, a dataset of observed samples can be used to rule out possible causal graphs.

Causal discrimination [20]: this notion implies that a classifier should produce exactly the same prediction for individuals who differ only in their sensitive attribute A while possessing identical attributes X. At a first glance, causal discrimination can be seen as an extreme case of conditional statistical parity when conditioning on all non-sensitive attributes $(E = X)$. However, conditional statistical parity is a group fairness notion which is satisfied if the proportion of individuals having the same non-sensitive attribute values and predicted accepted in both groups (e.g. male and female) is the same. This is why the mathematical formulation of conditional statistical parity (Table 1) is expressed in terms of conditional probabilities. Causal discrimination, however, considers every individual separately regardless of its contribution to sub-population proportions. Causal discrimination is suitable to use in decision making scenarios where it is very common to find individuals sharing exactly the same attribute values. For example, admission decision making based mainly on test scores and categorical attributes. The result of applying causal discrimination is the percentage of violations in the entire population (i.e. how many individuals are unfairly treated).

Fairness through awareness [15]: this notion is a generalization of causal discrimination which implies that similar individuals should have similar predictions. Let i and j be two individuals represented by their attributes values vectors v_i and v_j. Let $d(v_i, v_j)$ represent the similarity distance between individuals i and j. Let $M(v_i)$ represent the probability distribution over the outcomes of the prediction. For example, if the outcome is binary (0 or 1), $M(v_i)$ might be $[0.2, 0.8]$ which means that for individual i, $P(\hat{Y} = 0) = 0.2$ and $P(\hat{Y} = 1) = 0.8$. Let D be a distance metric between probability distributions. Fairness through awareness is achieved iff, for any pair of individuals i and j:

$$D(M(v_i), M(v_j)) \leq d(v_i, v_j)$$

In practice, fairness through awareness assumes that the similarity metric is known for each pair of individuals [28]. That is, a challenging aspect of this approach is the difficulty to determine what is an appropriate metric function to measure the similarity between two individuals. Typically, this requires careful human intervention from professionals with domain expertise [30].

5. DIAGRAM AND DISCUSSION

With the large number of fairness notions and the subtle resemblance between MLDM scenarios, deciding about which fairness notion to use is not a trivial task. More importantly, selecting and using a fairness notion in a scenario inappropriately may detect unfairness in an otherwise fair scenario, or the opposite, i.e., fail to identify unfairness in an unfair scenario.

One of the objectives of this survey is to systemize the selection procedure of fairness notions. This is achieved by identifying a set of fairness-related characteristics (Section 3) of the scenario at hand and then use them to recommend the most suitable fairness notion for that specific scenario. The proposed systemized selection procedure is illustrated in the decision diagram of Figure 1. The diagram is called "decision diagram" and not "decision tree" for the following reason. In typical decision trees, every leaf corresponds to a single decision, which is a fairness notion that *should* be used. However, the diagram in Figure 1 is designed such that every node indicates which notions are recommended, which notions should be avoided, and which notions must not be used.

The diagram is composed of four types of nodes:

- **Decision node (diamond):** based on fairness-related characteristics (Section 3)
- **Recommended node (rectangle):** a leaf node indicating that the fairness notion is suitable to be used given all fairness-related characteristics in the path to that node.
- **Warning node (triangle):** indicates that the fairness notion(s) is/are not recommended in all the branch in the right of the node. This node can appear in the middle of the edge between two decision nodes.
- **Must-not node (circle):** the fairness notion must not be used.

[11] Pearl et al. [37] use the term *surgical modification*.

[12] Kusner et al. [30] identify some exceptions, but guaranteeing that they will *not happen in general*.

To illustrate how the diagram should be interpreted, consider the recommended node predictive parity (34). According to the diagram, predictive parity is recommended in the scenario where intersectionality and/or masking are unlikely (decision node 1), standards do not exist (decision node 2), ground-truth is available or outcome Y is reliable (decision node 6), fairness is more sensitive to precision rather than recall (decision node 14), the prediction threshold is typically fixed (decision node 20) and the emphasis is on false positives rather than false negatives (decision node 24). In that particular scenario, equal opportunity must not be used (must-not node 42) because fairness in this scenario is particularly sensitive to false positives, while equal opportunity is completely insensitive to false positives. The warning node 9 along the same path indicates that statistical parity is not suitable in this scenario. Finally, any fairness notion for which there is no warning node or must-not node along the path of the scenario can be used in this scenario. For instance, all individual fairness notions can be used, which is indicated by the link to node 4, i.e., the square with a "4" inside at the end of several paths, as will be discussed below.

The lower part of the diagram corresponding to the "yes" branch of decision node 1 deals with individual fairness notions. In that branch all group fairness notions are not recommended (warning node 3) because they are not suitable when intersectionality or masking are likely.The part between decision nodes 7 and 15 is the only part with a non-tree-like structure. It expresses the fact that, typically, several individual fairness notions can be suitable at the same time. This indicates also that currently, the tensions between the various individual fairness notions are not well understood in the literature.

The diagram may be misleading if it is interpreted very categorically. This occurs when a user of the diagram navigates it and ends up using the recommended fairness notion without considering other important elements specific to the scenario at hand. The diagram can be misleading also when it is not clear which branch to take in a decision node. For example, the question in decision node 14 (emphasis on precision or recall?) is difficult to answer categorically in several scenarios. The decision nodes 13, 19, 22, and even 1, are typically easier to navigate, but can be challenging to settle in a number of scenarios. A potential solution would be to label one of the branches as default (to be followed when the answer is not clear), but this can, often result in a suboptimal decision. In summary, the diagram should be considered as guide and should never be used to supersede important elements specific to the scenario at hand.

6. CONCLUSION

With the increasingly large number of fairness notions considered in the relatively new field of fairness in ML, selecting a suitable notion for a given MLDM (machine learning decision making) becomes a non-trivial task. There are two contributing factors. First, the boundaries between the defined notions are increasingly fuzzy. Second, applying inappropriately a fairness notion may report discrimination in an otherwise fair scenario, or vice versa, fail to identify discrimination in an unfair scenario. This survey tries to address this problem by identifying fairness-related characteristics of the scenario at hand and then use them to recommend and/or discourage the use of specific fairness notions. Hence, the survey is an attempt to bridge the gap between the real-world use case scenarios of automated (and generally unintentional) discrimination and the mostly technical tackling of the problem in the literature.

7. AKNOWLEDGMENTS

The work of Catuscia Palamidessi was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme. Grant agreement № 835294.

8. REFERENCES

[1] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. propublica. See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016)

[2] Barocas, S., Hardt, M., Narayanan, A.: Fairness in machine learning. NIPS Tutorial (2017)

[3] Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019), `http://www.fairmlbook.org`

[4] Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. L. Rev. **104**, 671 (2016)

[5] Bellin, J.: The inverse relationship between the constitutionality and effectiveness of new york city stop and frisk. BUL Rev. **94**, 1495 (2014)

[6] Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research p. 0049124118782533 (2018)

[7] Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., Mullainathan, S.: Productivity and selection of human capital with machine learning. American Economic Review **106**(5), 124–27 (2016)

[8] Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)

[9] Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018)

[10] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 797–806 (2017)

[11] Crenshaw, K.: Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stan. L. Rev. **43**, 1241 (1990)

[12] Dehghan, A., Ortiz, E.G., Shu, G., Masood, S.Z.: Dager: Deep age, gender and emotion recognition using convolutional neural network. arXiv:1702.04280 (2017)

[13] Dieterich, W., Mendoza, C., Brennan, T.: Compas risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc (2016)

[14] Dodson, M.K., Cliby, W.A., Keeney, G.L., Peterson, M.F., Podritz, K.C.: Skene's gland adenocarcinoma with increased serum level of prostate-specific antigen. Gynecologic oncology **55**(2), 304–307 (1994)

[15] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)

[16] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)

[17] Eubanks, V.: Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press (2018)

[18] Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236 (2016)

[19] Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 329–338 (2019)

[20] Galhotra, S., Brun, Y., Meliou, A.: Fairness testing: testing software for discrimination. In: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. pp. 498–510 (2017)

[21] Garg, P., Villasenor, J., Foggo, V.: Fairness metrics: A comparative analysis. arXiv preprint arXiv:07864 (2020)

[22] Garvie, C.: The perpetual line-up: Unregulated police face recognition in America. Georgetown Law, Center on Privacy & Technology (2016)

[23] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. arXiv preprint arXiv:02413 (2016)

[24] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems. pp. 3315–3323 (2016)

[25] Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender systems: an introduction. Cambridge University Press (2010)

[26] Kamiran, F., Zliobaite, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and information systems **35**(3), 613–644 (2013)

[27] Kilbertus, N., Carulla, M.R., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: Advances in Neural Information Processing Systems. pp. 656–666 (2017)

[28] Kim, M., Reingold, O., Rothblum, G.: Fairness through computationally-bounded awareness. In: NIPS. pp. 4842–4852 (2018)

[29] Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807 (2016)

[30] Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in Neural Information Processing Systems. pp. 4066–4076 (2017)

[31] Lipton, Z., McAuley, J., Chouldechova, A.: Does mitigating ml's impact disparity require treatment disparity? In: Advances in Neural Information Processing Systems. pp. 8125–8135 (2018)

[32] Lowry, S., Macpherson, G.: A blot on the profession. British medical journal (Clinical research ed.) **296**(6623), 657 (1988)

[33] Marope, P.T.M., Wells, P.J., Hazelkorn, E.: Rankings and accountability in higher education: Uses and misuses. Unesco (2013)

[34] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)

[35] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K.: Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. arXiv preprint arXiv:1811.07867 (2020)

[36] O'Neill, C.: Weapons of math destruction. How Big Data Increases Inequality and Threatens Democracy (2016)

[37] Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)

[38] Santelices, M.V., Wilson, M.: Unfair treatment? the case of freedle, the sat, and the standardization approach to differential item functioning. Harvard Educational Review **80**(1), 106–134 (2010)

[39] Shrestha, Y.R., Yang, Y.: Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. Algorithms **12**(9), 199 (2019)

[40] Spanakis, E.K., Golden, S.H.: Race/ethnic difference in diabetes and diabetic complications. Current diabetes reports **13**(6), 814–823 (2013)

[41] Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). pp. 1–7. IEEE (2018)

[42] Wu, X., Zhang, X.: Automated inference on criminality using face images. arXiv preprint arXiv:1611.04135 pp. 4038–4052 (2016)

[43] Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th international conference on world wide web. pp. 1171–1180 (2017)

[44] Zliobaite, I.: A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148 (2015)

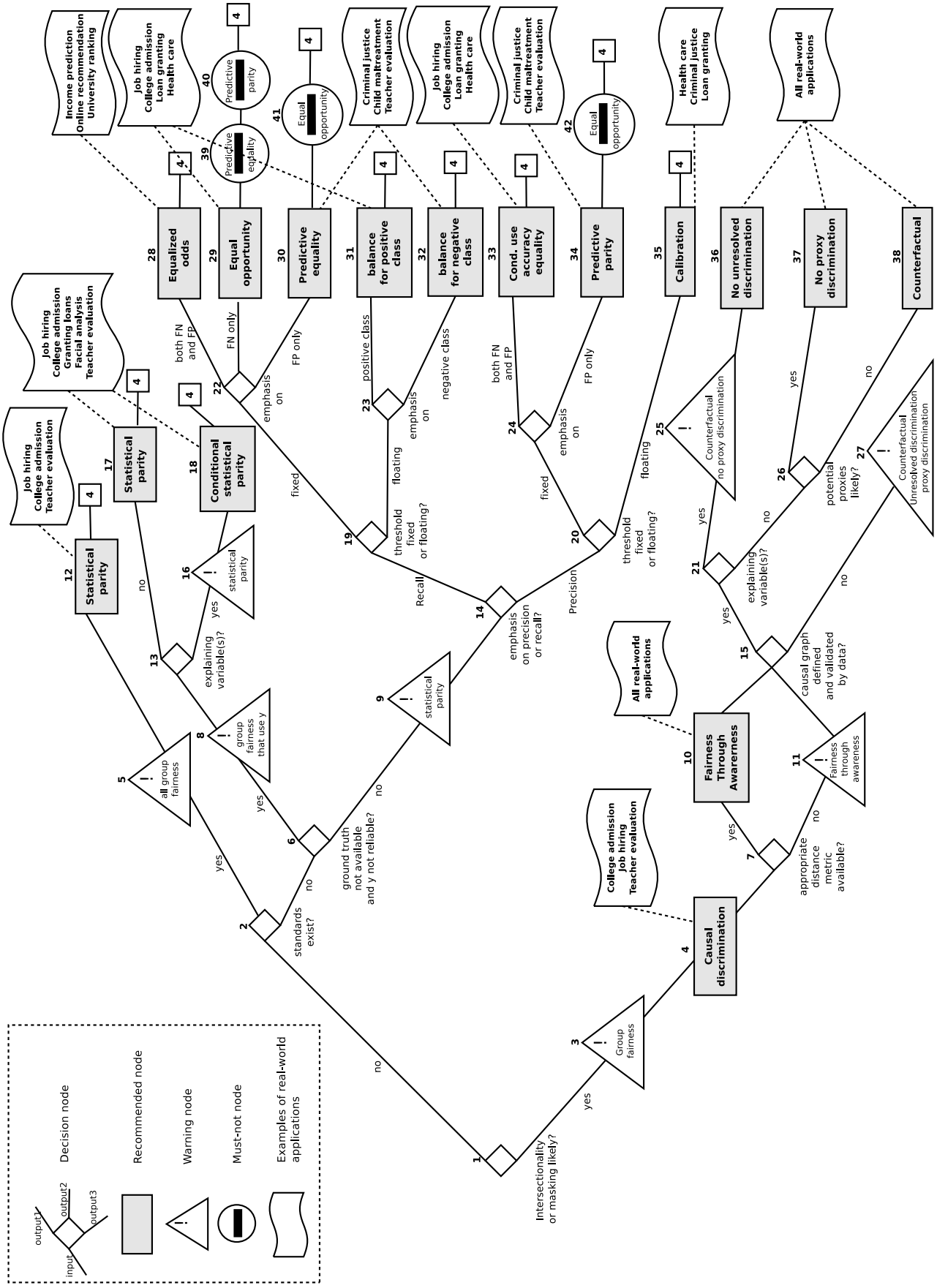


Figure 1: Fairness notions applicability decision diagram

Gendering algorithms in social media

Eduard Fosch-Villaronga
eLaw Center for Law and Digital Technologies, Leiden University
Leiden, The Netherlands
e.fosch.villaronga@law.leidenuniv.nl

Adam Poulsen
School of Computing and Mathematics, Charles Sturt University, Australia
apoulsen@csu.edu.au

Roger A. Søraa
Department of Interdisciplinary Studies of Culture, Norwegian University of Science and Technology (NTNU), Norway
roger.soraa@ntnu.no

Bart Custers
eLaw Center for Law and Digital Technologies, Leiden University
Leiden, The Netherlands
b.h.m.custers@law.leidenuniv.nl

ABSTRACT

Social media platforms employ inferential analytics methods to guess user preferences and may include sensitive attributes such as race, gender, sexual orientation, and political opinions. These methods are often opaque, but they can have significant effects such as predicting behaviors for marketing purposes, influencing behavior for profit, serving attention economics, and reinforcing existing biases such as gender stereotyping. Although two international human rights treaties include express obligations relating to harmful and wrongful stereotyping, these stereotypes persist both online and offline, and platforms often appear to fail to understand that gender is not merely a binary of being a 'man' or a 'woman,' but is socially constructed. Our study investigates the impact of algorithmic bias on inadvertent privacy violations and the reinforcement of social prejudices of gender and sexuality through a multidisciplinary perspective including legal, computer science, and queer media viewpoints. We conducted an online survey to understand whether and how Twitter inferred the gender of users. Beyond Twitter's binary understanding of gender and the inevitability of the gender inference as part of Twitter's personalization trade-off, the results show that Twitter misgendered users in nearly 20% of the cases (N=109). Although not apparently correlated, only 8% of the straight male respondents were misgendered, compared to 25% of gay men and 16% of straight women. Our contribution shows how the lack of attention to gender in gender classifiers exacerbates existing biases and affects marginalized communities. With our paper, we hope to promote the online account for privacy, diversity, and inclusion and advocate for the freedom of identity that everyone should have online and offline.

Keywords

Gender; Twitter; Inference; Gender Classifier; Privacy; Algorithmic Bias; Discrimination; LGBTQAI+; Gender Stereotyping; Social Media

1. INTRODUCTION

Online and social media platform providers use users' traits, including name, age, and gender, to improve user experience and personalize online behavioral advertising. By knowing users' characteristics, corporations can target or exclude certain groups more efficiently, tailor their services to users, and increase the time they spend on the platform [61]. In such a way, profiling makes marketing more precise and effective. However, a growing concern is the increasing use of opaque inferential analytics that reveal sensitive user attributes that serve attention economics [15] and that may reinforce existing biases which, although not explicit, can be very influential [10, 14].

A recurrent bias is gender stereotyping. Gender stereotyping "refers to the practice of ascribing to an individual 'woman' or 'man' specific attributes, characteristics, or roles by reason only of their membership in the social group of 'women or men'" [59]. However, gender stereotyping is a complex process that, although grounded in strong beliefs of what a gender is and should be, is both used and understood in a too simplistic manner. For instance, gay men are hyper-sexualized in e.g., the masculine promiscuity stereotype, or feminized, e.g., gay men who are perceived to be feminine fall into traditional female stereotypes. Two international human rights treaties include express obligations relating to harmful and wrongful stereotyping. Art. 5 of the Convention on the Elimination of All Forms of Discrimination against Women mandates States Parties to "take all appropriate measures to modify the social and cultural patterns of conduct of men and women, to achieve the elimination of prejudices and customary and all other practices which are based on the idea of the inferiority or the superiority of either of the sexes or stereotyped roles for men and women" [1]. Art. 8(1)(b) of the Convention on the Rights of Persons with Disabilities stresses that "States Parties undertake to adopt immediate, effective and appropriate measures to combat stereotypes, prejudices and harmful practices relating to persons with disabilities, including those based on sex and age, in all areas of life" [2]. However, these stereotypes are preserved both online and offline [22, 26]; platforms often appear to fail to grasp that gender is not limited to the simple binary of being solely a "man" or a "woman," but socially constructed [8].

Given that gender stereotypes persist online, and that the social media platform Twitter infers gender from a wide variety of sources,[1] we address the research question (RQ): *How accurate are Twitter's inferences of its users' gender identities?* Addressing this RQ brings into view concerns of discrimination, misgendering, and exacerbation of existing biases that online platforms persist in replicating that has already been highlighted by existing literature [23, 33]. Our goal is to investigate misgendering on Twitter and

[1] *See* https://help.twitter.com/en/rules-and-policies/data-processing-legal-bases

illustrate the impact of algorithmic bias on inadvertent privacy violations and how such biases reinforce social prejudices of gender and sexuality through a multidisciplinary perspective including legal, computer science, and queer media-studies viewpoints.

The reason behind our contribution lies in the idea that gender is a co-shaped, changing part of human identity tied into the socio-materiality of gendered relations often treated as a binary dichotomy. For instance, trans and non-binary users have recently claimed that they are being misgendered on Twitter because the categories "female" and "male" do not match who they are [16]. Second, platform providers no longer have to learn sensitive details about a particular user or correctly group users into categories for advertising to be effective, as advertising has a high tolerance for classification errors [62]. Nonetheless, not considering a broader understanding of gender in platforms can be socially harmful and costly, as technology usage and implementation may lead to further exacerbation of existing biases, including those relating to gender, race, and minorities [5, 24, 54].

In Section 2 of this article, we provide background information on inferential analytics to elucidate how companies infer specific user attributes, including gender, and how these techniques may harm users' rights. In Section 3, we explain the methods for this study, and in Section 4 we introduce the results. Our findings suggest that Twitter's binary understanding of gender excludes those not fitting the category "male" and "female." The results also show that inferring gender is part of Twitter's personalization trade-off and misgenders users in nearly 20% of the cases. Out of these cases, LGBTQIA+ individuals and straight women were misgendered more frequently than straight men. In Section 5 we discuss the lack of diversity in social media platforms and the role designers play in accounting for inclusivity and diversity. We conclude by presenting our future work, which includes a more extensive and refined survey to investigate this issue and the user's impressions further.

2. GENDERING ALGORITHMS

2.1 Profiling, inference analytics, and discrimination

Profiling techniques like regression, classification, or clustering mainly ascribe properties to people [9]. These methods infer distinct people's traits from different inputs of data, originating either from the person themself (i.e., predicting recidivism based on someone's criminal record) or others (i.e., others who ordered these shoes also like these shoes). Organizations use inferential analytics to induce user preferences using sensitive attributes such as race, gender, sexual orientation, political interests, and opinions [30, 56, 63]. These techniques can predict behaviors for marketing purposes and influence behavior for profit [68]. A critical feature of inferential analytics is that companies infer information from data not directly or indirectly provided by data subjects [14]. These inferences may be precise (like inferring age from the date of birth) or estimates (like inferring emotional states, e.g. happiness, or even intelligence from Facebook likes) [35]. In this way, data analytics can predict qualities that a data subject may not want to disclose and attributes that a data subject does not even know about themselves and ascribe them to an individual person.

One of the parameters used to infer attributes from people is the "like" button on many social media platforms [53]. In other words, what users like online tells something about who they are, such as their income [42] with a high degree of accuracy. Gender can also be inferred from Facebook likes with very high accuracy [35]. With approximately 250 Facebook likes, gender could be predicted with accuracy rates of 93%. Although this may seem like a high number, gender could be predicted with accuracy rates of about 70% when using only five Facebook likes. Moreover, when using only one single Facebook like, the accuracy rates were approximately 60% for gender predictions. According to Kosinski *et al.*, predictions for homosexuality were about 88% accurate for gays and 75% for lesbians, and predictions on being single versus in a relationship were about 67% accurate [35]– showing the complexity of inferring gender and sexuality through likes alone.

Inferential analytics may have some benefits. For instance, it can be a tool to fill gaps in fragmentary datasets or check the accuracy of available data by matching inferred data with the contested data. In this way, datasets enriched with many inferred attributes are likely to have higher levels of completeness and precision. In big data analytics, completeness and correctness of data is not a strict condition but can contribute to getting more well-defined and reliable results. Companies can identify that a particular customer prefers to consume video instead of text content, or is interested in learning about particular topics, like travel, fashion, or food. Companies use this information to personalize the user experience to fit the preferences of that particular individual.

However, inferential analytics has some drawbacks. When people's attributes are predicted, privacy is at stake, especially if people did not want to disclose specific personal information. Furthermore, these inferences may contain errors, leading to biased and unfair decisions and may lead to self-fulfilling prophecies [13]. These effects may amplify inequality, undermine democracy, lead to opinion echo chambers, and further push people into categories that are hard to break out of [49].

Machine learning and data mining tools can be developed so that they do not grant discriminating patterns such as gender stereotypes or profiles, a practice called discrimination-aware data mining [31]. The underlying idea is not to limit the data input (such as gender data), but to prevent the algorithms from yielding gender-based patterns, since not using gender data may still allow for predicting gender and thus result in indirect discrimination (discrimination by proxy). Focusing on the algorithms' design can prevent this when using gender in the development of data-driven decision models [67].

2.2 Gender inferences

Gender classification systems (GCS) are trained using a training dataset (or corpus) of structured and labeled data. These labels categorize data, and the features within, as either masculine or feminine [51]. Training a GCS builds a classification algorithm (or classifier) that categorizes features—such as body movements, physiological and behavioral characteristics, and facial features [51]—found in new data by comparing it to labeled features in the dataset. A GCS uses a feature extraction algorithm, classifier, and a dataset to make an inference [39].

Classifiers are trained in machine learning models. Exemplary models include neural networks [51], K-nearest neighbor [34], support vector machine [38], and Adaboost [41]. A classifier infers gender from video, images, or text, and the process is usually straightforward. First, data such as video or images are parsed into a GCS. Using a feature extraction algorithm, it then extracts features from the data, such as static body features, dynamic body features, apparel features, and biometrics [36, 38, 39]. Finally, it compares those features using a classifier to a feature dataset, which

is categorized by gender, and maps them to either category, inferring gender based on similarities in features [34, 51].

Similarly, a text-based GCS infers gender using features such as language, vocabulary, and frequency of words [39]. Text-based GCSs extract features using text mining from content found in forums, chat rooms, and social media [39, 50]. Beyond language, Corney *et al.* extended text-based feature extraction further into the typography field, training a classifier to make gender inferences based on style markers, structural characteristics, and gender-preferential language [12].

In the literature, developers have used classifiers to support text analysis techniques (e.g., sentiment and content analysis). Park *et al.* developed a GCS that supports sentiment analysis to identify the gender of persons making posts found on an online AIDS-related bulletin board [50]. The authors' GCS used a feature dataset that paired gender with the frequency of sentiment-driven words. During training, the GCS learned that women tended to use the words "thank," "bless," "scary," and, "illness" about twice as often as men, who themselves used "accurate," "important," "issue," and "aches" twice as often as women [50].

Several studies have made use of freely available Twitter user posts (or tweets) to train a GCS and infer the gender of other users [17, 19, 40, 45]. Lopes Filho *et al.* utilized a dataset categorizing gender by 60 textual meta-attributes associated with characters, syntax, words, structure, and morphology for the extraction of gender expression linguistic cues in tweets [40]. The authors compared different classifiers, finding that each accurately determined the gender of Twitter users, 63.5%, 61.96%, and 68.08% of the time. Using word unigrams, hashtags, and psychometric properties as features, the GCS developed by Fink *et al.* predicted the gender of Twitter users with 80% accuracy [17].

Gender recognition can be useful to support applications, such as face recognition and smart human-computer interface aid in other domains [51]. Developers use algorithmic gender classification in human-computer interaction, the security and surveillance industry, law enforcement, psychiatry, demographic research, education, commercial development, telecommunication, and mobile application and video games [34, 39, 51]. Depending on the application and dataset, developers may also use vision-based and biological information-based methods to make inferences [39].

However, "sex," "gender," and "sexuality" are often confused and used in overlapping ways, both by laypeople and experts. In this paper, we draw on the following definitions: "sex" usually refers to the assigned gender at birth based on medical factors (e.g. genitalia, chromosomes, and hormones), usually "m[20]ale" or "female" although in some cases "intersex." Sex can also be changed through medical intervention. "Gender" is both a "person's internal held sense of their gender"—also called gender identity—but is also tied to social, cultural and legal factors. "Sexuality" we take to mean the "physical, romantic, and/or emotional attraction to another person" [60]. We take into account that these definitions are also socially constructed through societal demands and norms.

3. METHODS

Available scientific literature focuses on how gender can be inferred from user attributes [17, 19, 40, 45]. However, there are not many studies that have compared the users' reported gender, the inferred gender from those attributes, and its correctness, although this avenue of research has been of an increasing interest in social sciences [23]. How algorithms exacerbate existing biases and affect marginalized communities is also a nascent area of specialization [23, 46, 64]. Our work contributes to the literature on algorithmic bias and discrimination by exploring misgendering on social media platforms like Twitter.

Given that gender stereotypes persist online, and that the social media platform Twitter infers gender from a wide variety of sources, we wondered how accurate Twitters' gender inferences of its users' gender identities are and, with the support of survey data, we explore what implications this social media practice has—such as the reinforcement of gender binarism and exacerbation of gender stereotypes [23]. We also refer to privacy and discrimination law, focusing on the impact of online behavioral advertising on inadvertent privacy violations [63] and the reinforcement of social prejudices.

We conducted a short survey disseminated using Twitter. For four days, from 22 to 26 May 2020, N=109 Twitter users responded. The online survey was prepared in Qualtrics and included five specific questions revolving around whether Twitter algorithms inferred users' gender and whether it was correct. In particular, we asked the user's sexual orientation (Q1), their gender identity (Q2), the pronouns they use (Q3), whether they provided Twitter with their gender information (Q4), and, if not, whether that was correctly assigned (Q5).[2] We gave the users instructions on how to find their assigned gender on Twitter,[3] and we processed anonymous data and surveyed the adult population only.

At the end of the survey, we exported, tabulated, and analyzed the data using Microsoft Excel Spreadsheet Software. The lead author analyzed the survey data, and the remaining authors examined the tabulated data and analysis to discuss discrepancies and ensure the reliability of the results. Empirically ascertaining if Twitter (mis)genders users lays the foundation for future work into the potential impacts in an extensive survey.

All of the respondents completed the survey in its entirety. However, the online survey has some limitations, including the small number of the respondents, which only amounts to N=109. This is due to the quick nature of our survey, within a limited, four-day timeframe. Another limitation may be the limited representativeness of the sample, which seems to over-represent the LGBTQIA+ community compared to the number of straight people in society in general. This potential bias may be due to one or more of the following reasons. First, the LGBTQIA+ community may be overrepresented among Twitter users (Twitter does not provide data on this) or be overrepresented in the authors' Twitter networks. Moreover, people from the LGBTQIA+ community may be more inclined to complete the survey, perhaps because the survey topic

[2] The exact wording of the questions were: 1) What is your sexual orientation?; 2) What is your gender identity?; 3) What pronouns do you use?; 4) Did you at one point provide Twitter with your gender?; 5) If you did not include your gender in your profile, the gender that appears in your profile may have been assigned by Twitter, is the gender appearing here correct?

[3] To know the gender assigned by Twitter, go to picture > settings and privacy > account > your Twitter data > password > account > confirm password > gender.

appealed to them, as it may relate to past experiences of gender stereotyping or misrepresentation, on Twitter or elsewhere.

4. RESULTS

Based on the conducted online survey, we identify the following data:

	Self-reporting				Incorrect by Twitter				% Incorrect by Twitter			
	Female	Male	Non-binary	All	Female	Male	Non-binary	All	Female	Male	Non-binary	All
Straight	37	25		62	6	2		8	16%	8%		13%
Gay		24		24		6		6		25%		25%
Lesbian	2			2								
Bisexual	4	5	1	10	1	1	1	3	25%	20%	100%	30%
Asexual	3		2	5			2	2	0%		100%	40%
Questioning	2	1		3					0%	0%		
Other	2		1	3	2			2	100%			67%
Sum	**50**	**55**	**4**	**109**	**9**	**9**	**3**	**21**	**18%**	**16%**	**100%**	**19%**

Table 1. Twitter gender inference accuracy in a N=109 sample data.

Out of N=109 respondents, 19% had their gender wrongly assigned, whereas Twitter inferred users' gender correctly in 81% of the cases. Our central hypothesis revolved around differences between the self-reported gender identity (male), the sexual orientation of users (gay), and the correctness of the Twitter assigned gender (female).

Twitter infers their users' identity from a wide variety of sources, such as information from the account, interactions with links, and cookie data,[4] but not from their sexual orientation. However, how apparently fair algorithmic designs and categorizations have ulterior and unintended consequences in specific communities is well-known in the literature [10, 21, 28]. For instance, our collected data shows that, out of the misgendered Twitter users that we analyzed, only 38% were straight. Only 8% of the straight men respondents were misgendered, compared to 25% of gay men and 16% straight women. Individuals that self-reported as bisexuals were misgendered in 25% of the cases for bisexual women and 20% for bisexual men. Respondents identifying as non-binary were misgendered in all cases. These results show that the LGBTQIA+ community and straight women were more often misgendered than straight men in our sample. Therefore, misgendering was, contrary to our hypothesis, not only limited to gay men compared to straight men. Moreover, women and non-binary are usually more misgendered by Twitter than men.

The findings also seem to suggest that lesbian and questioning people are less likely to be misgendered, although the numbers are small in our sample (two lesbian and three questioning participants)—more studies are needed for this sample population. One questioning and one lesbian participant answered that they had provided their gender to Twitter, meaning the gender of the remaining were inferred correctly by Twitter. The findings also show that non-binary participants (N=2) were misgendered, both of whom were also asexual participants. However, there were other asexual participants (N=3) whose gender (female in all cases) was correctly inferred by Twitter (each answered 'I do not know' when asked if they provided Twitter with their gender).

Of the 109 participants, only 15% provided Twitter with their gender, whereas 24% did not, and 61% did not remember doing so. 42% of those who did not provide Twitter with gender were from the LGBTQAI+ community. Of the 16 participants who provided their gender, all but one answered "Yes" about whether or not the gender appearing on their Twitter profile was correct. An outlier was an asexual, nonbinary person. This may indicate that either (1) some of those 16 participants were mistaken and had entered their gender into their Twitter profile previously or (2) Twitter may infer gender and change the one entered by the user.

Other findings resulted from discussions over Twitter, where we shared the online survey. Some respondents openly reported that Twitter used to misgender them, but that now Twitter gendered them correctly, probably due to their increasing interest in gender equality. Other respondents mentioned they had two profiles, but that Twitter misgendered the profile they used the most. A respondent suggested that, although gay, Twitter assigned his gender correctly, while another was surprised to be considered "female" while being a "male."

5. DISCUSSION

5.1 Misgendering in social media is discriminatory

Research affirms that gender identity is primarily subjective and internal, which juxtaposes with the idea that gender can be recognized automatically, at least with the state of art GCS [23]. Moreover, misgendering users via automated gender recognition systems have adverse implications, some of those being that they reinforce gender binarism, undermine autonomy, are a tool for surveillance, and threaten safety [23].

They also exacerbate existing stereotypes. Classifiers trained on real-world datasets are often biased because the data used to train them contains racial and gender stereotypes [7, 18, 43, 57]. Female names are more associated with family than career words, with arts more so than mathematics and science [47, 48]. Datasets imSitu and MS-COCO are significantly gender-biased and "models trained to perform prediction on these datasets amplify the existing gender bias when evaluated on development data" [65]. For example, the verb "cooking" is heavily biased towards women in a classifier trained using the imSitu dataset, amplifying existing gender stereotypes [65]. The same gender biases have been shown in natural language processing [55, 66], another method used to support gender classifiers [11].

To be misgendered reinforces also the idea that society does not consider or recognize a person's gender as "real," causing rejection, impacting self-esteem and confidence, felt authenticity, and increasing one's perception of being socially stigmatized [33]. If not addressed carefully, these gender biases in the offline world may propagate to artificial intelligence [10]. This is especially concerning given that available research suggests that many individuals perceive automatic misgendering as more harmful than human misgendering [23].

Moreover, when the tools used to extract patterns and profiles from data are not transparent, it may be hard for people to contest any decisions resulting from this, which may impede their freedom and autonomy and may inadvertently affect their privacy. In the EU, the collecting and processing of personal data are protected under the General Data Protection Regulation (GDPR), which also addresses discrimination issues in datasets. However, enforcing legislation in

[4] *See* https://help.twitter.com/en/rules-and-policies/data-processing-legal-bases

such cases is very challenging. For data protection, scholars note that information about a person's gender, age, financial situation, geolocation, and online profiles are not sensitive data according to Article 9 of the GDPR, despite often being grounds for discrimination [63]. Not being "sensitive data" translates into not enjoying the extra protection (such as users' informed and explicit consent) that categories of information deemed sensitive such as race, religion, or sexual orientation have. Discrimination in (patterns and profiles extracted from) large datasets can be hard to detect. Indirect discrimination takes place unintentionally when users are unaware of any harm profiles may be doing. However, it may also be the case that companies use profiles precisely to conceal discrimination, a process called masking [13]. Because direct discrimination in data is hard to detect, and indirect discrimination is nearly impossible to detect, it can be challenging to enforce equal treatment acts and data protection legislation.

Many forms of discrimination are illegal in most Western jurisdictions. Not hiring someone based on their gender, ethnicity, or sexual orientation, or because they have a criminal record, is prohibited for most professions. Not every decision based on the sensitive characteristics mentioned is forbidden, however. Legislation that forbids discrimination based specific characteristics lists the characteristics that may not serve as a basis for making decisions (including gender, ethnicity, political preferences, trade union membership, or sexual orientation). Nonetheless, "softer" forms of discrimination, in the form of stigmatization of specific population groups may occur, for example, in the formation of friendships. On a larger scale, this could lead to social polarization and segregation. For now, misgendering or addressing someone with the wrong pronoun is not sufficiently grave to be considered harassment under certain specific legal provisions (although there has been advocacyfor remedies for these acts [4]).

5.2 Inferences Organizations controversially infer gender for legitimate interests

Twitter makes inferences about users' accounts, including interests, age, and gender, to provide features such as account suggestions (e.g., suggested contacts, promoted accounts for the user to follow), advertising, recommendations, and timeline ranking.[5] Twitter uses users' content, activity, relationships, and interactions to genderize content production patterns [52], infer gender, and make these suggestions.[6] Twitter justifies making inferences about interests, age, and gender, stating that it helps tailor content to users, keeps the platform safe and enjoyable for all users, and enables Twitter to provide compelling, targeted advertising. In other words, users have to accept the trade-off if they want to have a personalized Twitter account.

The GDPR lists a limited number of legal grounds for data processing, including consent, the performance of a contract, or legitimate interests. Twitter states that it makes "inferences about your account - such as interests, age, and gender" for "legitimate purposes." The appeal to legitimate interests as a legal basis for data processing is controversial, as legitimate purposes are only a solid legal basis if there is a necessity. It is questionable, however, whether gender inferences are necessary for Twitter. Although the legitimate interest seems less constraining than other grounds for data processing, it should not be considered a "last resort" when all other grounds for lawful data processing fail [3].

Legitimate interest is the most appropriate legal ground for data processing if the data controller uses people's data in ways they would reasonably expect and have a minimal privacy impact, or where there is a compelling justification for the processing. If controllers choose this legal ground, they "should take on extra responsibility for considering and protecting people's rights and interests" [27]. Thus, three elements configure the basis for legitimate interest: identifying the legitimate interest, showing that the processing is necessary to achieve it, and balancing it against the individual's interests, rights, and freedoms.

Our survey findings highlight a significant number of misgendered users and question whether Twitter did balance their interests against individuals' interests. First, out of the 109 participants, only 15% provided Twitter with their gender, while Twitter inferred their gender anyway. Second, our results suggest that the LGBTQIA+ community and straight women may be more often misgendered than straight men. Third, remedies for opposing the processing seem not to correspond in magnitude to the subsequent impact of being misgendered. A user can modify or rectify the inferred gender but cannot escape that inference unless she actively opts out of Twitter's personalization features. Making users choose between these two is as if, in times of COVID-19, developers made users choose between health or privacy [25]. Moreover, it results in a privacy paradox: the gender inference causes a privacy issue (i.e., disclosing information people may want to keep to themselves), but to address this, users have to provide additional information, disclosing even more (or more detailed) information about themselves [13]. This is particularly problematic for communities that society has been historically discriminated against and in which gender is a sensitive part of their identity [16, 44].

5.3 Accounting for diversity in social media

Platforms exclude and misrepresent a large number of potential users if they are not respectful and inclusive towards their gender identity or sexual orientation. The assumption that gender is physiologically-rooted harms trans people overall by essentializing the body as the source of gender, and also harms non-binary people, who cannot be accurately classified [33]. As Fergus highlighted, transgender and non-binary users reported being misgendered by Twitter, which we found to be the case in our survey (100% of the non-binary participants reported being misgendered) [16]. These findings may result from the fact that Twitter gender classifiers do not account for diversity and work on male/female binary categorization that, although it does represent some people's gender expression, does not do justice to the freedom of identity that everyone should have.

Our study shows that when it comes to diversity and more inclusive engagement, social media platforms like Twitter still have a long way to go to become a more open and welcoming platform for a wide variety of users. Misgendering users in the background is not good practice, and beyond echoing deeply rooted stereotypes, can lead to privacy and discrimination issues,. The lack of diversity in marketing strategies is apparent when users can be gendered as male or female only. However, making strategies for a diverse

[5] *See* https://help.twitter.com/en/rules-and-policies/data-processing-legal-bases; *see* also https://help.twitter.com/en/using-twitter/account-suggestions.

[6] *See* https://help.twitter.com/en/using-twitter/account-suggestions

engagement with the "queer rainbow economy" can make for more affluent and more diverse revenue streams [6, 32, 58].

From all this, it is clear that digital identity and participatory culture play a massive role in the sense of self in the modern world and that there should be more effort to realize diversity and inclusion in the online world [29] to not perpetuate the normative view that certain groups of people, such as trans or non-binary people, do not exist [33].

6. SUMMARY

An online survey showed that, out of N=109 respondents, Twitter correctly inferred users' gender in 81% of the cases, and 19% were misgendered. A close look at the results shows that only 8% of the straight men respondents were misgendered, compared to 25% of gay men and 16% of straight women, while non-binary users were misgendered in all the cases.

Social media platforms like Twitter have economic incentives to know users' genders for commercialization and targeted advertisements. However, our investigation shows that inferring a user's gender with automated means clashes with the understanding that gender is subjective and internal. Misgendering has also broader consequences, leading to serious privacy, discrimination, autonomy, and self-identity issues. Misgendering reinforces gender stereotypes, accentuates gender binarism, undermines autonomy, and leads to toxic cultures and algorithmic bias [23, 37]. Moreover, misgendering causes a feeling of rejection, impacting one's self-esteem, confidence, and authenticity, increasing social stigmatization [33].

If users do not provide a gender parameter choice themselves, platforms may infer the user's gender from a wide variety of data sources, including personal data. Therefore, gender classifiers should account for diversity and inclusion, using a more accurate understanding of gender to represent contemporary society fully. Otherwise, inferential analytics may reinforce existing biases about gender stereotyping. By including diverse users early on, during the design, and with the possibility to provide feedback afterward, the technology can be experienced as more just and fairer. Inclusive engagement that reflects on the users as not homogeneous can have a positive impact on technology.

By identifying how inferential analytics may reinforce gender stereotyping and affect marginalized communities, we hope to continuously contribute to promoting the online account for privacy, diversity, and inclusion and advocate for the freedom of identity that everyone should have online and offline [69]. Looking forward, a more robust survey ought to be undertaken to further explore the social implications of gender inference on Twitter, such as discrimination and diversity in social media.

7. ACKNOWLEDGMENTS

Part of this project was funded by the LEaDing Fellows Marie Curie COFUND fellowship, a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 707404.

8. REFERENCES

[1] *Convention on the elimination of all forms of discrimination against women*. https://www.ohchr.org/en/professionalinterest/pages/cedaw.aspx#:~:text=On%2018%20December%201979%2C%20the,twentieth%20country%20had%20ratified%20it, 1979.

[2] *Convention on the rights of persons with disabilities*. https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/article-8-awareness-raising.html, 2008.

[3] Article 29 Data Protection Working Party. *Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC*, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf, 2014.

[4] Ashley, F. *No, pronouns won't send you to jail: The misunderstood scope of Bill C-16*. Medium, https://medium.com/@florence.ashley/no-pronouns-wont-send-you-to-jail-43c268cffd55, 2017.

[5] Bray, F. Gender and technology. *Annual Review of Anthropology*, 36(1): 37-53, 2007.

[6] Brown, G. Thinking beyond homonormativity: Performative explorations of diverse gay economies. *Environment and Planning A: Economy and Space*, 41(6): 1496-1510, 2009.

[7] Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the First Conference on Fairness, Accountability and Transparency*, 77–91. PMLR, 2018.

[8] Butler, J. *Gender trouble, feminist theory, and psychoanalytic discourse*. Routledge, 1990.

[9] Calders, T. and Custers, B. *What Is data mining and how does it work?* In: *Discrimination and privacy in the information society: Data mining and profiling in large databases* (eds. Custers et al.), pages 27–42. Springer, 2013.

[10] Caliskan, A., Bryson, J. J. and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183-186, 2017.

[11] Campa, S., Davis, M. and Gonzalez, D. *Deep & machine learning approaches to analyzing gender representations in journalism*.https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15787612.pdf, 2019.

[12] Corney, M., Vel, O. d., Anderson, A. and Mohay, G. Gender-preferential text mining of e-mail discourse. In *Proceedings of the 18th Annual Computer Security Applications Conference, 2002*, pages 282-289, 2002.

[13] Custers, B. Data dilemmas in the information society: Introduction and overview. In B. Custers, T. Calders, B. Schermer and T. Zarsky, editors, *Discrimination and privacy in the information society*, pages 3–26, Springer, 2013.

[14] Custers, B. *Profiling as inferred data: Amplifier effects and positive feedback loops*. Amsterdam University Press, http://www.jstor.org/stable/j.ctvhrd092.23, 2018.

[15] Davenport, T. D. and Beck, J. C. *The attention economy: Understanding the new currency of business*. Harvard Business School Press, 2002.

