Designing Fair Decision-Making Systems

A dissertation submitted towards the degree
Doctor of Engineering
of the Faculty of Mathematics and Computer Science of
Saarland University

by
Junaid Ali

Saarbrücken
2024

Date of Colloquium:	March 25, 2025
Dean of Faculty:	Univ.-Prof. Dr. Sebastian Hack
Chair of the Committee:	Prof. Dr. Ingmar Weber, Ph.D.
Reporters	
First Reviewer:	Prof. Dr. Krishna P. Gummadi, Ph.D.
Second Reviewer:	Prof. Dr. Adish Singla, Ph.D.
Third Reviewer:	Prof. Dr. Isabel Valera, Ph.D.
Academic Assistant:	Dr. Soumi Das

Abstract

The impact of algorithmic decision-making systems on individuals has raised significant interest in addressing fairness concerns within such systems. Designing fair systems entails several critical components, which have garnered considerable attention from the research community. However, notable gaps persist in three key components. Specifically, in this thesis, we address gaps in following components: i) evaluating existing approaches and systems for (un)fairness, ii) updating deployed algorithmic systems fairly, and iii) designing new decision-making systems from scratch. Firstly, we evaluate fairness concerns within foundation models. The primary challenge is that fairness definitions are task-specific while foundation models can be used for diverse tasks. To address this problem, we introduce a broad taxonomy to evaluate the fairness of popular foundation models and their popular bias mitigation approaches. Secondly, we tackle the issue of fairly updating already deployed algorithmic decision-making systems. To this end, we propose a novel notion of update-fairness and present measures and efficient mechanisms to incorporate this notion in binary classification. However, in cases where there is no deployed system or updating an existing system is prohibitively complex, we must design new fair decision-making systems from scratch. Lastly, we develop new fair decision-making systems for three key applications scenarios. Major challenges in designing these systems include computational complexity, lack of existing approaches to tackle fairness issues and designing human-subject based studies. We develop a computationally efficient mechanism for fair influence maximization to make the spread of information in social graphs fair. Additionally, we address fairness concerns under model uncertainty, i.e., uncertainty arising due to lack of data or the knowledge about the best model. We propose a novel approach for training nondiscriminatory systems that differentiate errors based on their uncertainty origin and provide efficient methods to identify and equalize errors occurring due to model uncertainty in binary classification. Furthermore, we investigate whether algorithmic decision-aids can mitigate inconsistency among human decision-makers through a large-scale study testing novel ways to provide machine advice.

Zusammenfassung

Der Einfluss algorithmischer Entscheidungssysteme auf das Leben von Menschen hat ein großes Interesse daran geweckt, dass solche Systeme fair sind. Die Entwicklung fairer Systeme umfasst mehrere kritische Komponenten, mit denen sich die Forschung intensiv beschäftigt hat. Bei drei Schlüsselkomponenten gibt es allerdings noch erhebliche Lücken: i) die Bewertung bestehender Ansätze und Systeme im Hinblick auf (Un-)Fairness, ii) faire Updates von Systemen im laufenden Betrieb und iii) der Entwurf neuer Entscheidungssysteme von Grund auf. Diese Arbeit befasst sich mit diesen drei Themen. Zum Ersten bewerten wir Fairnessbedenken in Basismodellen (Foundation Models). Die größte Herausforderung besteht dabei darin, dass Fairnessdefinitionen anwendungsspezifisch sind, während Basismodelle für unterschiedliche Anwendungen verwendet werden können. Um dieses Problem zu lösen, führen wir eine umfassende Taxonomie ein, um die Fairness gängiger Basismodelle und typischer Ansätze zur Vermeidung von Verzerrungen zu bewerten. Zum Zweiten befassen wir uns mit fairen Updates bereits laufender algorithmischer Entscheidungssysteme. Zu diesem Zweck entwickeln wir das Konzept der Update-Fairness, sowie Maßnahmen und effiziente Mechanismen, um das Konzept in der binären Klassifizierung zu nutzen. Zum Dritten, in Fällen in denen es noch kein System gibt oder die Aktualisierung eines bestehenden Systems zu komplex ist, müssen wir neue, faire Entscheidungssysteme von Grund auf entwickeln. Zu den größten Herausforderungen bei der Entwicklung fairer Systeme gehören dabei i) die Komplexität der Berechnungen, ii) der Mangel an bestehenden Ansätzen zur Lösung von Fairnessproblemen und iii) die Konzeption von Studien mit menschlichen Probanden. Deshalb entwickeln wir einen rechnerisch effizienten Mechanismus zur fairen Einflussmaximierung, um die Verbreitung von Informationen in sozialen Graphen fair zu gestalten. Darüber hinaus befassen wir uns mit Fairness bei Modellunsicherheiten, d.h. Unsicherheiten, die sich aus dem Mangel an Daten oder dem Wissen über das beste Modell ergeben. Dazu schlagen wir einen neuen Ansatz für das Training nicht- diskriminierender Systeme vor, der Fehler aufgrund der Art ihrer Unsicherheit unterscheidet, und entwickeln effiziente Methoden zur Identifizierung

