Approximating Optimal Transport via GANs for Recourse Disparity Analysis

Zachary Novack
znovack@andrew.cmu.edu

Qi Xuan Teo
qteo@andrew.cmu.edu

Ryan Steed
rsteed@andrew.cmu.edu *

1 Introduction

1.1 Motivation

In algorithmic systems, *recourse*—the ability of someone affected by a model's decision to obtain their desired outcome—is essential for preserving procedural justice and safeguarding against unfairness [Ustun et al., 2019]. Recourse is especially important for the increasing number of high-stakes decisions left under bias-prone algorithmic control, including lending [Siddiqi, 2012], hiring [Ajunwa et al., 2016], and public service administration [Chouldechova et al., 2018, Schwartz et al., 2017].

However, little work examines whether recourse is fairly distributed across social groups. For example, if a White and Black person with similar credit histories are both rejected for a loan, will their opportunities for recourse be the same?

The study of recourse in algorithmic systems aligns closely with the study of explainable models. In black box systems, explanations help model subjects understand why a particular decision was reached, provide grounds to contest adverse decisions, and give guidance for what could be changed to receive a desired result in the future [Wachter et al., 2017]. In this project, we focus on *input* recourse, which involves altering the model subject's situation without altering the decision process or providing any non-algorithmic remedies. We assume the model is fixed and immutable, so only avenue for a model subject to altering a decision is through feature alteration. Of course, some features are practically or morally immutable (e.g., someone's ethnicity)—instead, we focus on an *actionable* subspace of features over which a model subject reasonably has control.

We attempt to empirically quantify disparities in recourse between protected groups in a stylized hiring classification setting [Lipton et al., 2018]. We imagine an machine learning model in production that has already generated decisions for a set of model subjects with specific features, and we would like to evaluate disparities in the recourse available to subjects of this model. Our approach comprises a two stage process relying on probabilistic graphical models: in the first stage, we use a generative adversarial network (GAN) to approximate a mapping between similar individuals in each group; in the second stage, we produce counterfactual explanations generated by Structural Causal Model (SCM)-based algorithms and compare them at the group and matched-individual levels. We do find evidence of disparities in recourse available to individuals, and compare fairness for three counterfactual explanation methods.

1.2 Datasets

1.2.1 Lipton Synthetic Dataset

We use the synthetic dataset created by Lipton et al. [2018]. This is a dataset consisting of 2 features, hair length and work experience, labelled by whether or not a candidate with those attributes was hired. The data were generated probabilistically according to the following assumptions: (1) historical hiring processes are based solely on work experience, (2) women have on average fewer years of work experience than men, and (3) women have longer hair than men. There are 2000 observations in total.

*Audit only - limited contribution.

1.2.2 Chicago SSL Dataset

We also use the city of Chicago's Strategic Subject List (SSL) dataset Chicago [2017]. The dataset comprises of arrest data from August 1 2012 to July 31 2016, and was used to create a risk assessment score (the SSL). SSL scores measured the likelihood of an individual being involved in a shooting accident either as a victim or an offender, and range from 0 (low risk) to 500 (high risk).

SSL is calculated using 8 predictors, but the dataset includes 2 others (Sex and Race). The total dataset comprises 398582 observations of these 11 variables, which we list below:

Table 1: SSL Dataset Features

Name	Type	Range	Used in SSL
SSL	Numeric	0-500	NA
Age at Latest Arrest	Categorical Ordered	8 Categories	Yes
# Times Victim of Shooting Incident	Numeric	0 - 4	Yes
# Times Victim of Battery/Assault	Numeric	0 - 10	Yes
# Times Arrests for Violent Offenses	Numeric	0 - 12	Yes
Gang Affiliation	Categorical	0, 1	Yes
# Times Narcotic Arrests	Numeric	0 - 29	Yes
Trend in Criminal Activity	Numeric	-10 - 10	Yes
# Times Unlawful Use of Weapon Arrests	Numeric	0 - 4	Yes
Sex	Categorical	3 Categories	No
Race	Categorical	7 Categories	No

We use the cutoff detailed in [Black et al., 2020], wherein SSL values > 345 are classified as "at risk" and all others are "not at risk". Additionally, in order to provide a different qualitative example from the Lipton dataset, we use Race as our protected attribute, and group it the race category broadly into "white" and "non-white", and convert "Age at Latest Arrest" to numeric values, where the age is sample uniformly from the given range. As the GAN-approximation of the optimal transport mapping does not work well with categorical data, we use only the numeric columns in our analysis.