[16] Fergus, J. *Twitter is guessing users' genders to sell ads and often getting it wrong*. Input, https://www.inputmag.com/tech/twitter-guesses-your-gender-to-serve-you-ads-relevant-tweets-wrong-misgendered, 2020.

[17] Fink, C., Kopecky, J. and Morawski, M. Inferring gender from the content of tweets: A region specific example. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Medi*, pages 459–452, AAAI, 2012.

[18] Font, J. E. and Costa-jussà, M. R. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, pages 147–154. Association for Computational Linguistics, 2019.

[19] Garibo i Orts, O. A big data approach to gender classification in Twitter. In *Proceedings of the Ninth International Conference of the CLEF Association*, 2018.

[20] GLAAD *GLAAD media reference guide – transgender*. https://www.glaad.org/reference/transgender, n.d.

[21] Gomes, A., Antonialli, D. and Olivia, T. D. *Drag queens and Artificial Intelligence: Should computers decide what is 'toxic' on the internet?* , https://www.internetlab.org.br/en/freedom-of-expression/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/, 2019.

[22] Grant, A., Grey, S. and van Hell, J. G. Male fashionistas and female football fans: Gender stereotypes affect neurophysiological correlates of semantic processing during speech comprehension. *Journal of Neurolinguistics*, 53: 100876, 2020.

[23] Hamidi, F., Scheuerman, M. K. and Branham, S. M. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Paper 8, ACM, 2018.

[24] Hao, K. *Facebook's ad-serving algorithm discriminates by gender and race*. MIT Technology Review, https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias/, 2019.

[25] Harari, Y. N. *The world after coronavirus*. https://www.ft.com/content/19d90308-6858-11ea-a3c9-1fe6fedcca75, 2020.

[26] Hentschel, T., Heilman, M. E. and Peus, C. V. The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. *Frontiers in Psychology*, 10(11), 2019.

[27] Information Commissioner's Office *Legitimate interests*. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/legitimate-interests/, n.d.

[28] Ito, J. *Supposedly 'fair' algorithms can perpetuate discrimination*. Wired, https://www.wired.com/story/ideas-joi-ito-insurance-algorithms/, 2019.

[29] Jenkins, H., Ito, M. and boyd, d. *Participatory culture in a networked era: A conversation on youth, learning, commerce, and politics*. Polity Press, Cambridge, UK, 2016.

[30] Jernigan, C. and Mistree, B.F.T. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009.

[31] Kamiran, F., Calders, T. and Pechenizkiy, M. Techniques for discrimination-free predictive models. In B. Custers, T. Calders, B. Schermer and T. Zarsky, editors, *Discrimination and privacy in the information society*, pages 223–239, Berlin, Springer, Berlin, Heidelberg. 2013.

[32] Keating, A. and McLoughlin, D. Understanding the emergence of markets: A social constructionist perspective on gay economy. *Consumption Markets & Culture*, 8(2): 131-152, 2005.

[33] Keyes, O. The misgendering machines: Trans/HCI Implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW): Article 88, 2018.

[34] Khan, S. A., Ahmad, M., Nazir, M. and Riaz, N. A comparative analysis of gender classification techniques. *International Journal of Bio-Science and Bio-Technology*, 5(4): 223–244, 2013.

[35] Kosinski, M., Stillwell, D. and Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15): 5802-5805, 2013.

[36] Kumar, D., Gupta, R., Sharma, A. and Saroj, S. K. Gender classification using skin patterns. In *Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*, 2019.

[37] Lambrecht, A. and Tucker, C. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7): 2966-2981, 2019.

[38] Li, B., Lian, X.-C. and Lu, B.-L. Gender classification by combining clothing, hair and facial component classifiers. *Neurocomputing*, 76(1): 18-27, 2012.

[39] Lin, F., Wu, Y., Zhuang, Y., Long, X. and Xu, W. Human gender classification: A review. *International Journal of Biometrics*, 8(3-4): 275–300, 2016.

[40] Lopes Filho, J. A. B., Pasti, R. and de Castro, L. N. Gender classification of Twitter data based on textual meta-attributes extraction. In Á. Rocha, A. M. Correia, H. Adeli, L. P. Reis and M. Mendonça Teixeira, editors, *New advances in information systems and technologies*, pages 1025–1034, Cham, Springer International Publishing, 2016.

[41] Mathivanan, P. and Poornima, K. Biometric authentication for gender classification techniques: A review. *Journal of The Institution of Engineers (India): Series B*, 99(1): 79-85, 2018.

[42] Matz, S. C., Menges, J. I., Stillwell, D. J. and Schwartz, H. A. Predicting individual-level income from Facebook profiles. *PLOS ONE*, 14(3): e0214369, 2019.

[43] McDuff, D., Song, Y., Kapoor, A. and Ma, S. Characterizing bias in classifiers using generative models. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 2019.

[44] McLemore, K. A. Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity*, 14(1): 51-74, 2015.

[45] Nieuwenhuis, M. and Wilkens, J. Twitter text and image gender classification with a logistic regression n-gram model. In *Proceedings of the Ninth International Conference of the CLEF Association*, pages, 2018.

[46] Noble, S. U. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, New York, 2018.

[47] Nosek, B. A., Banaji, M. R. and Greenwald, A. G. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1): 101-115, 2002.

[48] Nosek, B. A., Banaji, M. R. and Greenwald, A. G. Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology*, 83(1): 44-59, 2002.

[49] O'Neil, C. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, New York, 2016.

[50] Park, S. and Woo, J. Gender classification using sentiment analysis and deep learning in a health web forum. *Applied Sciences*, 9, 2019.

[51] Rai, P. and Khanna, P. Gender classification techniques: A review. In D.C. Wyld et al. (editors), *Advances in computer science, engineering & applications*, pages 51–59, Berlin, Springer, 2012.

[52] Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schulz, J., Hale, T. M. and Stern, M. J. Digital inequalities and why they matter. *Information, Communication & Society*, 18(5): 569-582, 2015.

[53] Roosendaal, A. Facebook tracks and traces everyone: Like this! *Tilburg Law School Legal Studies Research Paper Series*, 2010.
[54] Schiebinger, L. Scientific research must take gender into account. *Nature*, 507(7490): 9, 2014.
[55] Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W. and Wang, W. Y. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, 2019.
[56] Thorson, K., Cotter, K., Medeiros, M. and Pak, C. Algorithmic inference, political interest, and exposure to news and politics on Facebook. *Information, Communication & Society*: 1-18, 2019.
[57] Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) 2011*, pages 1521–1528, 2011.
[58] Um, N.-H. Seeking the holy grail through gay and lesbian consumers: An exploratory content analysis of ads with gay/lesbian-specific content. *Journal of Marketing Communications*, 18(2): 133-149, 2012.
[59] UN Office of the High Commissioner Human Rights *Gender stereotyping*. https://www.ohchr.org/EN/Issues/Women/WRGS/Pages/GenderStereotypes.aspx, n.d.
[60] University of Washington Human Resources Office *Terminology*. https://hr.uw.edu/ops/transgender-resources/terminology/, n.d.
[61] Ur, B., Leon, P. G., Cranor, L. F., Shay, R. and Wang, Y. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, Article 4. ACM, 2012.
[62] Wachter, S. Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Journal*, 35(2), 2020.
[63] Wachter, S. and Mittelstadt, B. A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review*, 2019(2): 494–620, 2019.
[64] Willson, M. Algorithms (and the) everyday. *Information, Communication & Society*, 20(1): 137-150, 2017.
[65] Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K. W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017.
[66] Zhou, P., Zhao, J., Huang, K.-H. and Shi, W. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5279-5287, 2019.
[67] Žliobaitė, I. and Custers, B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2): 183-201, 2016.
[68] Zuboff, S. Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1): 75-89, 2015.
[69] Fosch-Villaronga, E., Poulsen, A., Søraa, R. A., & Custers, B. H. M. (2021). A little bird told me your gender: Gender inferences in social media. *Information Processing & Management*, *58*(3), 102541.

About the authors:

Dr. Eduard Fosch-Villaronga is an Assistant Professor at the eLaw Center for Law and Digital Technologies at Leiden University (NL). Fosch-Villaronga investigates the legal and regulatory aspects of robot and Artificial Intelligence (AI) technologies and is the leader of the Robotics and Autonomous Systems Working Group, which is an interdisciplinary Working Group at eLaw dedicated to advance the understanding of the legal, regulatory, and ethical implications of robots and autonomous systems. Previously, he was a Marie Sklodowska-Curie postdoctoral researcher at the same group and served the European Commission in the Sub-Group on Artificial Intelligence (AI), connected products and other new challenges in product safety to the Consumer Safety Network (CSN) to revise the General Product Safety directive. He is also the co-leader of the Working Group on the Ethical, Legal and Societal Aspects for Wearable Robots at the H2020 COST Action CA16116. Among his publications, Fosch-Villaronga published a book entitled *Robots, Healthcare, and the Law: Regulating Automation in Personal Care* with Routledge.

Mr. Adam Poulsen is a computer scientist & PhD candidate at Charles Sturt University, Australia. Poulsen's research focus is on value sensitive robots, LGBTQIA+ elders and care, and robot and machine ethics. Poulsen is a recipient of the Australian Government Research Training Program Scholarship. Among other initiatives, Poulsen has contributed actively to the Australian Association of Gerontology's Assistive Technology Special Interest Group as a co-convenor.

Dr. Roger A. Søraa is a researcher at the Department of Interdisciplinary Studies of Culture, NTNU Norwegian University of Science and Technology. He also works as a project leader and project developer for the Department of Neuromedicine and Movement Science at the same university. An ongoing collaboration between these departments is the Immersive Technologies and Robotics laboratory (ImRo), where Søraa serves as the deputy manager. Next to his positions at NTNU, Søraa is affiliated with RURALIS, where he researches smart technology for sustainable agriculture. Søraa completed his PhD dissertation in Studies of Science and Technology in 2018. Søraa's current research interests include robotization of gerontechnologies-, transport and agriculture, automation of work, and practices and digitalization of society and social media.

Prof. Dr. Bart H. M. Custers is a full professor of law and data science and director of eLaw, the Center for Law and Digital Technologies at Leiden University, the Netherlands. With a background in both law and physics, his research is focused on law and digital technologies. Research interests include discrimination and privacy issues of new technologies, particularly data mining and profiling. Dr. Custers has published seven books, including four books on discrimination and privacy in the context of Big Data. On a regular basis, he gives lectures on profiling and privacy issues of new technological developments. He presented his work at international conferences in the United States, Canada, China, Japan, the Middle East, Africa, and throughout Europe. He has published his work, over 100 publications, in both scientific and professional journals and newspapers.

Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning

Pieter Delobelle
Department of Computer Science, KU Leuven, Leuven.AI
Leuven, Belgium
pieter.delobelle@kuleuven.be

Paul Temple, Gilles Perrouin, Benoit Frénay, Patrick Heymans
PReCISE, NaDi, University of Namur
Namur, Belgium
firstname.lastname@unamur.be

Bettina Berendt
Department of Computer Science, KU Leuven, Leuven.AI
Leuven, Belgium
Faculty of Electrical Engineering and Computer Science, TU Berlin
Berlin, Germany
bettina.berendt@kuleuven.be

ABSTRACT

Machine learning is being integrated into a growing number of critical systems with far-reaching impacts on society. Unexpected behaviour and unfair decision processes are coming under increasing scrutiny due to this widespread use and its theoretical considerations. Individuals, as well as organisations, notice, test, and criticize unfair results to hold model designers and deployers accountable. We offer a framework that assists these groups in mitigating unfair representations stemming from the training datasets. Our framework relies on two inter-operating adversaries to improve fairness. First, a model is trained with the goal of preventing the guessing of protected attributes' values while limiting utility losses. This first step optimizes the model's parameters for fairness. Second, the framework leverages evasion attacks from adversarial machine learning to generate new examples that will be misclassified. These new examples are then used to retrain and improve the model in the first step. These two steps are iteratively applied until a significant improvement in fairness is obtained. We evaluated our framework on well-studied datasets in the fairness literature — including COMPAS — where it can surpass other approaches concerning demographic parity, equality of opportunity and also the model's utility. We investigated the trade-offs between these targets in terms of model hyperparameters and also illustrated our findings on the subtle difficulties when mitigating unfairness and highlight how our framework can assist model designers.

Keywords

Adversarial machine learning, fairness, neural networks

1. INTRODUCTION

Machine learning eases the deployment of systems that tackles various tasks: spam filtering, image recognition, etc. One of the most trendy applications is decision support. These systems give recommendations on who should get a loan, predict who could commit subsequent offences, etc. based on data describing individuals. Such systems have a desirable property: they provide objective, supposedly consistent decisions based on a collection of data. At first glance, this could counteract unfair decisions made by humans.

However, they still exhibit unfair behaviour. Such behaviours can impact individuals belonging to a specific social group. Well-studied examples include the COMPAS system that predicts the recidivism of pre-trial inmates [2, 10] and keep taking decisions in favor of Caucasian people compared with African-Americans. We consider fairness where the impact on individuals can be categorized as either *allocational harm* or *representational harm* [8]. With allocational harm, the favorable outcome (e.g. bail being granted) differs between social groups. Representational harm is more subtle, and include *differences in performance* between social groups, and *stereotyping*. We focus on allocational harm in this work, as decision support systems with different outcomes can affect social groups far beyond the outcome itself.

If an allocational harm exists when advising for a favorable outcome for a group, the decision towards an other group to not receive the same outcome can, ultimately, create a feedback loop where unfair behaviours are amplified [19, 32]. For example, consider a system that imposes more expensive loans to African-American people, who then fail to repay them, that will lead them to ask for another loan, etc.

To avoid such consequences, researchers increasingly focused on incorporating fairness as objectives in their systems. In *discrimination-aware data mining* (DADM), modifications were developed and applied to data, learning algorithms, or resulting patterns and models [25]. Recently, adversarial fairness continues in this direction with, for instance, research on learning representations [31, 43] and task-specific fair models [1, 36, 38].

Adversaries are also used when assessing the security of machine learning based systems. Biggio and Roli [5] synthesised a decade of research in adversarial machine learning. This domain aims at finding or creating examples that are problematic for a machine learning model, *e.g.* Biggio et al. [7], Papernot et al. [33, 34]. These examples can be injected directly into the training phase to perturb the training of the model, known as *poisoning attacks*, or they can simply be used to bypass the model that is supposed to act as a filter, in this case, they are called *evasion attacks*.

In this paper, we propose a new framework implementing

a gray-box fairness scenario coupling evasion attacks and fair machine learning using gradient reversal. We evaluate our framework on three datasets: (i) COMPAS, (ii) German Credit, and (iii) Adult. Our framework improves demographic parity and equal opportunity when comparing to the state of the art while globally improving the model's utility. We thus reconciles fairness and model performance.

This paper is an extension of our previous work [12], which was presented at the first workshop on Bias and Fairness in AI (BIAS 2020) held jointly with the ECML-PKDD conference. With respect to the original contribution, we provide: *(i)* a detailed analysis of tradeoffs between fairness and utility by computing and charting a Pareto front; *(ii)* a more nuanced interpretation of our results based on 30 random hyperparameter trials, where we calculated the expected validation accuracy, and *(iii)* a discussion on the role of prior domain knowledge to create both valid and realistic examples to feed to the model.

This paper is organised as follows: Section 2 discusses related work on adversarial machine learning techniques but also on measuring and mitigating unfairness. Section 3 presents our new framework, followed by its evaluation on the COMPAS, German Credit and Adult datasets in Section 4. Section 7 concludes and gives an outlook on future work.

2. BACKGROUND AND RELATED WORK

2.1 Poisoning and Evasion Attacks

Adversarial machine learning assesses the required effort to make a classifier unusable by forcing it to perform so many errors that users will not trust its predictions anymore [5]. The generation of adversarial attacks follows this black-box process: (i) probe an existing target model to gain information about it, (ii) copy an existing example, (iii) apply an adversarial technique that will modify the example depending on the desired goal.

Various models can be attacked including support vector machines (SVMs), linear models and even neural networks (NNs) [7, 33, 34]. Since all machine learning models are based on a similar set of assumptions, including the fact that they statistically approximate data distributions, adversarial machine learning leverage on these assumptions to train a surrogate classifier to start the attack on. After the attack finishes, because of these assumptions, newly generated data examples can be transferred back to the original target model [13]. Only one restriction remains on the surrogate classifier, attacks are gradient-based techniques requiring the discriminant function to be differentiable. We distinguish between *poisoning attacks* and *evasion attacks*. In the former, malicious examples are introduced in the training set in order to significantly and permanently affect the model to be trained [4, 6]. In the latter, malicious examples are provided at test time to harm the model without changing it [7].

In related work by Solans et al. [39], poisoning attacks have been used to influence the fairness of machine learning models in a black-box manner. The authors have also linked their poisoning attack to demographic parity, an evaluation metric that will be introduced in Section 2.2.

Kulynych et al. [29] also used poisoning attacks, specifically for countering effects of credit scoring systems. In addition, they provide an outline of how users can affect optimization systems to mitigate negative external impacts, called Personal Optimization Technologies (POTs). This framework could also be used to ground the adversarial attacks generated by the *Feeder* from our framework.

In this paper, we consider evasion attacks that are performed on *an already trained model*. We craft adversarial examples that are supposed to belong to a class while the model will assign them with a different one because of specific characteristics, highlighting an unfair behavior regarding a certain population. By carefully reintroducing these examples during retraining, we hypothesize that the retrained model will be fairer. While we rely on a similar example generation technique, our exploitation goal differs from [39].

2.2 Evaluating Fairness

There exist several measures of fairness in the literature, *e.g.* demographic parity [16], equalized odds and equalized opportunity [26], statistical parity [21, 43], disparate impact [10, 21], and threshold testing [35]. They are categorized into different family of measures, presented in the FairML book [3], depending on their mathematical expression. In the following, we focus on two different measures: (i) demographic parity which is probably the most popular but also controversial and (ii) equalized opportunity also quite popular but from a different family of measure. We define all measures via the predicted values of the classifier $\hat{Y}$ and the protected attribute A. We identify the disadvantaged group with $A = 1$ and the privileged group with $A = 0$. The similarities of predictions are described for $\hat{Y} = 1$.

Since the focus of most fairness measures is on the disadvantaged group having fewer (desired) opportunities, $\hat{Y} = 1$ is generally the desired outcome. Measures usually come in two forms. The first one expresses the requirement that the predicted values of the classifier $\hat{Y}$ conditioned on the protected attribute be equal [9]. Most of the time, this strict equality in the outcome predictions between both groups is unrealistic. The *relaxed* form accepts the previous equality not to be strict within a range that is to be determined. For instance, in a legal setting, the US Equal Employment Opportunity Commission (EEOC) uses the Demographic Parity ratio (as defined by Def. 2.1 and 2.2) with $\tau = 0.8$ ("*80% rule*" [21]), stating that disparate impact caused by employment-related decisions or structures can only be ascertained if $DPR \leq 0.8$.

Definition 2.1. Demographic parity (DP). DP is the equality or similarity of prediction outcomes as an absolute difference [16, 36]:

$$DP = \left| P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1) \right| \leq \epsilon. \quad (1)$$

Definition 2.2. Demographic parity ratio (DPR). DPR is the equality or similarity of prediction outcomes as a ratio:

$$DPR = \frac{P(\hat{Y} = 1 \mid A = 1)}{P(\hat{Y} = 1 \mid A = 0)} \geq \tau. \quad (2)$$

Demographic parity has received some criticisms, since the measure does not necessarily report on what many would define as fairness [16]. This issue stems from ignoring both the true outcome and individual merits. For instance, consider a selection procedure where two individuals apply and both belong to the same protected group. The entity that has to make the selection needs to select one individual from this

protected group, regardless of their qualification, in order to achieve demographic parity. Let say that one individual is qualified (*i.e.*, with high chances to get a positive true outcome $Y = 1$) and the other one is not. Deliberately selecting the unqualified individual would not be considered as fair regarding the other (qualified) applicant; however, the entity would still satisfy demographic parity. So these *token* individuals are not guaranteeing fairness since qualified individuals from the protected group are still mistreated.

Addressing the criticisms of demographic parity, Hardt et al. [26] presented two other metrics that extend the aforementioned ones. By including the true outcome Y, the authors show that this variable can serve as a *justification* for the predicted outcome. For example, in the case of COMPAS, this is the recidivism rate as measured by violent crimes in a two-year window. Conditioning by the true outcome is a justification that the authors consider to be a suitable interpretation of the *task-specific similarity measure* from Dwork et al. [16], which can otherwise be difficult to come up with. This is also very similar to *disparate mistreatment* [3, 41] used as an evaluation metric by Adel et al. [1].

Definition 2.3. Equal opportunity (EO). EO requires an independence $\hat{Y} \perp\!\!\!\perp A \mid Y$ of $\hat{Y}$ and A conditioned on the true outcome Y. Expressed as a difference, this yields:

$$\left| P(\hat{Y} = 1 \mid A = 0, Y = 1) - P(\hat{Y} = 1 \mid A = 1, Y = 1) \right| \leq \nu. \tag{3}$$

"Equality of opportunity" is satisfied if $\nu = 0$, and larger absolute values are indicative of unfairness in the model or data.

2.3 Fair Neural Networks

Several works tried to reduce the impact of data imbalance [20, 30] and data representations over ML models [24, 27]. In addition, fair models have been studied for a variety of learning algorithms, such as Naive Bayes classifiers [9] or SVMs [42]. Nowadays, the focus is also on neural networks due to their prediction performance [1, 31, 37].

Some researchers mitigate unfairness in neural networks using white-box adversaries [1, 17, 31, 37, 44]. In all these instances, a new model architecture is proposed with two goals: (i) predicting the main attribute Y (which we will refer to as the *utility of the model*; with $Y = 1$ being the positive outcome); (ii) not being able to predict the protected attribute A (with $A = 1$ considered as belonging to the protected group). The joint goals can be formally defined as a min max optimization problem [17] over the loss function L, i.e., $\min_\theta \max_\phi L(\theta, \phi)$, with an adversary ϕ and an encoder with parameters θ. We use this representation to predict both Y and A via a white-box adversary and a neural network. Adel et al. [1], Ganin et al. [22], Raff and Sylvester [36] all proposed to optimize a variant of the following loss function following

$$L(\theta, \phi) = E_{\theta,\phi}(X, Y) - \lambda D_{\theta,\phi}(X, A), \tag{4}$$

with $D_{\theta,\phi}$ the loss for predicting A from X, and $E_{\theta,\phi}$ the loss for the target prediction Y also from X and λ a hyper-parameter.

Gradient reversal was introduced by Ganin et al. [22] for domain adaptation, and later adapted by Raff and Sylvester [36] and Adel et al. [1] who treated the protected attribute A as a domain label. The gradient reversal strategy assumes that multiplying by a negative sign will increase the loss $D_{\theta,\phi}(X, A)$ of the branch $h_a : X \to \hat{A}$ and yields a representation X^* that is maximally invariant to changes in A [1, 36].

Using gradient reversal for fairness is based on the intuition that the inability to predict A is a suitable fairness goal. This differs slightly from the fairness evaluations presented in Section 2.2, but a similar loss function from Equation 4 based on demographic parity led to the architecture of FAD [1], which leverages gradient reversal specifically for fairness.

However, there is no guarantee that gradient descent with flipped gradients does guarantee the maximal invariance required for fairness. In the worst case, maximizing the loss $D_{\theta,\phi}(X, A)$ can even result in the opposite optimum for the shared layers with regard to A, because flipping the gradients with regard to A makes it perform gradient ascent for A. With the shared layers performing gradient ascent w.r.t. A followed by gradient descent in the adversarial branch, this creates a discrepancy between the parameters defining both components for predicting A. This means that the model is not only not maximally invariant on the last shared layer, but that the shared layers are still explicitly learning to predict the protected attribute A.

This is one of the major limitations of using GRL for fair models, as predictions of main attribute Y are not made on 'fair' representations. Elazar and Goldberg [18] made an empirical observation on *leakage* of protected attributes specifically for text-based classifiers that can also be traced back to this. In Section 3, we clarify how our ethical adversaries framework mitigates this issue, thus allowing GRL to be used for training fair models.

3. ETHICAL ADVERSARIES

Our main contribution is a framework that joins evasion attacks (see Section 2.1) and fair neural networks (see Section 2.3) to improve the overall fairness of the system. Thus, it relies on two types of ethical adversaries: (i) a *Feeder* that uses evasion attacks to create examples highlighting unfair representation of a certain population and (ii) an *adversarial Reader* that tries to predict the protected attributes of interest (age, gender, race, etc.). In addition of exhibiting fairness issues in the data and in the trained model, our framework leverages gradient reversal to minimise the ability of the reader to guess protected attributes ultimately yielding a fairer ML model without sacrificing utility.

Figure 1 presents the global architecture. Our network follows a typical architecture with a GRL (discussed in Section 2.3 and is represented by the Reader). The Feeder, on the left part, performs evasion attacks as discussed in Section 2.1. Both adversaries interact with each other in an iterative manner, forming the main difference between our framework and GANs [23]. To achieve better fairness and utility outcomes, our process –that consists of two steps– can be performed multiple times.

The first step starts with a trained neural network (target label in Figure 1) predicting a main attribute Y. In this network, the adversarial Reader adds a second branch that tries to predict a protected attribute A while the gradient reversal layer strives to minimise the confidence of the Reader to predict A. Additionally, as we discussed in Section 2.3,

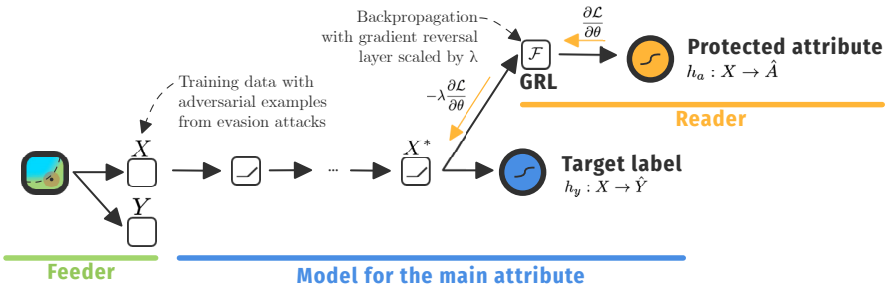


Figure 1: Ethical adversaries architecture: adversarial feeder on the left, and integrated adversarial reader on the right.

during the backward pass, a hyperparameter λ contributes to prioritize the utility versus the adversarial branch of the network. The model is trained with the joint loss of the original prediction target and the protected attribute.

In a second step, the Feeder, on the left part, performs evasion attacks as discussed in Section 2.1. The Feeder creates a set of adversarial points from an approximation of the target model, a.k.a a surrogate model, that is constructed on the same dataset as the model under attack. Our surrogate model is an SVM which it uses a radial basis (RBF) kernel function to cope with different level of model complexity. We selected this kernel since preliminary results on COMPAS showed that it is expressive enough and Biggio et al. [7] detailed how evasion attacks can be directly applied to SVMs with RBF kernels. The Feeder performs multiple evasion attacks on the surrogate function to generate adversarial examples that are similar to the training examples, but are wrongly classified.

For each iteration of this two-step process, adversarial examples are generated and included in the training set for adversarial retraining. Each adversarial example is added to the training set with the same label as the original example from which it was generated. The effect of the ratio of adversarial points in the dataset—the adversarial fraction—is further analyzed empirically in Section 4.3.

The Feeder is constrained to a maximum perturbation, so the adversarial examples that are generated are still similar to the original training examples. The perturbation distance d_{max} is the same for all input features and is applied after normalizing the data. Because the adversarial examples are within d_{max} of the training examples and thus similar, we assume all generated examples are also valid examples. The Feeder is ambivalent to possible—hard and soft—constraints. We opted to keep the Feeder ambivalent to eliminate another possible source of unfairness. We assume that the data on which the models are trained exhibits unfair patterns, something which we aim to correct with the Ethical Adversaries framework. It is therefore challenging to introduce constraints that would limit that adversarial examples to a limited set, especially when unfair patterns could discriminate minorities. However, we recognize that this unconstrained approach can be at odds with business logic and future work could focus on including constraints.

In terms of performance, constructing a surrogate classifier is the limiting factor. Using SVMs implies that the time complexity of the entire framework is $\mathcal{O}(n^3)$ with n the number of data points. The impact of adversarial attacks is linear on the overall complexity. But we should notice that adversarial retraining may drastically increase time to compute a separating function since included adversarial examples make the separation more difficult to find, or on the contrary, may not affect the function at all, if too few adversarial examples are included.

Both reading and feeding steps are run successively until we achieve better fairness and utility outcomes, which we demonstrate in Section 4.4. A key benefit of this process is that we prevent the Reader from learning biased representations, since these features cannot be used as proxies for the protected attribute anymore.

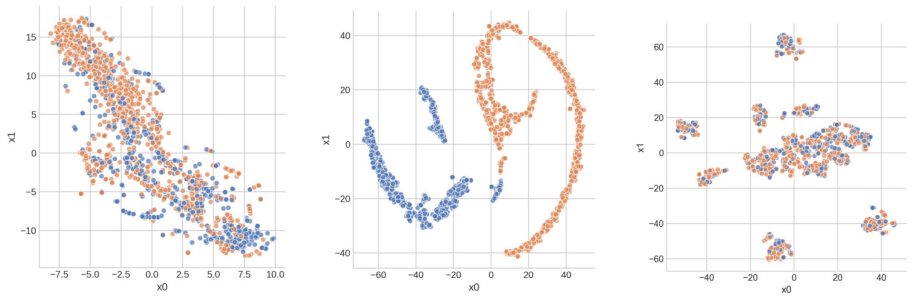
4. EVALUATION

We evaluate our model on three popular datasets: COMPAS [2], German Credit, and the Adult Census [28]. The COMPAS dataset was originally a sample of outcomes from the COMPAS system that predicted the risk of recidivism. This caused a debate about whether or not this score was disadvantaging African Americans [2, 10, 11, 14]. The dataset, therefore, includes the race of individuals. In line with previous research [1, 2, 42], we will only use individuals from *Caucasian* or *African-American* descent. As other groups are clearly less represented (e.g., only 31 instances for people of Asian descent), this raises issues during training and evaluation. It implies that there are minorities that are excluded from many studies; more datasets would be needed to study whether patterns of unfairness are similar and mitigation measures can be transferred, or whether these affect different demographics differently. COMPAS is composed of 5,278 instances and represented by 12 features. The target variable is whether a person has recidivated within two years. The race is used as a protected attribute.

The Adult dataset gathers 32,000 instances represented by 9 features. We use gender as a protected attribute and the binary target variable is income, whether someone earns more than 50,000 USD. German Credit is the smallest dataset, with only 1,000 instances and 20 features. There is a class imbalance, with 70% of all samples good credits and only 30% bad credits. The protected attribute is age, with a threshold at 25 years.

For reproducibility purposes, we have publicly released our code and provided users with a template that they can incorporate in their projects. It is compatible with all PyTorch models with only minor modifications, i.e., adding an adversarial branch and replacing the training loop. We recall that we have used the secML package[1] (v0.11) for running evasion attacks.

[1]https://secml.gitlab.io/



(a) Naive model (b) Model trained with a GRL ($\lambda = 50$) (c) Model trained with our framework

Figure 2: T-SNE dimensionality reduction of the activations in the last hidden layer on the held-out COMPAS test set. Distinct colors are used for the reported race of individuals in the dataset: either African-American ● or Caucasian ● .

4.1 Training setup

The model under attack. We start from a neural network of 3 hidden layers with 32 hidden units for COMPAS and German Credit and 128 for Adult, due to its larger encoded input. Each of the hidden units has a ReLU activation. This activation function is computationally efficient and mitigates the issue of vanishing gradients since the function never saturates, which makes it one of the most popular activation functions. For the output units, a softmax activation was used to get the classification and a linear activation for COMPAS. The network, including the adversarial reader, is trained with the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.9999$ and an initial learning rate $l_r = 0.01$, which is adjusted by a factor of 0.1 when reaching a plateau.

The adversarial reader. The adversarial reader is part of the model under attack and therefore follows the same training regime. The joint loss follows Equation 4 by including the GRL. The individual losses for both h_A and h_y are binary cross-entropy loss, except for COMPAS. In that case, the risk score is predicted as a regression problem with the MSE loss and then thresholded at 4 (low *vs* medium and high risk).

The adversarial feeder. In our setting, we can use the same training set for both the feeder and reader since they are part of the same, unique architecture.

We also approximate—relying on the earlier discussed transferability of attacks [13]—the attacked model by an SVM with a radial basis function kernel. We set the hyperparameters C and γ with a grid search with a reduced number of values: respectively $\{0.0001; 0.001; 0.01; 0.1; 1.0\}$ and $\{0.01; 0.1; 1; 10; 100; 1000\}$. We performed 10-fold cross validation.

4.2 Mitigating unfair representations

For each individual for the COMPAS test set, all three models derive a representation in the last hidden layer, on which we applied a t-SNE dimensionality reduction for a two-dimensional visualisation.

The model without fairness constraints (Figure 2a) has slight separation with regard to the protected attribute, but it is clearly separable in the representation from the model trained with a GRL (Figure 2b). This is also shown by retraining a one-layer perceptron on these representation. The model that was originally trained to predict only recidivism could be used to classify the protected attribute race with $AUC = 0.71$. The adversarial branch h_a that was trained simultaneously has an $AUC = 0.44$ As we mentioned earlier, this branch can be limited in predicting the protected attribute A. Which is the case here, as an independent perceptron has $AUC = 0.92$.

Here, we demonstrated that the hidden representation obtained by gradient reversal, not only still contains information about the protected attribute, but contains a stronger signal. Our architecture that joins 'adversarial fairness', also called the Reader, and 'adversarial learning', or the Feeder, (see Figure 1) leverages utility- *and* fairness-focused methods in a better way than the modification of the model alone. By injecting noise with the adversarial Feeder, our framework successfully mitigated this unfair representation, as shown in Figure 2c.

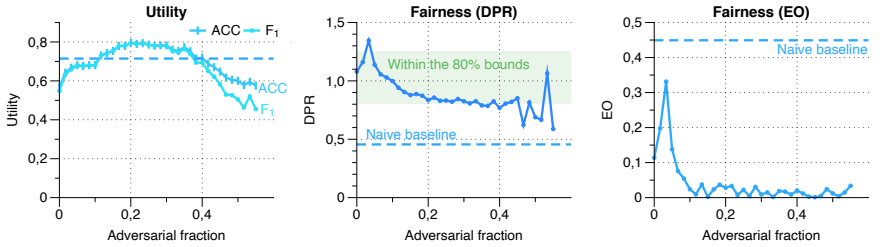
4.3 Effect of adversarial fraction

Figure 3 displays the effect of the adversarial fraction in the training dataset on COMPAS. When adversarial examples (equivalent to 25% of the training set size) are added to the training set, the utility is maximal. With higher fractions, the utility decreases and the development of the DP ratio fluctuates. This could stem from the minimax formulation, where a small fraction (i.e., 25%) helps optimize better for this saddle point, but higher fractions only add noise. We use this fraction for all further experiments, in future work this could be automated with a custom stopping criterion.

4.4 Benchmark results

Table 1 presents our results on the three datasets. We compare them with (i) a baseline without fairness goals, *i.e.*, a neural network without any particular control on fairness aspects, (ii) a re-implementation of the GRL [1, 22, 36] and (iii) the reported results from other works that incorporate fairness and cover a wide range of learning algorithms: Naive Bayes [9], random forests [37], SVMs [42] and neural networks [36, 43]. The models' utility was evaluated by binary classification accuracy and macro-averaged F_1 score; the latter highlights some issues when dealing with class imbalances, as is the case for German Credit.

Fairness is evaluated with demographic parity, both as an absolute difference (DP) and as a ratio (DPR), and equal opportunity (EO). Adel et al. [1] also report results on



(a) Utility (b) DPR (c) EO

Figure 3: Fairness and utility measures after each attack iteration on COMPAS (Batch size of 1024, $\lambda = 100$, epochs=100, 50 adversarial points per iteration)

Table 1: Results on the three datasets. An obelisk (†) show results reported by original papers. Results of classifiers without fairness constraints are reported as a baseline. Best results are in bold typeface. An asterisk (∗) indicates a division by zero.

Model	ACC	F1	DP	DPR	EO
Adult					
Baseline without fairness constraints	**0.839** ± 0.009	**0.763**	0.173	0.296	0.096
GRL	0.612 ± 0.012	0.518	0.059	1.931	**0.061**
NBF (NB) [9]	0.773†	—	**0.000**†	—	—
NBF (EM) [9]	0.801†	—	0.001†	—	—
Grad-Pred [36]	0.754†	—	**0.000**†	—	—
FF [37]	0.753†	—	**0.000**†	—	—
LFR [43]	0.702†	—	0.001†	—	—
Ours	0.814 ± 0.009	0.689	0.031	**0.784**	0.179
German Credit					
Baseline without fairness constraints	0.705 ± 0.063	0.624	0.018	0.929	0.198
GRL	0.710 ± 0.063	0.415	**0.000**	∗	**0.000**
Grad-Pred [36]	0.675†	—	0.001†	—	—
FF [37]	0.700†	—	**0.000**†	—	—
LFR [43]	0.591†	—	0.004†	—	—
Ours	**0.730** ± 0.062	**0.640**	0.006	**0.971**	0.175
COMPAS					
Baseline without fairness constraints	0.715	0.709	0.466	2.192	0.449
GRL	0.567	0.549	0.057	**0.926**	0.114
COMPAS risk predictions [2]	0.655 ± 0.029	0.654	0.289	1.829	**0.000**
Preference-based fairness [42]	0.675†	—	0.380†	—	—
Ours	**0.794**	**0.793**	**0.026**	0.840	0.008

both COMPAS and Adult but use a different setup for the Adult dataset. For COMPAS, the reported results (as well as their unfair baseline) are significantly higher than in our experiments, which we could replicate only when classifying high-risk individuals. To make a meaningful comparison, we also include our replication of *FAD* [1] as *GRL*.

The utility of our framework is the highest on the German Credit and COMPAS datasets, even surpassing the baseline model. On Adult, we achieve the highest utility of any model with fairness constraints. These results show that our model has only a very limited impact on the utility of the classifier, and it can even contribute to the training as shown in Figure 3. Note that on German Credit, a majority classifier would achieve 70% accuracy already, hence the inclusion of the F_1 score.

Regarding fairness evaluation, our framework gives the best results for COMPAS when considering DP. It also increases fairness as measured by DPR, which is the only one of the considered measures that indicates the "direction" of unfairness. More fairness is sometimes given by an *increase* towards parity (DPR=1) for the disadvantaged group: for the German Credit dataset, their chances of getting a loan increase. In COMPAS, the baseline has a EO of 2.192, the "bias against blacks" [2] *decreases* substantially with our model. For GRL, the near-equality of DPR (0.926) appears fairer, but this is not the case for DP and EO, where we observe an EO of 0.449 for GRL versus 0.008 for our model.

Figure 5 also reports the accuracy on COMPAS, however in function of the number of hyperparameter trials to illustrate how randomized hyperparameter assignments will affect the results [15]. The hyperparameter λ was varied between the interval $[0.1, 200]$ and the batch size selected from the set $\{256, 512, 1024, 2048\}$. Based on these results in relation to both (i) a model without any fairness constraints and (ii) a model with a GRL, we can conclude that our framework does perform better on this task compared to GRL. In ad-

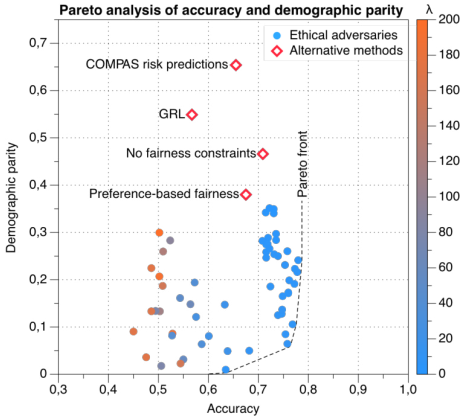


Figure 4: Pareto front of the utility, measured by accuracy, and demographic parity (lower is better) for COMPAS.

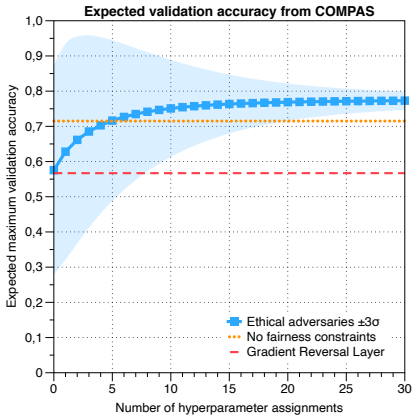


Figure 5: Expected maximum validation accuracy in function of the number of hyperparameter trials for COMPAS.

dition, given a sufficiently large compute budget it will even surpass the naive model without fairness constraints, since our model likely benefits from additional training examples. Nevertheless, it remains sensitive to optimal or suboptimal hyperparameter assignments, as is indicated by the large confidence interval. This sensitivity is also confirmed by the Pareto front in Figure 4, where many trials with large (> 50) values for λ performed quite poor on both utility and fairness.

5. CODE

We release an open source implementation— under the MIT licence—of our framework at: `https://github.com/iPieter/ethical-adversaries`

6. TOWARDS DOMAIN KNOWLEDGE INTEGRATION

While adversarial machine learning has been applied to the world of images and videos, we applied it in a different context. Our experiments aim at using evasion attacks to attempt generating new instances that represent personas, profiles or persons' representations. This context therefore forms a new domain space bringing its own set of constraints and value boundaries. In this section, we discuss the challenges this context brings.

The first challenge relates to *constraints' heterogeneity*. Indeed, in the image domain, the representation of pixels is homogeneous and thus easy to constrain while running evasion attacks. Yet, when it comes to different features representing categorical information, the problem is more complex.

Consider age category or highest obtained diploma, this information is part of the global description of the profile of a person and are usually categorical as only one choice is important/allowed. Some machine learning algorithms (*e.g.* SVMs) do not support easily such kind of categorical feature. Still, to be able to use such algorithms with this kind of features, categories can be decoupled into different features so that are mutually exclusive. For instance, regarding the highest diploma, one would create the following features: "has a Bachelor's Degree","has a Master's Degree", "has a PhD Degree", etc. While evasion attacks will bring perturbations to features independently, one needs to check the validity of the attacks [40].

There are different strategies: check constraints after each iteration and reject the latest modifications if any of these constraints is violated, wait until the attack has finished to reject the example in case of constraint violation, or even tradeoffs between these two extreme strategies. In this paper we did not applied a constraint enforcement strategy (see Section 3), but rather focused on limited perturbations that would generate similar examples to instances in the datasets, therefore likely to be valid. A stricter constraint enforcement strategy is left for future work.

The second challenge is linked to the *prior domain knowledge* that should also be taken into account in order to make a generated example *realistic*. For instance, in the context of the German credit dataset, it would be unlikely (and most probably illegal) that a kid (let's say under 10 years old) surfs to an online banking website to ask for a credit. We can easily transform this prior legal knowledge into a hard constraint and therefore consider that case invalid. As previously discussed, such constraints differ from the image domain. We would need the help of solvers which can take time to run and check all of them. Furthermore, some engineering may be needed to transform features' value into interpretable constraints for solvers and then back to the feature space. We were able to check some boundary constraints for features' value, but incorporating richer domain-knowledge is left to future work.

Our kid would be even more unusual if they additional stated that they had a PhD. While the fictional Sheldon Cooper from the Big Bang Theory show got his PhD at the age of sixteen and there exist actual cases of people obtaining their PhDs at very young ages[2], these are the kind of examples we may still want to avoid since there not representative. Because we relied on machine learning and statistical approaches, we did not see the case of one of our attack generating a kid holding a PhD. This can be explained by the fact that no examples were given looking alike it, thus the data distribution of the examples are not showing that this

[2] `https://en.wikipedia.org/wiki/Kim_Ung-yong`

is a possibility (the stochastic gradient descent procedure has no interest in trying to generating examples toward this direction of the feature space).

7. CONCLUSION AND FUTURE WORK

In this paper, we presented a novel architecture for integrating fairness constraints in machine learning models. Our architecture consists of two adversaries: (i) an adversarial reader that evaluates fairness constraints during model training and attempts to enforce them, and (ii) an adversarial feeder that performs iterative evasion attacks to discover previously uncovered regions in the input space. We evaluated our architecture on three well-studied datasets and showed that it can deliver high utility to models while satisfying fairness constraints. On COMPAS, we illustrated that our architecture yields a model that surpasses an unfair baseline regarding the utility (accuracy and F_1 score) and fairness. We provide evidence that gradient reversal alone is not sufficient (it might even be detrimental) but that our combination of adversaries leads to intrinsically fairer models.