und zum Ausgleich von Fehlern, die aufgrund von Modellunsicherheit in der binären Klassifikation auftreten. Darüber hinaus untersuchen wir, ob algorithmische Entscheidungshilfen die Inkonsistenz zwischen menschlichen Entscheidungsträgern reduzieren können, indem wir in einer groß angelegten Studie neuartige Wege maschinelle Unterstützung zu kommunizieren, testen.

Publications

Parts of this thesis have appeared in the following publications.

- "Evaluating the Fairness of Discriminative Foundation Models in Computer Vision".
  J. Ali, M. Kleindessner, F. Wenzel, V. Cevher, K. Budhathoki and Chris Russell.
  AAAI/ACM Conference on AI, Ethics, and Society (AIES) 2023.

- "(De)Noise: Moderating the Inconsistency Between Human Decision-Makers".
  J. Ali, N. Grgic-Hlaca, K. P. Gummadi, and J. W. Vaughan.
  NeurIPS workshop on Human-centered AI, 2022.

  Also accepted at:
  – N. Grgic-Hlaca, **J. Ali**, K. P. Gummadi, and J.W. Vaughan.
  ACM Conference On Computer-Supported Cooperative Work (CSCW), 2024.

- "Accounting for Model Uncertainty in Algorithmic Discrimination".
  J. Ali, P. Lahoti and K. P. Gummadi.
  AAAI/ACM Conference on AI, Ethics, and Society (AIES) 2021.

- "On the Fairness of Time-Critical Influence Maximization in Social Networks".
  J. Ali, M. Babaei, A. Chakraborty, B. Mirzasoleiman, K.P. Gummadi and A. Singla
  Transactions on Knowledge and Data Engineering (TKDE) 2021
  Also accepted at:
  – *NeurIPS workshop on Human-Centric Machine Learning workshop* 2019.
  – *International Conference on Data Engineering (ICDE)* 2022 as extended abstract.

- "Loss-Aversively Fair Classification".
  J. Ali, M.B. Zafar, A. Singla and K.P. Gummadi.
  AAAI/ACM Conference on AI, Ethics, and Society (AIES) 2019.

Additional publications while at Saarland University.

- *Unifying Model Explanability and Accuracy Through Reasoning Labels*.
  V. Nanda, **J. Ali**, K. P. Gummadi, M.B. Zafar.
  NeurIPS Workshop on Safety and Robustness in Decision Making (SRDM), 2019