1.2.3 German Credit Dataset

Lastly, we use the German Credit Dataset from Dua and Graff [2017], a commonly used dataset containing German credit data. It comprises 1000 instances of credit from the years 1973-1975 from a large regional bank in South Germany, with 300 bad credits and 700 good credits. We note that bad credits are actually over-represented here, with real-world averages being around 5 percent. There are 20 attributes in total, 13 categorical and 7 quantitative. We summarize the dataset in table 2. As we did for the SSL dataset, we only focus on the numeric features of the dataset, as the categorical features are not readily usable by the GAN approximation. Here for the protected attribute we focus on the "Foreign worker" column, to get a sense of whether we can capture discriminatory behavior based on nationality.

Table 2: German Credit Dataset Features

Quantitative	Categorical
	Status of existing checking account
	Credit history
	Purpose
Duration (months)	Savings account/bonds
Credit amount	Present employment
Installment rate (percent)	Personal status and sex
Present residence since	Other debtors/guarantors
Age (years)	Property
Number of existing credits	Other installment plans
Number of people liable	Housing
	Job
	Telephone
	Foreign worker

2 Background

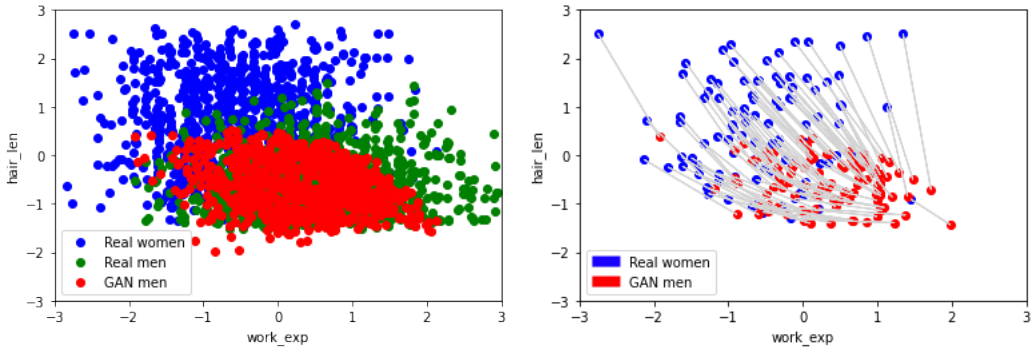

Figure 1: Left image: distribution of women and men in Lipton dataset with final trained GAN-generated men. Right image: optimal pairings between real women and GAN men

We report our initial results from the midway report in Figure 1, which depicts the results of training our Wasserstein GAN on the synthetic Lipton dataset, with the our approximation of the distribution of men show in red in the left plot. The optimal mappings between real women and GAN men (which is our estimation of the Optimal Transport mapping) is shown in the right plot. For our initial baseline result training for relatively few epochs, we see that our GAN is able to recover a reasonable estimation of the underlying distribution of men.

Our baseline approach for recourse from the midway report was to use FACE detailed in Section 4.2.1. Additionally, we defined the *Matchsets* $M_{\overline{S}}$ and $M_{\overline{S'}}$ to be the sets of points in S and S' where both they and their optimal matching received the negative outcome ($G(M_{\overline{S}}) = M_{\overline{S'}}$). Let S^- and S'^- be the subsets that just received the negative outcome (i.e. with no restriction on their match's behavior). Additionally, let $\widetilde{x}$ represent the counterfactual explanation provided for input point x. For a given point and attribute, we define $\delta(x_i) = |x_i - \widetilde{x}_i|$. Thus, the *Group-Level* Aggregate Recourse Disparity for a given attribute i (denoted $\Delta_G(i)$) is:

$$\Delta_G(i) = \frac{1}{|S^-|} \sum_{x \in S^-} \delta(x_i) - \frac{1}{|S'^-|} \sum_{x \in S'^-} \delta(x_i)$$

Table 3: Gender disparity in average recourse $\Delta_G(i)$ in the Lipton dataset