There is room for future work. First, we may optimize the runtime execution of the technique via faster learning of surrogate models. Second, we could use the target model directly instead of a surrogate classifier to support adversarial attacks and assess if transferability properties hold for fairness constraints. This requires heavyweight modifications of the secML framework to allow multiple output values in neural networks. Third, one could define constraints involving multiple features. Enforcing these *domain-specific* constraints during attack generation raises questions on the representation of the feature space and optimal convergence of the algorithms. Fourth, our framework is evaluated against allocational harms. More subtle differences— like a difference in the model's performance—are also affecting social groups. With some minor modifications, we suspect that these types of unfairness can be addressed with our framework. Finally, we would like to generate the most dissimilar examples possible to ensure good coverage of the unseen feature space with a minimal number of attacks.

Acknowledgements

Pieter Delobelle was supported by the Research Foundation - Flanders (FWO) under EOS No. 30992574 (VeriLearn). Pieter Delobelle also received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. Paul Temple is also supported by the EOS VeriLearn project (Fonds de la Recherche Scientifique, FNRS). Gilles Perrouin is an FNRS Research Associate. We also want to thank the secML developers from the PRALab (Pattern Recognition and Applications Laboratory, University of Cagliari, Sardegna, Italy) for having answered our numerous questions and helping us in using their newly developed library.

References

1. Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-Network Adversarial Fairness. In *AAAI Conference on Artificial Intelligence*, 2019.

2. Julia Angwin and Jeff Larson. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

3. Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

4. B. Biggio, L. Didaci, G. Fumera, and F. Roli. Poisoning attacks to compromise face templates. In *2013 International Conference on Biometrics (ICB)*, pages 1–7, 2013.

5. Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition Journal*, 84:317–331, 2018. ISSN 00313203. doi: 10.1016/j.patcog.2018.07.023.

6. Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, pages 1467–1474, 2012.

7. Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *ECML/PKDD*, pages 387–402, 2013.

8. Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv:2005.14050 [cs]*, May 2020.

9. Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, 21(2):277–292, 2010. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-010-0190-x.

10. Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv:1703.00056*, 2017.

11. Sam Corbett-Davies and Sharad Goel. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023*, 2018.

12. Pieter Delobelle, Paul Temple, Gilles Perrouin, Benoît Frénay, Patrick Heymans, and Bettina Berendt. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning, 2020.

13. Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In *28th USENIX Security Symposium*, pages 321–338, 2019.

14. William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. 2016.

15. Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics.

16. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255.

17. Harrison Edwards and Amos Storkey. Censoring Representations with an Adversary. *arXiv:1511.05897*, 2015.

18. Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, 10. ACL. doi: 10.18653/v1/D18-1002.

19. Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway Feedback Loops in Predictive Policing. *arXiv:1706.09847*, 2017.

20. Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, pages 15–19, 2020.

21. Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *21th ACM SIGKDD International Conference*, pages 259–268, 2015.

22. Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

23. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *NIPS*, pages 2672–2680. Curran Associates, 2014.

24. Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *AAAI*, volume 18, pages 51–60, 2018.

25. Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. Knowl. Data Eng.*, 25(7):1445–1459, 2013. doi: 10.1109/TKDE.2012.72. URL https://doi.org/10.1109/TKDE.2012.72

26. Moritz Hardt, Eric Price, ecprice, and Nati Srebro. Equality of Opportunity in Supervised Learning. In *NIPS*, pages 3315–3323. Curran Associates, 2016.

27. Gareth P Jones, James M Hickey, Pietro G Di Stefano, Charanpal Dhanjal, Laura C Stoddart, and Vlasios Vasileiou. Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *arXiv preprint arXiv:2010.03986*, 2020.

28. Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

29. Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. POTs: Protective Optimization Technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 177–188, January 2020. doi: 10.1145/3351095.3372853.

30. Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems*, 33, 2020.

31. David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning Adversarially Fair and Transferable Representations. *arXiv*, abs/1802.06309, 2018.

32. Rebekah Overdorf, Bogdan Kulynych, Ero Balsa, Carmela Troncoso, and Seda Gürses. Questioning the assumptions behind fairness solutions. *arXiv:1811.11293*, 2018.

33. N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy*, pages 372–387, 2016.

34. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Asian Conference on Computer and Communications Security*, pages 506–519. ACM, 2017. doi: 10.1145/3052973.3053009.

35. Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting discrimination. In Amos Storkey and Fernando Perez-Cruz, editors, *21st International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 96–105, Playa Blanca, Lanzarote, Canary Islands, 2018. PMLR.

36. E. Raff and J. Sylvester. Gradient Reversal against Discrimination: A Fair Neural Network Learning Approach. In *IEEE 5th International Conference on Data Science and Advanced Analytics*, pages 189–198, 2018. doi: 10.1109/DSAA.2018.00029.

37. Edward Raff, Jared Sylvester, and Steven Mills. Fair forests: Regularized tree induction to minimize model bias. In *Conference on AI, Ethics, and Society*, pages 243–250, 2018.

38. Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness GAN: Generating Datasets With Fairness Properties Using a Generative Adversarial Network. *IBM Journal of Res. and Dev.*, page 12, 2019.

39. David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. *arXiv preprint arXiv:2004.07401*, 2020.

40. Paul Temple, Mathieu Acher, Gilles Perrouin, Battista Biggio, Jean-Marc Jézéquel, and Fabio Roli. Towards quality assurance of software product lines with adversarial configurations. In Thorsten Berger, Philippe Collet,

Laurence Duchien, Thomas Fogdal, Patrick Heymans, Timo Kehrer, Jabier Martinez, Raúl Mazo, Leticia Montalvillo, Camille Salinesi, Xhevahire Tërnava, Thomas Thüm, and Tewfik Ziadi, editors, *Proceedings of the 23rd International Systems and Software Product Line Conference, SPLC 2019, Volume A, Paris, France, September 9-13, 2019*, pages 38:1–38:12. ACM, 2019. ISBN 978-1-4503-7138-4. doi: 10.1145/3336294.3336309. URL https://doi.org/10.1145/3336294.3336309.

41. Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *26th International Conference on World Wide Web*, pages 1171–1180, 2017. doi: 10.1145/3038912.3052660.

42. Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From Parity to Preference-Based Notions of Fairness in Classification. *arXiv:1707.00010*, 2017.

43. Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

44. Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of Conference on AI, Ethics, and Society*, pages 335–340. ACM Press, 2018. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278779.

Blind Spots in AI: the Role of Serendipity and Equity in Algorithm-Based Decision-Making

Cora van Leeuwen
imec-SMIT-VUB
Pleinlaan 9
1050 Brussels, Belgium
cora.van.leeuwen@vub.be

Annelien Smets
imec-SMIT-VUB
Pleinlaan 9
1050 Brussels, Belgium
annelien.smets@vub.be

An Jacobs
imec-SMIT-VUB
Pleinlaan 9
1050 Brussels, Belgium
an.jacobs@vub.be

ABSTRACT

Decisions support systems (DSS) are used more and more to offer right information at the right time. Serendipity has been pro- posed to ensure that the experience is broad and engaging. However, only designing for serendipity might not be enough to avoid historical discrimination affecting your DSS. For this reason we argue to include equity when designing for serendipity.

1. INTRODUCTION

Managing knowledge is an important skill, it can reduce or enhance the power of individuals and organizations [44]. Nowadays, systems based on big data and algorithmic processing are increasingly applied to support knowledge management and resulting activities. For example, algorithmic matchmaking systems are being used to match job seekers and potential employers, and several approaches have been put forward to implement these complex tasks [3]. These decision support systems (DSS) are trained to offer the right information at the right time with high accuracy, making use of (big) data. However, we are increasingly becoming aware that there is a problem of bias and that it is a difficult problem to address. It emerges in the system before the data is collected as well as in the other stages of the deep learning processes [28]. The question of "how standards of unbiased attitudes and non-discriminatory practices can be met in (big) data analysis and algorithm-based decision-making" is consequently a timely one. Today, there are many tools focusing on technical solutions mitigating bias built up by choices in training data and modelling methods [39]. However, this only partially solves the challenge to enable the creation of both a performant fair and engaging, interesting system. After all, these tools focus on the known social or statistical biases. What with the unknown unknown, the blind spots? We agree with scholars like Reviglio [44] and Ge et al. [26] and argue that designing for serendipity could help to overcome these blind spots in DSS. However, we want to debunk the idea that designing for serendipity is a guarantee to develop a system free from any bias. We therefore argue that the implementation and operationalization of serendipity should take into account additional principles, such as equity. Hence, the particular question this paper aims to address, is why and how serendipity and equity can help to overcome blind spots in the design of algorithmic decision support systems (DSS). In the remainder of this paper, we will first elaborate on the notion of blind spots and discuss why we consider them to be a result of existing paradigms to reduce information overload. Second, we address the notion of serendipity to overcome these blind spots. Following this, we put forward three guiding principles to introduce equity in the design of DSS. Throughout this work, we demonstrate these principles by means of an hypothetical case example of a job-matching system that aims to help job seekers find interesting vacancies.

2. INFORMATION OVERLOAD

The increasing amount of available data and digital information systems provide great opportunities for these kinds of decision support systems. However, as discussed in related literature [29, 11], this also comes with the problem of information overload. Although existing recommender systems have shown to be an efficient remedy by applying personalization and filtering techniques, there are growing concerns about the potential drawbacks of these systems (e.g. [11]). Indeed, while the current paradigm has shown to be effective to reduce information overload, it is also being criticized to exploit convergent system behavior rather than cater divergent behavior [44]. This aligns with the emerging call from scholars for additional and/or alternative metrics to optimize these information systems beyond accuracy or relevance (e.g. [33]).

Perhaps the most contentious discussion in this regard is the one about filter bubbles in online (social) media. Here, the hypothesis is that algorithmic personalization focused on accuracy or relevance results in a diminished exposure diversity. The latter implies that users are only being exposed to information that confirms their beliefs or properly aligns with their preferences. In this way, users of the system are literally blind to any information outside their bubble. While this is an useful feature in some cases - a dog owner does not want to get recommendations for cat food - in other situations it might be detrimental. For example, when you look online for information about a particular topic, e.g. the usefulness of vaccines, you might only find information that confirms your existing beliefs and thus result in a confirmation bias [30].

While this is a striking example that many like to associate with notable events in our contemporary society, the problem of information filtering and algorithmic curation exceeds this single application domain. Indeed, as decision support systems are increasingly being used in healthcare, financial

systems or juridical settings, these domains suffer from these biases too. Moreover, these blind spots are not only due to the filtering techniques themselves. They also result from the available information in the first place: the data sources and training data. In the case of job matching systems, for example, there is a recurrent observation that women are more often shown vacancies for part-time jobs, even though gender is not a variable that is taken into account in the model [51]. This appears to be, however, due to the actual situation in the job market in which women seem to more often apply to part-time jobs and consequently this behaviour is reinforced in the matchmaking model. The question then arises about how to deal with this: should the system simply keep this imbalance or should it try to remediate? And if we decide upon remediating, how can we make sure the system indeed includes items that might score less on relevance or accuracy, and thus represent the blind spots? The challenge is hence how to design these systems with attention to these blind spots? In the next section we illustrate how the concept of serendipity could be a first step in that direction.

3. SERENDIPITY AGAINST BLINDSPOTS

3.1 Serendipity in digital environments

The idea of having a set of guidelines that helps us to discover the unknown sounds appealing to many. Unexpected discoveries are a key driver for innovation and growth, and the study of how people encounter information accidentally has therefore been an important field in information science over the last decade. In this domain, the notion of serendipity is used to refer to the "unplanned ways to encounter resources that we find interesting" [10, p.7] . Serendipity is associated with a line of groundbreaking discoveries such as Alexander Fleming's discovery of Penicillin or Archimedes' principle. Indeed, this 'eureka moment' reflects the key characteristic of serendipity: an unexpected discovery.

Despite the importance of serendipity in epistemology, more recently the concept started to gain attention in digital information environments as well. Indeed, while the Web could be considered to be the ultimate serendipity engine [32], the problem of information overload and resulting algorithmic filtering techniques seem to diminish this serendipitous potential. Scholars therefore argue to apply serendipity as a key design principle for digital environments [44]. Serendipity is indeed concerned with discovering the unknown unknown and has been proposed to be applied in recommender systems to improve their quality and resulting user satisfaction [26]. Similarly, serendipity can play an important role in decision support systems to overcome the previously described blind spots. For example, adding serendipity to a job matching system, might result in recommending vacancies that the job seeker would never have thought of by him/herself. However, the question remains how to design these serendipitous encounters?

3.2 Designing for serendipity

While the notion of serendipity has gained attention in several scientific disciplines, ranging from psychology to sociology of science and computer science, there is a recurrent misconception that turns it into an ill-defined buzzword [45]. Indeed, serendipity is often referred to as a happy accident, an exceptional situation in which the right person was in the right place. However, it is important to understand that serendipitous discoveries are more complex: they are the result of a favourable combination of an individual's characteristics and so-called environmental affordances. The latter are "opportunities for action offered by the real world" [47, p.117] that allow individuals to engage in information-seeking activities leading up to serendipitous encounters.

3.3 The key affordances of serendipity

This approach allows one to overcome the paradox of 'designing accidents' as we will not design serendipity itself, but design for serendipity and thus a so-called serendipity potential. As mentioned before, there is an emerging line of research focusing on cultivating this serendipity potential in digital environments and stresses its importance and ethical value [43, 10, 44] . Björneborn [10] identifies three key affordances for serendipity: diversifiability, transferability and sensoriability. These affordances represent the "key aspects of human interactions with environments" [10, p.7]. The first one, diversifiability, refers to the extent to which such an environment can be diversified in terms of content. In the case of our job matching system, this would mean a set of recommended job vacancies across different sectors and/or with multiple modalities (both part-time and full-time for example). In this same example, the notion of transferability would refer to the capacity of the system to explore the possible vacancies. How easy can one navigate in between several vacancies; are there multiple ways to end up with one particular vacancy (e.g. through multiple search words), etc. For example, a vacancy for a 'store assistant' could be retrieved both when looking for jobs in retail as well as jobs with people. This illustrates one of the benefits of having divergent systems because they allow to create new semantic knowledge, in this case for the end-user. Finally, sensoriability deals with the sensory stimuli in the environment. In digital environments, this often refers to a combination of images, colored hyperlinks, explicit keywords, notifications or particular suggestions that are displayed.

3.4 Interaction of the system

An important additional aspect of these affordances is the identification of re- lated personal characteristics that are considered to be "the actoral components of these [environmental] affordances" [10, p.7]. More specifically, Björneborn [10] identifies three key personal factors that correspond to each of the previously described affordances: curiosity (diversifiability), mobility (traversability) and sensitivity (sensoriability). Without going into detail about each of these personal characteristics, it is obvious to also acknowledge the role of the user. This is important to keep in mind when thinking of the possible outcomes of decision support systems. Indeed, in our example of the job matching system, it is still up to the job seeker to engage with the proposed vacancies. Here, it has indeed been found that people's levels of extroversion and conscientiousness influence their job seeking attitudes [24].

3.5 It ain't a silver bullet

Although we argue that in the design of a DSS one should aim to incorporate these dimensions, we acknowledge that the application of this affordance reasoning is not straightforward. As with many principles, these concepts need to be operationalized and adjusted to the system's particular

context. In this operationalization, one should be equally aware of the potential biases that might enter in the design. Implementing these serendipity affordances is in no way an exemption to any other biases. In this paper, we therefore want to particularly pay attention to the affordance of diversifiability. After all, the mere requirement of having a diverse (information) environment, does not sufficiently take into account existing historic biases towards certain information sources, which is due to societal power dynamics very difficult to not replicate [21]. The above-mentioned example of women getting more part-time offers is such a bias, related to the gender stereotypical role of the women as the central home and childcare provider, where paid employment is assumed of second order importance. To overcome this historical bias, we suggest that the operationalization of diversity should take into account the notion of equity as it is more adequate to deal with this type of bias.

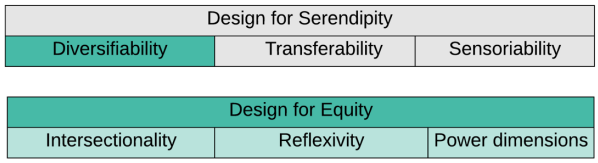


Figure 1: Conceptual connection between design for serendipity affordances and the dimensions to enable design for equity

4. MITIGATING HISTORIC BIAS

4.1 Historic bias & equity

Any kind of decision-support system relies on one or multiple models that are trained on data. A model is an abstraction of a process which makes predictions on historic data points [40]. Consequently, the predictions and recommendations are based on data that may contain (past) societal prejudices that were (unintentionally) encoded in these historic data points. This is called historical bias, it occurs when the world as it was or is influences the model to make unwanted biased outcomes [50]. Friedman and Nissenbaum [25] coined the term "pre-existing bias" to describe the same effects. According to them, this is a bias that can be traced back to institutions, practices and attitudes. An example of an historical bias can be found in Amazon's discontinued human resources machine learning algorithm. This was used to identify possible job candidates by predicting their success based on previous employees' success. The model is no longer in use as it downgraded CVs from women due to an inferred preference for male candidates [19]. Penā Gangadharan and Niklas [41] recognise that in order to resolve discrimination it is necessary to investigate the normative practices and institutions that have contributed to this difference in treatment, as technology is created within a society's laws and practices. This means that it is necessary to examine the socio-technical dynamics of gathering data by not discounting the historical choices that created the data [41,38]. One element that is proposed to stop discrimination and to obtain equality is to design for fairness.

Historical discrimination or data provenance (the availability of data) is sometimes explicitly ignored [14] when trying to achieve fairness in models. Gilbert and Mintz [27] have made the case that in order to counter bias in data it is important to place the data within a context. This is confirmed by Binns'[8] examination of the unfairness of discrimination in relation to political philosophy's role in algormorithic bias. He concluded that fairness needs to be contextualized in order to be truly fair. He proposes to use luck egalitarianism to achieve fairness, this doctrine states that to be fairly judged a person can only bear the burden of their own choices and those burdens caused by outside influences should be taken out of the equation. He acknowledges that this contextualisation of choices is not an easy task as what is truly a choice and what is a result of the historical circumstances is in most cases hard to distinguish. Research has been done to use machine learning to discover and use causal relationships that can be found in historical data to show the bias [37]. Cardaso et al. [36] examined if it was possible to use self created biased data to validate if there was bias in the system. What these two approaches have in common is that they use data to examine or contextualize historical inequality. A shortcoming here is that this can only be established when there is data available.

Currently many of the technical solutions to bias or unfairness are focused on creating fair outcomes by algorithm [1, 2, 22]. For example, they compare if all groups are treated equally by the model. This approach, however, does not address if it is fair to the individual to be judged on the basis of a group. Moreover, the studies focused on algorithmic fairness do not address that the perception of fairness might not be the same for everyone and that the promoted fairness is not determined within a vacuum. For example, Wang et al. [52] found that fairness perception differs greatly among the 579 participants of their experiment and is not easily determined or feelings of unfairness are not easily resolved. O'Neil [40] argues that fairness is hard to grasp for models and a focus on data means that unfairness is perpetuated and often unaddressed. She [40] witnessed this in her study of police predictive modeling and these findings are echoed by the ethnographic studies of algorithms by Eubanks [23]. Data collection, or lack there off, is inherently political and is part of structural oppression [21]. One example to illustrate both the need for an historical perspective and the influence of normative practices, is in the lack of data available on the lived experience of women [42], which has impacted many fields from urban planning to health. In the context of a DSS, this kind of historical bias limits the available data and resulting interpretations and support. To remediate this, it has been suggested to design for equity [16, 17, 21].

4.2 Designing for Equity

D'ignazio and Klein [21] argue that fairness as a concept is not sufficient to address inequality. This is because fairness is judged from a current moment in time without reflection on past advantages. Therefore it would be advantageous for those that have been privileged from the start. The decisions involved in creating fairness within DSS are numerous as fairness as a concept comes in many different iterations with each determining the type of fairness that is provided to the subject of a DSS (e.g. group fairness, individual fairness, group parity and others). Determining and documenting the choice in fairness alone does not solve bias. This is echoed by Hoffmann [31] who uses anti-discrimination discourse to explain that using fairness alone does not solve bias in technological solutions as it will replicate normative

structures and does not take into account the full lived experience of marginalized people. Equity on the other hand would take into account the context of a person which in turn enables an equitable outcome. The purpose of equity is to ensure that advantages and disadvantages are taken into consideration and to offer an alternative to the fairness principle. There have been initiatives to introduce what is called data justice into design practice. Data justice brings together multiple disciplines which are focused on the role of datafication within a wide variety of topics ranging from democratic procedures to the dehumanisation of decision-making [20].

When we look for inspiration on how to put this into practice, we can rely on the recent increased interest to integrate data justice in design and practice [16, 20]. The focus of these initiatives is to ensure that a plurality of voices and lived experiences are introduced within the design process, which is currently too focused on an universal experience [17]. Others have argued that integrating critical feminist theory originating from the social sciences in design of Human Computer Interaction (HCI) would ensure that there would be challenges to normative thinking [7]. According to Bardzell and Bardzell [7] critical feminist theory is uniquely qualified for this as resisting and critiquing the status quo is one of its tenets. Although the feminist theory originally set out from a gender inequality point of view, applying systematic reflexivity and a high awareness of historical power inequalities, created insight in the intersection of different inequalities coming along other social categories (e.g ethnicity, age, sexuality, ability, economical background) guiding social interaction. This lead to a body of work interested in investigating the different practices involved in creating organisations and technologies. Young [53] used their article as the basis to create a design practice for a feminist chatbot by proposing practical methods to include stakeholders within the creation process of a chatbot. Using insights from both these fields we argue to integrate the following into the diversifiablity affordance for serendipity: awareness of power dynamics, intersectionality and reflexivity.

4.2.1 Powerdynamics

Ignoring the structures of discrimination by focusing solely on single instances of blind spots disregards how these structures of power can have a wider impact [31]. D'ignazio and Klein [21] make use of the 4 domains of power (based on Patricia Hill Collins'[15] work) to explain structural oppression. First, there is the structural domain which organizes oppression via laws and policies. Second, the disciplinary domain which manages oppression by enforcing the laws and policies of the first domain. The third domain is the hegemonic domain which circulates oppression and creates acceptance via cultural activities and the media. And finally, there is the interpersonal domain which consists of the personal experience of oppression. Structural oppression can also result in a disproportionately privilege for a dominant group [21]. Robinson et al. [46] argue that disregarding structural racism while designing a system will result in it insidiously infiltrating your system on every level.

Applying these domains on the design for algorithm-decision making-processes in a labor context, there are laws and policies in place that are sensitive to equal access to jobs (structural domain of power). However, the disciplinary domain is often insufficiently equipped to enforce these laws and policies in daily practice. There are, however, cases where unequal access to jobs is challenged. It is on the third domain, the hegemonic domain, where a lot of change is still necessary. It is also in this domain a new DSS should be situated as it is a cultural expression captured in a technology: what is a good or a wrong decision. Finally, in the interpersonal domain, any oppression supported by the DSS is only shown to the person that is discriminated against. The end-users from the dominant group are not aware of their privilege, which thus remains a blind spot for them.

Designing for equity would mean to be aware of the power structures that made the decisions regarding the model and the manner of data gathering. In other words make visible who was allowed to be involved in the design process [16]. Once the power structures have been examined it becomes possible to challenge these decisions [21].

4.2.2 Complexity of intersectionality

Solely focusing on structural racism in the design of a DSS would ignore the intersectionality of most people's identities, and the related exclusion mechanisms - for example ageism, sexism, ableism, or marginalisation by wealth. The concept of intersectionality was proposed [18] to explain that race, gender and class do not operate in a silo. They intersect and on those points of intersection people's identity is constructed [16, 17]. According to Collins [15] people receive benefits and penalties based on their position on the intersections within systems of oppression. An example can be found in data sets created for training facial recognition software. Buolamwini and Gebru [12] found that the algorithm was unable to recognise black female faces because it was not only trained on a lack of black faces but it specifically lacked black female faces. There has been interest in re-examining the principles of fairness, for example, Burke [13] proposes multi-sided fairness. In this case, however, the multi-sided aspect is not considered within an individual itself as intersectionality proposes, but is related to fairness for multiple stakeholders within an algorithmic decision. Designing for equity would mean to design a system that is aware of the complex intersections of peoples' identity and to go beyond a universal design [17].

4.2.3 Reflexivity of data scientists

The third principle is reflexivity which is "the ability to reflect on and take responsibility for one's own position within the multiple, intersecting dimensions of the matrix of domination" [21, p.64]. Reflexivity is then achieved by being aware of your own position within the power dimensions and actively acknowledging the benefits and disadvantages of this position. The reflexivity achieves a transparency on the differences between data subject and data gatherer. This is important because it highlights possible gaps in the knowledge base of the data scientist.

5. IN (DESIGN) PRACTICE

On an abstract level, the trajectory of a DSS can be modelled as depicted in Figure 2. This trajectory is based on the simplified life cycle of AI as proposed by Binns and Gallo [9]. To further accommodate early consideration of ethical issues, we have elaborated the framework to reflect the necessity to consider the team who develops the AI. In addition, we have added a deeper layer to the training and test data phase to be able to differentiate between the different

steps that need to be taken to procure the training and test data. A lot of the blind spots are built up in this phase, but not all as discussed earlier. Applying ethical considerations at an early stage has been found to be the most beneficial and cost effective method [25]. It is therefore important to ensure that these phases can be described as detailed as possible. We will use this framework to place the different steps for integrating equity and serendipity within all the phases of the design process. While the role of diversifiability in serendipity has been acknowledged in the design of machine learning models [35], it is rarely considered as an important aspect of the design process itself. However, what is clear from the previous discussion, is that incorporating diversifiability should not be a mere feature of the algorithm itself. Diversifiability also emerges from related design activities such as data collection or developing measures of success because they are subject of the previously described design principles for equity.

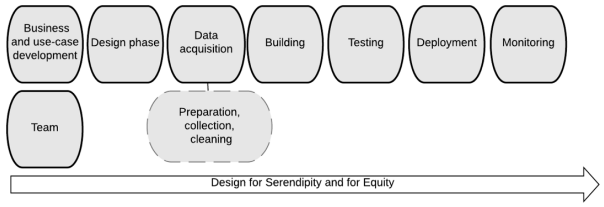


Figure 2: Based on the simplified life cycle of AI [9] an ideal typical process of the creation of a DSS is depicted, connected with the overall design for serendipity and equity as key drivers in the process

5.1 Living lab approach

We propose combining both the principles of data justice and critical feminist theory with a living lab approach. The latter is a research method that involves "multiple stakeholders, including users, in the exploration, co-creation and evaluation of (usually ICT-related) innovations within a realistic setting" [5, p. 1].

At the start of the project, in the business and use-case development phase, one assesses the stakeholders, the purpose of the DSS system and its context. Which stakeholders are under-served? A thorough stakeholder mapping should be conducted, keeping in mind the intersectionality of stakeholders' identity. In order to compensate for structural disadvantages, it will be necessary to examine where people might be penalized or privileged based on intersections of identity [15] . Moreover, it needs to be ensured that the system design is also informed by a representation (e.g. persona's) and involvement (e.g. user research, co-creation) of the most disadvantaged. In our earlier example of the job-matching system, those who have been identified as experiencing difficulties in being matched with a job should be involved in the design process. This will create additional semantic knowledge for the system to work with and consequently reduce potential biases. This could be, for example, people older than 50, educationally disadvantaged and minoritized people [51]. Involving them can be achieved by conducting user research and co-creation with representatives, not only focusing on the (historic and current) disadvantages, but also their strengths and experiences. During these sessions, system-designers should emphatically listen to the epistemic knowledge of stakeholders themselves, as they are uniquely qualified to critique the status quo [7]. Here, it is important to note that the involvement of these stakeholders needs to be a touch stone throughout the entire design process. Involving them throughout the project will ensure that their actual experiences impact the design, rather than the designers' interpretation of these experiences. This means that they would be involved in every stage of the life cycle (Figure 2). Another step where a living lab approach will benefit equity is within step 3 the data collection phase. The input is essential as interpretation of data can be tricky. Involving domain experts is essential to understand which data needs to be included to be able to make a decision. The involvement of marginalized stakeholders here would enable a new perspective on the data from a lived experience point of view. As said before the involvement of the stakeholders is throughout the lifecycle as they will be involved in the testing of the resulting model and can be involved in creating an inclusive deployment strategy.

The incorporation of equity within diversifiability can also be realized by examining the purpose of solutions together with the stakeholders identified as most impacted by the solutions. This means that the systems' beneficial impact on society is determined and possible harms for their peer group are identified. Furthermore, by interacting with these stakeholders we can learn from their experiences on how to avoid the harm and improve the solution for all. This includes an examination of the broader context by analysing the stakeholder mapping and the conducted sessions of co-creation and user research to present an overview of the possible positive and negative outcomes of the proposed solution. This can be implemented by conducting a domain analysis using classic scientific methods combined with the epistemological knowledge of the marginalized stakeholders.

5.2 Reflexivity

Next to listening and incorporating other viewpoints in the design process, there is also a need to focus on the design team. The design team should be aware of their own intersectional identities, values and position in society. Both on an individual team member level and as a group. This can be achieved by practicing reflexivity. In our example of labor mitigation, the team should first reflect on which intersection they would be placed (e.g. a team member could be a hetero-sexual 30 year old white man) and think about what this would mean for their own unconscious preferences and assumptions. Another aspect of this reflexive exercise involves studying how these intersectional identities reflect in or differ from the marginalized stakeholders. Subsequently, what kind of involvement of the marginalized stakeholders is needed to ensure that the design will have the desired impact. How are you going to realise this: an additional team member, a soundboard of experience experts, . . . Finally, if particular design choices are based on assumptions of use, these need to be disclosed in order to facilitate reflexivity on the possible impact of these assumptions later in the process. For example at the time of deployment or monitoring of the performance of the DSS. Leaflet or fact sheet approaches are commonly used as a tool to create that transparency [4, 49].

5.3 Challenges

These three principles of intersectionality, reflexivity and awareness of power dimensions help designers to ensure equity. Applying them also contributes to the diversifiability affordance that has been defined to design for serendipity. Indeed, as has been illustrated in the previous section, including diversity can only be considered as beneficial when it does not come at the cost of equality. We therefore argue that the practical methods put forward in this section should become an essential part of the design workflow of any algorithm-based decision-making system. In this way, they could be considered as complementary research approaches and practices to rather technical-oriented solutions that have been suggested before in order to deal with serendipity and diversity.

Although implementing these principles and corresponding methodologies sounds evident in theory, we acknowledge that there is still a lack of practical tools and procedural knowledge as to how to implement them in an actual design process. This is not only an open challenge related to the subjects presented in this work, but is applicable more generally to the discourse of ethical artificial intelligence. Recent works all explicitly point out that there is a need for domain appropriate ethical tools (e.g.[6, 39, 34]. From our experience as social scientists involved in the design of many of these systems, we believe that this challenge arises from (at least) two bottlenecks that need to be addressed collectively in order to be able to move forward. First of all, experience has taught us that ethical considerations do not have the attention that they should have to truly have ethics by design. They are seen as a nice to have instead of a key consideration throughout the design process. Secondly, there is a lack of a proper articulation of the actual steps that take place within the development of a DSS. This means that we do not know for sure if the trajectory in Figure 2 depicts a version of reality or is indeed solely a representation of an ideal. This hampers our ability to develop and provide more specific and concrete tools to implement the principles of serendipity and equity. We hope that the arguments in this paper demonstrate the importance of notions such as serendipity and equity in the design of algorithmic-based decision-making systems and invite scholars and practitioners from other domains to work collectively towards putting these principles into practice.

6. CONCLUSION

The availability of high volumes of data and intelligent decision-making systems present both opportunities and challenges to various actors. In the current research paradigm, the most important system objective of decision-making systems is accuracy. The relevance and applicability of the system is informally evaluated by the user and determines if they will continue to use application. While these systems allow for an enhanced informed decision-making process, the question arises to which extent the information they present is not flawed by biases and blind spots. In this work, we outlined how design principles based on serendipity and equity could help to re-mediate some of these weaknesses. The underlying rationale is that these principles will broaden the available information and semantic knowledge and in this way allow for divergent rather than convergent systems. We aimed to put forward a design rationale that would help to incorporate the principles of serendipity (diversifiability) and equity (inter- sectionality, reflexivity and power balances) in the development of DSS.

The main challenge, however, relates to the actual implementation of these methodological tools within an actual design process of a DSS. What is evident from the rationale presented in this paper, is the fact that design principles such as serendipity and equity shouldn't be limited to the activities related to the mere design or training of the system. Rather, it should be part of the entire trajectory from start to end, including testing and evaluating it continuously with several user groups. While we presented an ideal trajectory of this design and development process (cfr. Figure 2), we are aware of the fact that in reality this might not always be the case. Future research will need to be conducted to examine and elaborate on the process presented in figure 2 in order to present a realistic representation of the design of a DSS. We therefore call for an inter-disciplinary approach that considers these design principles such as serendipity and equity not merely as a nice to have, but as an essential component of the design of a DSS, and starting from this, discusses how these practices can be met in the actual design of these systems.

APPENDIX

A. ADDITIONAL AUTHORS

Additional authors: Pieter Ballon(imec-SMIT-VUB, email: `Pieter.Ballon@vub.be`)

B. REFERENCES

[1] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. arXiv preprint arXiv:1803.02453.

[2] Ajunwa, I., Friedler, S., Scheidegger, C. E., & Venkatasubramanian, S. (2016). Hiring by algorithm: predicting and preventing disparate impact. Available at SSRN.

[3] Al-Otaibi, S. T., & Ykhlef, M. (2012). A survey of job recommender systems. International Journal of the Physical Sciences, 7(29), 5127-5142.

[4] Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Reimer, D. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6-1.

[5] Ballon, P., Van Hoed, M., & Schuurman, D. (2018). The effectiveness of involving users in digital innovation: Measuring the impact of living labs. Telematics and Informatics, 35(5), 1201-1214.

[6] Ballon, P., Duysburgh, P., Fanni, R., Franck, G., Heyman, R., & Laenens, W. (2019). D2.2. Raamwerk voor ethische validatie van AI. Kenniscentrum Data & Maatschappij, Brussel, België (Authors alphabetically placed)

[7] Bardzell, S., & Bardzell, J. (2011, May). Towards a feminist HCI methodology: social science, feminism, and HCI. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 675-684).

[8] Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. arXiv preprint arXiv:1712.03586.

[9] Binns, R., & Gallo, V. (2019). An overview of the Auditing Framework for Artificial Intelligence and its core components. Retrieved February 17, 2020, from https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/

[10] Björneborn, L. (2017). Three key affordances for serendipity: Toward a framework connecting environmental and personal factors in serendipitous encounters. Journal of Documentation, 73(5), 1053-1081. 10.1108/JD-07-2016-0097

[11] Bozdag, E. (2013). Bias in algorithmic filtering and personalization. Ethics and information technology, 15(3), 209-227.

[12] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91).

[13] Burke, R. (2017). Multisided fairness for recommendation. arXiv preprint arXiv:1707.00093.

[14] Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory?. In Advances in Neural Information Processing Systems (pp. 3539-3550)

[15] Collins, P. H. (2002). Black feminist thought: Knowledge, consciousness, and the politics of empowerment. Routledge.

[16] Costanza-Chock, S. (2018). Design Justice: towards an intersectional feminist framework for design theory and practice. Proceedings of the Design Research Society.

[17] Costanza-Chock, S. (2018). Design Justice, A.I., and Escape from the Matrix of Domination. Journal of Design and Science. 10.21428/96c8d426

[18] Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stan. L. Rev., 43, 1241.

[19] Dastin, J. (2018) "Amazon scraps secret AI recruiting tool that showed bias against women" Reuters

[20] Dencik, L., Hintz, A., Redden, J., & Treré, E. (2019). Exploring data justice: Conceptions, applications and directions.

[21] D'Ignazio, Catherine, and Lauren F. Klein. Data feminism. MIT Press, 2020.

[22] Dwork, C., Immorlica, N., Kalai, A. T., & Leiserson, M. (2017). Decoupled classifiers for fair and efficient machine learning. arXiv preprint arXiv:1707.06613.

[23] Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

[24] Fort, I., Pacaud, C., & Gilles, P. Y. (2015). Job search intention, theory of planned behavior, personality and job search experience. International Journal for Educational and Vocational Guidance, 15(1), 57-74.

[25] Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3), 330-347.

[26] Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010). Beyond accuracy: Evaluating recommender systems by coverage and serendipity. Proceedings of the Fourth ACM Conference on Recommender Systems - RecSys '10, 257. 10.1145/1864708.1864761

[27] Gilbert, T. K., & Mintz, Y. (2019, January). Epistemic Therapy for Bias in Automated Decision-Making. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 61-67).

[28] Hao, K. (2019). This is how AI bias really happens—and why it's so hard to fix. https://www. technologyreview. com/s/612876/this-is-how-ai-bias-really-happensand-whyits-so-hard-to-fix.

[29] Ho, J., & Tang, R. (2001, September). Towards an optimal resolution to information overload: an infomediary approach. In Proceedings of the 2001 international ACM SIGGROUP conference on supporting group work (pp. 91-96).

[30] Holone, H. (2016). The filter bubble and its effect on online personal health information. Croatian medical journal, 57(3), 298

[31] Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society, 22(7), 900-915.

[32] Johnson, S., & From, W. G. I. C. (2010). The Natural History of Innovation.

[33] Kaminskas, M., & Bridge, D. (2016). Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. ACM Transactions on Interactive Intelligent Systems, 7(1), 1–42. 10.1145/2926720

[34] Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. Big Data & Society. 10.1177/2053951720915939

[35] Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. Knowledge-Based Systems, 111, 180-192.

[36] L. Cardoso, R., Meira Jr, W., Almeida, V., & J. Zaki, M. (2019, January). A framework for benchmarking discrimination-aware models in machine learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 437-444).

[37] Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2019, January). Fairness through causal awareness: Learning causal latent-variable models for biased data. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 349-358).

[38] Milan, Stefania, and Emiliano Treré. 2019. "Big Data from the South(s): Beyond Data Universalism." Television & New Media 20 (4): 319–35.

[39] Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Science and Engineering Ethics, 1-28.

[40] O'neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

[41] Peña Gangadharan, S., & Niklas, J. (2019). Decentering technology in discourse on discrimination. Information, Communication & Society, 22(7), 882-899.

[42] Perez, C. C. (2019). Invisible Women: Exposing data bias in a world designed for men. Random House.

[43] Race, T. M., & Makri, S. (Eds.). (2016). Accidental information discovery: cultivating serendipity in the digital age. Elsevier

[44] Reviglio, U. (2019a). Serendipity as an emerging design principle of the infosphere: challenges and opportunities. Ethics and Information Technology, 21(2), 151-166.

[45] Reviglio, U. (2019b). Towards a Taxonomy for Designing Serendipity in Personalized News Feeds. http://informationr.net/ir/24-4/colis/colis1943.html

[46] Robinson, W R, A Renson, and A I Naimi. 2020. "Teaching Yourself about Structural Racism Will Improve Your Machine Learning." https://academic.oup.com/biostatistics/article-abstract/21/2/339/5631851.

[47] Sadler, E. B., & Given, L. M. (2007). Affordance theory: a framework for graduate students' information behavior. Journal of documentation, 63(1), 115-141.

[48] Smets, A., Walravens, N. & Ballon, P. (2020). Designing Recommender Systems for the Common Good. In Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct), July 14–17, 2020, Genoa, Italy. ACM, New York, NY, USA, 3 pages. 10.1145/3386392.3399570

[49] Sokol, K., & Flach, P. (2020, January). Explainability fact sheets: a framework for systematic assessment of explainable approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 56-67).

[50] Suresh, H., & Guttag, J.V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. ArXiv, abs/1901.10002.

[51] Van Humbeeck, G. (2020, April). AI VDAB. Presentation presented during Data Date 2, Kenniscentrum Data & Maatschappij. https://data-en-maatschappij.ai/nieuws/data-date-2-ai-en-rekrutering

[52] Wang, R., Harper, F. M., & Zhu, H. (2020, April). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

[53] Young, J. 2017. "Designing Feminist Chatbots" 2017. https://www.ellpha.com/list/2017/9/23/designing-feminist-chatbots.

Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics

Tim Draws
Delft University of Technology
The Netherlands
t.a.draws@tudelft.nl

Nava Tintarev
Delft University of Technology
The Netherlands
n.tintarev@tudelft.nl

Ujwal Gadiraju
Delft University of Technology
The Netherlands
u.k.gadiraju@tudelft.nl

ABSTRACT

The way pages are ranked in search results influences whether the users of search engines are exposed to more homogeneous, or rather to more diverse viewpoints. However, this viewpoint diversity is not trivial to assess. In this paper, we use existing and novel ranking fairness metrics to evaluate viewpoint diversity in search result rankings. We conduct a controlled simulation study that shows how ranking fairness metrics can be used for viewpoint diversity, how their outcome should be interpreted, and which metric is most suitable depending on the situation. This paper lays out important groundwork for future research to measure and assess viewpoint diversity in real search result rankings.

1. INTRODUCTION

Search result rankings strongly influence user attitudes, preferences, and behavior [11; 13; 18; 19]. Underlying this effect are cognitive biases such as *position bias*, which describes users' tendency to pay more attention to documents at higher ranks [13; 18]. Recent research has demonstrated that these biases go to such an extent that rearranging search result rankings to favor different stances on the same topic can affect users' personal opinions [11; 19]. To mitigate such unintentional biases, it is important to maintain a strong viewpoint diversity in search result rankings – especially when they relate to disputed topics.

Viewpoint diversity in search result rankings is closely related to the notion of ranking fairness. The aim in fair ranking is to measure and adapt ranked lists in terms of their fairness concerning a given characteristic [7; 22; 27]. For example, a ranked list of candidates on a job seeking platform could be evaluated with respect to gender fairness. A fair ranking is then considered to be one in which gender does not affect the ranking of candidates. Analogously in this paper, a search result ranking is evaluated with respect to *viewpoint* – to the best of our knowledge, a novel application of ranking fairness. Such a viewpoint can for example convey different stances on a topic, or different underlying reasons for a given stance. A search result ranking that is fair (or unbiased) with respect to viewpoints would give each viewpoint its fair share of coverage, contributing to viewpoint diversity in the search results[1].

[1]Note that here we are thus looking at fairness in the *outcome* of a ranking algorithm; i.e., not at procedural fairness.

One major building block of studying viewpoint fairness in search result rankings is deciding how to measure it. Several metrics have been developed that assess fairness in rankings [27; 22] (see Section 2). These metrics evaluate fairness in terms of *statistical parity*, which is satisfied in a ranking if a given variable of interest – here, the expressed viewpoint – does not influence how documents are ranked. In this paper, we investigate whether ranking fairness metrics can be used to assess *viewpoint diversity* in search results.

We generate a range of synthetic search result rankings with varying degrees of ranking bias and explore the behavior of existing and novel ranking fairness metrics on these rankings. For our use case of viewpoint diversity, we consider two fundamental scenarios: *binomial viewpoint fairness*, in which the task is to measure viewpoint diversity with respect to one specific *protected viewpoint*, and *multinomial viewpoint fairness* where the aim is to protect all available viewpoints simultaneously. We make the following contributions:

1. We present a simulation study that illustrates how existing ranking fairness metrics can be used to assess viewpoint diversity in search result rankings. We show how these metrics behave under varying conditions of viewpoint diversity and provide a guide for their use (Section 4.2).
2. We propose a novel ranking fairness metric for assessing multinomial viewpoint fairness (Section 3.4) and also analyze its behavior (Section 4.2).

We find that all the considered ranking fairness metrics can distinguish well between different levels of viewpoint diversity in search results. However, which specific metric is most sensitive to a lack of viewpoint diversity depends on how many viewpoint categories there are, the distribution of advantaged and disadvantaged items in the ranking, and the severity of the ranking bias.

All code and supplementary material related to this research are openly available at `https://osf.io/nkj4g/`.