Table of contents

List of figures xv

List of tables xxv

1 Introduction 1

1.1 Motivation 1

1.1.1 Evaluating unfairness of existing approaches/systems 1

1.1.2 Fairly updating an already deployed ADMS 2

1.1.3 Designing new fair ADMSs 3

1.2 Challenges 6

1.2.1 Evaluating unfairness of existing approaches/systems 6

1.2.2 Fairly updating an already deployed ADMS 6

1.2.3 Designing new fair ADMSs 7

1.3 Overview of thesis contributions 8

1.3.1 Evaluating the Fairness of Discriminative Foundation Models . . . 8

1.3.2 Fairly updating an already deployed ADMS: Loss-Aversively Fair Classification 8

1.3.3 Designing new fair ADMSs 9

1.4 Thesis outline 11

2 Background 13

2.1 Different notions of fairness in ADMSs 13

2.2 Approaches of achieving fairness 15

2.2.1 Pre-processing . . . 15
2.2.2 In-processing . . . 17
2.2.3 Post-processing . . . 18
3 Fairness Evaluation of Discriminative Foundation Models . . . 20
3.1 New Taxonomy for (Un)Fairness Evaluation of Discriminative FMs . . . 21
3.1.1 Measures based on our taxonomy . . . 24
3.2 Foundation Models, CLIP, and Fairness of CLIP . . . 28
3.2.1 Contrastive Language Image Pretraining (CLIP) . . . 30
3.2.2 Existing (Un)Fairness Evaluations of CLIP . . . 31
3.2.3 Bias Mitigation Methods for CLIP . . . 33
3.3 Expected Behaviour and Evaluation Criteria . . . 36
3.3.1 Binary Zero-Shot Classification . . . 36
3.3.2 Image Retrieval . . . 37
3.3.3 Image Captioning . . . 38
3.3.4 Performance Measures . . . 39
3.4 Metrics Based on Our Taxonomy . . . 40
3.4.1 Human-Centric (Un)Fairness Metrics . . . 40
3.4.2 Non-Human-Centric Labelings: Performance Metrics . . . 45
3.5 Evaluation . . . 46
3.5.1 Datasets . . . 47
3.5.2 Experimental Details . . . 50
3.5.3 Zero-shot Classification . . . 51
3.5.4 Image Retrieval . . . 52
3.5.5 Image Captioning . . . 54
3.5.6 OpenCLIP Results . . . 56
3.5.7 In-processing Fairness for CLIP-Like Models . . . 57
3.6 Related Work . . . 58

3.6.1 Text Embeddings and Bias . . . 58
3.6.2 Further (Fairness) Aspects of CLIP . . . 59
3.7 Conclusion . . . 59
4 Fairly updating an ADMS: Loss-Aversively Fair Classification . . . 61
4.1 Related work . . . 62
4.2 New notion of fairness: Loss-aversive updates . . . 62
4.2.1 Formalizing Notion of Loss-Averse Updates . . . 63
4.3 Updating Classifiers Loss-Aversively . . . 65
4.4 Evaluation on Synthetic Dataset: SP . . . 68
4.4.1 Dataset and Experimental Set up . . . 68
4.4.2 Loss-aversively Fair Updates . . . 69
4.5 Evaluation on Synthetic Dataset: EOP . . . 70
4.5.1 Dataset and Experimental Setup . . . 70
4.5.2 Loss-Aversively Fair Updates . . . 71
4.6 Evaluation on Real-World Dataset: SP . . . 72
4.6.1 Dataset and Experimental Setup . . . 72
4.6.2 Loss-Aversively Fair Updates . . . 73
4.7 Evaluation on Real-World Dataset: EOP . . . 74
4.7.1 Dataset and Experimental Setup . . . 74
4.7.2 Loss-Aversively Fair Updates . . . 74
4.8 Conclusion . . . 75
5 Designing a new fair ADMS: Time-Critical Influence Maximization . . . 77
5.1 Background on Time-Critical Influence Maximization (TCIM) . . . 78
5.1.1 Influence Propagation in Social Network . . . 78
5.1.2 Utility of Time-Critical Influence . . . 79
5.1.3 TCIM as Discrete Optimization Problem . . . 80
5.1.4 Submodularity and Approximate Solutions . . . 81

5.2 Measuring Unfairness in TCIM . . . 82
5.2.1 Socially Salient Groups and Their Utilities . . . 82
5.2.2 Disparity in Utility Across Groups . . . 82
5.2.3 Measure of Unfairness . . . 83
5.3 Achieving Fairness in TCIM . . . 83
5.3.1 Fair TCIM-Budget . . . 83
5.3.2 Fair TCIM-Cover . . . 87
5.4 Evaluation on Synthetic Datasets . . . 91
5.4.1 Dataset and Experimental Setup . . . 91
5.4.2 TCIM under Budget Constraints . . . 92
5.4.3 TCIM under Coverage Constraints . . . 95
5.5 Experiments on Real-World Datasets . . . 97
5.5.1 Dataset and Experimental Setup . . . 97
5.5.2 TCIM under Budget Constraint . . . 99
5.5.3 TCIM under Coverage Constraint . . . 100
5.6 Related Work . . . 102
5.7 Conclusion . . . 103
6 Designing a new fair ADMS: Model Uncertainty . . . 104
6.1 A proposal to differentiate between types of errors . . . 105
6.2 Preliminaries and Background . . . 107
6.2.1 Binary Classification . . . 107
6.2.2 Background on Predictive Multiplicity . . . 108
6.3 Proposed approach . . . 110
6.3.1 Scalable Methods for Predictive Multiplicity . . . 111
6.3.2 Leveraging Predictive Multiplicity towards Fairness under Model Uncertainty . . . 113
6.4 Experiments . . . 114