Feature	$\Delta_G(i)$
Work Experience	0.015
Hair Length	0.429

In words, $\Delta_G(i)$ is the disparity in average recourse between two groups, where a positive value indicates that more recourse is needed for S rather than S' (throughout the paper, we use S to denote the protected class and S' the majority class). Our results from the Lipton dataset are summarized in Table 3, where the average disparity is assessed between the women (S) and the men (S'). These results highlight that at the group level, the effort necessary for women to flip their outcome is higher than that of the men for both features. Clearly, there is group-level unfairness in the opportunities for recourse provided—in the remainder of our report, we will explore whether this unfairness is exacerbated or alleviated at the individual level, and explore changes in fairness when we use other recourse mechanisms on other datasets.

3 Related Work

Our approach is built primarily on work in transport-based fairness and counterfactual recourse. The first component of our approach is inspired by FlipTest [Black et al., 2020], which estimates optimal transport to match similar in-distribution individuals from different protected groups (e.g., a man and woman with similar credit histories) through a generative adversarial network (GAN) described further in

§ 4.1. The second component of our approach is a method to generate counterfactual explanations that lead to actionable recourse—essentially, what decision a model subject would receive if they altered some aspect of the model input. Several methods exist [Ustun et al., 2019, Poyiadzi et al., 2020, Mothilal et al., 2020, Mahajan et al., 2019, Ross et al., 2021, Williams et al., 2022]; we test the three methods detailed in section 4.2.

Notably, none of the aforementioned papers on counterfactual explanations for recourse mention the possibility of disparities in recourse across protected groups. In the only papers on fairness in recourse we could find, Gupta et al. [2019] and von Kügelgen et al. [2022] propose group and individual fairness criteria for recourse and study their theoretical properties. In empirical experiments, Gupta et al. [2019] optimize for group fairness in the model-agnostic setting (as the average distance to the decision boundary for a given protected group), but do not evaluate individual fairness. von Kügelgen et al. [2022] evaluate individual-level fairness with counterfactual "twins"—out-of-distribution points used for counterfactual testing.

Our approach improves on these studies by utilizing a fairness metric that is model-agnostic and assumes nothing about the underlying causal model, and by defining individual fairness in recourse to prioritize feasibility (rather than simply flipping the protected attribute of an individual to assess individual fairness). In other words, our approach could be used to evaluate a set of historical decisions for disparity in recourse opportunities entirely by comparing outcomes between real individuals.

4 Methods and Model

Our study proceeds as follows: first, we train a GAN to approximate a mapping between similar individuals in each group (§ 4.1); second, we produce counterfactual opportunities for recourse using one of three methods (§ 4.2.1); third, we compare those counterfactuals at the individual- and group-level to evalute fairness in recourse (§ 5). In § 2, we showed that some group-level disparities in recourse may exist for recourse methods; in our remaining report, we explore disparities at the individual level and across different counterfactual explanation methods applied to additional datasets.

4.1 GAN Model

Inspired by the findings in Black et al. [2020], we seek to find the optimal mapping between points in the protected class to the non-protected class. Specifically, for any distributions $\mathcal{S}, \mathcal{S}'$ over the same feature space $\mathcal{X}$, we want to find the optimal transport mapping that minimizes the cost $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ of moving between a point in $\mathcal{S}$ to a point in $\mathcal{S}'$. We estimate $\mathcal{S}, \mathcal{S}'$ through sets of individual samples S, S' respectively and $|S| = |S'| = N$.

As iterative solutions for finding the exact optimal transport mapping do not scale for large N, we choose instead to approximate this mapping via a Wasserstein GAN [Arjovsky et al., 2017]. Given a generator model G (which we hope will capture the optimal transport mapping in one direction) and discriminator model D, we have the following modified loss functions for each model:

$$L_G = \frac{1}{N} \sum_{x \in S} D(G(x)) + \frac{\lambda}{N} \sum_{x \in S} c(x, G(x)), \qquad L_D = \frac{1}{N} \sum_{x' \in S'} D(x') - \frac{1}{N} \sum_{x \in S} D(G(x))$$

We trained this modified Wasserstein GAN using RMSProp with $\lambda = 0.0001$, learning rate 0.00005, and batch-size 32. D and G are both 2-layer linear networks with 128 hidden units per layer and ReLU activation.