2. BACKGROUND AND RELATED WORK

Diversity in search result rankings is not a novel topic. Several methods have been proposed to measure and improve diversity in ranked lists of search results [1; 2; 9; 20; 21]. Unlike previous methods, which aim to *balance* relevance (e.g., in relation to a user query) and diversity (e.g., in relation to user intent), we delve deeper into the notion of diversity. Specifically, we focus on ranking *fairness* for assessing

Table 1: The viewpoint label taxonomy we consider in this paper. Labels are denoted by s and represented as ordinal values.

s	Description	Example
-3	strongly opposing	"Horrible places! All zoos should be closed ASAP."
-2	opposing	"We should strive towards closing all zoos."
-1	somewhat opposing	"Despite the benefits of zoos, overall I'm against them."
0	neutral	"These are the main arguments for and against Zoos."
+1	somewhat supporting	"Although zoos are not great, they benefit society."
+2	supporting	"I'm in favor of zoos, let's keep them."
+3	strongly supporting	"There is nothing wrong with zoos – open more!"

viewpoint diversity, which originates from the field of fair machine learning.

Fairness and the mitigation of bias in machine learning systems extends into several different sub-fields [5, 6, 17]. One of these sub-fields – *fair ranking* – has received increasing attention recently, following calls for dealing with bias on the web [4]. This has led to the development of methods to increase ranking fairness [7, 8, 23, 28] as well as evaluative frameworks [3, 7, 10, 22] and metrics [15, 27] for assessing bias and fairness in ranked lists. Measuring ranking fairness requires deciding which notion of fairness to handle (i.e., defining a *fair ranking scenario*) [10, 7] and discounting the metric computation by rank to account for differences in attention over the ranks [7, 22].

Previously proposed ranking fairness metrics commonly presuppose that a *fair* or *unbiased* ranking is one in which *statistical parity* is present [27]. A machine learning algorithm satisfies statistical parity when an item's probability of receiving a given outcome is not affected by belonging to a *protected group* [24]. Such a protected group can be any subset of the overall population of items that share some characteristic that is not supposed to affect the algorithm outcome. In the context of ranking, statistical parity holds when membership in a protected group has no influence on a document's position in the ranking [27]. Suppose a user enters the query *Should Zoos Exist?* into a web search engine, which then returns a ranked list of search results. Each document in the ranking corresponds to *some* viewpoint concerning zoos or is neutral towards the topic. A ranking assessor could define the *opposing* side of the zoo-argument as the protected viewpoint. Statistical parity would then be satisfied if expressing the protected viewpoint does not affect the ranking of documents.

Yang and Stoyanovich [27] introduce three metrics that assess statistical parity in rankings. These metrics compare the group membership distribution (i.e., share of protected and non-protected items) in a ranking at different cut points (e.g., 10, 20, ...) with the ranking's overall group membership distribution. Aggregating the results of these comparisons in a discounted manner incorporates the intuition that an absence of bias is more important among higher ranks. We formalize the metrics introduced by Yang and Stoyanovich [27] in Section 3.4.

Ranking fairness has also been assessed in at least two other notable ways. First, Kulshrestha et al. [15] introduce a metric that quantifies ranking bias related to continuous attributes as opposed to group membership. Their metric considers the mean of a continuous variable of interest at each step of its computation. Despite this promising groundwork, measuring continuous ranking bias remains limited; for instance, by considering only the mean, other important characteristics of continuous distributions (i.e., such as the standard deviation or distribution type) are ignored. Second, recent work has defined *criteria* that a ranking has to fulfill to be considered *fair* [28, 22]. Whether the ranking fulfills these criteria is assessed using null hypothesis significance testing. Our aim, however, is to quantify the *degree* of viewpoint diversity in search result rankings.

3. MEASURING FAIRNESS IN RANKINGS

Viewpoint diversity in search results can best be illustrated by a running example. Consider that a user wants to form an educated opinion on the topic '*Should Zoos Exist?*', and turns to web search to gather information. Let us assume that each document that the user encounters in the search result list will express a viewpoint concerning zoos or be neutral towards the topic.[2] These viewpoints can be represented in an ordinal manner, as illustrated in Table 1. We thus categorize the different viewpoints related to zoos by placing them on a 7-point scale ranging from *strongly opposing* to *strongly supporting* the existence of zoos.[3]

3.1 Preliminaries and Notation

We are given a set of documents D and a set of viewpoint labels S. Both sets contain the same number of elements N. Each document $d \in D$ is uniquely associated with one label $s_d \in S$. Here, s_d reflects the viewpoint of document d towards a given disputed topic, rated on a 7-point scale ranging from *extremely opposing* to *extremely supporting*. The viewpoint labels in S are integers ranging from -3 to 3, where negative values indicate an *opposing viewpoint*, 0 indicates a *neutral viewpoint*, and positive values indicate a *supporting viewpoint* towards the debated topic (see Table 1 for an example). A ranked list of D is denoted as τ. We denote the number of items that belong to a subset p of S as S^p, which becomes $S^p_{1..i}$ when constrained to the top i ranked documents. Table 2 presents an overview of the notation introduced here.

3.2 Defining Fairness and Viewpoint Diversity

There are many definitions of fairness, and so before describing fairness metrics, we first identify which type of fairness to handle. In this paper, we focus on the notion of *statistical parity* (also commonly referred to as *group fairness*; see Section 2). This notion allows us to define several fairness

[2]Here, *neutral* could mean that a document is not opinionated, provides a balanced overview of the different viewpoints, or is irrelevant to the topic.

[3]Note that this is just one possible way to categorize existing viewpoints on a topic.

Table 2: Notation used throughout this paper.

Notation	Description
d	document
D	set of documents
s_d	viewpoint label of document d
S	set of viewpoint labels
S^p	number of items in set S that belong to subset p
τ	ranked list of set D
$S^p_{1\ldots i}$	S^p in the top i ranked documents
N	number of elements in D, S, and τ

aims for assessing viewpoint diversity. We consider two such aims, which we call *binomial viewpoint fairness* and *multinomial viewpoint fairness*. Below we describe these aims and align them with the notion of statistical parity in rankings.

Binomial viewpoint fairness. One aim for viewpoint diversity may be to treat one specific viewpoint, e.g., a minority viewpoint, fairly. For example, if a search result ranking on the query *Should Zoos Exist?* is dominated by arguments *supporting* zoos, the ranking assessor may want to evaluate whether the minority viewpoint (i.e., *opposing* zoos) gets its fair share of coverage. The assessor may consider a binomial classification of documents into one of two groups: expressing the minority viewpoint or not expressing the minority viewpoint. Here, expressing the minority viewpoint is analogous to a protected group. Statistical parity in a ranking of such documents is satisfied when expressing the minority viewpoint does not affect a document's position in the ranking.

Multinomial viewpoint fairness. Another aim when evaluating viewpoint diversity may be that *all* viewpoints are covered fairly. For example, a search result ranking on the query *Should Zoos Exist?* could be assessed without explicitly defining a specific viewpoint as the protected group but instead considering the distribution over several existing viewpoints. Here the assessor thus considers a multinomial classification of documents into some viewpoint taxonomy (e.g., into seven categories depending on polarity and severity of the viewpoint; see Section 3.1). In this case, we say that statistical parity is satisfied when for each viewpoint, the choice of viewpoint does not influence a document's position in the ranking. Multinomial viewpoint fairness is thus more fine-grained than binomial viewpoint fairness: whereas binomial viewpoint fairness focuses on fairness towards one protected viewpoint, multinomial viewpoint fairness requires being fair to all viewpoints simultaneously.

3.3 Desiderata and Practical Considerations for Metrics

Evaluating statistical parity. In this paper, we use ranking fairness metrics to assess viewpoint diversity in search result rankings. These are based on the notion of statistical parity, which is present in a ranking when the viewpoints that documents express do not affect their position in the ranking. However, we are only given the ranking and viewpoint per document and cannot assess the ranking algorithm directly. Statistical parity thus needs to be approximated. We choose to approximate statistical parity in the same way as previously developed ranking fairness metrics [27]. These metrics measure the extent to which the document distribution over groups (e.g., the protected and non-protected group) is the same in different top-i portions of the ranking compared to the overall ranking (see Section 2). The more dissimilar the distribution at different top-i is from the overall distribution, the less fair the ranking.

Discounting the ranking fairness computation. User attention depletes rapidly as the ranks go up [13, 18]. For example, in a regular web search, the majority of users may not even view more than 10 documents. This means that a measure of viewpoint diversity needs to consider the rank of documents, and not just whether viewpoints are present. More specifically, fairness is more important at higher ranks. A practical way to incorporate this notion into a ranking fairness metric is to include a discount factor. Sapiezynski et al. [22] point out that such a discount depends on the user model related to the particular ranking one is assessing. Similar to the ranking fairness metrics introduced by Yang and Stoyanovich [27], we choose the commonly used $\log_2$ discount for each metric we introduce below. Yang and Stoyanovich [27] suggest discounting in steps of 10 (see Section 2). Such a binned discount nicely incorporates the notion that ranking fairness is more important in the top 10 documents than it is in the top 20 documents. However, especially on the first page of search results, individual ranks matter a lot [13, 18]. We therefore decide to discount by individual rank and consider the top 1, 2, ... N documents at each step of the aggregation.

Normalization. When evaluating and comparing metrics, it is useful if they all operate on the same scale. We thus only consider normalized ranking fairness metrics.

3.4 Ranking Fairness Metrics

In this section, we describe the metrics that we use to assess viewpoint diversity in search result rankings. These metrics are partly based on existing ranking fairness metrics and partly novel. We adapt each metric that we use to fit the practical considerations outlined in Section 3.3. Taking these practical considerations into account, we define a template that each normalized ranking bias (nRB) metric that we use will follow:

$$\text{nRB}(\tau) = \frac{1}{Z}\sum_{i=1}^{N}\frac{F(i)}{\log_2(i+1)}. \quad (1)$$

Here, F is a function that quantifies the ranking bias in the ranked list τ. All metrics that we describe in the following subsections will only differ in terms of how they define F. The function F is iteratively computed for the top i documents and subsequently aggregated by using a $\log_2$ discount. Finally, Z is a normalizing constant that takes on the value for F given the maximally unfair permutation of τ.[4]

3.4.1 Metrics to assess binomial viewpoint fairness

Yang and Stoyanovich [27] propose three ranking fairness metrics to assess statistical parity in rankings (see Section 2). We interpret these metrics to fit binomial viewpoint

[4] A description of how we normalize each metric can be found at https://osf.io/nkj4g/.

fairness and adapt them to fit the considerations outlined in Section 3.3.

Note that, although we define a protected and a non-protected viewpoint before using any of these metrics, the metrics are in principle agnostic as to which of the two viewpoint categories (i.e., "protected" and "unprotected") is advantaged in the ranking. That is, they do not only measure when the protected viewpoint is treated unfairly but also capture if a ranking is biased *towards* the protected viewpoint. The categorization into protected and non-protected viewpoints should thus be viewed as a binary classification of documents that – in a fair scenario – does not affect how documents are ranked.

Normalized Discounted Difference (nDD). This metric computes the difference between the proportion of items that belong to the protected group at different top-i subsets of the ranking with and overall proportion:

$$\text{nDD}(\tau) = \frac{1}{Z}\sum_{i=1}^{N}\frac{1}{\log_2(i+1)}\left|\frac{S^p_{1\ldots i}}{i} - \frac{S^p}{N}\right|. \quad (2)$$

Here, S^p is the number of documents in the protected group and N is the total number of ranked documents.

Normalized Discounted Ratio (nDR). This metric measures the difference between the ratio of documents that express the protected viewpoint, and those who do not, at different top-i portions of the ranking with the overall ratio:

$$\text{nDR}(\tau) = \frac{1}{Z}\sum_{i=1}^{N}\frac{1}{\log_2(i+1)}\left|\frac{S^p_{1\ldots i}}{S^u_{1\ldots i}} - \frac{S^p}{S^u}\right|. \quad (3)$$

Here, S^u refers to the number of documents that do not express the protected viewpoint. Here we set the value of fractions to 0 if their denominator is 0 [27].

Normalized Discounted Kullback-Leibler Divergence (nDKL). This metric makes use of the *Kullback-Leibler divergence* (KLD), an asymmetric measure of difference between probability distributions [14]. For two discrete probability distributions P and Q that are defined on the same probability space $\mathcal{X}$, KLD is given by

$$\text{KLD}(P||Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right). \quad (4)$$

To measure binomial viewpoint fairness in a ranking, P and Q can be defined as

$$P = \left(\frac{S^p_{1\ldots i}}{i}, \frac{S^u_{1\ldots i}}{i}\right), Q = \left(\frac{S^p}{N}, \frac{S^u}{N}\right).$$

This way, KLD measures the divergence between the proportion of protected items at rank i and in the ranking overall [5]. We can insert KLD in Equation 1:

$$\text{nDKL}(\tau) = \frac{1}{Z}\sum_{i=1}^{N}\frac{\text{KLD}(P||Q)}{\log_2(i+1)}. \quad (5)$$

[5]Note that KLD is not defined for $P = (0, 1)$. In this case, we smooth to $P = (0.001, 0.999)$.

3.4.2 *Metric to assess multinomial viewpoint fairness*

To the best of our knowledge, no metrics have so far been proposed that explicitly assess ranking fairness for multiple categories at once. The previously introduced nDKL metric can in principle be expanded to assess multinomial viewpoint fairness. KLD measures the distance between two discrete probability distributions P and Q. In the multinomial case, we can define P and Q as multinomial distributions over the available viewpoint categories. For instance, in our use case of viewpoints rated on a 7-point scale, P and Q may be given by:

$$P = \left(\frac{S^{-3}_{1\ldots i}}{i}, \frac{S^{-2}_{1\ldots i}}{i}, \frac{S^{-1}_{1\ldots i}}{i}, \frac{S^{0}_{1\ldots i}}{i}, \frac{S^{+1}_{1\ldots i}}{i}, \frac{S^{+2}_{1\ldots i}}{i}, \frac{S^{+3}_{1\ldots i}}{i}\right),$$
$$Q = \left(\frac{S^{-3}}{N}, \frac{S^{-2}}{N}, \frac{S^{-1}}{N}, \frac{S^{0}}{N}, \frac{S^{+1}}{N}, \frac{S^{+2}}{N}, \frac{S^{+3}}{N}\right),$$

where $S^{-3,-2,\ldots,3}$ refer to the number of items in each viewpoint category.

A problem with using KLD for multinomial distributions is that its normalization becomes extremely complex. To normalize KLD, the maximally divergent distribution of items needs to be computed at each step. Whereas this is rather straightforward in the binomial case [6] finding the maximally divergent distribution becomes extremely expensive when more categories are added.

To resolve the normalization issue that comes with KLD, we propose a new metric that uses the Jensen-Shannon divergence (JSD) as an alternative distance function. Similarly to KLD, JSD measures the distance between two discrete probability distributions P and Q that are defined on the same sample space $\mathcal{X}$ [12]. JSD can in fact be expressed using KLD:

$$\text{JSD}(P||Q) = 0.5 * \Big(KLD(P||R) + KLD(Q||R)\Big).$$

Here, $R = 0.5 * (P + Q)$ is the mid-point between P and Q. In contrast to KLD (which can go to infinity), JSD is bound by 1 as long as one uses a base 2 logarithm in its computation [16]. Knowing this maximally possible value for JSD, also an aggregated, discounted version of JSD is easily normalized. We thus propose *Normalized Discounted Jensen-Shannon Divergence* (nDJS) as given by

$$\text{nDJS}(\tau) = \frac{1}{Z}\sum_{i=1}^{N}\frac{\text{JSD}(P||Q)}{\log_2(i+1)}, \quad (6)$$

where $\text{JSD}(P||Q)$ is the JSD between P and Q. Although we here propose nDJS specifically for assessing multinomial viewpoint fairness, note that it can be used to assess binomial viewpoint fairness as well.

4. SIMULATION STUDY

In this section, we show how the metrics introduced in Section 3.4 behave in different ranking scenarios. Our code to implement the metrics and simulation is openly available [7].

[6]A description of how we normalize each metric can be found at `https://osf.io/nkj4g/`.

[7]See our repository at `https://osf.io/nkj4g/`.

Table 3: Examples of sample weight allocations for the simulation of binomial (left-hand table, $\alpha = 0.5$) and multinomial viewpoint fairness (right-hand table; $\alpha = -0.8$).

viewpoint	-3	-2	-1	0	+1	+2	+3
weight	w_1	w_1	w_1	w_2	w_2	w_2	w_2
(rounded)	0.5	0.5	0.5	1.5	1.5	1.5	1.5

viewpoint	-3	-2	-1	0	+1	+2	+3
weight	w_2	w_2	w_1	w_2	w_2	w_2	w_2
(rounded)	0.2	0.2	1.8	0.2	0.2	0.2	0.2

4.1 Generating Synthetic Rankings

To simulate different ranking scenarios, we first generate three synthetic sets $S1$, $S2$, and $S3$ to represent different viewpoint distributions. The items in each set simulate viewpoint labels for 700 documents (i.e., to enable a simple balanced distribution over seven viewpoints) and are distributed as shown in Table 4. Whereas $S1$ has a balanced distribution of viewpoints, $S2$ and $S3$ are skewed towards supporting viewpoints[8]. We use $S1$, $S2$, and $S3$ to simulate both binomial and multinomial viewpoint fairness[9].

Table 4: Viewpoint distributions of the sets $S1$, $S2$, and $S3$.

	-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**
$S1$	100	100	100	100	100	100	100
$S2$	80	80	80	115	115	115	115
$S3$	60	60	60	130	130	130	130

Sampling. We create rankings of the viewpoint labels in $S1$, $S2$, and $S3$ by conducting a weighted sampling procedure. To create a ranking, viewpoint labels are gradually sampled from one of the three sets without replacement to fill the individual ranks. Each viewpoint label in the set is assigned one of two different sample weights that determine the labels' probability of being drawn. These two sample weights are controlled by the ranking bias parameter α and given by: $w_1 = 1.0001 - 1 \times \alpha$; $w_2 = 1.0001 + 1 \times \alpha$.

Alpha. For our simulation of binomial and multinomial viewpoint fairness, ranking bias is controlled by the continuous parameter $\alpha = [-1, 1]$. More specifically, α controls the sample weights w_1 and w_2 that are used to create the rankings. Whereas a negative α will result in higher ranks for viewpoints that are assigned w_1, a positive α will advantage viewpoints that are assigned w_2. The further away α is from 0, the more extreme the ranking bias. If α is set to exactly 0, no ranking bias is present: here it does not matter whether a viewpoint label is assigned w_1 or w_2, the sample weights are the same. In each simulation, we try 21 degrees of ranking bias for $\alpha = -1$ to $\alpha = 1$ in steps of 0.1.

4.1.1 *Simulating binomial viewpoint fairness*

To simulate binomial viewpoint fairness, we create ranked lists from $S1$, $S2$, and $S3$ with different degrees of ranking bias. Ranking bias – controlled by α – in this scenario refers to the degree to which expressing a protected viewpoint influences a document's position in the ranking. We define all *opposing* viewpoints (i.e., -3, -2, and -1) together as the protected viewpoint and assign them the sample weight w_1. All other viewpoints (i.e., 0, 1, 2, and 3) are thus non-protected and assigned the other sample weight w_2 when generating the rankings. Table 3 (left-hand table) shows an example of this sample weight allocation for $\alpha = 0.5$. In this example, the non-protected viewpoint is more likely to be drawn compared to the protected viewpoint.

Our weighted sampling procedure (see above) will produce slightly different rankings even when the same α is used. To get reliable results, we therefore create 1000 ranked lists for each α and aggregate the results.

4.1.2 *Simulating multinomial viewpoint fairness*

We simulate multinomial viewpoint fairness by again sampling rankings from $S1$, $S2$, and $S3$ with different degrees of ranking bias. This time the ranking bias α is defined as how much the expressed viewpoint generally affects a document's position in the ranking.

Since there are many scenarios in which one (or more) of several viewpoint categories could be preferred over others in a ranking, we focus on just one specific case: our simulation prefers *one* of the seven viewpoints over the other six. For example, this could be the case if a search result list is biased towards an extremely opposing viewpoint. We randomly assign the sample weight w_1 to one of the opposing viewpoints (i.e., -3, -2, or -1) and the sample weight w_2 to all remaining viewpoints for each ranking we create. This means that each ranked list we create prefers a different viewpoint, reflecting the idea that we do not know which viewpoint might be preferred before evaluating the ranking and we have no specific, pre-defined protected viewpoint. Table 3 (right-hand table) shows an example of this sample weight allocation for $\alpha = -0.8$. In this example, the ranked list will prefer the viewpoint -1 over all other viewpoints. We again compute 1000 ranked lists for each α and aggregate the results.

4.2 Metric Behavior

Here, we explore the behavior of the ranking fairness metrics introduced in Section 3.4 using the synthetic rankings from Section 4.1.

4.2.1 *Binomial viewpoint fairness*

Binomial viewpoint fairness can be assessed using nDD, nDR, or nDKL. Each of these metrics measures the degree to which expressing a protected viewpoint affects the ranking of documents. The ranking in our running example is considered fair if documents opposing zoos (i.e., -3, -2, and -1) get a similar coverage throughout the ranking compared to other viewpoints (i.e., 0, $+1$, $+2$, and $+3$). A fair scenario should lead to a low score on each of the three metrics.

Figure 1 shows the mean outcome of nDD, nDR, and nDKL from 1000 ranked lists per data set (i.e., $S1$, $S2$, and $S3$) and

[8] Due to symmetry we do not include similar distributions for opposing viewpoints.

[9] Because we are only interested in rankings with respect to viewpoint labels, we do not generate any actual documents here. Instead, we rank the labels themselves.

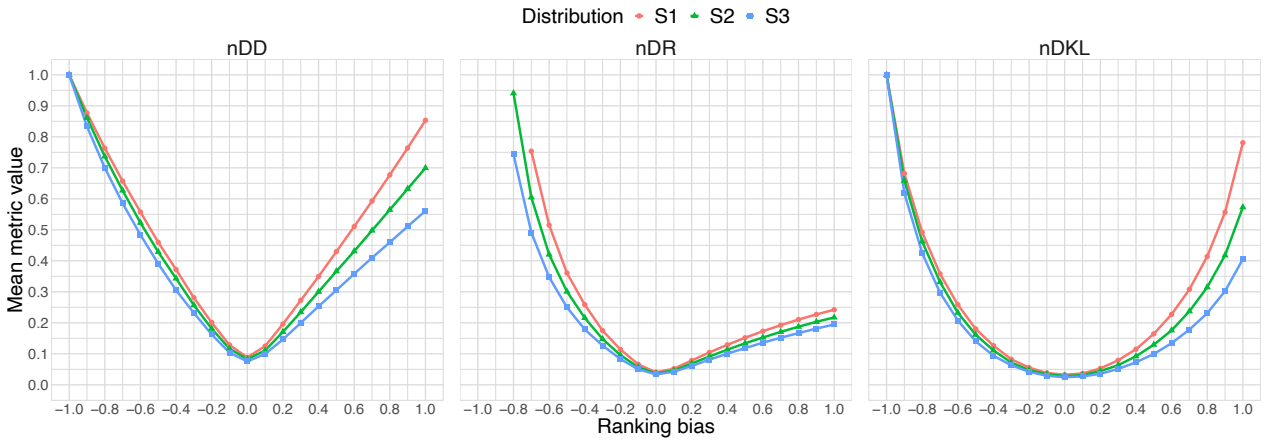


Figure 1: Behavior of the metrics nDD (left-hand plot), nDR (center plot), and nDKL (right-hand plot) on the sets $S1$ ($S^p = 300$), $S2$ ($S^p = 240$), and $S3$ ($S^p = 180$) across different α (ranking bias) settings.

α (i.e., ranking bias) setting. Each set represents a different overall distribution of viewpoints (see Table 4).

We note three characteristics that all three metrics share. First, each of the three metrics is lowest for low bias ($\alpha = 0$) and increases from there as the absolute value of α increases. This means that all three metrics function as expected: they produce higher values as ranking bias becomes more extreme. Second, each metric shows a steeper curve as the data sets contained fewer items that express the minority viewpoint (here, the protected opposing viewpoint) increases; i.e, $S1 > S2 > S3$. Different levels of ranking bias thus become easier to detect when the distribution of protected and non-protected items is more balanced. Third, each metric produces higher values for $\alpha = -1$ (protected viewpoint is *advantaged*) than for $\alpha = 1$ (protected viewpoint is *disadvantaged*). The reason behind this is that unfair treatment becomes increasingly harder to detect as the number of items in the disadvantaged group shrinks: if one group only encompasses around 25% of items (e.g., such as in $S3$), it is less odd to see several items of the other group ranked first than if the distribution is more balanced. That is also why each metric produces higher values at $\alpha = 1$ as the number of protected items increases.

Next to these general characteristics that are shared by all metrics, below we discuss differences that distinguish the metrics in terms of their behavior.

Normalized Discounted Difference. For each of the three data sets, nDD reaches its maximum value of 1 when $\alpha = -1$ and is at its lowest with mean values of approximately 0.08 when $\alpha = 0$. Depending on the number of items that express the protected viewpoint, nDD reaches mean values between 0.55 and 0.85 when $\alpha = 1$ for the three data sets in our simulation. The curves for nDD in Figure 1 are also comparatively steep. This indicates that nDD is especially useful for distinguishing low levels of ranking bias.

Normalized Discounted Ratio. The lowest mean nDR values in our simulation (reached at $\alpha = 0$ for each of the three data sets) all approximate 0.04. Even more so than nDD, nDR reaches mean values far below 1 when the protected viewpoint is disadvantaged in the ranking. The mean values for this form of extreme ranking unfairness range from approximately 0.19 to 0.24 in our simulation, depending on the number of protected viewpoint items. In comparison to the other two metrics, nDR is less steep than nDD but steeper than nDKL. It could thus be useful for detecting medium levels of ranking bias. However, if a ranking is unfair towards the minority viewpoint, nDR does not distinguish different levels of ranking bias well. We also find that our normalization procedure (i.e., dividing each metric outcome by the outcome for a maximally unfair ranking) does not normalize nDR correctly. Thus, the maximal mean values for nDR (which it reaches at $\alpha = -1$) lie above 1 and are therefore not displayed in Figure 1 (which has 1 as its upper limit).[10]

Normalized Kullback-Leibler Divergence. Similar to the other metrics, nDKL reaches its maximum value of 1 at $\alpha = -1$. In our simulation, the lowest mean values for nDKL (reached at $\alpha = 0$) approximated 0.03. Extremely positive α settings (i.e., disadvantaging the minority viewpoint) produce mean nDKL values between 0.40 and 0.78, depending on the number of items that express the minority viewpoint. Furthermore, nDKL has a more parabolic shape compared to nDD and nDR. Whereas nDKL can thus not distinguish low values of ranking bias well, it is useful for differentiating between high levels of ranking bias.

4.2.2 *Multinomial viewpoint fairness*

To assess multinomial viewpoint fairness, we use nDJS. This metric measures the degree to which the viewpoint that documents express is a factor for a ranking in general. For example, in a search result ranking related to the topic *Should Zoos Exist?*, a range of viewpoints may exist, some of which may be advantaged in the ranking over other viewpoints.

[10] We explore the reason behind this (including an alternative way to normalize nDR) in a supplementary document on our normalization procedures. This document can be found at https://osf.io/nkj4g/.

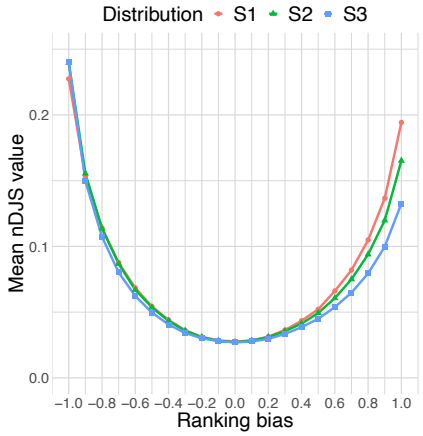


Figure 2: Behavior of nDJS on the sets $S1$, $S2$, and $S3$ across different α (ranking bias) settings. The number of items with sample weight w_1 for rankings from the sets $S1$, $S2$, and $S3$ are 100, 80, and 60, respectively.

That is why we cannot use binomial ranking fairness metrics here: we do not have a specific viewpoint to protect, but instead wish to protect all viewpoints equally. A maximally fair ranking scenario would give all viewpoints a coverage across the ranking that is proportional to their share in the overall distribution. For (approximately) fair rankings, nDJS should return a low value.

We test nDJS on synthetic rankings that simulate varying degrees of bias on three different sets of items ($S1$, $S2$, and $S3$, see Section 4.1). Figure 2 shows the mean outcome of nDJS from 1000 ranked lists per set and α (i.e., ranking bias) setting. Similar to the binomial ranking fairness metrics, nDJS does what it is expected to do: it produces its highest values at extreme α (ranking bias) settings and its lowest values at $\alpha = 0$. This means that nDJS can pick up the nuanced multinomial viewpoint fairness in our synthetic rankings. We observe, however, that due to its normalization, the maximum values for nDJS are much lower than for the metrics that assess binomial viewpoint fairness. When $\alpha = -1$ (i.e., when one random viewpoint is *disadvantaged* compared to others), nDJS produces mean values between approximately 0.18 and 0.21. Due to the different normalization, it is therefore not possible to compare results from nDJS directly to results from the binomial ranking fairness metrics. For low values of ranking bias, the mean nDJS values approximate 0.03 on all three data sets. The mean nDJS value lies between approximately 0.07 and 0.09 when $\alpha = 1$ (i.e., when one viewpoint is *advantaged* compared to others).

Similar to the binomial fairness metrics, the values that nDJS produces is again influenced by the proportion of advantaged items in the ranking. The more balanced this ratio, the easier it is to detect a ranking bias (i.e., the higher nDJS). Note that in this simulation, the distribution of advantaged and disadvantaged items was far from balanced, as we only treated one viewpoint label differently per ranking.

Table 5: Recommended metrics for different scenarios of ranking bias and overall viewpoint distribution (i.e., protected and non-protected items) in a ranked list.

		Ranking Bias		
		Low	Medium	High
Distribution	Low balance	nDD	nDD	nDD
	Medium balance	nDD	nDD	nDKL
	High balance	nDD	nDKL	nDKL

5. DISCUSSION

In this section, we summarize our findings, provide a guide to using the metrics we examined, and discuss the limitations and implications of this research.

5.1 Binomial Viewpoint Fairness

Each of the three metrics we tested in our simulation can measure binomial viewpoint fairness (nDD, nDR, nDKL; see Section 4.2). However, depending on the distribution of protected and non-protected items, as well as the direction and level of ranking bias, a different metric might be suitable. Table 5 shows which metric we recommend using in which scenario. In sum, we suggest taking the following considerations when assessing binomial viewpoint fairness:

1. Generally, the more balanced the overall distribution of protected and non-protected items in the ranking, the better the metrics are able to distinguish different levels of ranking bias. When ranking bias is disadvantaging a protected group that only contains a small number of items, nDR appears to be the most suitable metric because it is the least vulnerable in this case.

2. Which metric is most suitable also depends on how severe the bias in the ranking is estimated to be. Whereas nDD outputs the most divergent values for mild cases of ranking bias, nDKL distinguishes more severe cases of ranking bias better. Although nDR is slightly better in distinguishing medium levels of negative ranking bias, we do not recommend using it at all due to its normalization issues and weak performance when ranking bias is positive.

3. If the minority viewpoint is preferred in the ranking, ranking bias is well detected by all three metrics. However, when the minority group is *disadvantaged*, all metrics show a decrease in performance. In this case, we suggest using either nDD or nDKL, depending on how strong the ranking bias is.

5.2 Multinomial Viewpoint Fairness

We find that our novel metric nDJS can assess multinomial ranking fairness. Similarly to the binomial fairness metrics, nDJS can distinguish different levels of ranking bias best when the overall distribution of advantaged and disadvantaged viewpoints is balanced. A weakness of nDJS is that its normalization causes its outcome values to be much lower in general compared to binomial fairness metrics. We note that nDJS cannot be directly compared to nDD, nDR, or nDKL and recommend to interpret nDJS carefully when ranking bias is mild.

5.3 Caveats and Limitations

We note that our simulation study is limited in at least three important ways. First, we consider a scenario in which documents have correctly been assigned multinomial viewpoint labels. This allows us to study their behavior in a controlled setting. In reality, existing viewpoint labeling methods are prone to biases and issues of accuracy. Current opinion mining techniques are still limited in their ability to assign such labels [25] and crowdsourcing viewpoint annotations from human annotators can be costly and also prone to biases and variance [26].

Second, we assume that any document in a search result ranking can be assigned some viewpoint label concerning a given disputed topic. It is realistically possible for a document to contain several, or even all available viewpoints (e.g., a debate forum page). In these cases, assigning an overarching viewpoint label might oversimplify the nuances in viewpoints that exist *within* rankings and thereby not leading to a skewed assessment of viewpoint diversity in the search result ranking. Future work could look into best practices of assigning viewpoint labels to documents.

Third, our simulation of multinomial viewpoint fairness included only one specific case in which one viewpoint is treated differently compared to the other six. There are other scenarios where multinomial viewpoint fairness could become relevant. These scenarios differ in how many viewpoint categories there are, how many items are advantaged in the ranking, and to what degree. Simulating all of these potential scenarios is beyond the scope of this paper. Future work could however explore how metrics such as nDJS behave in such scenarios.

6. CONCLUSION

We adapted existing ranking fairness metrics to measure binomial viewpoint fairness and proposed a novel metric that evaluates multinomial viewpoint fairness. We find that despite some limitations, the metrics reliably detect viewpoint diversity in search results in our controlled scenarios. Crucially, our simulations show how these metrics can be interpreted and their relative strengths.

This lays the necessary groundwork for future research to assess viewpoint diversity in *actual* search results. We plan to perform such evaluations of existing web search engines concerning highly debated topics and upcoming elections. Such work would not only provide tremendous insight into the current state of viewpoint diversity in search result rankings but pave the way for a greater understanding of how search result rankings may affect public opinion.

Acknowledgements

This activity is financed by IBM and the Allowance for Top Consortia for Knowledge and Innovation (TKI's) of the Dutch ministry of economic affairs.

We also thank Agathe Balayn, Shabnam Najafian, Oana Inel, and Mesut Kaya for their comments on an earlier draft of this paper.

7. ADDITIONAL AUTHORS

Additional authors: Alessandro Bozzon (Delft University of Technology, The Netherlands, email: `a.bozzon@tudelft.nl`) and Benjamin Timmermans (IBM, The Netherlands, email: `b.timmermans@nl.ibm.com`).

8. REFERENCES

[1] A. Abid, N. Hussain, K. Abid, F. Ahmad, M. S. Farooq, U. Farooq, S. A. Khan, Y. D. Khan, M. A. Naeem, and N. Sabir. A survey on search results diversification techniques. *Neural Comput. Appl.*, 27(5):1207–1229, 2015.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. *Proc. 2nd ACM Int. Conf. Web Search Data Mining, WSDM'09*, pages 5–14, 2009.

[3] A. Asudeh, H. V. Jagadish, J. Stoyanovich, and G. Das. Designing fair ranking schemes. In *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pages 1259–1276, 2019.

[4] R. Baeza-Yates. Bias on the web. *Commun. ACM*, 61(6):54–61, 2018.

[5] A. D. Barocas, Solon and Selbst. Big data's disparate impact. *Calif. Law Rev.*, 104(671):671–732, 2016.

[6] R. K. Bellamy, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, and S. Mehta. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.*, 63(4-5), 2019.

[7] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. *41st Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 2018*, pages 405–414, 2018.

[8] L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. *Leibniz Int. Proc. Informatics, LIPIcs*, 107:1–32, 2018.

[9] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. *ACM SIGIR 2008 - 31st Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, Proc.*, pages 659–666, 2008.

[10] A. Das and M. Lease. A Conceptual Framework for Evaluating Fairness in Search. *arXiv Prepr. arXiv1907.09328*, 2019.

[11] R. Epstein and R. E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proc. Natl. Acad. Sci. U. S. A.*, 112(33):E4512–E4521, 2015.

[12] B. Fuglede and F. Topsøe. Jensen-Shannon divergence and Hubert space embedding. In *IEEE Int. Symp. Inf. Theory - Proc.*, page 31, 2004.

[13] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. *SIGIR 2005 - Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, 51(1):154–161, 2005.

[14] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22(1):79–86, 1951.

[15] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios. Search bias quantification: investigating political bias in social media and web search. *Inf. Retr. J.*, 22(1-2):188–227, 2019.

[16] J. Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory*, 37(1):145–151, 1991.

[17] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, and S. Staab. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 10(3):1–14, 2020.

[18] B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka. In Google we trust: Users' decisions on rank, position, and relevance. *J. Comput. Commun.*, 12(3):801–823, 2007.

[19] F. A. Pogacar, A. Ghenai, M. D. Smucker, and C. L. Clarke. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proc. ACM SIGIR Int. Conf. Theory Inf. Retr.*, ICTIR '17, pages 209–216, New York, NY, USA, 2017. Association for Computing Machinery.

[20] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. *SIGIR'11 - Proc. 34th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pages 1043–1052, 2011.

[21] T. Sakai and Z. Zeng. Which Diversity Evaluation Measures Are " Good "? In *SIGIR'19*, pages 595–604, 2019.

[22] P. Sapiezynski, W. Zeng, R. E. Robertson, A. Mislove, and C. Wilson. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proc. 2019 World Wide Web Conf.*, WWW '19, pages 553–562, New York, NY, USA, 2019. Association for Computing Machinery.

[23] A. Singh and T. Joachims. Fairness of exposure in rankings. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pages 2219–2228, 2018.

[24] S. Verma and J. Rubin. Fairness definitions explained. In *Proc. Int. Work. Softw. Fairness*, FairWare '18, pages 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

[25] R. Wang, D. Zhou, M. Jiang, J. Si, and Y. Yang. A survey on opinion mining: From stance to product aspect. *IEEE Access*, 7:41101–41124, 2019.

[26] F. L. Wauthier and M. I. Jordan. Bayesian bias mitigation for crowdsourcing. In *Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011*, pages 1–9, 2011.

[27] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proc. 29th Int. Conf. Sci. Stat. Database Manag.*, SSDBM '17, pages 1–6, New York, NY, USA, 2017. Association for Computing Machinery.

[28] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. FA*IR: A Fair Top-k Ranking Algorithm. In *Admit One*, pages 1–18, 2017.

Generative Counterfactuals for Neural Networks via Attribute-Informed Perturbation

Fan Yang, Ninghao Liu, Mengnan Du, Xia Hu
Department of Computer Science and Engineering, Texas A&M University
{nacoyang, nhliu43, dumengnan, xiahu}@tamu.edu

ABSTRACT

With the wide use of deep neural networks (DNN), model interpretability has become a critical concern, since explainable decisions are preferred in high-stake scenarios. Current interpretation techniques mainly focus on the feature attribution perspective, which are limited in indicating *why* and *how* particular explanations are related to the prediction. To this end, an intriguing class of explanations, named *counterfactuals*, has been developed to further explore the "what-if" circumstances for interpretation, and enables the reasoning capability on black-box models. However, generating counterfactuals for raw data instances (i.e., text and image) is still in the early stage due to its challenges on high data dimensionality and unsemantic raw features. In this paper, we design a framework to generate counterfactuals specifically for raw data instances with the proposed **A**ttribute-**I**nformed **P**erturbation (**AIP**). By utilizing generative models conditioned with different attributes, counterfactuals with desired labels can be obtained effectively and efficiently. Instead of directly modifying instances in the data space, we iteratively optimize the constructed attribute-informed latent space, where features are more robust and semantic. Experimental results on real-world texts and images demonstrate the effectiveness, sample quality as well as efficiency of our designed framework, and show the superiority over other alternatives. Besides, we also introduce some practical applications based on our framework, indicating its potential beyond the model interpretability aspect.

1. INTRODUCTION

The past decade has witnessed the success of deep neural networks (DNN) in many application domains [31]. Despite the superior performance, DNN models have been increasingly criticized due to its black-box nature [6]. Interpretable machine learning techniques [7] are thus becoming significantly vital, especially in those high-stake scenarios, such as medical diagnosis. To effectively interpret black-box DNNs, most approaches investigate the feature attributions between input instances and output predictions through correlation analysis, so that humans can have a sense of which part of the instance contributes most to the model decision. A typical example is the heatmaps employed for image classification [36], where saliency scores are capable of indicating the feature importance for one prediction label.

However, existing correlation-based explanations are neither discriminative nor counterfactual [29], since they are not able to help understand why and how particular explanations are relevant to model decisions. Thus, to further explore the decision boundaries of black-box DNN, *counterfactuals* have gradually come to the attention of researchers, as an emerging technique for model interpretability. Counterfactuals are essentially some synthetic samples within data distribution, which can flip the prediction. With counterfactuals, humans can understand how input changes affect the model and conduct reasoning under "what-if" circumstances. Take a loan applicant who got rejection for instance. Correlation-based explanations may simply indicate those most contributed features (e.g., income and credit) for rejection, while counterfactuals are capable of showing how the application could be accepted with certain changes (e.g., increase the monthly income from $\$5,000$ to $\$7,000$).

Recent work have made some initial attempts on conducting counterfactual analysis. The first line of research [19; 3] employed the prototype and criticism samples in the training set as the raw ingredients for counterfactual analysis, though those selected samples are not counterfactuals in nature. Some other work [12; 1] utilized feature replacement techniques to create hypothetical instances as counterfactuals, where a query instance and a distractor instance are typically needed for counterfactual generation. Besides, contrastive intervention [5; 43] on the query instance is another way to generate counterfactuals. By reasonably perturbing input features, counterfactuals can be obtained in the form of modified data samples.

Despite the existing efforts, generating valid counterfactuals for raw data instances is still challenging due to the following reasons. First, effective counterfactuals for certain label are not guaranteed to be existed in training set, so the selected prototypes and criticisms are not always sufficient for counterfactual analysis. The related sample selection algorithms are highly possible to select some "unexpected" instances due to data constraints [19], which would largely limit the reasoning on model behaviors. Second, efficient feature replacement for raw data instances could be very hard and time-consuming [12]. Also, relevant distractor instances for replacement may not be available in particular scenarios considering privacy and security issues, such as loan applications. Third, modifying query samples with intervention can simply work on a limited types of data, such as tabular data [43] and naive image data [5]. For general raw data like real-world texts or images, intervention operation in data space can be intractable, which makes it difficult to be used in practice.

To handle the aforementioned challenges, the high-dimension data space and unsemantic raw features are the two obstacles ahead. In this paper, we design a framework to generate counterfactuals specifically for raw data instances with the proposed **A**ttribute-**I**nformed **P**erturbation (**AIP**) method. By utilizing the power of generative models, we can obtain useful hypothetical instances within the data distribution for counterfactual analysis. Essentially, our proposed AIP can guide a well-trained generative model to generate valid counterfactuals by updating representations in the attribute-informed latent space, which is a concatenated coding space for both raw features and semantic attributes. Due to the different data formats, raw features and attributes are typically encoded in different ways to construct such attribute-informed space. Compared with the input space, attribute-informed latent space has two merits for counterfactual generation: (1) raw features are encoded as low-dimension ones which are more robust and efficient for generation; (2) data attributes are modeled as joint latent features which are more semantic for conditional generation. As for the construction of attribute-informed latent space, we employ two losses to conduct the training of generative models, where the reconstruction loss is used to guarantee the quality of raw feature embedding and the discrimination loss is used to ensure the correct attribute embedding. Through the adaptive gradient-based optimization, AIP can iteratively derive valid counterfactuals which are able to flip the prediction. The main contributions of this paper are summarized as follows:

- We design a general framework to derive counterfactuals for raw data instances by employing generative models;
- We develop AIP to iteratively update the parameters of generative models in an attribute-informed latent space;
- We evaluate the designed framework with AIP on several real-world datasets including raw texts and images, and demonstrate the superiority over other alternatives.

2. PRELIMINARIES

In this section, we briefly introduce related contexts to our problem, as well as some basics of the employed techniques.

2.1 Counterfactual Explanation

Counterfactual explanation is a natural extension under the framework of example-based reasoning [34], where particular data samples are provided to promote the understandings on model behaviors. Nevertheless, counterfactuals are not common examples, since they are typically generated under the "what-if" circumstances which may not necessarily exist. According to the theory proposed by J. Pearl [30], three distinct levels of cognitive ability are needed to fully master the behaviors of a particular model, i.e., *seeing*, *doing* and *imagining* from the easiest to the hardest. In fact, counterfactual explanation is just raised to meet the imagining-level cognition for model interpretation.