6.4.1 Datasets . . . 114
6.4.2 Experimental Setup . . . 116
6.4.3 Benchmarks and Metrics . . . 116
6.4.4 Synthetic Experiments . . . 119
6.4.5 Evaluation on Real-World Datasets . . . 120
6.5 Related Work . . . 121
6.6 Conclusion . . . 122
7 Designing a new fair ADMS: Human decision-makers . . . 123
7.1 Background . . . 124
7.2 Methodology . . . 128
7.2.1 Experimental Design . . . 128
7.2.2 Stimulus Material . . . 133
7.2.3 Data Collection . . . 136
7.2.4 Decision Aids . . . 138
7.3 Confirmatory Results . . . 142
7.3.1 H1: Overall Change in Decisions . . . 142
7.3.2 H1′: Propensity to Change Particular Decisions . . . 144
7.3.3 H2: Accuracy of Respondents' Decisions . . . 145
7.3.4 H3: Consistency Between Respondents' Decisions . . . 146
7.4 Exploratory Results . . . 147
7.4.1 Other Measures of Accuracy . . . 148
7.4.2 Other Measures of Inconsistency . . . 151
7.4.3 Agreement with Machine Advice . . . 155
7.5 Conclusion . . . 156
8 Discussion, Limitations & Future Work . . . 157
8.1 Evaluating (un)fairness of existing approaches/systems . . . 157
8.1.1 Discussion and Implications . . . 157

8.1.2 Limitations and Future Works . . . 158
8.2 Fairly updating an already deployed ADMSs . . . 159
8.2.1 Discussion and implications . . . 160
8.2.2 Limitations and Future works . . . 161
8.3 Designing new fair ADMS . . . 162
8.3.1 Time-Critical Influence Maximization . . . 162
8.3.2 Model Uncertainty . . . 166
8.3.3 Human decision-makers . . . 169
Appendices . . . 177
A Fairness Evaluation of Discriminative Foundation Models . . . 178
A.1 Experimental details . . . 178
B Designing new fair ADMS: Model Uncertainty . . . 180
B.1 Training details . . . 180
B.2 Non-linear Classifiers . . . 182
C Designing new fair ADMS: Human Decision-Makers . . . 183
C.1 Effect Sizes . . . 183
Bibliography . . . 185

List of figures

3.1 Classification Task - Demographic disparity - Subjective task - FairFace dataset 23

3.2 [Classification - DTPR - Objective - CelebA] The plots show the TPR disparity, given by Eq. (3.7), between men and women for three zero-shot classification tasks using the CelebA dataset on top and the accuracy on the bottom. *The results demonstrate that mutual information and fair PCA based methods reduce disparity. However where the dimension of the CLIP embeddings is reduced significantly, using mutual information based methods, accuracy can also lower significantly.* . . 24

3.3 [Classification - DDP - Subjective - Flickr30k] Using Flickr30K dataset, this figure shows box plots of DDP, given by Eq. (3.5), for several subjective zero-shot classification tasks. *Most methods effectively reduce classification bias, except for the prompt based method. One reason could be that the model provided by the authors was trained to have a higher importance for maintaining representational powers of the embedding (itc loss: Section 3.2.3.2) as opposed to reducing bias.* 25

3.4 [Retrieval - DDP - Subjective - FairFace] These figures show the average DDP, given by Eq. (3.6), for gender (left) and race (right) attributes averaged over several image retrieval tasks, given in Appendix A.1, using the FairFace dataset. *The results demonstrate that protected attribute specific queries and fair PCA based methods do well in removing bias for image retrieval tasks. Mutual information based methods also perform well for the gender attribute.* 26

3.5 [Retrieval - DDP - Subjective - Flickr30k] The plot shows the DDP, given by Eq. (3.6), for gender attribute using Flickr30K dataset. *All the methods, except the prompt based method, decrease the disparity between men and women for the retrieval tasks.* 27