4.2 Recourse Mechanisms

In an effort to compare how different recourse mechanisms differ in terms of recourse disparity, we implement three algorithms: FACE [Poyiadzi et al., 2020], DiCE [Mothilal et al., 2020], and BAYES [Williams et al., 2022]. Note that for all algorithms, we learn the same simple logistic regression model (using LBFGS and L2 regularization with $\lambda = 1$) which generates the labels for our data (access to such a model is explicitly needed for DiCE and BAYES). We refer to these algorithms as "resource mechanisms," imagining that a model subject could use these mechanisms as a guide to deliberately change their features and achieve a desired outcomes.

4.2.1 FACE

For the purposes of our initial benchmarks, we use FACE (Feasible and Actionable Counterfactuals) [Poyiadzi et al., 2020], which generates a K-Nearest Neighbor ($k = 20$) graph of the dataset and proposes a counterfactual explanation as the closest point in the graph from our input that has the desired outcome. FACE places a particular emphasis on feasibility and actionability (e.g. not asking one to quadruple their salary, or allowing one to change their race). The aforementioned graph is modified based on these constraints by removing appropriate edges or by generating a new graph. FACE was selected as it represents a simple baseline approach that maximizes representativeness (i.e. counterfactuals must exist in the dataset).

4.2.2 DiCE

We also use DiCE (Mothilal et al. [2020])as another method for recourse. This method of recourse adds another dimension into counterfactual generation, positing that explanations should also be diverse. If all explanations only propose one class of methods to achieve an alternate outcome, then participants would have no choice but to attempt that method regardless of its feasibility or attractiveness, and participants would have no alternatives should the primary option fail.

DiCE captures diversity using Determinal Point Processes (DPP) Kulesza [2012]. Given some counterfactuals c_i, c_j and a kernel matrix $\mathbf{K}$:

$$\text{dpp\_diversity} = det(\mathbf{K}),$$

where $\mathbf{K}_{ij} = \frac{1}{1+dist(c_i,c_j)}$. DiCE optimizes a combined loss function over all the generated counterfactuals

$$\underset{c_1,\ldots,c_k}{\arg\min} \frac{1}{k}\sum_{i=1}^{k} L(f(c_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} dist(c_i, x) - \lambda_2 \text{dpp\_diversity}(c_1, \ldots, c_k),$$

where k is the number of counterfactuals to be generated, $f(\cdot)$ is the black box being evaluated, and x is the original input. For all datasets, we use $\lambda_1 = 2$, $\lambda_2 = 1$

4.3 BAYES

(a) Our PGM (b) Standard PGM

Figure 2: Left: Proposed PGM for BAYES. Right: Standard PGM for counterfactual generation.

We denote the last method BAYES, proposed in Williams et al. [2022]. This method focuses more on the *prior* of the data distribution, which helps to address infeasible or inactionable outcomes. Standard methods tend to generate counterfactuals x' around the reference point x, treating the prior as $p(x'|x)$. Instead, the BAYES represents the connection between x and x' with an undirected edge. This allows us to reframe the problem as sampling from the posterior, or counterfactual, distribution $p(x'|x, y, y')$, with:

$$p(x, x') = \mathcal{N}\left(\begin{bmatrix}\mu\\ \mu\end{bmatrix}, \begin{bmatrix}\Lambda^{-1} & W\\ W^T & \Lambda^{-1}\end{bmatrix}\right)$$

$$p(y'|x') = \mathcal{N}(Ax' + b, L^{-1})$$

$$p(x'|x, y, y') \propto p(x|y', x')p(y'|x')p(x') = p(x|x')p(y'|x')p(x')$$

In this model, we can leverage a pre-constructed Structural Causal Model (SCM) of a given dataset—or, if there is no obvious causal model, assume the covariates are independent. For our datasets, we are in the latter case (i.e. we have no underlying causal DAG for any of the datasets).