Within the contexts of this paper, we only discuss counterfactuals under the assumption of "closest possible world" [42], where desired outcomes can be obtained through the smallest changes to the world. To be specific and simple without loss of generality, consider a binary classification model $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \{0, 1\}$, where 0 and 1 respectively indicate the undesired and desired output. The model input $\mathbf{x} \in \mathbb{R}^d$ is further assumed to be sampled from data distribution $\mathcal{P}(\mathbf{x})$. Then, given a query instance $\mathbf{x}_0$ with the undesired model output (i.e., $f_{\boldsymbol{\theta}}(\mathbf{x}_0) = 0$), the corresponding counterfactual $\mathbf{x}^*$ can be mathematically represented as:

$$\mathbf{x}^* = \underset{\mathbf{x}|\mathcal{P}(\mathbf{x})>\eta}{\arg\min}\, l(\mathbf{x}, \mathbf{x}_0) \quad \text{s.t. } f_{\boldsymbol{\theta}}(\mathbf{x}^*) = 1, \tag{1}$$

where $l : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is a distance measure in the input space, and $\eta > 0$ denotes the threshold which quantifies how likely the sample $\mathbf{x}$ is under the distribution $\mathcal{P}(\mathbf{x})$. The obtained counterfactual $\mathbf{x}^*$ is regarded to be valid if it can flip the target classifier $f_{\boldsymbol{\theta}}$ to the desired prediction.

Although finding counterfactuals is somewhat similar to generating adversarial examples (in terms that both tasks aim to flip the model decision by minimally perturbing the input instance), they are essentially different in nature. Following the previous settings, the adversarial sample $\mathbf{x}^{adv}$ for model $f_{\boldsymbol{\theta}}$, with query instance $\mathbf{x}_0$, can be generally indicated by:

$$\mathbf{x}^{adv} = \underset{\mathbf{x}=\mathbf{x}_0+\boldsymbol{\delta}}{\arg\min} \|\boldsymbol{\delta}\|_p \quad \text{s.t. } f_{\boldsymbol{\theta}}(\mathbf{x}^{adv}) \neq f_{\boldsymbol{\theta}}(\mathbf{x}_0), \tag{2}$$

where $\boldsymbol{\delta}$ denotes the adversarial perturbation on the query, $\|\cdot\|$ represents the norm operation and $p \in \{\infty, 1, 2, \cdots\}$. Comparing with Eq. 1, we note that counterfactual example has two significant differences from adversarial sample. First, counterfactual generation process is subject to the original data distribution, while adversarial samples are not constrained by the distribution. This difference brings about the fact that counterfactuals are all in-distribution samples, but adversarial examples are mostly out-of-distribution (OOD) samples. Second, counterfactual changes on the query need to be human-perceptible, while adversarial perturbations are usually inconspicuous [37]. Therefore, the key problem of counterfactual explanation actually lies in how to generate such in-distribution sample, with human-perceptible changes on the query, to flip the model decision as desired.

2.2 Generative Modeling

Generative modeling is a typical task under the paradigm of unsupervised learning. Different from discriminative ones, which involves discriminating input samples across classes, generative modeling aims to summarize the data distribution of input variables and further create new samples that plausibly fit into that distribution [28]. In practice, a well-trained generative model is capable of generating new examples that are not only reasonable, but also indistinguishable from real examples in the problem domain. Conventional examples of generative modeling include Latent Dirichlet Allocation (LDA) and Gaussian Mixture Model (GMM).

As emerging families of generative modeling, Generative Adversarial Network (GAN) [10] and Variational Auto-Encoder (VAE) [21] have been attracting lots of attentions due to their exceptional performance in a myriad of applications, especially for the task on image and text generation [41; 17]. By taking full advantage of their power on raw data with high dimensionality, we are able to better investigate how those data samples were created in the first place, which potentially benefits the generation of certain hypothetical example. To this end, we specifically employ some advanced generative models (i.e., GAN and VAE) to study the counterfactual explanation for black-box DNN on raw data instances, providing effective generative counterfactuals for better model understanding.

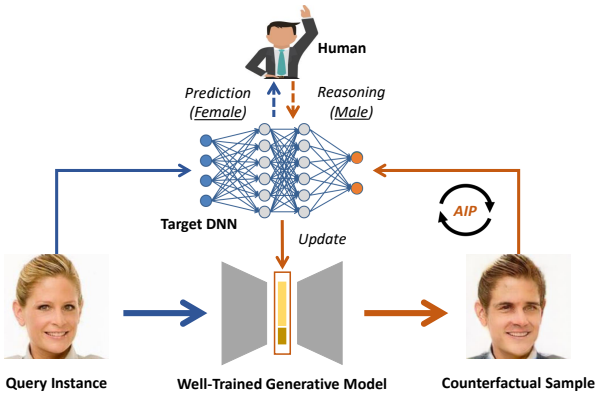


Figure 1: Designed framework for counterfactual generation.

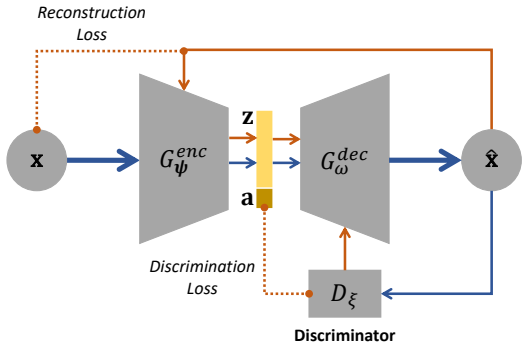


Figure 2: General illustration of the attribute-informed latent space in generative models.

3. COUNTERFACTUAL GENERATION

In this section, we first introduce the designed generative counterfactual framework for raw data instances. Then, we present how to specifically construct the attribute-informed latent space with generative models. Finally, we show the details of our proposed AIP method.

3.1 Generative Counterfactual Framework

We design a framework to create counterfactual samples for raw data instances, as illustrated by Fig. 1. To effectively handle the high dimensionality and unsemantic features, we utilize the generative modeling techniques to aid the counterfactual generation process. Consider a target DNN $F_{\phi}: \mathbb{R}^d \rightarrow \{1, \cdots, C\}$, which is a black-box model for counterfactual analysis, where $\mathbb{R}^d$ is the input data space and $\{1, \cdots, C\}$ denotes the model prediction space with C different outputs. Given a query instance $\mathbf{x}_0$, $F_{\phi}(\mathbf{x}_0) = \mathbf{y}_0$ outputs a one-hot vector. To effectively generate a valid counterfactual sample $\mathbf{x}^* \in \mathbb{R}^d$ that can flip the F_{ϕ} decision to $\mathbf{y}^* \in \{1, \cdots, C\}$ as desired, a generative model is trained to achieve this in the framework. The applied generative modeling plays two important roles in the counterfactual generation process: (1) guarantee that all created instances are in-distribution samples, since it can be regarded as a stochastic procedure that generates samples $\mathbf{x} \in \mathbb{R}^d$ under the particular data distribution $\mathcal{P}(\mathbf{x})$; (2) assume that underlying latent variables can be mapped to the data space under certain circumstances, which ensures the sufficient feasibility for hypothetical examples. Thus, a well-trained generative model is the basis for high-quality counterfactuals within the designed framework.

The employed generative model specifically serves two subtasks for counterfactual generation, i.e., data *encoding* and *decoding*. For raw data instances like images, the input space $\mathbb{R}^d$ could be extremely large, which makes it difficult and inefficient to directly create counterfactuals for the query. In our designed framework, data encoding is conducted to map the input data space to a low-dimension attribute-informed latent space, which is formulated as a joint embedding space for both raw features and data attributes. In this way, each data sample $\mathbf{x}$ can be effectively encoded through the function $G_{\psi}^{enc}: \mathbb{R}^d \rightarrow \mathbb{R}^k \oplus \mathbb{R}^t$, where $\mathbb{R}^k$ is the latent space for raw feature embeddings, $\mathbb{R}^t$ indicates the data attribute space, and $\oplus$ represents a concatenation operator. Reversely, the mapping for decoding is from the attribute-informed latent space to the original data space. The decoder function can be similarly indicated by $G_{\omega}^{dec}: \mathbb{R}^k \times \mathbb{R}^t \rightarrow \mathbb{R}^d$. Although G_{ψ}^{enc} and G_{ω}^{dec} typically have two different focuses, they are jointly trained as a whole generative model in an end-to-end manner. The issues about how to derive G_{ψ}^{enc} and G_{ω}^{dec} will be particularly discussed in Sec. 3.2.

To finally obtain the counterfactual $\mathbf{x}^*$ for model F_{ϕ} with query $\mathbf{x}_0$, we further need to modify the attribute-informed latent space of the deployed generative model. Specifically, we use the proposed AIP to update the attribute-informed latent vector of $\mathbf{x}_0$, according to the counterfactual loss calculated. Assuming $G_{\psi}^{enc}(\mathbf{x}_0) = \mathbf{z}_0 \oplus \mathbf{a}_0$ ($\mathbf{z}_0 \in \mathbb{R}^k, \mathbf{a}_0 \in \mathbb{R}^t$), AIP method can jointly update $\mathbf{z}_0$ and $\mathbf{a}_0$, so as to minimize the corresponding loss counter-factually. The overall counterfactual loss consists of two parts, i.e., *prediction loss* and *perturbation loss*. Prediction loss is set to ensure the flip of model decisions, and perturbation loss is involved to guarantee the "closest possible" changes on the query, which are both indispensable for counterfactual generation. For the prediction loss, we simply follow the common cross-entropy term, expressed as $L_d(F_{\phi}(\mathbf{x}), \mathbf{y}^*) = -\mathbf{y}^* \log(F_{\phi}(\mathbf{x})) - (1 - \mathbf{y}^*) \log(1 - F_{\phi}(\mathbf{x}))$. For the perturbation loss, we employ two l_2 norms respectively on $\mathbb{R}^k$ and $\mathbb{R}^t$, indicated by $L_b(\mathbf{z}, \mathbf{a}, \mathbf{z}_0, \mathbf{a}_0) = \|\mathbf{z} - \mathbf{z}_0\|_2 + \|\mathbf{a} - \mathbf{a}_0\|_2$ ($\mathbf{z} \in \mathbb{R}^k, \mathbf{a} \in \mathbb{R}^t$), to restrain the query changes, which can also be regarded as a regularization term. Further, the overall counterfactual loss can thus be represented as follows:

$$\begin{aligned} L_c(\mathbf{z}, \mathbf{a}, \mathbf{z}_0, \mathbf{a}_0, \mathbf{y}^*) =& L_d\left(F_{\phi}(G_{\omega}^{dec}(\mathbf{z}, \mathbf{a})), \mathbf{y}^*\right) \\ &+ \alpha L_b(\mathbf{z}, \mathbf{a}, \mathbf{z}_0, \mathbf{a}_0), \end{aligned} \tag{3}$$

where α is a balance coefficient between the two loss terms. With the proposed AIP method, the designed framework can generate the valid counterfactual example $\mathbf{x}^*$ with the aid of optimized $\mathbf{z}^*, \mathbf{a}^*$ through the decoder function (i.e., $\mathbf{x}^* = G_{\omega}^{dec}(\mathbf{z}^*, \mathbf{a}^*)$). The details of the proposed AIP method will be introduced in Sec. 3.3.

3.2 Attribute-Informed Latent Space

Constructing an appropriate attribute-informed latent space is the key part for generative modeling in our designed framework, which has direct influences on the quality of generated counterfactuals. To achieve this, we need to well train a gen-

erative model, better capturing the raw data features as well as relevant data attributes, where embedded features can bring about more robust bases for counterfactual analysis, and incorporated attributes are able to provide more semantics for conditional generation. Here, the data attributes mainly indicate those extra information from humans along with raw instances, such as annotations or labels, which can usually be represented as one-hot vectors.

In practice, it is common that different generative models are employed for different tasks or data. Since different models typically involve disparate architectures, their training schemes can totally differ from each other. Take the GAN and VAE for example, where GAN is usually trained to obtain an equilibrium between a generator and a discriminator function, while VAE is typically trained to maximize a variational lower bound of the data log-likelihood. Therefore, to better introduce how to specifically construct the attribute-informed latent space with generative models, we present a general illustration shown by Fig. 2, although it may not be fully representative for all kinds of models.

We generally introduce the data modeling process with an encoder-decoder structure, which corresponds to the data encoding and decoding in our designed framework. Essentially, the attribute-informed latent space can be regarded as an extended code space of auto-encoders, where attribute information is properly incorporated into the representation beyond the raw feature inputs. By concatenating attribute vector $\mathbf{a}$ to raw feature embedding $\mathbf{z}$, the decoder function aims to achieve the conditional generation based on $\mathbf{a}$. To ensure the attribute consistency between original sample $\mathbf{x}$ and generated sample $\hat{\mathbf{x}}$, discriminator D_ξ is particularly employed, which is trained separately and used to classify the attributes of $\hat{\mathbf{x}}$. To effectively train such generative model, two basic loss terms are required, which are the discrimination loss and reconstruction loss. The overall training can be indicated by:

$$\min_{\psi,\omega} \mathop{\mathbb{E}}_{\substack{\mathbf{x}\sim\mathcal{P}(\mathbf{x})\\ \mathbf{a}\sim\mathcal{P}(\mathbf{a})}} \sum_{i=1}^{t} -a_i \log D_\xi^i(\hat{\mathbf{x}}) - (1-a_i)\log\left(1-D_\xi^i(\hat{\mathbf{x}})\right) + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{P}(\mathbf{x})} \|\mathbf{x}-\hat{\mathbf{x}}\|_2, \tag{4}$$

where a_i denotes the i-th attribute in $\mathbf{a}$, and D_ξ^i indicates the prediction of D_ξ on the i-th attribute. After sufficient training, $G_{\boldsymbol{\psi}}^{enc}$ and $G_{\boldsymbol{\omega}}^{dec}$ can be effectively obtained, and the attribute-informed latent space can be constructed with the aid of $G_{\boldsymbol{\psi}}^{enc}$. For specific tasks and architectures, the generative modeling process could be further enhanced with other loss terms, such as the adversarial losses employed in [16] for better visual quality.

3.3 Attribute-Informed Perturbation

With the obtained $G_{\boldsymbol{\psi}}^{enc}$ and $G_{\boldsymbol{\omega}}^{dec}$ for generative modeling, we then introduce the proposed AIP method to finally derive the counterfactual for target DNN $F_{\boldsymbol{\phi}}$ with the query $\mathbf{x}_0$. To guarantee the quality of the generated counterfactuals, AIP needs to find the sample that can minimize the counterfactual loss indicated by Eq. 3. Under the "closest possible world" assumption, the corresponding counterfactual sample can be denoted as:

$$\mathbf{x}^* = G_{\boldsymbol{\omega}}^{dec}\left(\mathop{\arg\min}_{\mathbf{z}\in\mathbb{R}^k,\ \mathbf{a}\in\mathbb{R}^t} L_c(\mathbf{z},\mathbf{a},\mathbf{z}_0,\mathbf{a}_0,\mathbf{y}^*)\right). \tag{5}$$

Algorithm 1: Attribute-Informed Perturbation (AIP)

Input: $F_{\boldsymbol{\phi}}$, $G_{\boldsymbol{\psi}}^{enc}$, $G_{\boldsymbol{\omega}}^{dec}$, $\mathbf{x}_0$, $\mathbf{y}^*$, μ, γ, α, β, $n_{\max}$
Output: Counterfactual sample $\mathbf{x}^*$

```
Initialize μ, γ, α, β ;
Initialize n = 0, x = x_0 ;
Construct the latent space with z ⊕ a ← G_ψ^enc(x) ;
while F_φ(G_ω^dec(z, a)) ≠ y* or n ≤ n_max do
    Update z and a according to Eq. 6 ;
    Update step sizes with μ ← βμ and γ ← βγ ;
    n ← n + 1;
Reconstruct the sample with x* ← G_ω^dec(z, a) ;
if F_φ(x*) == y* then
    Return x* as the counterfactual for F_φ with x_0;
else
    Return None – No valid counterfactual exists;
```

To effectively solve Eq. 5, the proposed AIP method utilizes an iterative gradient-based optimization algorithm with dynamic step sizes (controlled by a decaying factor β), which helps the iteration process converge faster. In each iteration, the updated $\mathbf{z}$ and $\mathbf{a}$ can be derived as follows:

$$\begin{cases} \mathbf{z}^{(n+1)} = \mathbf{z}^{(n)} - \mu^{(n)} \nabla_{\mathbf{z}} L_c\left(\mathbf{z}^{(n)}, \mathbf{a}^{(n)},\ \mathbf{z}_0,\ \mathbf{a}_0,\ \mathbf{y}^*\right) \\ \mathbf{a}^{(n+1)} = \mathbf{a}^{(n)} - \gamma^{(n)} \nabla_{\mathbf{a}} L_c\left(\mathbf{z}^{(n)}, \mathbf{a}^{(n)},\ \mathbf{z}_0,\ \mathbf{a}_0,\ \mathbf{y}^*\right) \end{cases}, \tag{6}$$

where n indicates the iteration index, μ and γ respectively denotes the step sizes of updates on $\mathbf{z}$ and $\mathbf{a}$. Specifically, the proposed AIP method can be summarized in Algorithm 1. It is important to note that AIP only works on the optimization of $\mathbf{z}, \mathbf{a}$, and does not involve the parameter update on $F_{\boldsymbol{\phi}}$, $G_{\boldsymbol{\psi}}^{enc}$, $G_{\boldsymbol{\omega}}^{dec}$. Thus, the proposed AIP method should be less time-consuming and easily deployed for counterfactual generation task, compared with those generative frameworks which need extra model training [35; 38].

4. EXPERIMENTS

In this section, we evaluate the designed counterfactual generation framework with the proposed AIP on several real-world datasets, both quantitatively and qualitatively. Overall, we conduct two sets of experiments respectively on *text* and *image* counterfactual generation, by utilizing different data modeling techniques. With conducted experiments, we aim to answer the following four research questions:

- How *effective* is the designed framework in generating counterfactuals with AIP?
- How is the *quality* of created counterfactuals from our designed framework aided by AIP?
- How *efficient* is the counterfactual generation under the designed framework with AIP?
- Can we *benefit* other practical tasks with the counterfactuals generated from our design framework with AIP?

4.1 Experimental Settings

4.1.1 Real-World Datasets.

Throughout the whole experiments, we employ three real-world datasets to evaluate the performance of the designed

Table 1: Dataset statistics in experiments.

Datasets	#Instance	#Attribute	Type	Domain
Yelp	455,000	1	Texts	Sentiment
Amazon	558,000	1	Texts	Sentiment
CelebA	202,599	13	Images	Human Face

framework with AIP, including both raw texts and images. The relevant data attributes depend on the particular tasks, which are collected either from labels or annotations. The statistics of the involved datasets are shown in Table 1.

- **Yelp User Review Dataset**[1] [2]: This dataset consists of user reviews from the Yelp associated with relevant rating scores. We involve a tailored and modified version of this data for our experiments on text counterfactuals. Specifically, we consider the reviews with ratings higher than three as *positive* samples and regard the others as *negative* ones, and we further use these sentiment labels as the relevant attribute for data modeling. The vocabulary of our involved Yelp data contains more than 9,000 words, and the average review length is around 9 words.
- **Amazon Product Review Dataset**[2] [15]: This dataset is also involved as a raw textual dataset for our experiments. Similar to the Yelp data, we modify the original rating information of reviews into the sentiment categories (i.e., *positive* and *negative*), and further model these labels as an sentiment attribute of the raw textual reviews. Amazon dataset has more than 50,000 words in vocabulary, and the average length is around 15.
- **CelebFaces Attributes (CelebA) Dataset**[3] [24]: This is a large-scale face attributes dataset, containing tons of raw face images with human annotations. We employ this dataset for our experiments on image counterfactual generation, and select 13 representative face attributes (out of 40) for data modeling along with raw face images. The involved thirteen attributes include: *Male*, *Young*, *Blond_Hair*, *Pale_Skin*, *Bangs*, *Mustache* and etc.

4.1.2 Target Model for Interpretation.

Since we mainly discuss the counterfactuals of raw data instances, DNN is a better choice as our target model. For the target DNN F_ϕ in our experiments, we employ some regular structures for the corresponding tasks. Particularly, for the text sentiment classification, we use a common convolutional architecture in [20] to pre-train a DNN classifier for further counterfactual analysis. For the image attribute classification task, similarly, we utilize a simple convolutional network [13] to prepare a target classifier, where the model is trained with one of those attributes as the label. During the evaluations on counterfactual generation, target DNNs are fixed without further training.

4.1.3 Employed Generative Modeling Techniques.

In our experiments, different data modeling techniques are employed for different types of data. In particular, we use different generative models to construct the corresponding attribute-informed latent space, regarding to text and image. For textual reviews (Yelp, Amazon), we utilize the modeling techniques introduced in [17], and build a transformer-based VAE to effectively formulate the relevant attribute-informed latent space. For face images (CelebA), we mainly follow the modeling method of AttGAN [16], where more complicated training schemes are employed, compared with the general one shown in Sec. 3.2, for better visual quality of the generated images. Both of the employed generative models should be well-trained on the corresponding datasets before the counterfactual generation process, so as to guarantee the high quality of generated counterfactuals.

4.1.4 Alternative Methods and Baselines.

To effectively evaluate the performance of the designed framework with AIP, we incorporate following alternative methods and baselines for comparison.

- **TextBugger** [22]: This is a general method for adversarial text generation, which is built based on the word attribution and bug selection. The created text samples can effectively flip the prediction of the target classifier. We employ this method as a baseline to specifically compare with our generated text counterfactuals.
- **DeepWordBug** [9]: This is another method focusing on the adversarial text generation, where a token scoring strategy is utilized to guide the character-level adversarial perturbation. This method is employed as a baseline for text counterfactuals as well.
- **FGSM** [11]: Fast gradient sign method is a common way to generate image adversarial samples, by using the gradients of the loss with respect to the input. The sample created by this method can effectively maximize the loss, so as to flip the original prediction. We employ this method as a baseline specifically for our generated image counterfactuals.
- **Counter_Vis** [12]: This is a recent method in generating image counterfactuals, where particular image regions are replaced to flip the model decision. We employ this method as an alternative method for image counterfactual generation.
- **CADEX** [27]: This is a state-of-the-art method for counterfactual generation, where the gradient-based method is directly applied to modify the input space of query. This method is originally proposed for tabular data, and we modify it simply as an alternative for image counterfactuals, due to the particularity of texts.
- **xGEMs** [18]: This is a state-of-the-art method for generating counterfactuals, which also employs the generative modeling technique for sample generation. This method only involves the latent space modeling and cannot achieve the conditional generation with semantic attributes. We employ this method as an important alternative for both text and image counterfactuals.
- **AIP_R**: This is the random version of our proposed AIP method, which updates all parameters in a random way.

[1] `https://www.yelp.com/dataset`
[2] `http://jmcauley.ucsd.edu/data/amazon/index_2014.html`
[3] `http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`

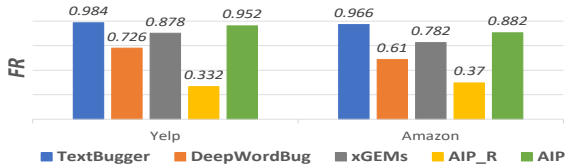


Figure 3: Effectiveness evaluation for text counterfactuals.

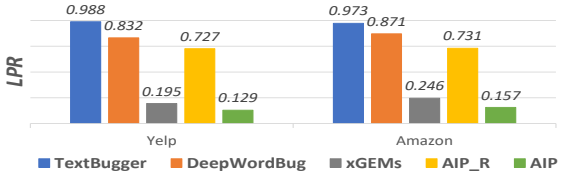


Figure 4: Quality evaluation for text counterfactuals.

4.2 Text Counterfactual Evaluations

In this part, we evaluate the designed framework with AIP in generating text counterfactuals, regrading to a convolutional neural net (CNN) built for sentiment classification. The involved raw texts for target DNN come from the user/product reviews in Yelp and Amazon datasets, where 90% are used for training, 5% for development and 5% for testing.

4.2.1 *Effectiveness Evaluation.*

In order to evaluate the effectiveness for text counterfactuals, we employ the metric *Flipping Ratio* (FR) to measure the relevant performance, which reflects how likely the generated text samples would flip the model decision to $\mathbf{y}^*$. Specifically, FR can be calculated as follows:

$$\text{FR} = |\mathcal{X}_f| \Big/ |\mathcal{X}_q| \quad (\mathbf{x}_0 \in \mathcal{X}_f \text{ if } F_{\phi}(\mathbf{x}_0) = \mathbf{y}^*), \tag{7}$$

where $\mathcal{X}_f$ indicates the set of query samples with which new flipping instances can be generated by particular methods, and $\mathcal{X}_q$ denotes the set of all testing queries. In our experiments, there are 500 testing queries in total (i.e., $|\mathcal{X}_q| = 500$), which are randomly selected from the test set. Fig. 3 illustrates our experimental results on both Yelp and Amazon datasets. According to the numerical results, we note that our designed framework with AIP can work well on both datasets, and has competitive performance among all other alternatives as well as baselines, although TextBugger achieves the highest FR score with better robustness (i.e., the performance variance across different datasets). Besides, we also observe that AIP_R does not effectively work for generating flipping samples, which indicates that random optimization in attribute-informed latent space cannot help for counterfactual sample generation.

4.2.2 *Quality Evaluation.*

As for the quality assessment of counterfactual samples, we employ the *Latent Perturbation Ratio* (LPR) metric to measure the latent closeness between the generated sample $\mathbf{x}^*$ and original query instance $\mathbf{x}_0$. Since high-quality counterfactual samples typically need to ensure sparse changes in the robust feature space, thus the smaller the LPR is, the better the counterfactual we have. To be specific, the LPR

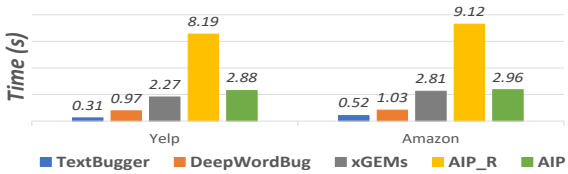


Figure 5: Efficiency evaluation for text counterfactuals.

Table 2: Case studies on generated text samples.

Counterfactual on *Negative* sentiment (Yelp)
Query: this is the worst walmart neighborhood market out of any of them
TextBugger: this is the worst walmart neighborho0d market out of a ny of them
DeepWordBug: this id the wosrt walmart neighobrhood market out of any of htem
xGEMs: that is good walmart market out of any neighborhood
AIP: this is the best walmart neighborhood market for all of them
Counterfactual on *Positive* sentiment (Amazon)
Query: this item works just as i thought it would
TextBugger: this item w0rks just as i tho ught it would
DeepWordBug: this item wroks just ae i thought it wolud
xGEMs: this item works out poorly just as i thought disappointed
AIP: this item works bad just as i thought it would not play

can be calculated by:

$$\text{LPR} = \left\|\mathbf{z}^* - \mathbf{z}_0\right\|_0 \Big/ k, \tag{8}$$

where $\|\cdot\|_0$ indicates the l_0 norm operation, $\mathbf{z}^*$ and $\mathbf{z}_0$ are the raw feature embeddings respectively for $\mathbf{x}^*$ and $\mathbf{x}_0$. To make a fair comparison, we use the same encoder function G_{ψ}^{enc} for all generated samples to obtain the corresponding latent representation vectors. In this set of experiments, the latent dimension is 256 (i.e., $k = 256$), and the final LPR value for a particular method is recorded with the average over 500 testing queries. The relevant numerical results are presented in Fig. 4. From the experiments, it is noted that xGEMs and the proposed AIP method significantly outperform other baselines, indicating that the corresponding generated samples actually maintain more robust features regarding to the query. Furthermore, the proposed AIP is noted to be slightly better than xGEMs, which may partially result from the conditional generation brought by attribute vector $\mathbf{a}$. This set of results also validate a fact that adversarial samples typically utilize some artifacts to flip the model decisions, instead of using some robust features.

4.2.3 *Efficiency Evaluation.*

To compare the efficiency, we record the time consumption for each method over 500 testing queries in the generation phase on the same machine. Specifically, we calculate the average time cost for one query, and further employ this as the metric to access the efficiency for particular methods. Fig. 5 shows the relevant experimental results. Based on the statistics, it is observed that adversarial related methods (i.e., TextBugger and DeepWordBug) consume less time per query on average, compared with the counterfactual generation methods, which is mainly due to the fact that adversarial methods do not need to conduct encoding computations before sample generation. As for our proposed AIP method, the time efficiency is roughly the same as the alternative xGEMs, but it is significantly better than its random version AIP_R which needs more iterations to converge.

4.2.4 *Qualitative Case Studies.*

Here, we present several representative case studies from

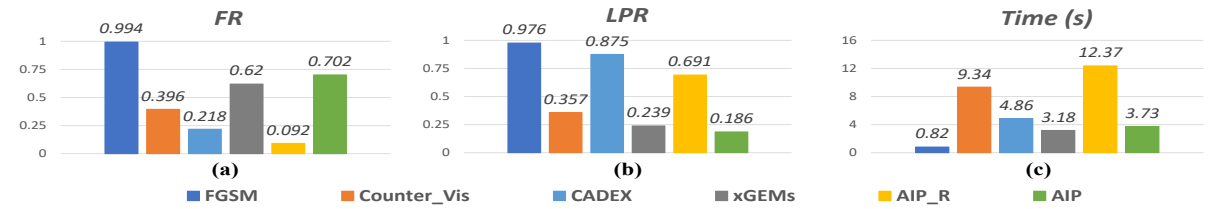


Figure 6: Evaluations for image counterfactual generation.

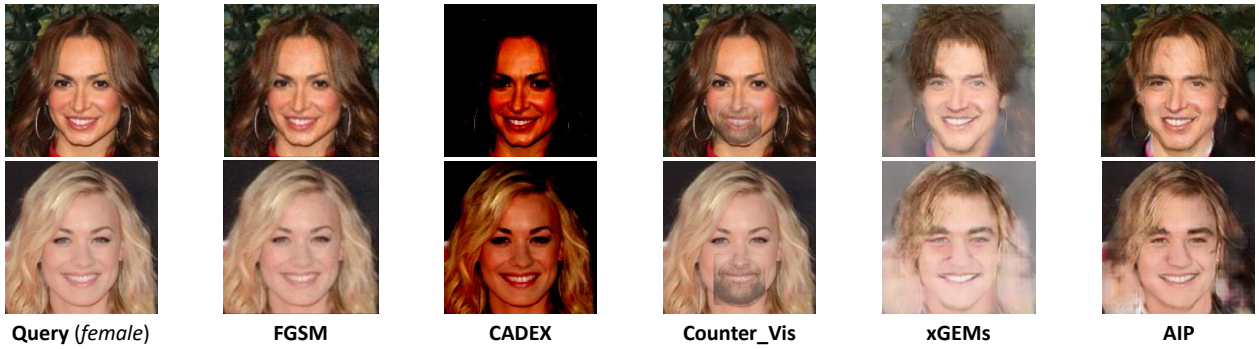


Figure 7: Qualitative case studies on generated image samples.

different methods, shown in Tab. 2, aiming to provide a qualitative comparison for generated text samples. Based on the Tab. 2, we can see that adversarial texts typically provide limited insights for humans on counterfactual analysis, since they mainly make use of the model artifacts to flip the prediction. Nevertheless, with the samples generated by xGEMs and AIP, we can easily observe some sentiment variation regarding to the query instance, which sheds light on model behaviors and facilitates further human reasoning on black-box models. Besides, compared with xGEMs, the proposed AIP method usually can generate more sensible counterfactuals with the aid of attribute conditions.

4.3 Image Counterfactual Evaluations

In this part, we specifically evaluate the designed framework with AIP on image counterfactual generation. Instead of simply considering one attribute for conditional generation in texts, we take multiple attributes into account for image counterfactuals. In this set of experiments, our target DNN follows the common CNN architecture and is trained as a gender classifier, which can classify an input image as *Male* or *Female*. All involved raw images for target DNN come from the CelebA dataset, and we use 90% data for training, 5% for development, 5% for testing. The relevant quantitative results are all illustrated by Fig. 6.

4.3.1 *Effectiveness Evaluation.*

For the effectiveness assessment, we still use the FR metric indicated by Eq. 7. In the experiments, we set $|\mathcal{X}_q| = 500$, and aim to test how many of them can be effectively flipped with particular methods. Fig. 6(a) illustrates the relevant numerical results, where adversarial method FGSM performs the best on FR and can flip nearly every testing query. We note that the proposed AIP method ranks the second, and outperforms other counterfactual generation methods. Besides, it is also observed that CADEX and AIP_R performs relatively bad for the image counterfactual task within certain iterations, even though CADEX is proved to work well for tabular instances [27].

4.3.2 *Quality Evaluation.*

Similar to text counterfactuals, we employ the LPR metric, shown as Eq. 8, to measure the quality of the generated image counterfactuals. In experiments, the latent dimension k constructed by G_{ψ}^{enc} is $1,024$ (i.e., $k = 1024$), and the LPR for particular method is recorded by calculating the average over 500 testing queries. Relevant experimental results are shown by Fig. 6(b). Based on the LPR comparison, we note that the samples generated by FGSM and CADEX change a lot in the latent feature space, because both methods directly rely on the input perturbation for sample generation. As for the proposed AIP, it achieves the lowest LPR among all the alternatives and baselines, and it is significantly better than its random version AIP_R.

4.3.3 *Efficiency Evaluation.*

We similarly employ the average time consumption per query to evaluate the efficiency aspect for image counterfactual generation. Specifically, the average time is obtained over the 500 testing queries randomly selected from the test set. Fig. 6(c) shows the relevant experimental results. According to the statistics and comparison, we note that FGSM is the most efficient one, and xGEMs consumes the least time on average among all other counterfactual-based methods. As for the proposed AIP, a competitive efficiency performance is observed, and is remarkably superior compared with that of Counter_Vis, CADEX and AIP_R.

4.3.4 *Qualitative Case Studies.*

To facilitate a qualitative comparison among different meth-

ods, we specifically show some case studies, illustrated by Fig. 7. We select several query instances whose model predictions are female, and then employ different methods to generate the corresponding image samples which flip the model decisions for counterfactual purpose. According to the results, we note that the samples generated by FGSM and CADEX do not have salient visual changes regarding to the query instances, which largely limits the human reasoning on model behaviors. Among other alternative methods, it is observed that the proposed AIP is capable of generating counterfactuals with better visual quality, which present much smoother transitions from female to male.

4.4 Influence of Hyper-parameter α

In this part, we show some additional results on the influence of hyper-parameter α in Eq. 3. Other experimental settings keep unchanged. The relevant results are shown by Fig. 8.

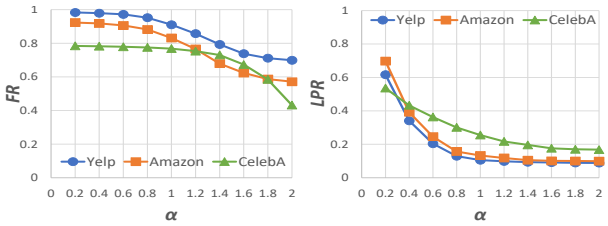


Figure 8: Influence of α on FR and LPR metrics.

Based on the results, we observe that α serves as a knob to control the effectiveness and sample quality of the designed framework. To select an appropriate α, we actually need to strike a balance between FR and LPR, where the larger the α is, the lower the effectiveness is and the higher the sample quality is. Different data types may also have different trade-off curves.

4.5 Applications

In this part, we focus on some practical scenarios which may benefit from the counterfactual samples generated by our designed framework. In particular, we show the applications of the framework respectively on *feature interaction* and *data augmentation*.

4.5.1 Feature Interaction

Understanding the feature interaction could be very important in lots of real-world domains. A typical example is the bias detection task, where humans aim to find out a related set of features which can significantly influence the correctness or fairness of model decision. Utilizing our designed framework for counterfactual analysis can partially help this practical task. By observing the perturbation scale on attribute vector $\mathbf{a}$ of the generated counterfactual, humans can have a sense on which semantic features contribute significantly to the flipping of model decision. To illustrate the point, we show another case result from the designed framework with AIP in Fig. 9. Here, we train an age classifier on the CelebA dataset as our target DNN, and aim to analyze the feature interaction of a query prediction as "Old". Based on the attribute perturbations of the generated sample, we note that the top semantic attributes are "Male", "Bushy_Eyebrows", "Black_Hair" and "Bangs", besides the target attribute. This result directly demonstrates

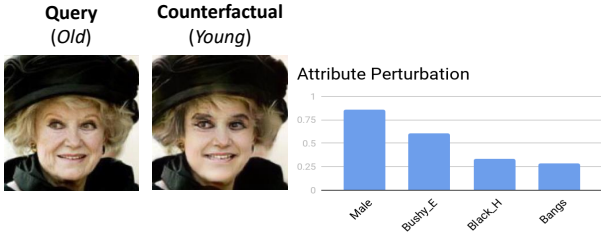


Figure 9: Feature interactions for the decision change.

Table 3: Model performance with data augmentation.

Dataset		CNN [20]	VDCNN [4]
Yelp	*Initial*	82.33% (± 0.61%)	88.79% (± 0.53%)
	Augmented	83.16% (± 0.57%)	89.95% (± 0.46%)
Amazon	*Initial*	81.96% (± 0.52%)	88.55% (± 0.63%)
	Augmented	82.41% (± 0.49%)	88.76% (± 0.55%)
Dataset		**CNN** [13]	**ResNet** [14]
CelebA	*Initial*	87.32% (± 0.22%)	90.96% (± 0.27%)
	Augmented	88.85% (± 0.21%)	91.35% (± 0.25%)

the fact that the "Male" attribute has a strong interaction with the predicted attribute for this particular query, and the target DNN exists potential gender bias for its age predictions. Please note that such feature interaction indicates the attribute sensitivity for flipping model decisions, instead of the attribution for predictions, which is different from the existing local interpretation methods (e.g., LIME [32]).

4.5.2 Data Augmentation

Another application of the designed framework is the data augmentation for model training. By taking full advantage of the generated counterfactual samples as new training instances, we aim to obtain better DNN models with higher performance and robustness. Specifically, to test the improvement, we train several DNN models on relatively smaller training sets, which are essentially the subsets of original data. For the sentiment classifiers on Yelp and Amazon, our initial training size is $20,000$, containing $10,000$ positive and $10,000$ negative reviews. The extra counterfactual training size is $2,000$ whose queries are randomly selected from the initial training set. For the binary age classifier on CelebA, we employ a similar setting for training, where each class includes $10,000$ initial samples, and $2,000$ generated counterfactual samples are further incorporated for augmentation. Relevant experimental results are shown in Tab. 3. Based on the statistics, we note that the augmented training with counterfactual samples typically achieves higher classification accuracies with smaller variances, which can also be observed under some advanced DNN structures.

5. RELATED WORK

Generating counterfactuals is just one of interpretation methods for black-box models, which generally belongs to the family of interpretable machine learning. According to the particular problems, interpretation methods can be divided into the following three categories in general.

The first category of methods aims to answer the "*What*"-type questions, i.e., what part of the input mostly contribute to the model prediction. A representative work in this cat-

egory is LIME [32], where authors employ linear models to approximate the local decision boundary and further formulate it as a sub-modular optimization problem for model interpretation. The feature importance in LIME is obtained by observing the prediction changes after perturbing input samples. Related methods can also be found in Anchors [33] and SHAP [25]. Another common methodology under this category is to utilize the model gradient information, where gradients are regarded as an indicator for perturbation sensitivity. Related methods can be found in GradCAM [36], Integrated Gradients [40], and SmoothGrad [39].

The second category aims to answer the "*Why*"-type questions, i.e., why the input is predicted as label A instead of B. The methods under this category can be quite different from the previous ones, since these methods need to consider two labels simultaneously. There are several different methods proposed for this problem. For example, the authors in [5] design a contrastive perturbation method to derive related positive and negative features of inputs regarding the concerned label. Besides, a general method based on structural causal models is proposed in [26] to tackle the problem in classification and planning scenarios. Also, a generative framework CDeepEx is designed in [8] to particularly investigate this problem for images by utilizing GAN.

The third category lies in the "*How*"-type questions, i.e., how to particularly modify the input so as to flip the model prediction to the preferred label. This problem is a natural extension of the "Why"-type, and it can somewhat to be handled by the second category of methods under some simple scenarios. However, for problems with high-dimension space, previous categories of methods typically fail due to the intractable computation for sample modification. Several particular methods are raised to solve this issue. For example, authors in [12] propose a straightforward solution with image region replacement, which is essentially a feature replacement process for input with the aid of a distractor. In work [1], authors novelly use the input itself as the distractor for feature replacement by utilizing GAN for inpainting. Besides, generative modeling is another potential way for this problem, and related methods can be found in [38; 18; 23]. Our work belongs to this branch of methodology.

6. CONCLUSION AND FUTURE WORK

In this paper, we design a framework to generate counterfactual explanation for black-box DNN models specifically with raw data instances. By taking advantage of the generative modeling, we effectively construct an attribute-informed latent space for particular data, and further utilize this space for counterfactual generation. To guarantee the validity of the generated samples, we propose the AIP method to iteratively optimize the specific attribute-informed latent vectors according to the counterfactual loss term, from which the counterfactuals can be finally obtained through data reconstruction. We evaluate the designed framework with AIP on several real-world datasets, including both texts and images, and demonstrate its effectiveness, sample quality as well as efficiency. Future extension of this work may possibly include the investigation under the "close possible worlds" assumption, where the goal is to find an optimal set of counterfactuals for a query instead of a single sample. Besides, employing causal models for counterfactual generation is another promising direction to explore.

7. REFERENCES

[1] C. Agarwal, D. Schonfeld, and A. Nguyen. Removing input features via a generative model to explain their attributions to classifier's decisions. *arXiv preprint arXiv:1910.04256*, 2019.

[2] N. Asghar. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*, 2016.

[3] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, pages 8928–8939, 2019.

[4] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, 2017.

[5] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in neural information processing systems*, pages 592–603, 2018.

[6] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[7] M. Du, N. Liu, and X. Hu. Techniques for interpretable machine learning. *Commun. ACM*, 63(1):68–77, 2020.

[8] A. Feghahati, C. R. Shelton, M. J. Pazzani, and K. Tang. Cdeepex: Contrastive deep explanations. 2018.

[9] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2018.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations*, 2015.

[12] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual visual explanations. In *ICML*, pages 2376–2384, 2019.

[13] T. Guo, J. Dong, H. Li, and Y. Gao. Simple convolutional neural network on image classification. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pages 721–724. IEEE, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] R. He and J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of WebConf*, 2016.

[16] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

[17] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1587–1596, 2017.

[18] S. Joshi, O. Koyejo, B. Kim, and J. Ghosh. xgems: Generating examplars to explain black-box models. *arXiv preprint arXiv:1806.08867*, 2018.

[19] B. Kim, R. Khanna, and O. O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in neural information processing systems*, pages 2280–2288, 2016.

[20] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the Conference on EMNLP*, pages 1746–1751, 2014.

[21] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] J. Li, S. Ji, T. Du, B. Li, and T. Wang. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*, 2018.

[23] S. Liu, B. Kailkhura, D. Loveland, and H. Yong. Generative counterfactual introspection for explainable deep learning. Technical report, Lawrence Livermore National Lab.(LLNL), 2019.

[24] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[25] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[26] T. Miller. Contrastive explanation: A structural-model approach. *arXiv preprint arXiv:1811.03163*, 2018.

[27] J. Moore, N. Hammerla, and C. Watkins. Explaining deep learning models with constrained adversarial examples. In *Pacific Rim International Conference on Artificial Intelligence*, pages 43–56. Springer, 2019.

[28] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[29] J. Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[30] J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[31] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys*, 51(5):1–36, 2018.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *32nd AAAI Conference on Artificial Intelligence*, 2018.

[34] E. L. Rissland. Example-based reasoning. *Informal reasoning in education*, pages 187–208, 1991.

[35] P. Samangouei, A. Saeedi, L. Nakagawa, and N. Silberman. Explaingan: Model explanation via decision boundary crossing transformations. In *Proceedings of ECCV*, pages 666–681, 2018.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[37] A. Sen, X. Zhu, L. Marshall, and R. Nowak. Should adversarial attacks use pixel p-norm? *arXiv preprint arXiv:1906.02439*, 2019.