3.6 [Classification - DTPR - Objective - MIAP] *The x-axis shows three classification tasks: i) 'inconspicuous photo of a person' vs 'prominent photo of a person', where ground truth was based on whether the bounding box of the person occupied more than 50% of the image. ii) 'child' vs 'adult' iii) 'one person' vs 'more than one person'. On top we show the disparity in the true positive rates across the gender attribute and in the bottom we show the accuracy. We see that mutual information based methods while in some cases do reduce the disparity but they incur a reduction in accuracy. On the other hand fair PCA based methods reduce the disparity while incurring almost no loss in accuracy.* 32

3.7 [Retrieval - Cosine similarity - Subjective - FairFace] These figures are heatmaps that show the absolute difference in cosine similarity, scaled up by a factor of 100, for different image retrieval queries using different methods for gender (left) and race (right) attributes on FairFace dataset. *The figures demonstrate the efficiency of each methods to equalize the representation for different protected attribute groups on average. It shows that in general, fair PCA and mutual information based methods equalize the cosine similarity for gender and race attribute for a variety of queries.* 40

3.8 [Retrieval - Cosine similarity - Subjective - Flickr30k] The figure is heatmap that show the absolute difference in cosine similarity, scaled up by a factor of 100, for different queries using different methods for gender attribute on Flickr30K dataset. *The figure demonstrates the efficiency of each methods to equalize the representation for different protected attribute groups on average. It shows that in general, fair PCA based methods and the mutual information based methods equalize the cosine similarity for gender attribute for a variety of queries.* 41

3.9 [Retrieval - Cosine similarity - Subjective - MSCOCO] The figure is a heatmap that shows the absolute difference in cosine similarity, scaled up by a factor of 100, for different queries using different methods for gender attribute on MSCOCO dataset. *The figure demonstrates the efficiency of each methods to equalize the representation for different protected attribute groups on average. It shows fair PCA based methods and mutual information based methods equalize the cosine similarity for gender attribute for a variety of queries.* 42

3.10 [Retrieval - Cosine similarity - Subjective - FairFace - OpenCLIP] These figures are heatmaps that show the absolute difference in cosine similarity, scaled up by a factor of 100, for different image retrieval queries using different methods for gender (left) and race (right) attributes using FairFace dataset with the OpenCLIP. *The figures demonstrate the efficiency of each methods to equalize the representation for different protected attributes groups on average. It shows that in general, fair PCA based methods reduce the difference in cosine similarity for gender and race attribute for a variety of queries.* 44

3.11 [Classification - DDP - Subjective - FairFace - OpenCLIP] These figures show DDP for classification, given by Eq. (3.5), using OpenCLIP using FairFace dataset. *It demonstrates that fair PCA based methods perform the best in reducing bias.* 45

3.12 [Retrieval - DDP - Subjective - FairFace - OpenCLIP] These figures show DDP for image retrieval, given by Eq. 3.6, using OpenCLIP on FairFace dataset. *It demonstrates that gender balanced queries and fair PCA are most effective in reducing demographic disparity in subjective image retrieval tasks.* 46

3.13 [Classification - DDP - Subjective - Flickr30K - OpenCLIP] These figures show DDP for classification, given by Eq. 3.5, using OpenCLIP on Flickr30K dataset. *It demonstrates that fair PCA based methods are the most effective in reducing bias in classification tasks.* 50

3.14 [Retrieval - DDP & Cosine similarity - Subjective - Flickr30K - OpenCLIP] These figures show DDP, given by Eq. (3.6), for retrieval task using OpenCLIP using Flickr30K dataset on the left, and absoulte differences in the cosine similarity between men and women for different queries on the right. 51

3.15 [Retrieval - DDP - Subjective - MSCOCO] The figure on the top shows DDP, given by Eq. (3.6), for retrieval tasks using MSCOCO dataset. *These results demonstrate bias in human-centric subjective tasks. At the bottom, we observe the fraction of query results that actually include a person. Surprisingly, for many human-related queries, the retrieved images do not feature any humans at all. Additionally, this demonstrates that the simple baseline of gendered queries perform very well in reducing disparity. However, the mutual information-based approaches, although effective in reducing disparity in some cases, fail to retrieve images containing humans. Interestingly, Fair PCA, trained on the inferred gender attribute, manages to return appropriate images while still reducing some disparity. One possible reason for this could be that the gender labels derived from the captions, which serve as ground truth, are quite noisy. In contrast, training fair PCA on on the inferred gender attribute directly from the CLIP model appears to yield better results in this context.* 54