5 Results

In § 2, we looked at group-level average recourse to examine group-level disparities. In this section, we utilize the GAN-based optimal transport to assess recourse disparity *at the individual level* using matched pairs of protected (women, people of color, or foreign workers) and non-protected individuals. Here, instead of using a point estimate (as we did in the group-Level case), what we care about is the *distribution* of recourse disparities between optimally paired individuals,

$$\Delta_I(i) = |x_i - \tilde{x}_i| - |G(x)_i - G(\tilde{x})_i|$$

$\Delta_I(i)$ describes the recourse disparity for similar individuals in different groups wishing to change attribute x. $\Delta_I(i)$ is positive when the protected group has to shift x a greater distance to change their outcome than the non-protected group.

Figure 3 plots the distribution of $\Delta_I(i)$ in the Lipton dataset. No recourse mechanism consistently prefers one group or another (as shown by the relative symmetry of the distributions at 0), though the spread and skew of these distributions depends heavily on the recourse mechanism. Namely, DiCE and FACE show disparity much more clustered around low values (in the worst case, a woman would need an extra year of work experience to change her outcome than a man), while BAYES has a much wider and spread out distribution (in the worst case, a woman would need nearly 2 years extra experience).

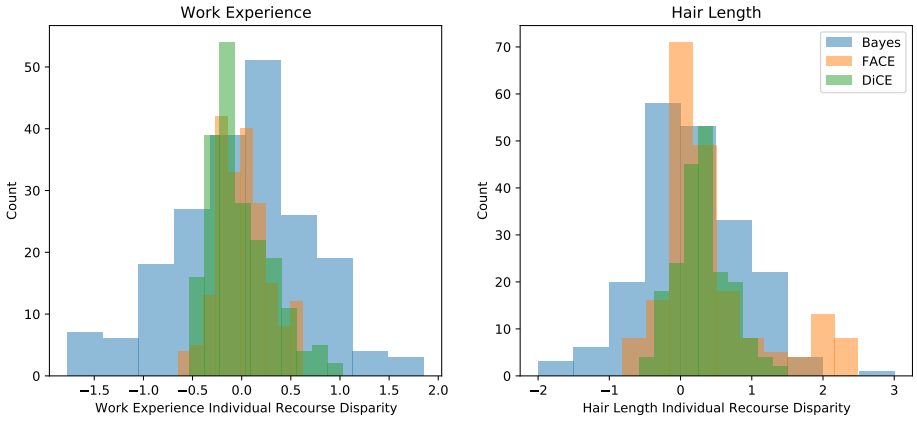

Figure 3: Distributions of Recourse Disparity between OT-matched pairs for Lipton dataset. Points less than zero are points where the non-protected group has to shift the feature by more to achieve a different outcome; points greater than zero indicate situations where recourse is more difficult for the protected group.

These trends, however, are somewhat dataset dependent. Figure 4 shows this, as within the SSL dataset (top plots) BAYES consistently has the tightest disparity distribution while both DiCE and considerably FACE show extreme disparity at the expense of the protected group (i.e. a right skew). In the German Credit dataset (bottom plots), we see that on DiCE consistently is centered about 0, though even then all models generally describe that though the average disparity at the group level may be moderately right-skewed, the distributions show clear non-Gaussianity that describe a large faction of users who receive unfair recourse.

One possible reason that FACE may show worse disparities is that in complex and relatively small datasets, using specifically in-distribution samples for recourse yields values that are wildly far from the original point. It is also unsurprising that DiCE generally performs well (in terms of minimizing recourse disparity), as DiCE does not prioritize the "realness" of the proposed counterfactual at all, and thus is not subject to the issues FACE finds (though it may suffer in terms of how well people prefer the counterfactuals generated by DiCE).

These results highlight that issues in recourse may vary drastically by the dataset and the individual feature. Still, as in § 2, the distributions of disparity have a clear right-ward skew, suggesting that the *protected* group (whether women, non-whites, or foreign workers) are generally disadvantaged by these decision-making models and recourse mechanisms. This skew is most notable for the criminal activities dataset, where DiCE reports that a person of color may need up to 15 fewer violent offenses to change their outcome than a White person.