[38] S. Singla, B. Pollack, J. Chen, and K. Batmanghelich. Explanation by progressive exaggeration. In *8th International Conference on Learning Representations*, 2020.

[39] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[40] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017.

[41] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

[42] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[43] A. White and A. d. Garcez. Measurable counterfactual local explanations for any classifier. *arXiv preprint arXiv:1908.03020*, 2019.

An Empirical Comparison of Bias Reduction Methods on Real-World Problems in High-Stakes Policy Settings

Hemank Lamba
Carnegie Mellon University
Pittsburgh, PA
hlamba@andrew.cmu.edu

Kit T. Rodolfa*
Carnegie Mellon University
Pittsburgh, PA
krodolfa@cmu.edu

Rayid Ghani*
Carnegie Mellon University
Pittsburgh, PA
rayid@cmu.edu

ABSTRACT

Applications of machine learning (ML) to high-stakes policy settings — such as education, criminal justice, healthcare, and social service delivery — have grown rapidly in recent years, sparking important conversations about how to ensure fair outcomes from these systems. The machine learning research community has responded to this challenge with a wide array of proposed fairness-enhancing strategies for ML models, but despite the large number of methods that have been developed, little empirical work exists evaluating these methods in real-world settings. Here, we seek to fill this research gap by investigating the performance of several methods that operate at different points in the ML pipeline across four real-world public policy and social good problems. Across these problems, we find a wide degree of variability and inconsistency in the ability of many of these methods to improve model fairness, but post-processing by choosing group-specific score thresholds consistently removes disparities, with important implications for both the ML research community and practitioners deploying machine learning to inform consequential policy decisions.

1. INTRODUCTION

There has been a recent increase in the use of machine learning models to support decisions in high stakes domains with societal impact, including informing bail decisions [20, 70, 6], hiring [61], healthcare delivery [57, 62] and social service interventions [9, 22, 60]. These decisions affect critical aspects of people's lives and if not done responsibly, can hurt already vulnerable and historically-disadvantaged communities. This combination of increased use, increased potential for improving social outcomes, and increased risk of harm has prompted questions from researchers, policymakers, citizens, and the media about the role these models can play in exacerbating (or reducing) existing inequities [36, 58, 54, 17], giving rise to a growing area, FairML, focused on dealing with issues of bias and fairness in building and using machine learning systems. FairML research has grown to span issues around defining bias in machine learning models, enumerating a variety of metrics that can be used to measure model bias [71], detecting instances of it through audit tools [67, 12], and methods for reducing (or mitigating the impact of) bias in ML models [17]. In this work, we focus on bias reduction methods, which can be broadly categorized into three groups based on the stage of analysis at which they are applied:

1. *Pre-processing methods* typically involve changing the data in some manner *before* building models.
2. *In-processing methods* typically involve using ML models/methods that are explicitly designed to deal with bias *in the process of building models*, such as using regularization approaches.
3. *Post-processing methods* typically involve adjusting scores, thresholds, or model selection *after the model predictions* have been generated.

Despite active research and the development of several new methods in the area of FairML in recent years, there has been a lack of extensive empirical evaluation across them to assess their effectiveness on real-world problems and data. The majority of this research has typically focused on achieving abstract, general-purpose definitions of fairness, and evaluated on benchmark data sets (such as the adult data set [23]) or limited data sets (such as COMPAS [50]). While that is a reasonable starting point, benchmark data sets often do not reflect the richness, nuances, and constraints of real-world problems, making it unclear for both researchers and practitioners how to assess the applicability and effectiveness of these methods or make decisions around which ones to use under given circumstances in real-world situations.

In this paper, we attempt to fill this empirical gap by presenting a **comprehensive empirical evaluation** of different bias reduction strategies over **four real-world problems** that come from public policy and social good settings. We want to emphasize that real-world problems are not just data sets but rather a combination of business/policy problems, the corresponding machine learning formulation and evaluation metrics, an extensive set of features generated in the feature engineering process, a large and varied set of ML models and hyperparameters, and a validation methodology and metric(s) that mirrors the deployment scenario. To that end, we describe the analytical formulation for each of these problems, the feature engineering process, parameters of temporal model selection using a wide variety of ML models and hyperparameters, and then evaluate the effectiveness of a variety of bias reduction strategies in reducing specific disparities while preserving as much of the original evaluation metric of interest as possible. We believe that this paper not only fills a critical gap today for researchers

*These authors contributed equally.

and practitioners of FairML but also provides a framework for researchers proposing new methods to follow when reporting the effectiveness of their work.

2. RELATED WORK

The focus of this paper is not on the entire process of building ML systems that lead to fair and equitable outcomes but more narrowly on methods that are used to reduce the bias in the predictions of ML models. With that focus, as mentioned earlier, bias reduction methods can be categorized broadly into three categories, based on the phase of the analysis pipeline to which they are applied: (a) Pre-processing, (b) In-processing and, (c) Post-processing.

2.1 Pre-processing

Pre-processing approaches assume that the bias in the ML models is caused by certain variables in the data or by the distribution of the data being used to train and validate the ML models. Most of the pre-processing approaches thus try to modify the data by either removing the sensitive variable (gender or race for example) or by changing the data distribution (with respect to the sensitive variable) by sampling.

Omission of sensitive variables has been widely explored in the past [29, 69]. This approach is based on the assumption that if machine learning model is not given the protected variable as a feature, the model that is trained will not be dependent on the protected variable, making the model unbiased. Unfortunately, this assumption is often overly-optimistic (and violated) in real-world problems where several other features, including ones relevant to the prediction problem, may be strongly correlated with the protected attribute. Recent work has described how omission of sensitive variables for training models often may not affect bias reduction (or even increase biases) despite decreasing model accuracy [16, 42, 24]. Despite these well-documented limitations, we included this strategy in the present exploration of fairness-enhancing methods because this notion of "fairness through unawareness" nevertheless persists and has commonly been posited by policymakers, decision-makers (in governments, non-profits, and corporations), and students we have worked with. Notably, other researchers have proposed more nuanced approaches to modifying the input data to remove correlations with the protected attribute in addition to the attribute itself. Although we do not explore this direction for pre-processing here, we refer the reader to [26] for an example of this approach in the context of disparate impact as a measure of fairness.

Resampling involves modifying the distribution of the training data by either over- or under-sampling examples to reduce disparities when the modified data is used in model training. Calders et al. [15] explored three different sampling techniques to fix existing bias in the data distribution to ensure that a model (in their case, Naive Bayes) trained on the modified data is more fair. Similarly, Iosifidis et al. [39] used clustering across sensitive attribute and labels to come up with representative training data to train models, and Kamiran et al. [42] explored multiple techniques involving sampling and re-weighing of training instances as pre-processing steps before applying machine learning models. Other popular preprocessing techniques involve relabelling and perturbation [44], details of which we omit from the paper.

2.2 In-processing

In-processing bias reduction methods generally include regularization or constrained optimization approaches to account for fairness metrics while solving their underlying classifier's optimization problem. Regularization adds penalty terms to the objective function of the classifier such that it is penalized for unfair solutions, whereas constrained optimization generally introduces fairness as a hard constraint in order to directly reject solutions that fail to satisfy fairness criteria. Kamishima et al. [43] proposed a regularization technique that uses mutual information of the sensitive attribute and prediction class, penalizing any increase in conditional probability on a specific subgroup. Zafar et al. [75] extended on this work by introducing fairness constraints into the objective function of the underlying classifier. One challenge faced by these approaches, however, is that these constraints often yield a non-convex objective function, making the optimization problem inherently difficult. To address this issue, Zafar proposed an efficient method for solving the resulting non-convex formulation. Similar techniques for different fairness metrics and even general classes of metrics have also been proposed in the literature [18]. Jiang et al. proposed an approach that minimizes Wasserstein-1 distances between classifier output and sensitive information [41]. Heidari et al. proposed a Rawlsian concept of fariness that can be introduced as a constraint into any convex loss-minimization algorithm [33]. Similar methods have also been extended to neural-network based models [52] as well as decision trees [3].

2.3 Post-processing

Post-processing methods are generally agnostic to the machine learning models used, and modify the outputs to improve fairness in predictions or classifications. This involves training meta-models with fairness constraints [18, 25] or directly thresholding or modifying model scores to improve fairness [32, 65]. Hardt et al. proposed methods for **direct post-hoc adjustments to scores** (or binary predicted classes) from trained classifiers to achieve either equalized odds or equality of opportunity by choosing group-specific thresholds that meet these fairness goals [32]. Recently, we have extended on this work, applying similar methods across a number of policy contexts and finding little or no trade-off in model accuracy in doing so [65, 64].

Another fairness-enhancing strategy that can be applied on top of a range of underlying machine learning methods involves **decoupling the training or selection of classifiers**, as proposed by Dwork and colleagues [25]. This approach starts from the hypothesis that a model trained to do well on the entire population might not fully capture differences in predictiveness of features or other important patterns across groups and posits that training separate models for each protected group might better pick up on these nuances. Because fully decoupling the models might significantly reduce the available training data (particularly for small groups), they also suggest exploring different levels of transfer learning between groups, giving a relative weight to training examples from the protected group or rest of the population (so, at the other extreme, one might train models across the full population, but select best-performing models for each group rather than a single overall model).

Other authors, including Celis et al [18] as well as Menon and Williamson [55], have proposed methods that perform a **constrained optimization to train a meta-model** to

improve the fairness of a prediction score generated by a model. These methods seem particularly useful where membership in the protected groups is not known apriori but can be estimated (for instance, [18] describes estimating a joint probability distribution over outcomes and sensitive attributes). However, when group membership is known, these methods will generally result in stretching or shifting within-group score distributions without reordering in a manner equivalent to choosing separate thresholds for each group (for more detail, see our discussion in the supplemental materials from [64]).

Finally, and perhaps most simply, fairness can be incorporated into the process of **model selection**. After training a large set of different model types and hyperparameter values, the validation set performance of these different trained models can be assessed both in terms of traditional accuracy metrics (such as AUC-ROC, precision@k, or other confusion matrix based metrics) as well as fairness metrics appropriate to the context. Choosing a model to deploy then becomes an optimization problem over two dimensions, with a Pareto frontier reflecting a menu of potential trade-offs between these two goals of accuracy and fairness [72]. In practice, the trade-offs presented by this frontier might be a function of inherent properties of the data and problem as well as the extent to which the grid search that was performed covers the possible space of model types and hyperparameters. Although relatively straightforward in nature and implementation, relying entirely on model selection is somewhat arbitrary as it relies entirely on finding a model specification that performs well on both fairness and accuracy metrics without taking active steps to ensure or improve fairness.

3. COMPARISON SETUP

This section describes our setup to conduct the empirical evaluation across bias reduction methods. We describe the specific methods we chose to compare, the policy contexts for the problems we use to conduct that empirical evaluation, and the specific experimental setup for each real-world problem (the data used, features generated, models built, evaluation metric, protected group, and bias metric).

3.1 Methods to Compare

While a large number of bias reduction methods exist in each category we describe in Section 2 (Pre-processing, In-processing, and Post-processing), in this paper, we focus on a few representative methods from each category to compare with each other. The methods chosen for this study are described below.

3.1.1 Pre-Processing Methods

Removing the Protected Attribute: For each problem domain, we define a set of protected attributes and remove those from the data before performing any ML modeling.

Sampling: We apply sampling to our training sets with respect to the protected group in three ways: a) changing the marginal distribution of the protected and non-protected subgroups, b) changing the label distribution within the protected and non-protected subgroups, and c) changing both simultaneously. Here, we implemented the six sampling strategies described in Table 1 reflecting a set of reasonable a priori hypothesis about how these distributions in the training data might influence model fairness.

To formalize our sampling approaches, we define $Protected$ as the protected value/group (such as Race=Black) and $NonProtected$ as the set of values that are considered Non-Protected (such as Race=White). The (binary) label variable is represented as Y with values 0 and 1. $P^0(\cdot)$ represents a probability distribution in the original dataset and $P'(\cdot)$ represents a probability distribution after resampling. With those definitions, each of our three sampling settings are:

(A) Balances the data by changing the ratio of Protected to Non-Protected while preserving the original label distribution within each group.

The goal is to achieve:

$$\frac{P'(NonProtected)}{P'(Protected)} = \alpha$$

while preserving the original label distribution within $Protected$ and $NonProtected$ such that

$$P'(Y = 1 \mid NonProtected) = P^0(Y = 1 \mid NonProtected)$$
$$\text{and } P'(Y = 1 \mid Protected) = P^0(Y = 1 \mid Protected)$$

In Table 1, Strategy 1 uses this approach with $\alpha = 1$.

(B) Balances the label distribution across each subgroup: Protected and NonProtected. The goal is to achieve:

$$P'(Y = 1 \mid NonProtected) = \beta_{NP}$$
$$P'(Y = 1 \mid Protected) = \beta_P$$
$$\text{such that } \frac{\beta_{NP}}{\beta_P} = \gamma$$

while preserving the original marginal distributions for Protected and NonProtected such that:

$$P'(NonProtected) = P^0(NonProtected)$$
$$\text{and } P'(Protected) = P^0(Protected)$$

In Table 1, Strategy 2 uses this approach (with $\beta_P = \beta_{NP} = 0.5$ and $\gamma = 1$), as does Strategy 3 (with $\beta_P = \beta_{NP} = P^0(Y = 1 \mid NonProtected)$ and $\gamma = 1$) and Strategy 4 (with $\beta_{NP} = P^0(Y = 1 \mid NonProtected)$ and $\beta_P = 0.5$).

(C) Adjusts the marginal distribution of Protected and Non-Protected as well as the label distributions by setting α, β_P, β_{NP}, and γ as described above.

In Table 1, Strategy 5 uses this approach (with $\alpha = 1$ and $\beta_P = \beta_{NP} = 0.5$) as does Strategy 6 (with $\alpha = 1$, and $\beta_P = \beta_{NP} = P^0(Y = 1 \mid NonProtected)$).

Note that in each strategy, in order to balance two distributions, we can either *undersample* from the majority distribution or *oversample* from the minority distribution. In case of oversampling, we randomly sample (with duplicates allowed) to generate more examples[1] increasing the total number of examples as little as possible while achieving the desired distributions. When undersampling, we remove as few examples as possible in order to achieve the desired distributions. Also note that we only sample in each training set while keeping the distribution of the validation sets the same as in the original data.

[1]For oversampling, we do not make use of methods such as SMOTE [19] as each feature might have a specific set of constraints and this method does not take into account the overall joint distribution.

Table 1: Sampling strategies used in this study.

	Ratio: Protected to Non-Protected	Label Dist. Protected	Label Dist. Non-Protected
1	1:1	Original	Original
2	Original	50-50	50-50
3	Original	Same as Non-Protected	Original
4	Original	50-50	Original
5	1:1	50-50	50-50
6	1:1	Same as Non-Protected	Original

3.1.2 *In-Processing Methods*

In this paper, we focus on in-processing through constrained optimization to reduce model disparities. This approach includes fairness metrics in the objective function and seeks to produce predictions that maximize accuracy while taking fairness into account.

Zafar and colleagues [75, 74] proposed a constrained optimization method centered on a fairness notion they described as "disparate mistreatment." A model can be said to have disparate mistreatment when misclassification rate for the protected and non-protected group are different, and their work described optimization problems using either False Positive Rate (FPR) or False Negative Rate (FNR) as a measurement of misclassification. Formally, this optimization problem (for FNR) is defined by:

$$\min L(\theta)$$
$$\text{s.t. } P(\hat{y} \neq y \mid z = 0, y = 1) - P(\hat{y} \neq y \mid z = 1, y = 1) \leq \epsilon$$
$$P(\hat{y} \neq y \mid z = 0, y = 1) - P(\hat{y} \neq y \mid z = 1, y = 1) \geq -\epsilon$$

where, L is the loss function (over model parameters θ), $\hat{y}$ prediction, y original label, z is the protected attribute, and ϵ denotes the tolerance boundaries for a fair output.

For our problem settings, we focus on True Positive Rate (TPR) disparities (also referred to "equality of opportunity" by Hardt [32]) as the appropriate metric of fairness (see the discussion on problem settings below, as well as in [66, 64]). However, because $TPR = 1 - FNR$, we make use of Zafar's method to equalize FNR. In doing so, we used a very small value of $\epsilon = 0.0001$ to find solutions which remove disparities entirely.

Recently, open source toolkits such as FairLearn [13] have also been introduced which try to reduce biases, according to a given metric in classification problems. However, we do not include FairLearn in this study setting because it only generates binary predicted class labels rather than a continuous score. This makes it poorly suited to our problem settings where we focus on choosing the k highest-risk entities for intervention based on an organization's resource constraints (as discussed in more detail below). In other work, we have explored heuristics such as sampling to select top k predictions from the output of FairLearn but found that it performed poorly since it wasn't designed for that purpose [68].

3.1.3 *Post-Processing Methods*

We define the post-processing class of methods as any method that is applied once the model has been built, typically in adjusting the scores that the models produced or using different thresholds to create classification decisions. We describe several such methods above and discuss here the methods we explored in the present work.

Post-Hoc Adjustments: Here we expand on some of our recent work [65, 64] using a method to equalize TPRs across groups while keeping the total number of individuals selected constant, reflecting the "top k" setting of the policy problems we consider (see the discussion on problem settings below for more details). In short, because TPR increases monotonically with depth in a predicted score, we can find a single solution (up to randomized tie breaking) with equalized TPR across groups by adjusting the score thresholds for each group while keeping the total number of individuals selected constant. In practice, these threshold adjustments are made on the model scores in one validation split (say, at time $t = 0$) to decompose the overall number of individuals to select by group,[2] then these group-specific target numbers are applied to a subsequent validation set to evaluate how well this fairness-enhancing strategy generalizes into the future. Note that, as mentioned above, some of the meta-model approaches such as those described in [18, 55] can be shown to be mathematically equivalent to choosing different score thresholds when protected group membership is known (rather than modeled) and a unique equitable solution exists, as is the case here. As such, we don't explore those methods separately from these post-hoc adjustments through group-specific thresholding.

Composite Models: Following the proposal of Dwork and colleagues [25], we investigated two options for building composite models from models trained or selected for their performance on subgroups. On the one extreme, we simply used the grid of models trained on the full population but performed model selection separately for each subgroup (reflecting the complete transfer learning approach described by Dwork). On the other extreme, we trained separate models just with examples from each subgroup (the fully decoupled approach in Dwork) and added these to the model grid for subgroup-specific model selection. One challenge with implementing these composite models, however, is that the scores from the separate models chosen for different subgroups have not been calibrated and cannot be assumed to be comparable. As such, one needs to determine how to appropriately choose a total "top k" set of individuals across these different models. Because we were making use of these composite models in the interest of improving fairness, a natural means of choosing these thresholds was to apply the same method choosing TPR-equalizing thresholds described above. It is somewhat challenging to determine whether fairness improvements seen from these composite strategies are more a result of the group-specific thresholds or decoupling the model building or selection itself. However, one hope here would be that the decoupling should improve the accuracy of model predictions on the subgroups, so success for these methods ideally would show not just similar disparity mitigation to post-hoc adjustments but also improved overall accuracy metrics at the same level of fairness.

[2]For instance, if a program can intervene on 100 individuals, this process might break that down into 75 Black individuals and 25 white individuals. Because score distributions are likely to change over time, group-specific "top k" values are used rather than score thresholds to ensure the total number of targeted individuals remains fixed.

Model Selection: As noted above, an additional simple approach that falls under our umbrella of post-processing strategies is to account for fairness metrics in the process of model selection. However, this approach is not only very sensitive to the machine learning method/hyperparameter grid explored but also relies on some degree of luck that specifications with favorable trade-offs will be found. Here, we explored two options by which fairness could be included in the model selection process:

- Setting a "Disparity Constraint" reflecting a largest acceptable disparity. Here, we only consider models with disparity no higher than a certain value, then choose the model with the highest precision among these. Note that it may be possible that no models have a low enough disparity to meet the criteria, in which case we choose the model closest to this cut-off (making it a soft constraint and guaranteeing a model will always be chosen).

- Setting an "Accuracy Constraint" reflecting a largest acceptable loss in accuracy to improve fairness. Here, we only consider models with precision@k within a given number of percentage points below the best model, then choose the model with lowest disparity among these. Note that because this constraint is relative to the performance of the most-accurate model, there will always be at least one meeting the criteria, so this is a hard constraint.

For each type, we explored eight levels of the constraint, from placing little or no weight on fairness to strongly selecting for fair models. For Disparity Constraints, these included allowing disparities up to 5.0, 2.0, 1.5, 1.3, 1.2, 1.1, 1.05, or 1.0 (that is, exact equity). For the Accuracy Constraints, these included allowing a decrease in precision of up to 0.0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.50, and 0.60 percentage points.

3.2 Problems, Data, and Experimental Setup

Our empirical evaluation of these methods was done on three real world problems that we have worked on in collaboration with various government agencies. These span mental health and criminal justice (with Johnson County, Kansas), housing safety inspections (with San Jose, CA), and education outcomes (with the Education Ministry of El Savador). Since the data for these problems is confidential and not available publicly, we also replicate this empirical evaluation on a crowdfunding problem from DonorsChoose[3] where the data is publicly available. This will allow other researchers and practitioners to replicate our work before applying it to their own problems. In general, these problem settings involve six elements:

1. **Features**: Each project we use in this study went through an extensive feature engineering process. As is typically done in real-world ML systems, the features generated included raw and transformed information about the entities of interest (such as demographics) as well as temporal and spatial aggregations (while respecting temporal boundaries in train and validation sets to avoid leakage).

[3]http://wwww.donorschoose.org

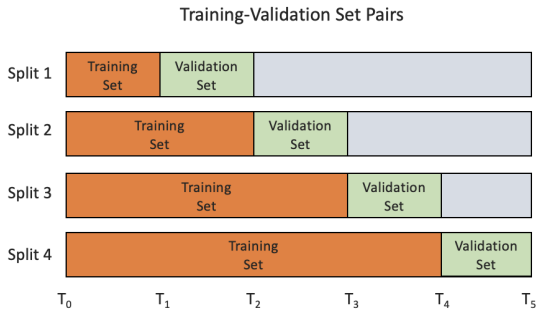


Figure 1: The temporal validation approach used in these settings to capture the non-stationary nature of the data and guard against leakage. Time is used to split the available data into a series of training and validation sets, testing for generalization performance on "future" data relative to model training.

2. **Label**: In each of the problem domains, the decision on the definition of the label is part of the formulation process and is done in collaboration with the partnering organization. In all of these problems, the label was determined by the occurrence of an event at some point in the future from the time of prediction, for example, an individual being booked into jail in the next 12 months or a crowdfunding project failing to get fully funded in the next 4 months.

3. **Train and Validation Splits**: Since most real-world prediction problems are temporal in nature and violate stationary distribution assumptions, we use temporal validation to split our datasets into train and validation sets [38]. These train and validation sets are usually temporally sequential in nature, where each candidate model is trained on data from "past" data and validated on "future" data (see Figure 1 for a diagram).

4. **Models**: We train a wide variety of model and hyperparameter combinations, including logistic regression, tree-based models, and ensembles such as random forests and boosted trees. The reasoning behind a wide grid was both to understand the effectiveness of different models along both the "accuracy" and bias dimensions as well as to provide the model selection process with as much diversity as possible. The model types and hyperparameters used for each problem are listed in Table 2.

5. **Choice of Bias Metric**: In all of these problems, a key decision to make is the choice of the appropriate bias metric(s). We use the Fairness Tree (Figure 2), a framework developed and used in [65] to inform that choice. Since in all the problems we describe below, we are supporting assistive interventions (i.e. reducing disparities in false negatives is more important than those in false positives), and have limited resources to intervene compared to the number of people that need support, the Fairness Tree framework leads us to choose Recall (True Positive Rate) Disparity as the primary bias metric.

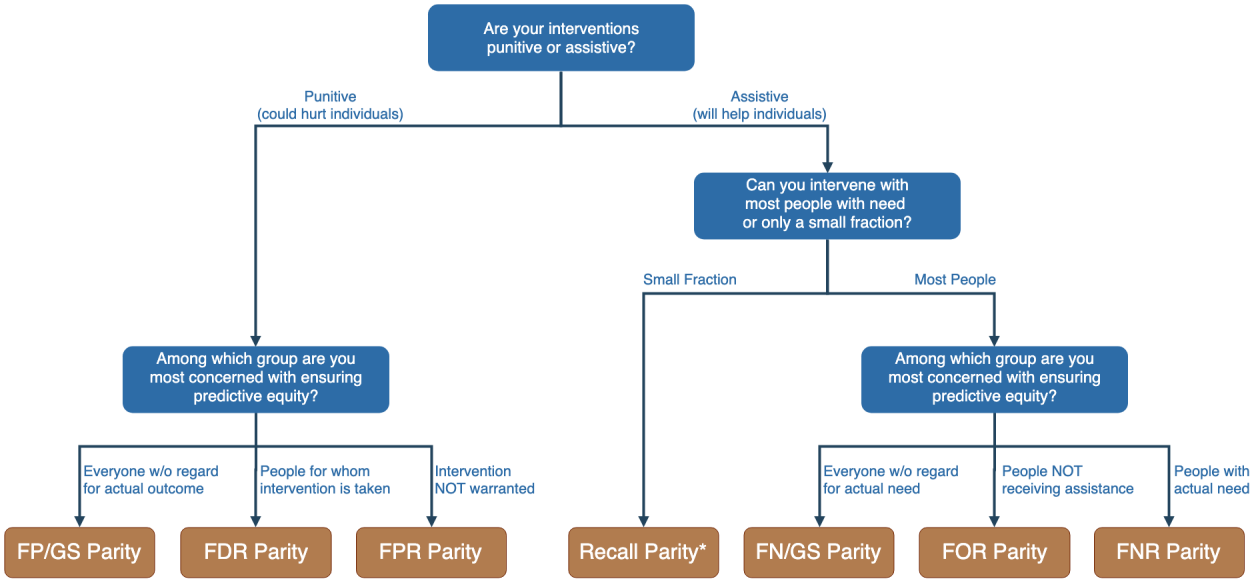


Figure 2: Fairness Tree framework to help identify appropriate fairness metrics based on the intended use. The metrics in the leaf nodes are: False Negative Rate (FNR), False Omission Rate (FOR), False Negatives Adjusted to Group Size (FN/GS), Recall/True Positive Rate (TPR), False Positive Rate (FPR), False Discovery Rate (FDR), and False Positives Adjusted to Group Size (FP/GS).

6. **Evaluation Methodology**: For each temporal validation set we calculate the evaluation metric as well as the bias metric (with respect to the protected group) for all models. These results are aggregated by calculating the mean and standard errors.

Each of the four policy problems used for the present empirical evaluation are described in detail below, including details about the underlying data set, the performance and fairness metrics of interest, and the protected group for bias and fairness analysis.

3.2.1 *Mental Health Outreach - Johnson County KS*

Untreated mental health conditions often result in a negative spiral, which can culminate in repeated periods of incarceration with long term consequences both for the affected individual and the community as a whole [30]. Surveys of inmate populations have suggested a high prevalence of multiple and complex needs, with 64% of people in local jails suffering from mental health issues and 55% meeting criteria for substance abuse or dependence [40]. The criminal justice system is poorly suited to address these needs, yet houses three times as many individuals with serious mental illness as hospitals [27].

Since 2016, Johnson County, KS, has partnered with our group to help them break this cycle of incarceration by identifying individuals who might benefit from outreach with mental health resources and are at risk for future incarceration. While the Johnson County Mental Health Center (JCMHC) currently provides services to the jail population, needs are generally identified reactively, for instance through screening instruments individuals fill out when entering jail. The new program being developed will supplement these existing approaches by adding a new automatic referral system for people who are at risk of being booked into jail, with the hope that they can be outreached to reduce their risk of returning to jail.

Through our partnership, we obtained administrative data from their mental health center, jail system, police arrests, and ambulance runs. ML modeling was focused on Johnson County residents with any history of mental health need who had been released from jail within the past three years. Early results from this work were described in [9]. A field evaluation of the predictive model is ongoing at the time of this writing, but validation on historical data demonstrated a 12% improvement over a baseline based on the number of bookings in the prior year and 4.8-fold increase over the population prevalence.

3.2.2 *Housing Safety Inspections - San Jose, CA*

The Multiple Housing team in San Jose's Code Enforcement Office is tasked with protecting the occupants of properties with three or more units, such as apartment buildings, fraternities, sororities, and hotels. They do so by conducting routine inspections of these properties, looking for everything from blight and pest infestations to faulty construction and fire hazards (see [35] and [45] for a discussion of the importance of housing inspections to public health). Although the city of San Jose inspects all of the properties on its Multiple Housing roster over time, and expects to find minor violations at many of them, it is important that they can identify and mitigate dangerous situations early to prevent accidents. With more than 4,500 multiple housing properties in San Jose, CA – many of which comprise multiple buildings and hundreds of units – it is not possible for the city to inspect every unit every year. San Jose recently instituted a tiered approach to prioritizing inspections, inspecting riskier properties more frequently and thoroughly. Although the tier system helped focus inspections on riskier

Table 2: Data and Experimental Setup for our four problems.

	Mental Health and Criminal Justice	**Housing Safety Inspections**	**Student Outcomes**	**Education Crowdfunding**
Prediction Task	Jail booking within the next 12 months	Housing unit having a violation within the next year	Student not returning to school next year	Project not getting fully funded within 4 months
Timespan	2013-01-01 to 2019-04-01	2011-01-01 to 2017-06-01	2009-01-01 to 2018-01-01	2010-01-01 to 2014-01-01
# of entities	61,192	4,593	801,242	210,310
Feature Groups	Demographics Mental Health History Past Diagnosis Mental Health Programs Police Interactions Past Jail Incarceration Jail Booking Details	Building Permits Past Citations Past Violations House Prices Census Data	Age Relative to Grade Repeated Grades Rural/Urban Academic History Dropout History Gender Illness Family Information	Funding Request Details Donation Details Past Funding Rates Project Description
# of Features	3,465	1,657	220	319
Base Rate	0.12	0.43	0.25	0.24
Evaluation Metric	Precision@500	Precision@500	Precision@10000	Precision@1000

Model Types and Hyperparameters (as specified by scikitlearn parameters used in the experiments)	`Decision Tree`
	Max Depth:
	(1,2,3)
	Min Samples Split:
	(10, 50, 100)
	
	`Random Forest`
	Num Estimators:
	(100, 1000, 5000)
	Min Samples Split:
	(10, 25, 100)
	Max Depth:
	(5, 10, 50)
	
	`Logistic Regression`
	Penalty:
	(l1, l2)
	C:
	(0.001, 0.01, 0.1, 1, 10)
		`Decision Tree`
		Criteria:
		(gini, entropy)
		Max Depth:
		(1,2,3,5,10,20,50)
		Min Samples Split:
		(10, 20, 50, 1000)
		
		`Random Forest`
		Max Features:
		(sqrt, log2)
		Criteria:
		(gini, entropy)
		Num Estimators:
		(100, 1000, 5000)
		Min Samples Split:
		(10, 20, 50, 100)
		Max Depth:
		(2, 5, 10, 20, 50, 100)
		
		`Extra Trees`
		Max Features:
		(sqrt, log2)
		Criterion:
		(gini, entropy)
		Num Estimators:
		(100, 1000, 5000)
		Min Samples Split:
		(10, 20, 50, 100)
		Max Depth:
		(2, 5, 10, 50, 100)
		
		`Logistic Regression`
		Penalty:
		(l1, l2)
		C:
		(0.001, 0.01, 0.1, 1, 10)
			`Decision Tree`
			Max Depth:
			(1, 5, 10, 20, 50, 100)
			Min Samples Split:
			(2, 5, 10, 100, 1000)
			
			`Extra Trees`
			Num Estimators: (100)
			Max Depth: (5, 50)
			
			`Logistic Regression`
			Penalty:
			(l1, l2)
			C:
			(0.0001, 0.001, 0.1, 1, 10)
			
			`Random Forest`
			Num Estimators: (100, 500)
			Min Samples Split: (2, 10)
			Class Weight:
			(Balanced Subsample, Balanced)
			Max Depth: (5, 50)
				`Random Forest`
				Num Estimators:
				(100, 500, 1000)
				Min Samples Split:
				(10, 50)
				Max Depth:
				(10, 50, 100)
				
				`AdaBoost`
				Num Estimators: (500, 1000)
				`Decision Tree`
				Max Depth:
				(1, 5, 10, 20, 50, 100)
				Min Samples Split:
				(2, 5, 10, 100, 1000)
				
				`Logisitic Regression`
				C:
				(0.0001, 0.001, 0.01, 0.1, 1, 10)
				penalty:
				(l1, l2)

Train and Validation Sets	Temporal Block: 4 months	Temporal Block: 2 months	Temporal Block: 1 year	Temporal Block: 3 months
Protected Group	Race	Median Income	Age Relative to Grade	Poverty Level

properties, the new system has its limitations. The city evaluates tier assignments for properties infrequently (every 3 to 6 years), and these adjustments require a great deal of expertise and manual work while leaving out a rich amount of information.

In order to provide a more nuanced view of properties' violation risk over time and allow for more efficient scheduling of inspections, the Code Enforcement Office partnered with us to develop a model to predict the risk that a serious violation would be found if a given property was prioritized for inspection (similar tools have been developed for allocating fire inspections in New York [7] and health inspections in Boston [28]). Evaluation of the model on historical data indicated that it could provide a 30% increase in precision relative to the current tier system and the model's predictive accuracy was confirmed during a 4-month field trial in 2017.

3.2.3 *Improving Educational Outcomes - El Salvador*

Each year from 2010 through 2016, 15-29% of students enrolled in school in El Salvador did not return to school in the following year. This high dropout rate is cause for serious concern, with significant consequences for economic productivity, workforce skill, inclusiveness of growth, social cohesion, and increasing youth risks [11, 8]. El Salvador's Ministry of Education has programs available to support students with the goal of reducing these high dropout rates, but the budget for these programs is not large enough to reach every student and school in El Salvador.

Predictive modeling has been deployed to help schools identify students at risk of dropping out in several contexts [49, 4, 14] and El Salvador partnered with us in 2018 to make use of these methods to focus their limited resources on the students at highest risk of not returning each year. Student-level data was provided by the Ministry of Education, including demographics, urbanicity, school-level resources (e.g., classrooms, computers, etc), gang and drug violence, family characteristics, attendance records, and grade repetition. For the present study, we focused on the state of San Salvador and identifying the 10,000 highest-risk students, considering annual cohorts of approximately 300,000 students and drawing on 5 years' of prior examples as training data.

3.2.4 *Education Crowdfunding - DonorsChoose*

Since the projects above used confidential and sensitive data and were done under data use agreements, we are not able to make that data publicly available. For our work to be easily reproducible, we include a fourth problem in this study where the data is available publicly, focused around crowdfunding for education by the organization DonorsChoose.

Many schools in the United States, particularly in poorer communities, face funding shortages [56]. Often, teachers themselves are left to fill this gap, purchasing supplies for their classrooms when they have the individual resources to do so [37]. The non-profit DonorsChoose was founded in 2000 to help alleviate these shortages by providing a platform where teachers post project requests focused on their classroom needs and community members can make contributions to support these projects. Since 2000, they have facilitated $970 million in donations to 40 million students in the United States [2]. However, approximately one third of all projects posted fail to reach their funding goal.

Here, we make use of a dataset DonorsChoose made publicly available for the 2014 KDD Cup (an annual data science competition) including information about projects, the schools posting them, and donations they received. Because the other case studies explored here focused on proprietary and often sensitive data shared with us under data use agreements that cannot be made publicly available, we included a case study surrounding this publicly-available dataset. While we have not partnered with DonorsChoose to deploy the machine learning system described, we otherwise treated this case study as we would any of our applied projects. Here, we consider a resource-constrained effort to assist projects at risk of going unfunded (for instance, providing a review and consultation) capable of helping 1,000 projects in a 2-month window, focusing on the most recent 2 years' of data available in the extract (earlier data had far fewer projects and instability in the baseline funding rates as the platform ramped up). This dataset is publicly available at kaggle.com [1].

4. RESULTS

Overall results across the different methods and problems we evaluated are shown in Figure 3. Each graph shows the relative performance of models with a given strategy in terms of the "performance metric" (namely precision@k on the x-axis) and fairness with respect to the protected group (namely True Positive Rate or Recall disparities on the y-axis), with error bars representing the 95% confidence interval over all temporal validation splits. The ideal model would have a value of 1.0 for both of these metrics – models appearing further to the right on the x-axis are more accurate while those appearing closer to the dashed y=1.0 line are more equitable (departures from this line in either direction reflect disparities favoring one or the other group). The blue circle in all of the graphs refers to the *Original* model — a term we use to specify the model built and selected only focused on maximizing the accuracy metric. All the other points are results from the bias reduction methods that we investigated. Note that in this figure we only include the best-performing sampling strategies (either in terms of fairness or accuracy) but discuss and show the wider range of sampling results in the graphs below. Likewise, we only show the composite models without decoupled training because the results from the two strategies were generally similar, but discuss and show these results in more detail below as well. Additionally, model selection approaches are omitted from Figure 3 because they generally required considerable decreases in precision@k to improve fairness, allowing us to focus the overall analysis on the nuance between the other methods (see Figure 7 and the related discussion for these results).

4.1 Overall Results

Across the four problems, a few general patterns seem to emerge from our experiments:

1. **Considerable disparities, ranging from 30-100%, were observed in the *baseline* models** for all four problems. That is, building models which optimize only for some measure of accuracy consistently resulted in appreciable biases if fairness was not actively pursued as an outcome. This is not a surprising outcome and a result that has been demonstrated in various

studies but an important point to keep in mind when building ML models.

2. There was **considerable variability across strategies and settings** in the effectiveness and ability of the fairness-enhancing methods considered here to remove these disparities, with **most methods showing only moderate success** or success only in a few settings. Comparisons across these methods is discussed in more detail below.

3. Only the two approaches which made use of **separate thresholds across groups (composite models and post-hoc adjustments) were consistently successful** in removing disparities and did so without any appreciable loss in model accuracy.

Below we discuss these results in more detail, examining the performance of each fairness-enhancing approach in turn.

4.2 Effect of Removing Sensitive Attribute

A common misconception in the context of algorithmic fairness is that simply omitting a sensitive attribute can help a model achieve fair predictions through "unawareness." Several authors [46, 59] have spoken to the fallacy of this concept, both as a result of correlations between protected attributes and other potentially relevant ones and because access to the sensitive attribute might help models pick up on patterns that improve accuracy for the protected group and result in lower disparities. However, we included this strategy here both for completeness as well as to understand and demonstrate how this approach might perform in practice. Unsurprisingly, then, the results in Figure 4 show **this strategy is inconsistent in the magnitude and direction of its impact across the four problems**. Although omitting the protected attribute did improve model fairness somewhat in the Education Crowdfunding and Student Outcomes contexts, in neither case did it fully remove the disparities from the initial model, and in the latter case these improvements came at the cost of a moderate decrease in precision@k. Moreover, in the Inmate Mental Health setting, removing the race attribute actually made the models somewhat less fair on average while in the Housing Safety context doing so had no effect on either fairness or accuracy. Taken together, these results are very consistent with the notion that "fairness through unawareness" by **removing the sensitive feature cannot be relied upon to improve the fairness of machine learning models**.

4.3 Effect of Sampling

The other pre-processing method we explored involved sampling of the training data. As discussed above, a number of parameters must be determined in choosing a sampling strategy: the relative distributions of the protected and non-protected subgroups, the label distributions within each group, and whether to over- or under-sample training examples to achieve the target distribution. Here, we explored six strategies (Table 1) that reflect combinations of three reasonable hypothesis:

- A 1:1 ratio between protected group and non-protected group training examples might tell the model to treat errors in each group as equally important, alleviating differential error rates.

- Equal label distributions within the two subgroups might tell the model to not treat protected group membership as important.

- A 50/50 label distribution in one or both subgroup might alleviate any issues arising from imbalance in the training set.

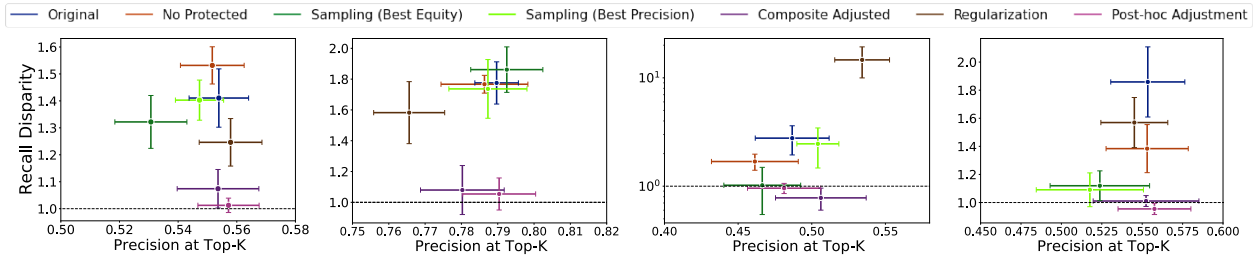
Figure 5 shows the results of applying these six strategies to the training data in each of the four policy settings. Although resampling of the training data had an impact on the models in many of the problem settings, there was a considerable inconsistency in the results both across settings and sampling strategies. In the Education Crowdfunding and Student Outcomes settings, many (but not all) of the strategies showed improvements in model fairness, while none of the strategies yielded fair results in the Housing Safety or Inmate Mental Health settings. Interestingly, this pattern reflects the results observed when removing the protected attribute described above, suggesting that both strategies may be accomplishing the same thing by effectively telling the model not to treat subgroup membership as important. Note, in particular, that in the Education Crowdfunding setting, sampling strategies 2, 3, 5, and 6 show improvements and each of these strategies equalizes the label distribution across the protected and non-protected subgroups.

Two more general patterns in Figure 5 do seem of note: First, over- and under-sampling approaches to the sample sampling strategy appear to yield similar results, suggesting that, at least in these four policy contexts, decreasing the total number of training examples through undersampling did not have an appreciable impact on model performance. Second, strategy 4 yielded particularly variable results, ranging from little impact to large disparities in either direction (note that in the Education Crowdfunding setting, both over- and under-sampling for strategy 4 resulted in no predicted positives in the protected class, yielding infinite disparities, so this strategy is omitted from the graph). However, this result might not be too surprising in light of the fact that this is the only strategy considered here where we adjusted the label distribution among the protected subgroup (to 50/50) without changing the distribution of non-protected subgroup. Depending on the baseline distribution, of course, this might (or might not) tell the model to see the protected attribute (or correlated features) as particularly important as a predictor of the outcome.

Taken together, these results suggest that **sampling of the training data can have an impact on disparities in the resulting models' predictions, but that these effects are both context and parameter dependent**. Without an obvious or consistent pattern for how a given sampling strategy will translate into changes in fairness metrics in a given modeling context, model developers are left to conduct a search over different values of these sampling parameters to explore this space in their setting. Even so, **there does not seem to be strong empirical evidence that a fairness-enhancing solution will be found in a particular context, or at what cost to model accuracy**.

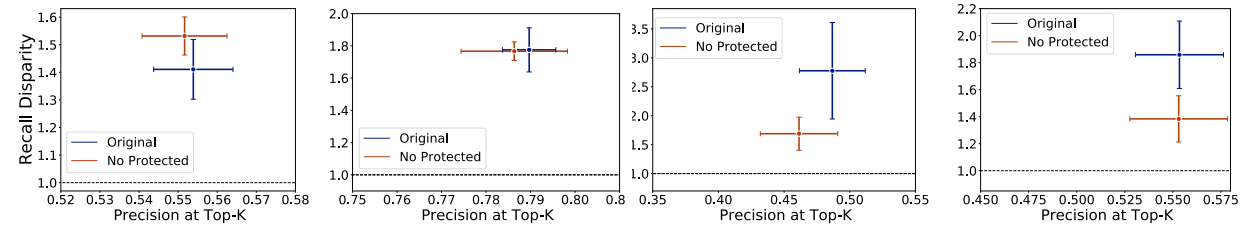
4.4 Effect of In-Processing Methods

The in-processing method considered here was proposed by Zafar and colleagues [75]. Models were trained with a con-



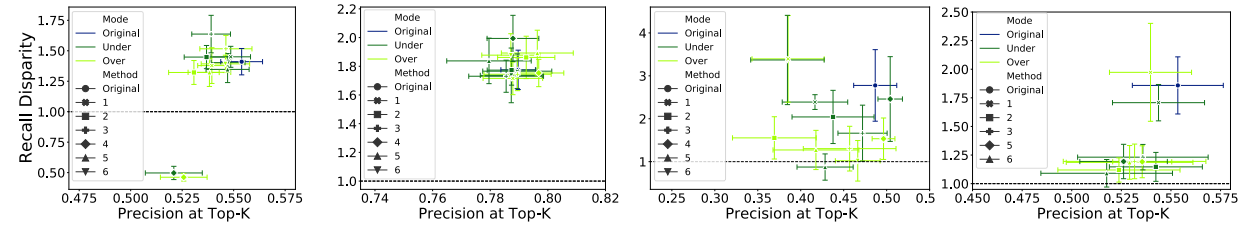
(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 3: Results from the different fairness-enhancing strategies considered here across the four policy settings, showing the relationship between model accuracy (as measured by precision@k) on the x-axis and fairness (as measured by recall disparities) on the y-axis. Ideal models would have high values of precision@k and be near a disparity value of 1.0. Note that the y-axis in (c) is on a log scale based on the large variation in performance across methods (performance for each method is shown separately on a linear scale in the figures below). Error bars show 95% confidence intervals over validation sets.