3.16 [Classification - DDP - Subjective - MSCOCO] The figure on the top shows DDP, given by Eq. (3.5), for classification tasks using MSCOCO dataset. *These results show bias for human-centric subjective tasks. They demonstrate that for most methods reduce disparity across gender in classification tasks.* 56

4.1 [Synthetic dataset. Enforcing statistical parity] These figures show a comparison between the solutions of Problem (P4.1), using SP proxies, and Problem (P4.3). Left panel shows the beneficial outcome rates, i.e., , positive class acceptance rates, for a classifier only enforcing SP constraint (solid lines), and a classifier additionally enforcing the "loss-averse" constraint (dotted lines). Right panel shows the nondiscrimination-accuracy tradeoff for both the classifiers. Enforcing "loss-averse" constraint, defined in Eq. (4.7), leads to significant additional loss in accuracy for the same level of discrimination. 67

4.2 [Synthetic dataset. Enforcing equality of opportunity] Figure on the left shows the beneficial outcome rates, i.e., , true positive rates, for a classifier only enforcing EOP constraint (solid lines) and a classifier additionally enforcing the "loss-averse" constraint, given in Eq. (4.8), is shown in dotted lines. Figure on the right shows nondiscrimination-accuracy tradeoff for both the classifiers. 70

4.3 [Adult dataset. Enforcing statistical parity] Left panel shows the beneficial outcome rates, i.e., , positive class acceptance rates, for a classifier only enforcing SP constraint, i.e., , solution of Problem (P4.1) using SP proxies (solid lines), and a classifier additionally enforcing the "loss-averse" constraint, i.e., , solution of Problem (P4.3) (dotted lines). Right panel shows the nondiscrimination-accuracy tradeoff for both the classifiers. Enforcing "loss-averse" constraint, defined in Eq. (4.7), leads to a significant additional loss in accuracy for the same level of discrimination. 72

4.4 [SQF dataset. Enforcing equality of opportunity] These figures show similar results as Figure (4.2) using SQF dataset. 74

5.1 An example to illustrate the disparity across groups in the standard approaches to TCIM. (Left) Graph with $|\mathcal{V}| = 38$ nodes belonging to two groups shown in "blue dots" ($|\mathcal{V}_1| = 26$) and "red triangles" ($|\mathcal{V}_2| = 12$). (Right) We compare an optimal solution to the standard TCIM-BUDGET problem P5.1 and an optimal solution to our formulation of TCIM-BUDGET with fairness considerations given by FAIRTCIM-BUDGET problem P5.4. For different time critical deadlines τ, normalized utilities are reported for the whole population $\mathcal{V}$, for the "blue dots" group $\mathcal{V}_1$, and for the "red triangles" group $\mathcal{V}_2$. As τ reduces, the disparity between groups is further exacerbated in the solution to TCIM-BUDGET problem P5.1. Solution to FAIRTCIM-BUDGET problem P5.4 achieves high utility and low disparity for different deadlines τ. 79

5.2 Demonstration of concave function encouraging picking seeds which influence under-represented group. X-axis represents group influence and y-axis represents the value of $\mathcal{H}$ for the corresponding group influence. In this example we have two groups, $\mathcal{V}_1$ and $\mathcal{V}_2$. $\mathcal{V}_1$ is under-influenced compared to $\mathcal{V}_2$, using the seed set S. In the next iteration we have an option to either include node a or b in our seed set, both of which add the same amount of total influence. Adding node a in our seed set influences $\mathcal{V}_1$which is the under-influenced group, while adding node b influences nodes from $\mathcal{V}_2$, as demonstrated in the figure. The traditional method, given by problem P5.1, would treat both of these nodes as equally good. However, since we are passing the group influences through a concave function the increase in the value of $\mathcal{H}(z)$ will be more if we pick node a, i.e., our method will pick node a because $\delta_1 > \delta_2$. 85