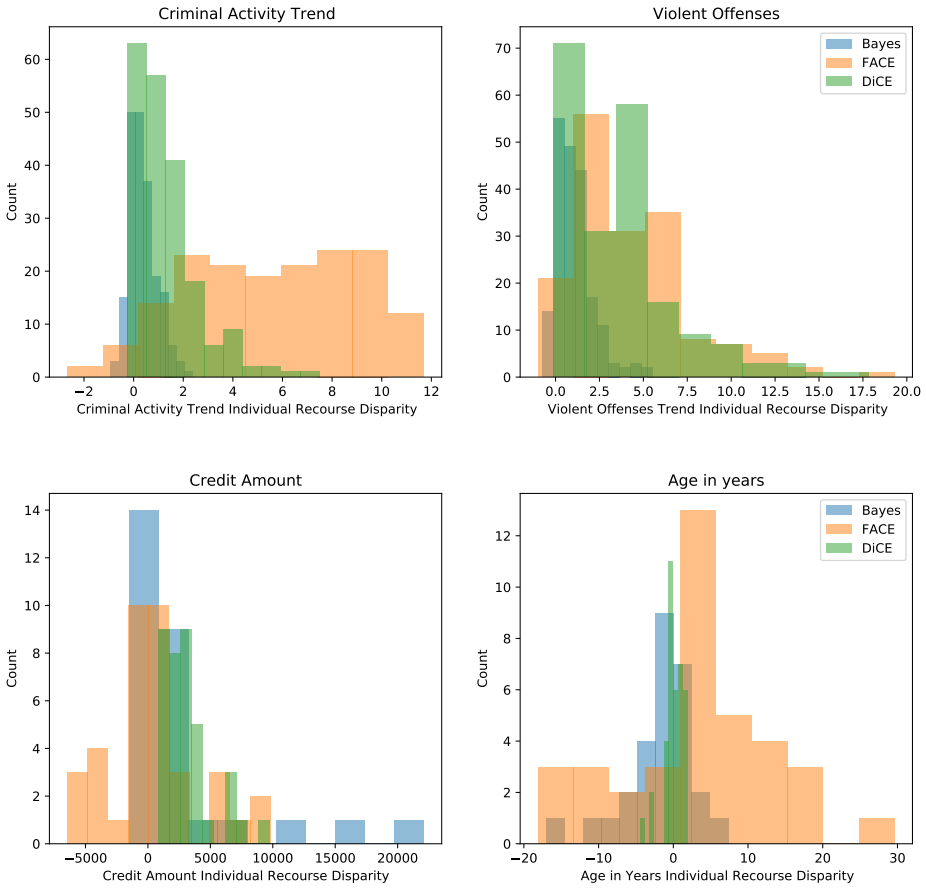

Figure 4: Top: Distributions of Recourse Disparity between OT-matched pairs for 2 features of SSL dataset. Bottom: German Credit dataset. Points less than zero are points where the non-protected group has to shift the feature by more to achieve a different outcome; points greater than zero indicate situations where recourse is more difficult for the protected group.

6 Discussion

This study presents an interesting new problem for machine learning and for counterfactual explanations. Our work shows that in addition to experiencing unfairness in the outcomes of a machine learning model, marginalized groups may find it more difficult to find recourse after an unwanted decision—in the cases we study, the marginalized group generally has to work harder to reverse an outcome than non-marginalized group. The extent to which we observe disparity in recourse depends on the choice of recourse mechanism. Some mechanisms show greater disparities than others; for instance, we speculate that FACE tends to exaggerate disparities because it relies on in-distribution points. In a practical setting where these mechanisms are used to guide individuals, these mechanism-induced disparities could mislead marginalized groups into thinking their situation is hopeless or excessively difficult.

Even for the most conservative recourse mechanism, disparities in available recourse usually still exist. In addition to fairness, accuracy, and explainability, developers should certainly consider fairness in recourse

to ensure that all subject to a model's decisions have a fair opportunity to control important decisions made about them.

6.1 Limitations

There are many clear limitations to our work, namely with regards to the optimal transport approximation. GAN training can be considerably unstable, and in the case that the distributions are not meaningfully separable (i.e. if features are similar across protected groups) any disparity captured on account of this GAN may only be fitting noise in the feature distribution. Additionally, we only looked at *numeric* features for the entire downstream task given limitations for conventional GANs to work with mixed data-types (and especially discrete data), and thus, our modeling results are somewhat only representative of cases when input data has no categorical features. This being said, further work could very well be done to modify our GAN loss function in order to account for these mixed-type distributions. On the recourse side, all of our validated recourse mechanisms may possibly be intensely sensitive to their initial hyperparameters, and thus different, more optimal behavior for a given recourse mechanism may be possible. In the future, we hope to conduct a more thorough hyperparameter search across recourse mechanisms and model specifications (in terms of the decision-making model).