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 4: Effect of removing the protected attribute from machine learning modeling on model accuracy (precision@k) and fairness (recall disparities). Error bars show 95% confidence intervals over validation sets.



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 5: Results from resampling of training data for machine learning on model accuracy (precision@k) and fairness (recall disparities). Each of the six strategies from Table 1 was performed either via under-sampling (dark green) or over-sampling (light green). Error bars show 95% confidence intervals over validation sets.

straint to equalize the false negative rate[4] between the protected and non-protected group in each setting, with results shown in Figure 6. In general, in-processing failed to appreciably improve the fairness of the models in any of the four settings, reducing disparities only slightly in three settings and making them appreciably worse in the fourth (Student Outcomes).

Importantly, these results should not be seen as an inherent critique of either Zafar's method specifically or in-processing in general, but rather as a mismatch between the currently available methods using this approach and the common context of resource-constrained problem settings in which a given number of highest-risk entities must be selected for an intervention. In-processing methods generally add a fairness constraint to a classifier that optimizes for overall accuracy around an implicit threshold of 0.5 (or best-partitioning decision boundary). To select the "top k" for intervention, we naively threshold the resulting score (or, equivalently, shift the decision boundary) to yield only k highest-risk predicted positives. Of course, the fairness constraints used during model training applied to the original boundary, not the shifted one. As such, it is not surprising that Zafar's method here failed to improve fairness of these models when applied to a "top k" setting, even if it might perform well in settings without such a constraint. Perhaps for this reason, other methods such as Microsoft's Fair Learn [13] only provide predicted class labels without a continuous score, but unfortunately those methods are also poorly suited to the "top k" setting where a small subset of k individuals would need to be randomly chosen from the predicted positive class at considerable cost to accuracy/precision.[5]

4.5 Effect of Model Selection

Results of applying fairness-aware model selection in these contexts are shown in Figure 7. Here, several of the settings suggest an often considerable trade-off between fairness and accuracy, with constraints that put more weight on fairness in the model selection process yielding sizable decreases in precision@k (note that the range of the x-axes for these graphs is generally much wider than for the results of using other methods). For instance, in the Education Crowdfunding setting, disparities could be removed through the model selection process, but at the expense of loosing nearly half of the model's precision. In other cases, even large fairness constraints in the model selection process failed to remove disparities effectively: even when sacrificing a large amount of precision in the Inmate Mental Health context, the resulting models still showed considerable disparities of 1.27 on average. Likewise, in the Housing Safety context, fairness-aware model selection failed to reduce the disparities in these models regardless of constraint type or size. Notably, across all four contexts, similar results could be obtained by placing either a soft constraint on the largest acceptable disparity or a hard constraint on the largest acceptable decrease in precision@k to improve fairness (represented by different colors in Figure 7).

Although on the surface, these results suggest some semblance of the "Pareto Frontier" one might anticipate could reflect an inherent trade-off between fairness and accuracy, it is important to keep in mind that the nature of this frontier is highly dependent on the model grid over which this selection process is taking place (that is, other model type/hyperparameter combinations may perform better on one or both metrics). Likewise, other approaches at improving model fairness (such as the other methods explored here) may expand this frontier and allow for considerably less drastic trade-offs between fairness and accuracy.

4.6 Effect of Post-Hoc Adjustments

Figure 8 shows the results of post-hoc adjustments to equalize TPR across groups by choosing separate, group-specific thresholds. Across all four policy settings, this approach consistently improved the fairness of the models, entirely removing the disparities in most cases. Notably, this improved fairness was achieved with negligible cost in terms of model accuracy in all four settings. While this lack of fairness-accuracy trade-off is somewhat surprising on its face, the "top k" setting likely plays a role here as well. With limited resources relative to needs, there are many ways to choose k individuals for intervention with equally high precision, making it possible to swap some high-risk individuals from one group with those from another in order to improve fairness without appreciably reducing accuracy. To the extent that any small trade-offs may exist when making these adjustments, they seem to be dominated by variation over time in the generalization performance of the models, yielding consistently fair adjusted models without sacrificing accuracy (for a more detailed discussion of the lack of trade-offs with this approach, see our recent work in [64]).

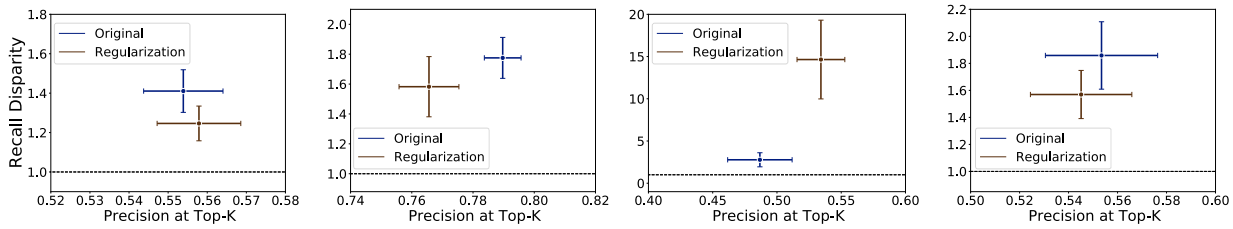
4.7 Effect of Composite Models

The final approach explored here follows Dwork's proposal [25] to build composite models, either through separate model selection or fully decoupled training for each subgroup. Figure 9 presents the results of these two strategies in each of the policy settings. In general, we find these approaches to perform quite well, consistently reducing the disparities across all four settings. As noted above, because the uncalibrated scores of these group-specific models cannot be assumed to be comparable, we combined the models across groups by making use of the same process of choosing TPR-equalizing thresholds as we used to make post-hoc adjustments to single models. As such, the fairness improvements seen here might either be a result of the composite strategy itself or the method for choosing thresholds, which, as seen above was itself very successful in reducing disparities here. However, if selecting (or training) separate models was appreciably improving model performance for the subgroups, we might hope to see increases in the overall accuracy of the composite models relative to the post-hoc adjusted ones in Figure 8, but the results here do not lend evidence to support this hypothesis. While there may be a slight improvement in precision@k for the composite model in the Student Outcomes setting, the difference in that setting is far from statistically significant and accuracy of the composite models in other settings is nearly identical to or somewhat lower than that of the post-hoc adjusted models.

To disentangle the effects of the composite modeling strategy itself from the TPR-equalizing group-specific thresholds used here, other strategies for choosing and combining the models across subgroups could be explored, although these

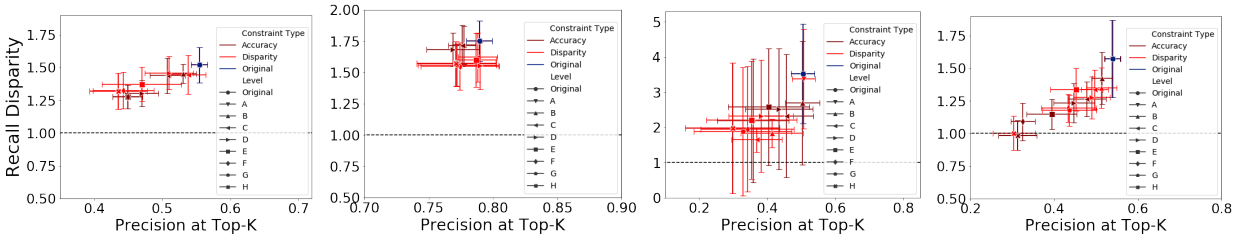
[4] Note that $FNR = 1 - TPR$, so this constraint is equivalent equalizing TPR across groups.

[5] We explored this package in particular in the Education Crowdfunding setting in a recent tutorial presented at the 2020 KDD and 2021 AAAI conferences [68].



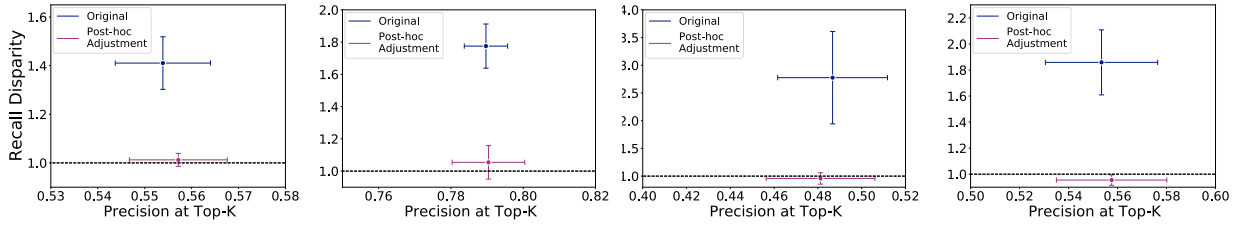
(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 6: Results of using the in-processing method proposed by Zafar and colleagues to perform fairness-constrained optimization during model training on model accuracy (precision@k) and fairness (recall disparities). Error bars show 95% confidence intervals over validation sets.



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 7: Effect of fairness-aware model selection on model accuracy (precision@k) and fairness (recall disparities). Model selection was performed either by setting a maximum acceptable disparity and choosing the model with the best precision@k among these (Disparity Constraint) or setting a maximum acceptable decrease in precision@k and choosing the lowest-disparity model among these (Accuracy Constraint). For each type, eight levels of constraint were explored (labeled A-H in the figure, from least to most weight on fairness). For Disparity Constraints, these are: A: 5.0, B: 2.0, C: 1.5, D: 1.3, E: 1.2, F: 1.1, G: 1.05, H: 1.0; for Accuracy Constraints, these are: A: 0.0, B: 0.05, C: 0.10, D: 0.15, E: 0.2, F: 0.25, G: 0.5, H: 0.6. Error bars show 95% confidence intervals over validation sets.



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 8: Effect of post-hoc adjustments on model accuracy (precision@k) and fairness (recall disparities). Separate score thresholds were chosen for the protected and non-protected subgroups to equalize recall across the groups. Error bars show 95% confidence intervals over validation sets.

are complicated somewhat by the requirement of the "top k" setting here that a total of k entities is selected across groups. The score threshold yielding the desired number of entities will vary considerably across pairs of models, so the appropriate cut-off at which to evaluate model performance for one subgroup depends on what models it will be combined with for other subgroups. As a preliminary experiment, we explored a simplified strategy in which we selected models for each subgroup based on their performance among the same number of highest-risk individuals that would have been selected from a single (non-composite) model. These group-specific models were combined and then the top k individuals with the highest scores in the composite model were chosen with a single score threshold.[6] In these initial experiments, composite models with a single score threshold failed to improve on either the accuracy or fairness of the original models, lending support to the conclusion that the improvements observed in Figure 9 are likely driven by the TPR-equalizing thresholds used to combine the models across subgroups.

In most of the problem settings considered here, the composite models with and without fully-decoupled training performed similarly, but the Housing Safety context provides a notable exception (Figure 9(c)). Although the composite approach performs well in this setting, the decoupled strategy shows a considerable loss in precision as well as overshooting the necessary adjustment to achieve a fair result (ending up with bias in the opposite direction). Notably, the Housing Safety dataset is considerably smaller than the others used here, with an order of magnitude fewer entities than the next-largest setting. As observed in [25], one potential disadvantage to decoupled model training is that the smaller number of training examples might degrade model performance, particularly if there are common patterns in the data that could be learned across groups. We would, of course, expect this issue to be exacerbated as the overall number of available examples decreases. Likewise, performing model selection on relatively small subgroups might be prone to over-fitting, choosing an overly-optimistic model specification whose performance reverts to a lower mean when measuring generalization performance on a future validation set. Such over-optimistic performance estimates for one subgroup could also affect the recall-balancing thresholds chosen across groups, leading to relatively too many individuals being chosen from one group and yielding disparities in the final composite model as well.

5. DISCUSSION

The goal of the current work was to build on the extensive recent work in algorithmic fairness by comparing how the wide variety of proposed fairness-enhancing approaches and methods perform in the context of real-world problems in high-stakes policy problems. While our aim was not to comprehensively include every existing approach, we sought to sample a wide range of techniques applied at different phases of the machine learning pipeline by pre-processing of the input data, in-processing during model training, and post-processing of trained models. Similarly, we focus here on resource-constrained assistive policy contexts where the optimization problem reflects a "top k" setting and we argue TPR disparities are an appropriate fairness metric (reflecting a concept of "equality of opportunity" as discussed in [32, 65]). While some of the results described here might not generalize beyond this problem setting, we note that it is very commonly encountered in high-stakes decisions across education [4, 49], healthcare [60], criminal justice [34], social services [10], as well as many other contexts [48, 73, 22, 7, 28], and has been the most common formulation encountered in the more than 100 projects we have been involved in applying machine learning to social good problems with government and non-profit partners.

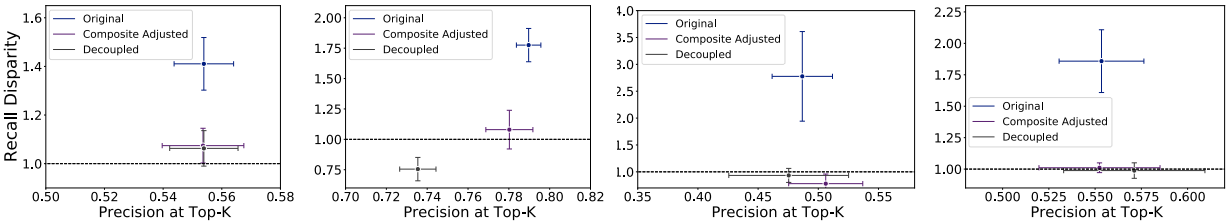
In this setting, **pre-processing methods showed decidedly mixed and inconsistent results**, with both sampling and omitting the protected attribute improving fairness in some contexts but not others. This inconsistency is perhaps not entirely surprising given the wide range of potential contributors to disparities at any stage of the machine learning pipeline [63], only some of which we might expect these pre-processing methods to address. Unfortunately, it seems unclear *a priori* whether these strategies will be effective in a given context (or, with sampling, what approach will work), making them unreliable as a fairness-enhancing approach.

Similarly, **we found little success with removing disparities through in-processing**, but, as noted above, existing methods to add fairness constraints in the process of model training seem particularly poorly suited to the "top k" setting. In principle, developing new in-processing methods better suited to the "top k" setting should be feasible, but poses particular technical challenges. As other work developing methods for this setting (without fairness constraints) has observed, the loss function in this setting is, in general, not only non-convex, but discontinuous (as disjoint regions of the parameter space yielding exactly k predicted positives must be connected by regions yielding either more or fewer than k) [51]. To our knowledge, no methods presently exist that seek to improve fairness through in-processing for "top k" models, but we believe this could be an interesting future research direction.

By contrast, we found **consistent success across the four problems with eliminating disparities using post-hoc methods.** Across all four policy settings considered here, these improvements in model fairness could be accomplished without a corresponding trade-off in accuracy. Although such trade-offs are often assumed to be an inherent aspect of reducing disparities in machine learning models [26, 76, 16] making this result somewhat surprising, the resource-constrained nature of these policy settings may contribute to the lack of trade-off as discussed above. Further, the consistency with which fair predictions could be obtained without cost to accuracy across the settings considered here may have important implications for policymakers and machine learning practitioners, reinforcing the moral imperative to ensure the fairness of models deployed in similar contexts (see our recent work in [64] for a more detailed discussion of these policy implications).

Given the success of these post-hoc adjustments across models and settings, we also investigated whether applying these adjustments on top of the pre-processing and in-processing strategies explored here could remove any residual (or newly

[6]Note that this approach will likely yield a different number of individuals in each group than was in process of selecting the group-specific models, so the assumption being made here is that these differences will be small enough that the model performance among each subgroup will not depart appreciably from what was used during selection.



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 9: Effect of composite model approaches on model accuracy (precision@k) and fairness (recall disparities). Composite models were developed by choosing separate models for each subgroup either with or without decoupled training. Error bars show 95% confidence intervals over validation sets.

introduced) disparities from those methods. Consistently, post-hoc adjustment by choosing thresholds to equalize TPR yielded more fair results, even when applied in combination with other strategies that failed to improve fairness themselves. While this result suggests a robustness of this strategy, we did not observe any improvement in model performance by combining post-hoc adjustments with other strategies, so we do not see any advantage to doing so in practice. Finally, we should note the importance of considering the broader context in which a machine learning model will be applied. While the work here has focused on improving the fairness of a model's predictions, doing so is only one step in the process of ensuring outcomes of the broader sociotechnical system are themselves equitable. In most policy contexts, these models are deployed in a manner intended to inform the decision-making process of a human expert such as a doctor, case worker, or school administrator, rather than being fully autonomous. As such, fairness in a model's recommendations is not necessarily a guarantee that interventions will be allocated fairly, depending on how and when these humans in the loop follow or override them. Further, the interventions themselves may not be equally effective for everyone. For instance, additional after-school tutoring might be difficult to attend for students who have work or family obligations in the afternoons, or programs offered only in English might not effectively serve individuals for whom it is not their first language. Likewise, when the labels themselves are measured in inaccurate and disparate ways, such as using arrests as a proxy for crime commission [5, 21, 53, 47, 31], measures of fairness that take these labels as "ground truth" will fail to capture these underlying disparities. Understanding the implications for fairness at each stage of the process — from label definition through modeling to decisions and interventions — is essential to understanding and mitigating biases in deployed machine learning systems that impact people's lives. The work here explores one key aspect of this process, but machine learning practitioners and the policymakers who deploy and act on the systems they build must be cognizant of these broader contextual aspects as well.

6. SUMMARY AND FUTURE WORK

In the present study, we explored the performance of several proposed fairness-enhancing methods on reducing bias and enhancing fairness in general and improving equality of opportunity (as measured by TPR disparities) in particular across four real-world policy contexts. Among the methods considered, we found that post-hoc adjustments to model scores by choosing TPR-equalizing group-specific score thresholds was capable of removing disparities without loss of accuracy in all four settings. Most directly, our results have implications for practitioners building and deploying machine learning systems in similar resource-constrained policy contexts for whom this post-hoc approach should be both straightforward to implement and likely to improve the fairness of their models. For the machine learning research community, we believe this work highlights the importance of evaluating new methods on real-world problems, in particular demonstrating a gap with how well-suited current in-processing methods are to this "top k" setting.

Although we focus here on characteristics of machine learning problems commonly encountered in high-stakes policy contexts, it will be important extend this work to other policy settings, particularly those for which other bias metrics beyond TPR disparity are of interest. In particular, we hope to understand whether the consistent improvements of the post-hoc adjustments employed here will generalize to other fairness metrics, especially those which are not guaranteed to be monotonically increasing or decreasing with the model score. Additionally, in all the settings considered here, the sensitive attribute was known exactly, but this is not always the case. Unfortunately, many of the approaches considered in this study (such as sampling, composite models, and post-hoc adjustment) cannot be directly applied where there is uncertainty around group membership, and more work will be required to both extend these methods to those contexts as well investigate the performance of methods that are inherently better-suited to them (such as those described in [18, 55]). Finally, although we sought to sample a range of fairness-enhancing methods across pre-, in-, and post-processing approaches, many more methods have been proposed than we could incorporate in the present work and continuing to extend upon these findings with additional methods will be an interesting avenue for future work.

7. ACKNOWLEDGEMENTS

This project was partially funded by the C3.AI Digital Transformations Institute. We would also like to thank the Data Science for Social Good Fellowship fellows, project partners, and funders as well as our colleagues at the Center for Data Science and Public Policy at University of Chicago for the initial work on projects that were extended and used in this study.

8. REFERENCES

[1] Data on donorschoose. https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose/data. Accessed: 2020-06-23.

[2] Statistics on donorschoose. https://www.donorschoose.org/about. Accessed: 2020-06-23.

[3] S. Aghaei, M. J. Azizi, and P. Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1418–1426, 2019.

[4] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. Addison. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, LAK '15, page 93–102, New York, NY, USA, 2015. Association for Computing Machinery.

[5] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias. Technical report, ProPublica, 5 2016.

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias risk assessments in criminal sentencing. *ProPublica, May*, 23, 2016.

[7] S. Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[8] M. N. Atwell, R. Balfanz, J. Bridgeland, and E. Ingram. Building a grad nation: Progress and challenge in raising high school graduation rates. annual update 2019. *Civic*, 2019.

[9] M. J. Bauman, K. S. Boxer, T.-Y. Lin, E. Salomon, H. Naveed, L. Haynes, J. Walsh, J. Helsby, S. Yoder, R. Sullivan, et al. Reducing incarceration through prioritized interventions. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–8, 2018.

[10] M. J. Bauman, R. Sullivan, C. Schneweis, R. Ghani, K. S. Boxer, T.-Y. Lin, E. Salomon, H. Naveed, L. Haynes, J. Walsh, J. Helsby, and S. Yoder. Reducing Incarceration through Prioritized Interventions. In *Proceedings of the Conference on Computing and Sustainable Societies (COMPASS)*, pages 1–8, New York, New York, USA, 2018. ACM.

[11] C. R. Belfield and H. M. Levin. *The price we pay: Economic and social consequences of inadequate education*. Brookings Institution Press, 2007.

[12] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[13] S. Bird, M. Dudık, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *URL https://www. microsoft. com/en-us/research/uploads/prod/2020/05/Fairlearn_whitepaper. pdf*, 2020.

[14] A. J. Bowers, R. Sprott, and S. A. Taff. Do we know who will drop out? a review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 96(2):77–100, 2012.

[15] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[16] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3995–4004, 2017.

[17] S. Caton and C. Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

[18] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[20] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[21] A. Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, 6 2017.

[22] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148. PMLR, 2018.

[23] D. Dua and C. Graff. UCI machine learning repository, 2017.

[24] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[25] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133. PMLR, 2018.

[26] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[27] E. Fuller Torrey, A. D. Kennard, D. Eslinger, R. Lamb, and J. Pavle. More Mentally Ill Persons Are in Jails and Prisons Than Hospitals: A Survey of the States. Technical report, Treatment Advocacy Center and National Sheriffs' Association, 2010.

[28] E. L. Glaeser, A. Hillis, S. D. Kominers, and M. Luca. Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review*, 106(5):114–18, 2016.

[29] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7):1445–1459, 2012.

[30] M. Hamilton. People with complex needs and the criminal justice system. *Current Issues in Criminal Justice*, 22(2):307–324, 2010.

[31] B. E. Harcourt. Risk as a Proxy for Race: The Dangers of Risk Assessment. *Federal Sentencing Reporter*, 27(4):237–243, 2015.

[32] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

[33] H. Heidari, C. Ferrari, K. P. Gummadi, and A. Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *arXiv preprint arXiv:1806.04959*, 2018.

[34] J. Helsby, S. Carton, K. Joseph, A. Mahmud, Y. Park, A. Navarrete, K. Ackermann, J. Walsh, L. Haynes, C. Cody, et al. Early intervention systems: Predicting adverse interactions between police and the public. *Criminal justice policy review*, 29(2):190–209, 2018.

[35] H. Holtzen, E. G. Klein, B. Keller, and N. Hood. Perceptions of physical inspections as a tool to protect housing quality and promote health equity. *Journal of health care for the poor and underserved*, 27(2):549–559, 2016.

[36] A. Howard and J. Borenstein. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24(5):1521–1536, 2018.

[37] M. Huzra. What do teachers spend on supplies, 2015.

[38] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[39] V. Iosifidis, B. Fetahu, and E. Ntoutsi. Fae: A fairness-aware ensemble framework. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1375–1380. IEEE, 2019.

[40] D. J. James and L. E. Glaze. Mental Health Problems of Prison and Jail Inmates. Technical report, US Department of Justice, Bureau of Justice Statistics, 2006.

[41] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR, 2020.

[42] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[43] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[44] T. Kehrenberg, Z. Chen, and N. Quadrianto. Tuning fairness by balancing target labels. *Frontiers in Artificial Intelligence*, 3:33, 2020.

[45] E. G. Klein, B. Keller, N. Hood, and H. Holtzen. Affordable housing and health: a health impact assessment on physical inspection frequency. *Journal of public health management and practice*, 21(4):368–374, 2015.

[46] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27, 2018.

[47] J. A. Kroll, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. Accountable Algorithms. *University of Pennsylvania Law Review*, 165:633–706, 2016.

[48] A. Kumar, S. A. A. Rizvi, B. Brooks, R. A. Vanderveld, K. H. Wilson, C. Kenney, S. Edelstein, A. Finch, A. Maxwell, J. Zuckerbraun, et al. Using machine learning to assess the risk of and prevent water main breaks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 472–480, 2018.

[49] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1909–1918, 2015.

[50] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1), 2016.

[51] L.-P. Liu, T. G. Dietterich, N. Li, and Z.-H. Zhou. Transductive optimization of top k precision. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1781–1787, 2016.

[52] P. Manisha and S. Gujar. A neural network framework for fair classifier. *arXiv preprint arXiv:1811.00247*, 10, 2018.

[53] S. G. Mayson. Bias In, Bias Out. *Yale Law Journal*, 128:2018–2035, 2019.

[54] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[55] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, 2018.

[56] I. Morgan and A. Amerikaner. Funding gaps 2018: An analysis of school funding equity across the us and within each state. *Education Trust*, 2018.

[57] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[58] O. A. Osoba and W. Welser IV. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation, 2017.

[59] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

[60] E. Potash, J. Brew, A. Loewi, S. Majumdar, A. Reece, J. Walsh, E. Rozier, E. Jorgenson, R. Mansour, and R. Ghani. Predictive modeling for public health: Preventing childhood lead poisoning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2039–2047, 2015.

[61] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020.

[62] A. Ramachandran, A. Kumar, H. Koenig, A. De Unanue, C. Sung, J. Walsh, J. Schneider, R. Ghani, and J. P. Ridgway. predictive analytics for retention in care in an urban hiv clinic. *Scientific reports*, 10(1):1–10, 2020.

[63] K. Rodolfa, P. Saliero, and R. Ghani. Bias and fairness. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, and J. Lane, editors, *Big data and social science*, chapter 13. CRC Press, 2020.

[64] K. T. Rodolfa, H. Lamba, and R. Ghani. Machine learning for public policy: Do we need to sacrifice accuracy to make models fair? *arXiv preprint arXiv:2012.02972*, 2020.

[65] K. T. Rodolfa, E. Salomon, L. Haynes, I. H. Mendieta, J. Larson, and R. Ghani. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 142–153, 2020.

[66] K. T. Rodolfa, E. Salomon, L. Haynes, I. H. Mendieta, J. Larson, and R. Ghani. Case study: Predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 142–153, New York, NY, USA, 1 2020. ACM.

[67] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[68] P. Saleiro, K. T. Rodolfa, and R. Ghani. Dealing with bias and fairness in data science systems: A practical hands-on tutorial. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3513–3514, 2020.

[69] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019.

[70] J. L. Skeem and C. T. Lowenkamp. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4):680–712, 2016.

[71] S. Verma and J. Rubin. Fairness Definitions Explained. *IEEE/ACM International Workshop on Software Fairness*, 18:7, 2018.

[72] P. Williams, W. Kendall, and M. Hooten. Model selection using multi-objective optimization. *arXiv preprint arXiv:1810.10669*, 2018.

[73] T. Ye, R. Johnson, S. Fu, J. Copeny, B. Donnelly, A. Freeman, M. Lima, J. Walsh, and R. Ghani. Using machine learning to help vulnerable tenants in new york city. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 248–258, 2019.

[74] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *26th International World Wide Web Conference*, pages 1171–1180, Perth, Australia, 2017. WWW.

[75] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 962–970, Fort Lauderdale, FL, 4 2017. PMLR.

[76] I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

Adversarial Attacks and Defenses: An Interpretation Perspective

Ninghao Liu[†], Mengnan Du[†], Ruocheng Guo[‡], Huan Liu[‡], Xia Hu[†]
Department of Computer Science and Engineering, Texas A&M University, TX, USA
Computer Science & Engineering, Arizona State University, Tempe, AZ, USA
[†]{nhliu43, dumengnan, xiahu}@tamu.edu, [‡]{rguo12, huanliu}@asu.edu

ABSTRACT

Despite the recent advances in a wide spectrum of applications, machine learning models, especially deep neural networks, have been shown to be vulnerable to *adversarial attacks*. Attackers add carefully-crafted perturbations to input, where the perturbations are almost imperceptible to humans, but can cause models to make wrong predictions. Techniques to protect models against adversarial input are called *adversarial defense* methods. Although many approaches have been proposed to study adversarial attacks and defenses in different scenarios, an intriguing and crucial challenge remains that how to really understand model vulnerability? Inspired by the saying that "if you know yourself and your enemy, you need not fear the battles", we may tackle the challenge above after interpreting machine learning models to open the black-boxes. The goal of *model interpretation*, or *interpretable machine learning*, is to extract human-understandable terms for the working mechanism of models. Recently, some approaches start incorporating interpretation into the exploration of adversarial attacks and defenses. Meanwhile, we also observe that many existing methods of adversarial attacks and defenses, although not explicitly claimed, can be understood from the perspective of interpretation. In this paper, we review recent work on adversarial attacks and defenses, particularly from the perspective of machine learning interpretation. We categorize interpretation into two types, feature-level interpretation, and model-level interpretation. For each type of interpretation, we elaborate on how it could be used for adversarial attacks and defenses. We then briefly illustrate additional correlations between interpretation and adversaries. Finally, we discuss the challenges and future directions for tackling adversary issues with interpretation.

Keywords

Adversarial attacks, adversarial defenses, interpretation, explainability, deep learning

1. INTRODUCTION

Machine learning (ML) techniques, especially recent deep learning models, are progressing rapidly and have been increasingly applied in various applications. Nevertheless, concerns have been posed about the security and reliability issues of ML models. In particular, many deep models are susceptible to adversarial attacks [1; 2]. That is, after adding certain well-designed but human imperceptible perturbation or transformation to a clean data instance, we are able to manipulate the prediction of the model. The data instances after being attacked are called *adversarial samples*. The phenomenon is intriguing since clean samples and adversarial samples are usually not distinguishable to humans. Adversarial samples may be predicted dramatically differently from clean samples, but the predictions usually do not make sense to a human.

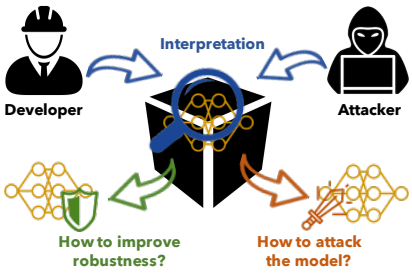


Figure 1: Interpretation can either provide directions for improving model robustness or attacking on its weakness.

The model vulnerability to adversarial attacks has been discovered in various applications or under different constraints. For examples, approaches for crafting adversarial samples have been proposed in tasks such as classification (e.g., on image data [3], text data [4], tabular data [5], graph data [6; 7]), object detection [8], and fraud detection [9]. Adversarial attacks could be initiated under different constraints, such as assuming limited knowledge of attackers on target models [10; 11], assuming higher generalization level of attack [12; 13], posing different real-world constraints on attack [14; 15]. Given the advances, several questions could be posted. First, are these advances relatively independent of each other, or is there an underlying perspective from which we can discover the commonality behind them? Second, should adversarial samples be seen as the negligent corner cases that could be fixed by putting patches to models, or are they deeply rooted in the internal working mechanism of models that it is not easy to get rid of?

Motivated by the idiom that "if you know yourself and your enemy, you need not fear the battles" from *The Art of War*, in this paper, we answer the above questions and review the recent advances of adversarial attack and defense approaches from the perspective of interpretable machine learning. The

relation between model interpretation and model robustness is illustrated in Figure 1. On the one hand, if adversaries know how the target model works, they may utilize it to find model weakness and initiate attacks accordingly. On the other hand, if model developers know how the model works, they could identify the vulnerability and work on remediation in advance. Interpretation refers to the human-understandable information explaining what a model has learned or how a model makes predictions. Exploration of model interpretability has attracted many interests in recent years, because recent machine learning techniques, especially deep learning models, have been criticized due to lack of transparency. Some recent work starts to involve interpretability in the analysis of adversarial robustness. Also, although not being explicitly specified, in this survey, we will show that many existing adversary-related work can be comprehended from another perspective as an extension of model interpretation.

Before connecting the two domains, we first briefly introduce the subjects of interpretation to be covered in this paper. *Interpretability* is defined as "the ability to explain or to present in understandable terms to a human [16]". Although a formal definition of interpretation still remains elusive [16; 17; 18; 19], the overall goal is to obtain and transform information from models or their behaviors into a domain that human can make sense of [20]. For a more structured analysis, we categorize existing work into two categories: feature-level interpretation and model-level interpretation, as shown in Figure 2. Feature-level interpretation targets to find the most important features in a data sample for its prediction. Model-level interpretation explores the functionality of model components, and their internal states after being fed with input. This categorization is based on whether the internal working mechanism of models is involved in interpretation.

Following the above categorization, the overall structure of this article is organized as below. To begin with, we briefly introduce different types of adversarial attack and defense strategies in Section 2. Then, we introduce different categories of interpretation approaches, and demonstrate in detail how interpretation correlates to the attack and defense strategies. Specifically, we discuss feature-level interpretation in Section 3 and model-level interpretation in Section 4. After that, we extend the discussion to additional relations between interpretation and adversarial aspects of models in Section 5. Finally, we discuss some opening challenges for future work in Section 6.

2. ADVERSARIAL MACHINE LEARNING

Before understanding how interpretation helps adversarial attack and defense, we first provide an overview of existing attack and defense methodologies.

2.1 Adversarial Attacks

In this subsection, we introduce different types of threat models for adversarial attacks. The overall threat models may be categorized under different criteria. Based on different application scenarios, conditions, and adversary capabilities, specific attack strategies will be deployed.

2.1.1 Untargeted vs Targeted Attack

Based on the goal of attackers, the threat models can be classified into targeted and untargeted ones. For *targeted* attack,

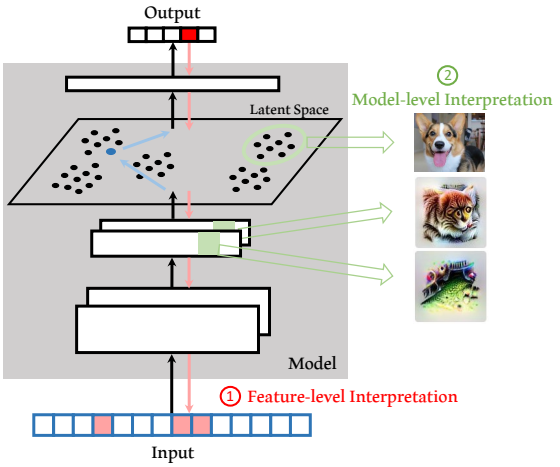


Figure 2: Illustration of feature-level interpretation and model-level interpretation for a deep model.

it attempts to mislead a model's prediction to a specific class given an instance. Let f denote the target model exposed to adversarial attack. A clean data instance is $\mathbf{x}_0 \in X$, and X is the input space. We consider classification tasks, so $f(\mathbf{x}_0) = c, c \in \{1, 2, ..., C\}$. One way of formulating the task of targeted attack is as below [2]:

$$\min_{\mathbf{x} \in X} \ d(\mathbf{x}, \mathbf{x}_0), \quad \text{s.t. } f(\mathbf{x}) = c' \tag{1}$$

where $c' \neq c$, and $d(\mathbf{x}, \mathbf{x}_0)$ measures the distance between the two instances. A typical choice of distance measure is l_p norms, where $d(\mathbf{x}, \mathbf{x}_0) = \|\mathbf{x} - \mathbf{x}_0\|_p$. The core idea is to add small perturbation to the original instance $\mathbf{x}_0$ to make it being classified as c'. However, in some cases, it is important to increase the confidence of perturbed samples being misclassified, so the task may also be formulated as:

$$\max_{\mathbf{x} \in X} \ f_{c'}(\mathbf{x}), \quad \text{s.t. } d(\mathbf{x}, \mathbf{x}_0) \leq \delta \tag{2}$$

where $f_{c'}(\mathbf{x})$ denotes the probability or confidence that $\mathbf{x}$ is classified as c' by f, and δ is a threshold limiting perturbation magnitude. For *untargeted* attack, its goal is to prevent a model from assigning a specific label to an instance. The objective of untargeted attack could be formulated in a similar way as targeted attack, where we just need to change the constraint as $f(\mathbf{x}) \neq c$ in Equation 1, or change the objective as $\min_{\mathbf{x} \in X} f_c(\mathbf{x})$ in Equation 2.

In some scenarios, the two types of attacks above are also called *false positive* attack and *false negative* attack. The former aims to make models misclassify negative instances as positive, while the latter tries to mislead models to classify positive instances as negative. False positive attacks and false negative attacks sometimes are also called Type-I attacks and Type-II attacks, respectively.

2.1.2 One-Shot vs Iterative Attack

According to practical constraints, adversaries may initiate one-shot or iterative attacks to target models. In *one-shot* attack, they have only one chance to generate adversarial samples, while *iterative attack* could take multiple steps to find the better perturbation direction. Iterative attacks can

generate more effective adversarial samples than one-shot attacks. However, it also requires more queries to the target model and more computation to initiate each attack, which may limit its application in some computational-intensive tasks.

2.1.3 Data-Dependent vs Universal Attack

According to information sources, adversarial attacks could be data-dependent or data-independent. In *data dependent* attack, perturbations are customized based on the target instance. For example, in Equation 1, the adversarial sample $\mathbf{x}$ is crafted based on the original instance $\mathbf{x}_0$. However, it is also possible to generate adversarial samples without referring to the input instance, and it is also named as *universal* attack [12; 21]. The problem can be abstracted as looking for a perturbation vector $\mathbf{v}$ so that

$$f(\mathbf{x}+\mathbf{v}) \neq f(\mathbf{x}) \text{ for "most" } \mathbf{x} \in X. \tag{3}$$

We may need a number of training samples to obtain $\mathbf{v}$, but it does not rely on any specific input at test time. Adversarial attacks can be implemented efficiently once the vector $\mathbf{v}$ is solved.

2.1.4 Perturbation vs Replacement Attack

Adversarial attacks can also be categorized based on the way of input distortion. In *perturbation* attack, input features are shifted by specific noises so that the input is misclassified by the model. In this case, let $\mathbf{x}^*$ denote the final adversarial sample, then it can be obtained via

$$\mathbf{x}^* = \mathbf{x}_0 + \Delta\mathbf{x}, \tag{4}$$

and usually $\|\Delta\mathbf{x}\|_p$ is small.

In *replacement* attack, certain parts of the input are replaced by adversarial patterns. Replacement attack is more natural in physical scenarios. For example, criminals may want to wear specifically designed glasses to prevent them from being recognized by computer vision systems [1]. Also, surveillance cameras may fail to detect persons wearing clothes attached with adversarial patches [14]. Suppose $\mathbf{v}$ denotes the adversarial pattern, then replacement attack can be represented by using a mask $\mathbf{m} \in \{0,1\}^{|\mathbf{x}_0|}$, so that

$$\mathbf{x}^* = \mathbf{x}_0 \odot (\mathbf{1}-\mathbf{m}) + \mathbf{v} \odot \mathbf{m} \tag{5}$$

where the symbol $\odot$ denotes element-wise multiplication.

2.1.5 White-Box vs Black-Box Attack

In *white-box* attack, it is assumed that attackers know everything about the target model, which may include model architecture, weights, hyper-parameters, and even training data. White-box attacks help to discover intrinsic vulnerabilities of the target model. It works in ideal cases representing the worst scenario that defenders have to confront. *Black-box* attack assumes that attackers are only accessible to the model output, just like regular end-users. This is a more practical assumption in real-world scenarios. Although a lot of detailed information about models is occluded, black-box attacks still pose a significant threat to machine learning systems due to the transferability property of adversarial samples discovered in [11]. In this sense, an attacker could build a new model f' to approximate the target model f, and adversarial samples created on f' could still be effective to f.

[1]https://www.inovex.de/blog/machine-perception-face-recognition/

2.2 Defenses Against Adversarial Attacks

In this subsection, we briefly introduce the basic idea of different defense strategies against adversaries.

2.2.1 Input Denoising

As adversarial perturbation is a type of human-imperceptible noise added to data, then a natural defense solution is to filter it out, or to use additional random transformation to offset adversarial noise. It is worth noting that f_m could be added prior to model input layer [22; 23; 24], or as an internal component inside the target model [25]. Formally, for the former case, given an instance $\mathbf{x}^*$ which is probably affected by adversaries, we hope to design a mapping f_m, so that $f(f_m(\mathbf{x}^*)) = f(\mathbf{x}_0)$. For the latter case, the idea is similar except that f is replaced by certain intermediate layer output h.

2.2.2 Model Robustification

Refining the model to prepare itself against a potential threat from adversaries is another widely applied strategy. The model refinement could be achieved from two directions: changing the training objective, or modifying the model structure. Some examples of the former one include adversarial training [2; 1], and replacing empirical training loss with robust training loss [26]. The intuition behind it is to consider in advance the threat of adversarial samples during model training, so that the resultant model gains robustness from training. Examples of model modification include model distillation [27], applying layer discretization [28], controlling neuron activations [29]. Formally, let f' denote the robust model, the goal is to make $f'(\mathbf{x}^*)) = f'(\mathbf{x}_0) = y$.

2.2.3 Adversarial Detection

Unlike the previous two strategies where we hope to discover the true label given an instance, adversarial detection tries to identify whether the given instance is polluted by adversarial perturbation. The general idea is to build another predictor f_d, so that $f_d(\mathbf{x}) = 1$ if $\mathbf{x}$ has been polluted, and otherwise $f_d(\mathbf{x}) = 0$. The establishment process of f_d could follow the normal routine of building a binary classifier [30; 31; 32].

Input denoising and model robustification methods proactively recover the prediction from influences of adversarial attacks by focusing on modifying the input data and model architectures, respectively. Adversarial detection methods passively decide whether the model should make predictions against the input in order not to be fooled. Implementations of proactive strategies are usually more challenging than passive ones.

3. FEATURE-LEVEL INTERPRETATION IN ADVERSARIAL MACHINE LEARNING

Feature-level interpretation is a widely used post-hoc method to identify feature importance for a prediction result. It focuses on the end-to-end relation between input and output, instead of carefully examining the internal states of models. Some examples include measuring the importance of phrases of sentences in text classification [33], and pixels in image classification [34]. In this section, we will discuss how

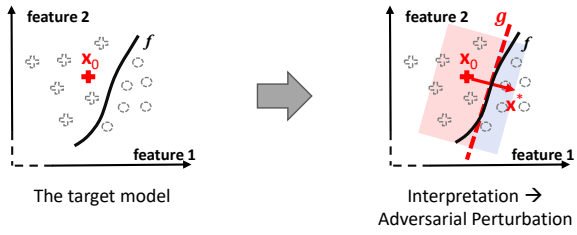


Figure 3: Interpretation naturally unveils the direction of adversarial perturbation (g denotes the local interpreter).

this type of interpretation correlates with the attack and defense of adversaries, given that many works on adversarial machine learning do not analyze adversaries from this perspective.

3.1 Feature-Level Interpretation for Understanding Adversarial Attacks

In this part, we will show that many feature-level interpretation techniques are closely coupled with existing adversarial attack methods, thus providing another perspective to understand adversarial attacks.