5.3 where $\mathcal{F}(z) = \min\left\{\frac{f_\tau(S;\mathcal{V}_i,\mathcal{G})}{|\mathcal{V}_i|}, Q\right\}$. Demonstration of the constraint in problem P5.6. X-axis represents the fraction of group influences and y-axis represents the value of per group constraint in problem P5.6 for the corresponding group influence. In this example we have two groups, $\mathcal{V}_1$ and $\mathcal{V}_2$ of roughly same size. $\mathcal{V}_1$ has not reached the prescribed quota, Q, while $\mathcal{V}_2$ has already been influenced up to the prescribed quota. In the next iteration we have an option to either include node a or node b in our seed set, both of which add the same amount of total influence. Adding node a in our seed set influences only $\mathcal{V}_1$, while adding node b influences nodes from only $\mathcal{V}_2$, as demonstrated in the figure. The traditional method, problem P5.2, would treat both of these nodes as equally good candidates for including in the seed set because they add equal fraction of total influence. However, since we require all the groups to be influenced up to the required quota, selecting node a will increase our constraint value, $\mathcal{F}(z)$, while by selecting node b the constraint value would stay the same as $\mathcal{V}_2$ has already reached the required quota of influence. 89

5.4 [Synthetic Dataset: Budget Problem] The figures show that solving TCIM-BUDGET problem P5.1 can lead to disparity in number of influenced nodes belonging to different groups, while FAIRTCIM-BUDGET problem P5.4 fares better in terms of achieving parity of influence, with marginally lower total influence. See Section 5.4.2 for further details. 90

5.5 [Synthetic Dataset: Budget Problem] These figures demonstrate that lower activation probabilities, uneven group sizes, and cliquishness can lead to higher disparity of influence between different groups with TCIM-BUDGET problem P5.1. In comparison our proposed method, FAIRTCIM-BUDGET given by problem P5.4, leads to solutions which yield lower disparity. For further details, see Section 5.4.2. 92

5.6 [Synthetic Dataset: Cover Problem] These figures show a comparison of TCIM-COVER problem P5.2, in red, and FAIRTCIM-COVER problem P5.6, in blue. They show that FAIRTCIM-COVER achieves lower disparity of influence between different groups with slightly bigger solution set sizes. See Section 5.4.3 for further details. 94

5.7 [Rice-Facebook Dataset: Budget Problem] Comparison of results solving TCIM-BUDGET problem P5.1 and FAIRTCIM-BUDGET P5.4. We experimented with 4 groups and total influence includes all the groups, but we show group influences and disparity for only two groups which showed the maximum disparity. The results demonstrate that our method, given by problem P5.4, yields seed set which propagate influence in a more fair manner, at the cost of a marginally lower total influence. See Section 5.5.2 for further details. 96

5.8 [Rice-Facebook Dataset: Cover Problem] These figures demonstrate the results of TCIM-COVER problem P5.2, in red, and FAIRTCIM-COVER problem P5.6, in blue. We experimented with 4 groups and total influence includes all the groups but we show group influences for the two groups which had maximum disparity. The results show that our method achieves a more equal coverage for all the groups at the expense of only slightly larger seed sets. See Section 5.5.3 for further details. 97

5.9 [Instagram-Activities Dataset] These figures demonstrate a comparison of TCIM-BUDGET (Problem P5.1)vs FAIRTCIM-BUDGET (Problem P5.3) and TCIM-COVER (Problem P5.2) vs FAIRTCIM-COVER (Problem P5.5) problems. The results show that our methods fare better compared to the traditional methods. Even though the fraction of influence seems small, since the graph comprises 0.5m nodes, the differences in fractions are significant in total numbers. 99

5.10 [Facebook-Snap dataset] These figures demonstrate a comparison of TCIM-BUDGET (Problem P5.1) vs FAIRTCIM-BUDGET (Problem P5.3)and TCIM-COVER (Problem P5.2) vs FAIRTCIM-COVER (Problem P5.5) problems. The results show that our method improves the disparity of the influence between different groups. The results for the budget problem show some improvement in the disparity. However, in comparison the reduction in the total influence is also small. One can consider a concave wrapper with a larger curvature to improve the disparity. The results for the cover problem, show a clear improvement in the disparity between the groups. 101

6.1 Illustrative example: Consider a binary classification task with two features and a sensitive feature represented by the shape of the points, i.e., circles and triangles. Green and red colors represent ground truth positive and negative labels, respectively. Classifiers $C1$ and $C2$ are equally accurate classifiers achieving 79% accuracy. The difference between false positives of triangles and circles for $C1$ is 22% and -12% with $C2$. However, these two classifiers disagree on their decision on 17% of the data, i.e., which lies in the ambiguous region shown in the shaded blue region. If we were to pick one of these classifiers it would be unfair to the points receiving a favorable decision with the other classifier. On the other hand, a fair classifier equalizing false positive rates, using [275], gives an accuracy of only 71%. However, it changes the decisions of several points that clearly belong to the positive cluster. 106