6.2 Future Work

Given the somewhat mixed results we saw between how different recourse mechanisms handle individual fairness, additional hyperparameter tuning and model specification may be done in order to look for any possible data-agnostic trend between these recourse mechanisms. Williams et al. [2022] notes a comparison between these methods in terms of *explainability* (and that BAYES is consistenly the most explainable), and thus, our work may help policy makers and domain experts consider the tradeoffs between user preference of explainability and downstream fairness concerns. Also, we initially tested BAYES because of its improvement over FACE and DiCE in terms of explainability in Williams et al. [2022], but we only used a very simple assumption of the causal structure; to leverage the full potential of this method, future work could explore domains with clearer causal priors. Additionally, future work could explore more bespoke recourse mechanisms based on generating optimal transport mappings between users in the same social group but with different outcomes, either to better assess recourse disparity or to modify the loss function for DiCE to directly minimize recourse disparity. Other questions remain: for what kinds of individuals does recourse disparity tend to be the worst? How much of our results are attributable to the model making recourse difficult, as compared to under-/over-exaggeration of difficulty by the recourse mechanism?

7 Teammates and Work Division

Our work division is slightly unique due to the composition of the group. Zachary was responsible for handling the code-base, while Qi Xuan was in charge of report writing. Ryan is only auditing, so he just helped with the problem motivation, final editing, and obtaining code/materials from colleagues. The group largely kept the same task allocation as the midway report, with the others assisting Zach in the code whenever possible, such as finding relevant reference materials like datasets and Github repositories.

8 Code

All our code is stored in the following Github repository: [link]. We have included a Readme file which contains instructions on how to run the code.

Acknowledgments and Disclosure of Funding

Special thanks to Josh Williams, who provided us with an advance copy of his paper in submission [Williams et al., 2022] and accompanying code, which we used to test their BAYES counterfactual explanation method.

References

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

Naeem Siddiqi. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, June 2012. ISBN 978-1-118-42916-7.

Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*, 2016.

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 134–148. PMLR, January 2018. URL https://proceedings.mlr.press/v81/chouldechova18a.html. ISSN: 2640-3498.

Ira M. Schwartz, Peter York, Eva Nowakowski-Sims, and Ana Ramos-Hernandez. Predictive and prescriptive analytics, machine learning and child welfare risk assessment: The Broward County experience. *Children and Youth Services Review*, 81:309–320, October 2017. ISSN 0190-7409. doi: 10.1016/j.childyouth.2017.08.020. URL https://www.sciencedirect.com/science/article/pii/S0190740917303523.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)*, 31(2):841–888, 2017. URL https://heinonline.org/HOL/P?h=hein.journals/hjlt31&i=859.

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? *Advances in neural information processing systems*, 31, 2018.

City of Chicago. Strategic subject list - historical: City of chicago: Data portal, May 2017. URL https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List-Historical/4aki-r3np.

Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan 2020. doi: 10.1145/3351095.3372845. URL http://dx.doi.org/10.1145/3351095.3372845.

D. Dua and C. Graff. Statlog (german credit data) data set, 2017. URL https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan 2020. doi: 10.1145/3351095.3372850. URL http://dx.doi.org/10.1145/3351095.3372850.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

Alexis Ross, Himabindu Lakkaraju, and Osbert Bastani. Learning models for actionable recourse. *Advances in Neural Information Processing Systems*, 34, 2021.

Joshua Williams, Anurag Katakkar, Hoda Heidari, and Zico J Kolter. Toward generating actionable counterfactual explanations via posterior inferene. 2022.

Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing Recourse across Groups. *arXiv:1909.03166 [cs, stat]*, September 2019. URL http://arxiv.org/abs/1909.03166. arXiv: 1909.03166.

Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the Fairness of Causal Algorithmic Recourse. *arXiv:2010.06529 [cs, stat]*, March 2022. URL http://arxiv.org/abs/2010.06529. arXiv: 2010.06529.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Alex Kulesza. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012. doi: 10.1561/2200000044. URL https://doi.org/10.1561%2F2200000044.