3.1.1 Gradient-Based Techniques

Following the notations in previous discussion, we let $f_c(\mathbf{x}_0)$ denote the probability that model f classifies the input instance $\mathbf{x}_0$ as class c. One of the intuitive ways to understand why such prediction is derived is to attribute prediction $f_c(\mathbf{x}_0)$ to feature importance in $\mathbf{x}_0$. According to [35], $f_c(\mathbf{x}_0)$ can be approximated with a linear function surrounding $\mathbf{x}_0$ by computing the first-order Taylor expansion:

$$f_c(\mathbf{x}) \approx f_c(\mathbf{x}_0) + \mathbf{w}_c^T \cdot (\mathbf{x} - \mathbf{x}_0) \tag{6}$$

where $\mathbf{w}_c$ is the gradient of f_c with respect to input at $\mathbf{x}_0$, i.e., $\mathbf{w}_c = \nabla_{\mathbf{x}} f_c(\mathbf{x}_0)$. From the interpretation perspective, $\mathbf{w}_c$ entries of large magnitude correspond to the features that are important around the current output.

However, another perspective to comprehend the above equation is that, the interpretation $\mathbf{w}_c$ also indicates the most effective direction to change the prediction result by perturbing input away from $\mathbf{x}_0$. If we let $\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}_0 \propto -\mathbf{w}_c$, we are attacking the model f with respect to the input-label pair $(\mathbf{x}_0, c)$. Such perturbation method is closely related to the Fast Gradient Sign (FGS) attacking method [1], where:

$$\Delta\mathbf{x} = \epsilon \cdot sign(\nabla_{\mathbf{x}} J(f, \mathbf{x}_0, c)), \tag{7}$$

except that (1) FGS computes the gradient of a certain cost function J nested outside f, and (2) it applies an additional $sign()$ operation on gradient for processing images. However, if we define J with cross entropy loss, and the true label of $\mathbf{x}_0$ is c, then

$$\nabla_{\mathbf{x}} J(f, \mathbf{x}_0, c) = -\nabla_{\mathbf{x}} \log f_c(\mathbf{x}_0) = -\frac{1}{f_c(\mathbf{x}_0)} \nabla_{\mathbf{x}} f_c(\mathbf{x}_0), \tag{8}$$

which points to exactly the opposite direction of interpretation $\mathbf{w}_c$. The high-level idea behind this case is that, if the interpretation of a model is known, a straightforward way to undermine the model is to remove the important information or components relevant to the interpretation.

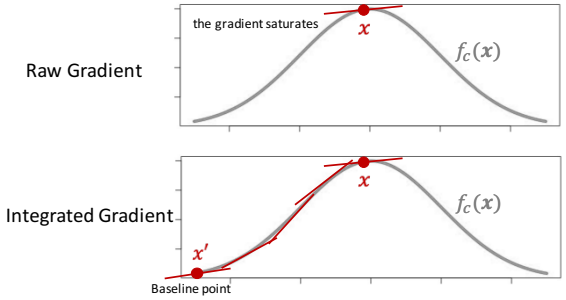


Figure 4: Raw gradients only consider the local sensitivity of output to input value changes, which could be limited in measuring the contribution of a feature to the prediction.

The traditional FGS method is proposed under untargeted attacks, where the goal is to impede input from being correctly classified. For targeted attack, where the goal is to misguide the model prediction towards a specific class, a typical way is Box-constrained L-BFGS (L-BFGS-B) method [2]. Assume c' is the target label, the problem of L-BFGS-B is formulated as:

$$\underset{\mathbf{x} \in X}{\text{argmin}} \quad \alpha \cdot d(\mathbf{x}, \mathbf{x}_0) + J(f, \mathbf{x}, c') \tag{9}$$

where d is considered to control the perturbation distance, and X is the input domain (e.g., $[0, 255]$ for each channel of image input). The goal of attack is to make $f(\mathbf{x}) = c'$, while making $d(\mathbf{x}, \mathbf{x}_0)$ to be small. Suppose we apply gradient descent to solve the problem, and $\mathbf{x}_0$ is the starting point. Similar to the previous discussion, if we define J as the cross entropy loss, then

$$-\nabla_{\mathbf{x}} J(f, \mathbf{x}_0, c')) = \nabla_{\mathbf{x}} \log f_{c'}(\mathbf{x}_0) \propto \mathbf{w}_{c'} . \tag{10}$$

On one hand, $\mathbf{w}_{c'}$ locally and linearly interprets $f_{c'}(\mathbf{x}_0)$, and it also serves the most effective direction to make $\mathbf{x}_0$ towards being classified as c'.

According to the taxonomy of adversarial attacks, the two scenarios discussed above can also be categorized into: (1) one-shot attack, since we only perform interpretation once, (2) data-dependent attack, since the perturbation direction is related with $\mathbf{x}_0$, (3) white-box attack, since model gradients are available. Other types of attack could be crafted if different interpretation strategies are applied, which will be discussed in later sections.

Improved Gradient-Based Techniques. The interpretation methods based on raw gradients, as discussed above, are usually unstable and noisy [36; 37]. The possible reasons include: (1) the target model's prediction function itself is not stable; (2) gradients only consider the local output-input relation so that its scope is too limited (Figure 4); (3) the prediction mechanism is too complex to be approximated by a linear substitute. Some approaches for improving interpretation (i.e., potential adversarial attack) are as below.

- **Region-Based Exploration:** To reduce random noises in interpretation, SmoothGrad is proposed in [38], where the final interpretation $\mathbf{w}_c$, as a sensitivity map, is obtained by averaging multiple interpretation results of instances sampled around the target instance $\mathbf{x}_0$, i.e., $\mathbf{w}_c = \sum_{\mathbf{x}' \in \mathcal{N}(\mathbf{x}_0)} \frac{1}{|\mathcal{N}(\mathbf{x}_0)|} \nabla f_c(\mathbf{x}')$. The averaged sensitivity map

will be visually sharpened. A straightforward way to extend it for adversarial attack is to perturb input by reversing the averaged interpretation. Furthermore, [39] designed a different strategy by adding a step of random perturbation before gradient computation in attack, to jump out of the non-smooth vicinity of the initial instance. Spatial averaging is a common technique to stabilize output. For example, [40] applies it as a defense method to derive more stable model predictions.

- **Path-Based Integration:** To improve interpretation and consider a broader input scope, [41] proposes Integrated Gradient (InteGrad). After setting a baseline point $\mathbf{x}^b$, e.g., an all-black image in classification tasks, the interpretation is defined as:

$$\mathbf{s}_c = \frac{(\mathbf{x}_0 - \mathbf{x}^b)}{D} \circ \sum_{d=1}^{D} [\nabla f_c](\mathbf{x}^b + \frac{d}{D}(\mathbf{x}_0 - \mathbf{x}^b)), \quad (11)$$

which is the weighted sum of gradients along the straight-line path from $\mathbf{x}_0$ to the baseline point $\mathbf{x}^b$. Let $\mathbf{s}_c(m)$ denote the m-th entry of $\mathbf{s}_c$, then the prediction function could be decomposed as below:

$$f_c(\mathbf{x}) \approx f_c(\mathbf{x}^b) + \sum_{m=1}^{M} \mathbf{s}_c(m), \quad (12)$$

which is different from the decomposition in Eq 6. Here $\mathbf{s}_c(m)$ denotes the contribution of the m-th feature to the prediction result. Therefore, a new type of adversarial attack could be conducted by deleting or removing those features with high contribution scores. This type of feature deletion or feature occlusion attack is different from FGS that perturbs feature values.

Interestingly, although in many cases gradient-based interpretation is intuitive as visualization to show that the model is functioning well, it may be an illusion since we can easily transform interpretation into adversarial perturbation.

3.1.2 Distillation-Based Techniques

The interpretation techniques discussed so far require gradient information $\nabla_{\mathbf{x}} f$ from models. Meanwhile, it is possible to extract interpretation without querying a model f more than $f(\mathbf{x})$. This type of interpretation method, here named as the distillation-based method, can also be used for adversarial attacks. Since no internal knowledge is required from the target model, they are usually used for black-box attacks.

The main idea of applying distillation for interpretation is to use an interpretable model g (e.g., a linear model) to locally mimic the behavior of the target deep model f [42; 43]. Once we obtain g, existing white-box attack methods could be applied to craft adversarial samples [5]. In addition, given an instance $\mathbf{x}_0$, to guarantee that g more accurately mimics the behaviors of f, we could further require that g locally approximates f around the instance. The objective is thus as below:

$$\min_g \mathcal{L}(f, g, \mathbf{x}_0) + \alpha \cdot C(g), \quad (13)$$

where $\mathcal{L}$ denotes the approximation error around $\mathbf{x}_0$. For example, in LIME [44]:

$$\mathcal{L}(f, g, \mathbf{x}_0) = \sum_{\mathbf{x}' \in \mathcal{N}(\mathbf{x}_0)} \exp(-d(\mathbf{x}_0, \mathbf{x}')) \|f(\mathbf{x}') - g(\mathbf{x}')\|_2^2, \quad (14)$$

and $\mathcal{N}(\mathbf{x}_0)$ denotes the local region around $\mathbf{x}_0$. In addition, LEMNA [45] adopts mixture regression models for g and fused lasso as regularization $C(g)$. After obtaining g, we can craft adversarial samples targeting g by removing important features or perturbing input towards the reversed direction of interpretation. According to the property of transferability [11], an adversarial sample that successfully fools g is also likely to fool f. The advantages are two-fold. First, the process is model-agnostic and does not assume availability to gradients. It could be used for black-box attacks or attacking certain types of models (such as tree-based models) that do not use gradient backpropagation in training. Second, one-shot attacks on g could be more effective thanks to the smoothness term $C(g)$ as well as extending the consideration to include the neighborhood of $\mathbf{x}_0$ [46]. Thus, it has the potential to make defense methods based on obfuscated gradients [47] to be less robust. However, the disadvantage is that crafting each adversarial sample requires higher computation cost.

In certain scenarios, it may be beneficial to make adversarial patterns understandable to humans as real-world simulation when identifying model vulnerability. For example, in autonomous driving, we need to consider physically-possible patterns that could cause misjudgment of autonomous vehicles [48]. One possible approach is to encourage adversarial instances to fall into the data distribution [49], which could be implemented through a regularization term $\|\mathbf{x}_0 + \Delta\mathbf{x} - AE(\mathbf{x}_0 + \Delta\mathbf{x})\|_2^2$, where $AE(\cdot)$ denotes an autoencoder. By minimizing the normalization term, the perturbed data $\mathbf{x}_0 + \Delta\mathbf{x}$ can be well modeled by the autoencoder, which implies that it is close to the data manifold. Another strategy is to predefine a dictionary, and then make the adversarial perturbation to match one of the dictionary tokens [48], or a weighted combination of the tokens [50].

3.1.3 Influence-Function Based Techniques

Instead of measuring feature importance (e.g., feature sensitivity, feature contribution) as explanations, influence functions provide a new perspective by measuring the importance of data instances. Suppose $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_N$ are the training instances, and θ denotes the model parameters. Let $\mathcal{L}(\mathbf{x}_n, \theta)$ be the loss on a single instance, and $\frac{1}{N}\sum_{n=1}^{N} \mathcal{L}(\mathbf{x}_n, \theta)$ be the overall empirical loss. The optimal parameters are given by $\hat{\theta} = \text{argmin}_\theta \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}(\mathbf{x}_n, \theta)$. According to [51], influence function could be used to answer several questions: (1) how model parameters θ would change if an instance $\mathbf{x}_n$ is removed, (2) how model prediction on a test point $\mathbf{x}_{test}$ would change if an instance $\mathbf{x}_n$ is removed, (3) how model prediction would change if an instance $\mathbf{x}_n$ is modified. By answering the third question, through experimental demonstration, the paper shows that after injecting perturbed data instances into the training set (i.e., data poisoning), the new model will make wrong predictions on some test points.

In explanations derived from influence functions, the fundamental unit is the data instance instead of the feature. Therefore, it seems difficult to directly utilize explanation results from influence functions to initiate adversarial attacks. However, in graph analysis, influence functions are

useful in studying the importance of graph components (e.g., nodes and edges) that can be regarded as the "features" of the graph. A graph can be denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ is the set of nodes, and $\mathcal{E}$ is the set of edges. Each edge is denoted as $(v_i, v_j) \in \mathcal{E}$. An important task in graph analysis is node embedding, where we learn an embedding vector for each node. The embedding vectors can be used in downstream tasks such as node classification and link prediction. One of the fundamental requirements for learning embeddings is that the embeddings of nodes that are connected or have similar contexts (e.g., similar neighbors) should be close to each other. By utilizing influence functions, it is possible to identify how adding or deleting an edge would change the node embeddings [52]. The addition or deletion of a small number of edges can be seen as adversarial attacks on graph data.

3.2 Feature-Level Interpretation for Adversarial Defenses

The feature-level interpretation could be used for defense against adversaries through adversarial training and detecting model vulnerability.

3.2.1 Model Robustification With Feature-Level Interpretation

The feature-level interpretation could help adversarial training to improve model robustness. Adversarial training [1; 3] is one of the most applied proactive countermeasures to improve the robustness of the model. Feature-level interpretation could help in crafting adversarial samples to unveil the weakness of the model. The adversarial samples are then injected into the training set for data augmentation. The overall loss function can be formulated as:

$$\min_f \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}} [\alpha J(f(\mathbf{x}), y) + (1-\alpha) J(f(\mathbf{x}^*), y)]. \quad (15)$$

In the scenario of adversarial training, feature-level interpretation helps in preparing adversarial samples $\mathbf{x}^*$, which may refer to any method discussed in Section 3.1.1 and Section 3.1.2. Although such an attack-and-then-debugging strategy has been successfully applied in many traditional cybersecurity scenarios, one key drawback is that it tends to overfit the specific approach that is used to generate $\mathbf{x}^*$. Therefore, the adversarial training is usually conducted for multiple rounds.

To train more robust models, some optimization based methods have been proposed. [26] argues that traditional Empirical Risk Minimization (ERM) fails to yield models that are robust to adversarial instances, and proposed a min-max formulation to train robust models:

$$\min_f \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}} [\max_{\delta\in\Delta X} J(\mathbf{x}+\delta, y)], \quad (16)$$

where ΔX denotes the set of allowed perturbations. It formally defines adversarially robust classification as a learning problem to reduce adversarial expected risk. This min-max formulation provides another perspective on adversarial training, where the inner task aims to find adversarial samples, and the outer task retrains model parameters. [39] further improves its defense performance by crafting adversarial samples from multiple sources to augment training data. [53] further identifies a trade-off between robust classification error and natural classification error, which provides a solution to reduce the negative effect on model accuracy after adversarial training.

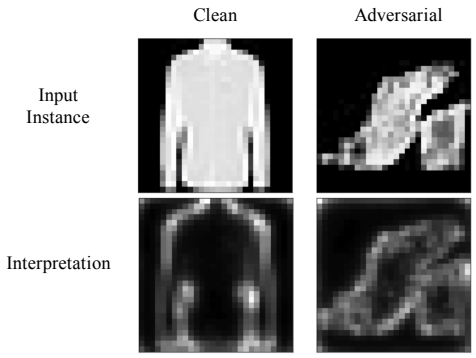


Figure 5: The interpretation of an adversarial sample may differ from the one of a clean sample. Top-left: a normal example from the shirt class of Fashion-MNIST dataset. Bottom-left: the explanation map for the classification. Top-right: an adversarial example, originally from the sandal class, that is misclassified as a shirt. Bottom-right: the explanation map for the misclassification.

Besides adversarial training, in more cases, feature-level interpretation plays the role of providing motivation for robust learning. For example, empirical interpretation results pointed out, that an intriguing property of CNN is its bias towards texture instead of shape information when making predictions [54]. To tackle this problem, [55] proposes Info-Drop, a plug-in filtering method to remove texture-intensive information during forward propagation of CNN. Feature map regions with low self-information, i.e., regions with texture patterns that contain less "surprise", tend to be filtered out. In this way, the model will pay more attention to regions such as edges and shapes, and be more robust under various scenarios including adversaries.

3.2.2 Adversarial Detection With Feature-Level Interpretation

In the scenario where a model is subject to adversarial attacks, interpretation may serve as a new type of information for directly detecting adversarial patterns. The motivation is illustrated in Figure 5. In the adversarial image which originally shows a shoe, although the model classifies it as a shirt, its interpretation result does not resemble the one obtained from the clean image of a shirt. A straightforward way to distinguish interpretations is to train another classifier as the detector trained with interpretations of both clean and adversarial instances, paired with labels indicating whether the sample is clean [56; 57; 58; 59]. Specifically, [59] directly uses gradient-based saliency map as interpretation, [58] adopts the distribution of Leave-One-Out (LOO) attribution scores, while [57] proposes a new interpretation method based on masks highlighting important regions. [60] proposes an ensemble framework called X-Ensemble for detecting adversarial samples. X-Ensemble consists of multiple sub-detectors, each of which is a convolutional neural network to classify whether an instance is adversarial or benign. The input to each sub-detector is the interpretation of the instance's prediction. More than one interpretation method

is deployed, so there are multiple sub-detectors. A random forest model is then used to combine sub-detectors into a powerful ensemble detector.

In more scenarios, interpretation serves as a diagnostic tool to qualitatively identify model vulnerability. First, we could use interpretation to identify whether inputs are affected by adversarial attacks. For example, if the interpretation result shows that unreasonable evidence has been used for prediction [61], then it is possible that there exist suspicious but imperceptible input patterns. Second, interpretation may reflect whether a model is susceptible to adversarial attack. Even given a clean input instance, if the interpretation of model prediction does not make much sense to humans, then the model is at the risk of being attacked. For example, in a social spammer detection system, if the model regards certain features as important, but they are not strongly correlated with maliciousness, then attackers could easily manipulate these features without much cost to fool the system [5]. Also, in image classification, CNN models have been demonstrated to focus on local textures instead of object shapes, which could be easily utilized by attackers [54]. An interesting phenomenon in image classification is that, after refining a model with adversarial training, feature-level interpretation results indicate that the refined model will be less biased towards texture features [62].

Nevertheless, there are several challenges that impede the intuitions above from being formulated to formal defense approaches. First, the interpretation itself is also fragile in neural networks. Attackers could control prediction and interpretation simultaneously via indistinguishable perturbation [63; 64]. Second, it is difficult to quantify the robustness of a model through interpretation [36]. Manual inspection of interpretation helps discover defects in model, but visually acceptable interpretation does not guarantee model robustness. That is, defects in feature-level interpretation indicate the presence but not the absence of vulnerability.

4. MODEL-LEVEL INTERPRETATION IN ADVERSARIAL MACHINE LEARNING

In this review, model-level interpretation is defined with two aspects. First, model-level interpretation aims to figure out what has been learned by intermediate components in a trained model [65; 35], or what is the meaning of different locations in latent space [66; 67; 68]. Second, given an input instance, model-level interpretation unveils how the input is encoded by those components as latent representation [66; 67; 23; 25]. In our discussion, the former does not rely on input instances, while the latter is the opposite. Therefore, we name the two aspects as *Model Component Interpretation* and *Representation Interpretation* respectively to further distinguish them.

4.1 Model Component Interpretation for Understanding Adversarial Attacks

In deep models, model component interpretation can be defined as exploring the visual or semantic meaning of each neuron. A popular strategy to solve this problem is to recover the patterns that activate the neuron of interests at a specific layer [69; 35]. Following the previous notations, let $h(\mathbf{x})$ denote the activation degree of neuron h given input $\mathbf{x}$, the perceived pattern of the neuron can be visualized by solving the problem below:

$$\underset{\mathbf{x}'}{\text{argmax}}\, h(\mathbf{x}') - \alpha \cdot C(\mathbf{x}'), \tag{17}$$

where $C(\cdot)$ such as $\|\cdot\|_1$ or $\|\cdot\|_2$ acts as regularization. Conceptually, the result contains patterns that neuron h is sensitive to. If we choose h to be f_c, then the resultant $\mathbf{x}'$ illustrates class appearances learned by the target model. Another discussion about different choices of h, such as neurons, channels, layers, logits and class probabilities, is provided in [70]. Similarly, we could also formulate another minimization problem

$$\underset{\mathbf{x}'}{\text{argmin}}\, h(\mathbf{x}') + \alpha \cdot C(\mathbf{x}'), \tag{18}$$

to produce patterns that prohibit activation of certain model components or prediction towards certain classes.

The interpretation result $\mathbf{x}'$ is highly related with several types of adversarial attacks, with some examples shown in Figure 6.

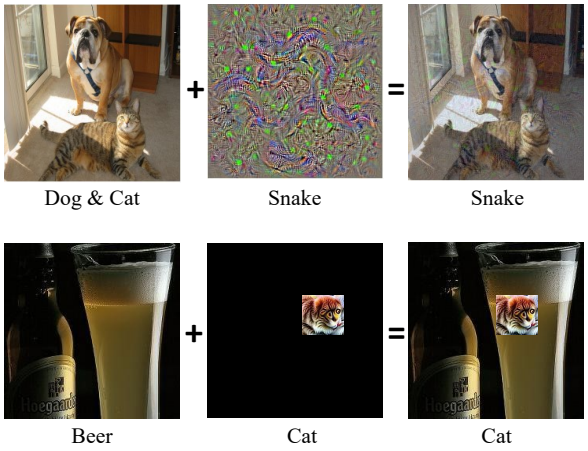


Figure 6: Examples of adversarial attacks after applying model-level interpretation. Upper: Targeted universal perturbation. Lower: Universal replacement attack.

- **Targeted-Universal-Perturbation Attack:** If we set h to be class relevant mapping such as f_c, and solve Eq. 17 to get the interpretation, then $\mathbf{x}'$ can be directly added to target input instance as targeted perturbation attack. That is, given a clean input $\mathbf{x}_0$, the adversarial sample $\mathbf{x}^*$ is crafted simply as $\mathbf{x}^* = \mathbf{x}_0 + \lambda \cdot \mathbf{x}'$ to make $f(\mathbf{x}^*) = c$. It belongs to universal attack, because the interpretation process in Eq.17 does not utilize any information of the clean input. An example is shown in the upper row of Figure 6. A clean image is classified as "dog" (or "cat") by the model. Meanwhile, an image is generated by solving Eq.17, by setting h as f_c where c denotes "snake". By adding the generated image to the clean image, the resultant image is recognized as "snake", although it still looks like a dog and a cat in human eyes.
- **Untargeted-Universal-Perturbation Attack:** If we set h to be the aggregation of a number of middle-level layer mappings, such as $h(\mathbf{x}') = \sum_l \log(h^l(\mathbf{x}'))$ where h^l denotes the feature map tensor at layer l, the resultant $\mathbf{x}'$ is expected to produce spurious activation to confuse the

prediction of CNN models given any input, which implies $f(\mathbf{x}_0 + \lambda \cdot \mathbf{x}') \neq f(\mathbf{x}_0)$ with high probability [13]. This can be seen as an untargeted and universal attack.

- **Universal-Replacement Attack:** Adversarial patches, which completely replace part of the input, represent a visually different attack from perturbation attack. Based on Eq.17, more parameters such as masks, shape, location and rotation could be considered in the optimization to control $\mathbf{x}'$ [71]. The patch is obtained as $\mathbf{x}' \odot \mathbf{m}$, and the adversarial sample $\mathbf{x}^* = \mathbf{x}_0 \odot (\mathbf{1} - \mathbf{m}) + \mathbf{x}' \odot \mathbf{m}$, where $\mathbf{m}$ is a binary mask that defines patch shape. Besides, recent research shows that, if we define h as the objective score function in person detectors [14] or as the logit corresponding to human class [72], by solving Eq.18, it is possible to produce real-world patches attachable to human bodies to avoid them being detected by surveillance camera. An example of adversarial patches is shown in the bottom row of Figure 6. A clean image is classified as "beer". Meanwhile, a small image patch is generated, which is recognized as a "cat". By attaching the generated patch to the clean image, the prediction on the new image will be affected by the patch.

4.2 Representation Interpretation for Initiating Adversarial Attacks

Representation learning plays a crucial role in recent advances of machine learning, with applications in vision [73], natural language processing [74] and network analysis [75]. However, the opacity of representation space also becomes the bottleneck for understanding complex models. A commonly used strategy toward understanding representation is to define a set of explainable bases, and then decompose representation points according to the bases. Formally, let $\mathbf{z}_i \in \mathbb{R}^D$ denote a representation vector, and $\{\mathbf{b}_k \in \mathbb{R}^D\}_{k=1}^K$ denote the basis set, where D denotes the representation dimension and K is the number of base vectors. Then, through decomposition

$$\mathbf{z}_i = \sum_{k=1}^{K} p_{i,k} \cdot \mathbf{b}_k, \tag{19}$$

we can explain the meaning of $\mathbf{z}_i$ through referencing base vectors whose semantics are known, where $p_{i,k}$ measures the affiliation degree between instance $\mathbf{z}_i$ and $\mathbf{b}_k$. The work of providing representation interpretation following this scheme can be divided into several groups:

- **Dimension-wise Interpretation**: A straightforward way to achieve interpretability is to require each dimension to have a concrete meaning [76; 77], so that the basis can be seen as non-overlapping one-hot vectors. A natural extension to it would be to allow several dimensions (i.e., a segment) to jointly encode one meaning [78; 79].
- **Concept-wise Interpretation**: A set of high-level and intuitive concepts could first be defined, so that each $\mathbf{b}_k$ encodes one concept. Some examples include visual concepts [67; 66; 80], antonym words [81], and network communities [68].
- **Example-wise Interpretation**: Each base vector can be designed to match one data instance [82; 83] or part of the instance [84]. Those instances are also called prototypes. For example, a prototype could be an image region [84] or a node in networks [83].

The extra knowledge obtained from representation interpretation could be used to guide the direction of adversarial perturbation. However, the motivation of this type of work usually is to initiate more meaningful adversaries and then use adversarial training to improve model generalization, but not for the pure purpose of undermining model performance. For example, in text mining, [50] restricts perturbation direction of each word embedding to be a linear combination of vocabulary word embeddings, which improves model performance in text classification after adversarial training. In network embedding, [85] restricts perturbation of a node's embedding towards the embeddings of the node's neighbors in the network, so that adversarial training improves node classification and link prediction performances.

4.3 Model-Level Interpretation for Adversarial Defenses

Model-level interpretation develops an internal understanding of a model, including its weakness. Defenders could either choose to improve model robustness or develop a detector using internal data representation.

4.3.1 Model Robustification With Model-Level Interpretation

Some high-level features learned by deep models are not robust, which are insufficient to train reliable models. A novel algorithm is proposed in [86] to build datasets of robust features. Given a robust model f_r, $\mathcal{H}_r$ denotes the set of activations of neurons in the robust model, and $h : X \to \mathbb{R}$, $h \in \mathcal{H}_r$ is a transformation function that maps input to a neuron activation. Each instance in the robust dataset $\mathcal{D}_r$ is constructed from the original dataset $\mathcal{D}$. The instances in the robust dataset are expected to satisfy:

$$\mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}_r}[h(\mathbf{x}) \cdot y] = \begin{cases} \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}}[h(\mathbf{x}) \cdot y], & \text{if } h \in \mathcal{H}_r \\ 0, & \text{otherwise} \end{cases}. \tag{20}$$

In this way, input information that corresponds to non-robust representations is suppressed. Instances in the robust dataset are expected to contain only the features that are relevant to the robust model.

Despite not being directly incorporated in model training, inspections of model-level interpretation, especially on latent representation, have motivated several defense methods. Through visualizing feature maps of latent representation layers, the noise led by adversarial perturbation can be easily observed [25; 23; 58]. With this observation, [25] proposes adding denoising blocks between intermediate layers of deep models, where the core function of the denoising blocks are chosen as low-pass filters. [23] observes that adversarial perturbation is magnified through feedforward propagation in deep models, and proposed a U-net model structure as denoiser. Furthermore, through neuron pattern visualization, [87] finds that the convolutional kernels of CNNs after adversarial training tend to show a more smooth pattern. Based on this observation, they propose to average each kernel weight with its neighbors in a CNN model, in order to improve the adversarial robustness.

4.3.2 Adversarial Detection With Model-Level Interpretation

Instead of training another large model as a detector using raw data, we can also leverage model-level interpretation to detect adversarial instances more efficiently. In this

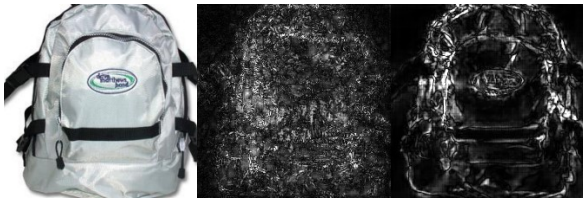


Figure 7: Explanations obtained from adversarially trained models focus less on textures and more on shape information [62]. Left: Input image. Middle: Gradient based explanation of a model without adversarial training. Right: Gradient based explanation of a model after adversarial training.

case, model-level interpretation plays a similar role as feature engineering, which helps distinguish between normal and adversarial instances. By regarding neurons as high-level features, readily available interpretation methods such as SHAP [88] could be applied for feature engineering to build adversarial detector [56]. After inspecting the role of neurons in prediction, a number of critical neurons could be selected. A steered model could be obtained by strengthening those critical neurons, while adversarial instances are detected if they are predicted very differently by the original model and steered model [29]. Nevertheless, the majority of work on adversarial detection utilizes latent representation of instances without inspecting their meanings, such as directly applying statistical methods on representations to build detectors [21; 89] or conducting additional coding steps on activations of neurons [28].

5. ADDITIONAL RELATIONS BETWEEN ADVERSARY AND INTERPRETATION

In the previous context, we have discussed how interpretation could be leveraged in adversarial attack and defense. In this section, we complement this viewpoint by analyzing the role of adversarial aspects of models in defining and evaluating interpretation. In addition, we specify the distinction between the two domains.

5.1 Improving Interpretation via Building Robust Models

In previous content, we have discussed the role of interpretation in studying model robustness. From another perspective, it has been found that, improving model robustness could also improve the understandability of explanations. First, the representations learned by robust models tend to align better with salient data characteristics and human perception [90]. Therefore, adversarially robust image classifiers are also useful in more sophisticated tasks such as generation, super-resolution, and translation [91], even without relying on GAN frameworks. Also, when attacking a robust classifier, the resultant adversarial samples are more likely to be recognized by humans [90]. In addition, retraining with adversarial samples [62], or regularizing gradients to improve model robustness [92], has been discovered to reduce noises from gradient-based sensitivity maps, and encourage CNN models to focus more on object shapes in making predictions. An example is shown in Figure 7. Finally, [93] presents the principle of "feature purification". The work discovers that dense mixtures of patterns exist in the weights of models trained with clean data using normal gradient descent. The dense pattern mixtures still generalize well when being used to predict normal data, but they are extremely sensitive to small perturbations in the input. After adversarial training, dense pattern mixtures could be removed, and the activation patterns of neurons will be easier to understand.

5.2 Defining Interpretation Approaches via Adversarial Perturbation

Some definitions of interpretation are inspired by adversarial perturbation. For feature-level interpretation, to understand the importance of a certain feature x, we try to answer a hypothetical question that "What would happen to the prediction Y, if x is removed or distorted?". This is closely related to *causal inference* [94; 95], and samples crafted in this way are also called *counterfactual explanations* [96]. For example, to understand how different words in sentences contribute to downstream NLP tasks, we can erase the target words from input, so that the variation in output indicates whether the erased information is important for prediction [97]. In image processing, salient regions could be defined as the input parts that most affect the output value when perturbed [57]. Considering that using traditional iterative algorithms to generate masks is time-consuming, Goyal et al. [98] develops trainable masking models that generate masks in real time.

Besides defining feature-level interpretation, the similar strategy can be used to define model component interpretation. Essentially we need to answer the question that "How the model output will change if we change the component in the model?". The general idea is to treat the structure of a deep model as a causal model [99], or extract human understandable concepts to build a causal model [100], and then estimate the effect of each component via causal reasoning. The importance of a component is measured by computing output changes after the component is removed.

As a natural extension from the discussion above, adversarial perturbation can also be used to evaluate the interpretation result. For example, after obtaining the important features, and understanding whether they are positively or negatively related to the output, we could remove or distort these features to observe the target model's performance change [5; 45]. If the target model's performance significantly drops, then we are likely to have the correct interpretation. However, it is worth noting that the evaluation will not be fair if the metric and interpretation methods do not match [101].

5.3 Uniqueness of Model Explainability

Despite the common techniques applied for acquiring interpretation and exploring adversary characteristics, some aspects of the two directions put radically different requirements. For example, some applications require interpretation to be easily understood by human especially by AI novices, such as providing more user-friendly interfaces to visualize and present interpretation [102; 103; 104], while adversarial attack requires perturbation to be imperceptible to human. Some work tries to adapt interpretation to fit human cognition habits, such as providing example-based interpretation [105], criticism mechanism [106] and counterfactual explanation [107]. The emphasis of understandability in interpretability, by its nature, is exactly opposite to

the main objective in adversarial attack, which is to add perturbation that is too subtle to be perceived by human.

6. CHALLENGES AND FUTURE WORK

We briefly introduce the challenges in leveraging interpretation to analyze the adversarial robustness of models. Meanwhile, we discuss the future research directions.

6.1 Models With Better Interpretability

Although interpretation could provide important directions against adversaries, interpretation techniques with better stability and faithfulness are needed before it could really be widely used as a reliable tool. As one of the challenges, it has been shown that many existing interpretation methods are vulnerable to manipulations [63; 64; 108]. A stable interpretation method, given an input instance and a target model, should produce relatively consistent results under the situation that the input may be subject to certain noises. As a preliminary work, [109] analyzes the phenomenon from a geometric perspective of decision boundary and proposed a smoothed activation function to replace ReLU. [110] proposes a sparsified variant of SmoothGrad [38] to produce saliency maps that is certifiably robust to adversaries.

Besides post-hoc interpretation, another challenge we are facing is how to develop models that are intrinsically interpretable [36]. With intrinsic interpretability, it may be more straightforward to identify and modify the undesirable aspects of model. Some preliminary work starts to explore applying graph-based models, such as proposing relational inductive biases to facilitate learning about entities and their relations [111], towards a foundation of an interpretable and flexible scheme of reasoning. Novel neural architectures have also been proposed, such as capsule networks [112] and causal models [113].

6.2 Adversarial Attacks in Real Scenarios

The most common scenario in existing work considers adversarial noises or patches in image classification or object detection. However, these types of perturbation may not represent the actual threats in the physical world. To solve the challenge, more realistic adversarial scenarios need to be studied in different applications. Some preliminary work include verification code generation[2], semantically or syntactically equivalent adversarial text generation [4; 114], and adversarial attack on graph data [6; 115]. Meanwhile, model developers need to be alerted to new types of attacks that utilize interpretation as the back door. For example, it is possible to build models that predict correctly on normal data, but make mistakes on input with certain secret attacker-chosen property [116]. Also, recently researchers found that it is possible to break data privacy by reconstructing private data merely from gradients communicated between machines [117].

6.3 Model Improvement Using Adversaries

The value of adversarial samples goes beyond simply serving as prewarning of model vulnerability. It is possible that the vulnerability to adversarial samples reflects some deeper generalization issues of deep models [118; 119]. Some preliminary work has been conducted to understand the difference between a robust model and a non-robust one. For example, it has been shown that adversarially trained models possess better interpretability [62] and representations with higher quality [91]. [120] also tries to connect adversarial robustness with model credibility, where credibility measures the degree that a model's reasoning conforms with human common sense. Another challenging problem is how to properly use adversarial samples to benefit model performance, since many existing works report that training with adversarial samples will lead to performance degradation, especially on large data [3; 25]. Recently, [121] shows that, by separately considering the distributions of normal data and adversarial data with batch normalization, adversarial training can be used to improve model accuracy.

[2]https://github.com/littleredhat1997/captcha-adversarial-attack

7. CONCLUSION

In this paper, we review the recent work of adversarial attacks and defenses by combining them with the recent advances of interpretable machine learning. Specifically, we categorize interpretation techniques into feature-level interpretation and model-level interpretation. Within each category, we investigate how the interpretation could be used for initiating adversarial attacks or designing defense approaches. After that, we briefly discuss other relations between interpretation and adversarial perturbation/robustness. Finally, we discuss the current challenges of developing transparent and robust models, as well as some potential directions to further study and utilize adversarial samples.

8. REFERENCES

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[3] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. 2017.

[4] Qi Lei, Lingfei Wu, Pin-Yu Chen, Alexandros G Dimakis, Inderjit S Dhillon, and Michael Witbrock. Discrete adversarial attacks and submodular optimization with applications to text classification. *Systems and Machine Learning (SysML)*, 2019.

[5] Ninghao Liu, Hongxia Yang, and Xia Hu. Adversarial detection with model interpretation. In *KDD*, 2018.

[6] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *KDD*, 2018.

[7] Jian Kang and Hanghang Tong. N2n: Network derivative mining. In *CIKM*, 2019.

[8] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018.

[9] Mary Frances Zeager, Aksheetha Sridhar, Nathan Fogal, Stephen Adams, Donald E Brown, and Peter A Beling. Adversarial learning in credit card fraud detection. In *2017 Systems and Information Engineering Design Symposium (SIEDS)*, 2017.

[10] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017.

[11] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[12] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017.

[13] Konda Reddy Mopuri, Aditya Ganeshan, and Venkatesh Babu Radhakrishnan. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*.

[14] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *CVPR Workshops*, 2019.

[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[16] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[17] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 2006.

[18] Frank C Keil. Explanation and understanding. *Annu. Rev. Psychol.*, pages 227–254, 2006.

[19] Carl G Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of science*, 1948.

[20] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 2018.

[21] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *ICLR*, 2017.

[22] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

[23] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *CVPR*, 2018.

[24] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[25] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *CVPR*, 2019.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[27] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE.

[28] Jiajun Lu, Theerasit Issaranon, and David Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *ICCV*, 2017.

[29] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In *NIPS*, 2018.

[30] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960*, 2017.

[31] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017.

[32] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

[33] Mengnan Du, Ninghao Liu, Fan Yang, Shuiwang Ji, and Xia Hu. On attribution of recurrent neural network predictions via additive decomposition. In *The World Wide Web Conference*, 2019.

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[35] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[36] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

[37] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

[38] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[39] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[40] Xiaoyu Cao and Neil Zhenqiang Gong. Mitigating evasion attacks to deep neural networks via region-based classification. In *ACSAC*, 2017.

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

[42] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Distilling knowledge from deep networks with applications to healthcare domain. *arXiv preprint arXiv:1512.03542*, 2015.

[43] Jun Gao, Ninghao Liu, Mark Lawley, and Xia Hu. An interpretable classification framework for information extraction from online healthcare forums. *Journal of healthcare engineering*, 2017.

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*, 2016.

[45] Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. Lemna: Explaining deep learning based security applications. In *CCS*, 2018.

[46] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 2013.

[47] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. 2018.

[48] Adith Boloor, Xin He, Christopher Gill, Yevgeniy Vorobeychik, and Xuan Zhang. Simple physical adversarial examples against end-to-end autonomous driving models. In *ICESS*. IEEE, 2019.

[49] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NIPS*, 2018.

[50] Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. Interpretable adversarial perturbation in input embedding space for text. In *IJCAI*, 2018.

[51] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, 2017.

[52] Yaojing Wang, Yuan Yao, Hanghang Tong, Feng Xu, and Jian Lu. Discerning edge influence for network embedding. In *CIKM*, 2019.

[53] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.

[54] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 2018.

[55] Baifeng Shi, Dinghuai Zhang, Qi Dai, Zhanxing Zhu, Yadong Mu, and Jingdong Wang. Informative dropout for robust representation learning: A shape-bias perspective. 2020.

[56] Gil Fidel, Ron Bitton, and Asaf Shabtai. When explainability meets adversarial learning: Detecting adversarial examples using shap signatures. *arXiv preprint arXiv:1909.03418*, 2019.

[57] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017.

[58] Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I Jordan. Ml-loo: Detecting adversarial examples with feature attribution. *arXiv preprint arXiv:1906.03499*, 2019.

[59] Chiliang Zhang, Zuochang Ye, Yan Wang, and Zhimou Yang. Detecting adversarial perturbations with saliency. In *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*, pages 271–275. IEEE, 2018.

[60] Jingyuan Wang, Yufan Wu, Mingxuan Li, Xin Lin, Junjie Wu, and Chao Li. Interpretability is a kind of safety: An interpreter-based ensemble for adversary defense. In *KDD*, 2020.

[61] Yinpeng Dong, Hang Su, Jun Zhu, and Fan Bao. Towards interpretable deep neural networks by leveraging adversarial examples. *arXiv preprint arXiv:1708.05493*, 2017.

[62] Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *International Conference on Machine Learning*, pages 7502–7511, 2019.

[63] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *AAAI*, 2019.

[64] Akshayvarun Subramanya, Vipin Pillai, and Hamed Pirsiavash. Fooling network interpretation in image classification. In *ICCV*, 2019.

[65] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *CVPR*, 2019.

[66] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2018.

[67] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *ECCV*, 2018.

[68] Ninghao Liu, Xiao Huang, Jundong Li, and Xia Hu. On interpretation of network embedding via taxonomy induction. In *KDD*, 2018.

[69] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, page 1.

[70] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

[71] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[72] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *CCS*, 2016.

[73] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

[74] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational intelligenCe magazine*, 2018.

[75] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[76] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

[77] Abhishek Panigrahi, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya. Word2sense: Sparse interpretable word embeddings. In *ACL*, 2019.

[78] Ninghao Liu, Qiaoyu Tan, Yuening Li, Hongxia Yang, Jingren Zhou, and Xia Hu. Is a single vector enough? exploring node polysemy for network embedding. In *KDD*, 2019.

[79] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. Learning disentangled representations for recommendation. In *Advances in Neural Information Processing Systems*, 2019.

[80] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *NeurIPS*, 2019.

[81] Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings. In *The World Wide Web Conference*, 2020.

[82] Been Kim, Cynthia Rudin, and Julie A Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *NIPS*, 2014.

[83] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[84] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *NeurIPS*, 2019.

[85] Quanyu Dai, Xiao Shen, Liang Zhang, Qiang Li, and Dan Wang. Adversarial training methods for network embedding. In *The World Wide Web Conference*, 2019.

[86] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

[87] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *CVPR*, 2020.

[88] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, 2017.

[89] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[90] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[91] Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *NeurIPS*, 2019.

[92] Ninghao Liu, Xiao Huang, Jundong Li, and Xia Hu. On interpretation of network embedding via taxonomy induction. In *KDD*, 2018.

[93] Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. *arXiv preprint arXiv:2005.10190*, 2020.

[94] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 2020.

[95] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 2020.

[96] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[97] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.

[98] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *NIPS*, 2017.

[99] Tanmayee Narendra, Anush Sankaran, Deepak Vijaykeerthy, and Senthil Mani. Explaining deep learning models using causal inference. *arXiv preprint arXiv:1811.04376*, 2018.

[100] Michael Harradon, Jeff Druce, and Brian Ruttenberg. Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*, 2018.

[101] Ninghao Liu, Yunsong Meng, Xia Hu, Tie Wang, and Bo Long. Are interpretations fairly evaluated? a definition driven pipeline for post-hoc interpretability. *arXiv preprint arXiv:2009.07494*, 2020.

[102] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009.

[103] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

[104] Fan Yang, Shiva K Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D Ragan, Shuiwang Ji, and Xia Ben Hu. Xfake: Explainable fake news detector with visualizations. In *WWW*, 2019.

[105] Isabelle Bichindaritz and Cindy Marling. Case-based reasoning in the health sciences: What's next? *Artificial intelligence in medicine*, 2006.

[106] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*, 2016.

[107] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *ICML*, 2019.

[108] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AAAI*, 2020.

[109] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *NeurIPS*, 2019.

[110] Alexander Levine, Sahil Singla, and Soheil Feizi. Certifiably robust interpretation in deep learning. *arXiv preprint arXiv:1905.12105*, 2019.

[111] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[112] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *NIPS*, 2017.

[113] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[114] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, 2018.

[115] Qinghai Zhou, Liangyue Li, Nan Cao, Lei Ying, and Hanghang Tong. Admiring: Adversarial multi-network mining. In *ICDM*, 2019.

[116] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

[117] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *arXiv preprint arXiv:1906.08935*, 2019.

[118] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, 2019.

[119] Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018.

[120] Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. Learning credible deep neural networks with rationale regularization. In *ICDM*, 2019.

[121] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan Yuille, and Quoc V Le. Adversarial examples improve image recognition. *arXiv preprint arXiv:1911.09665*, 2019.