6.2 [Synthetic dataset] Figure demonstrates that state of the art fairness methods are effected by label noise. 113

6.3 [Synthetic dataset] Figure shows the expected class while equalizing FPRs using the classifiers solving P6.4. It demonstrates that our method is stable under label noise, as it consistently identifies same regions as ambiguous for different levels of noise values. 114

6.4 [Synthetic dataset] This figure shows the ambiguous regions (in red) identified by the four methods discussed in the paper. It demonstrates that our methods identify similar ambiguous regions compared to the exact methods proposed by Marx et al. [191]. The results correspond to $\epsilon = 0.03$. We see similar results for different values of ϵ. 116

7.1 Graphical overview of experimental conditions T1-T5. In T1 and T2, respondents review their decisions one-by-one, while in T3-T5 they review decisions in randomly (T3) or meaningfully selected (T4 and T5) pairs. In T2 and T5 respondents are additionally provided with (different kinds of) explicit machine advice. 127

7.2 Description of the experimental design shown to participants at the beginning of the experiment. 133

7.3 Stimulus material. 134

7.4 Average duration of the experiment, per experimental condition, and per experimental phase. The experimental conditions T1–T5 are shown on the x-axis. The values for the pre-review experimental phase are shown in blue, while the post-review values are shown in orange. We report mean values calculated across respondents ± 1.96 standard errors of the mean (SEM). 137

7.5 H1: Effect of the interventions on people's propensity to update decisions, across all 30 apartments. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents ± 1.96 standard errors of the mean (SEM). 141

7.6 H1': Effect of the interventions on people's propensity to update decisions, across the subset of apartments that were shown in the review phase. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents ± 1.96 standard errors of the mean (SEM). 144

7.7 H2: Effect of the interventions on the accuracy of respondents' decisions. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents ± 1.96 standard errors of the mean (SEM). 146

7.8 H3: Effect of the interventions on the consistency between respondents' decisions. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents ± 1.96 standard errors of the mean (SEM). 147

7.9 Error in people's implicit relative judgments. The y-axis shows the fraction of instances where people's implicit relative ordering of apartments (>,< or =) did not match the ground truth ordering based on the listing price. We report mean values calculated across all respondents and pairs of apartments ± 1.96 standard errors of the mean (SEM). 148

7.10 Distribution of errors in respondents' estimates, across all treatments. The x-axis shows the magnitude of errors, i.e., the difference between the apartments' true prices and the respondents' estimates. The y-axis shows the number of responses in our dataset that exhibited a certain magnitude of error. 149

7.11 Directionality of response updates. The y-axis shows the fraction of revised responses that were updated to increase (blue) or to decrease (orange) the initial price estimates, for each of the experimental conditions T1-T5, shown on the x-axis. 150

7.12 Effect of the interventions on the consistency of people's absolute judgments. The experimental conditions T1–T5 are shown on the x-axis. . 151

7.13 Effect of the interventions on the consistency of people's implicit relative judgments. The experimental conditions T1–T5 are shown on the x-axis. 153

B.1 [Synthetic dataset-non-linear] The figure on the left shows the 2 moons dataset, the middle figure shows the best non-linear boundary with green regions classified as positive and red regions as negative and the one on the right shows the ambiguous regions identified using our method. The figure demonstrate that unlike Marx et al. [191] our methods can also be used to identify predictive multiplicity for non-linear classifiers. 181

List of tables

3.1 Taxonomy for evaluating fairness of discriminative foundation models. . . 22

3.2 [Classification - Accuracy - Objective - StanfordCars, Food-101, VOC objects & Imagenet] *The bias mitigation methods shown in the table were trained using the FairFace Dataset. We used the test splits for all the datasets. The results show that fair PCA based methods retain performance on non-human objective tasks. We would like to note that we only show results with a prompt of "a photo of a {label}", while the original CLIP paper aggregates results using several prompts, which they did not disclose. In some cases this can result in a difference in evaluation numbers that we are reporting compared to the original CLIP paper. However, our results are within the margin of improvement that the original CLIP paper claims to achieve using prompt engineering.* 28

3.3 [Retrieval - DDP & Precision - Objective - IdenProf] This table shows fairness evaluation for representational bias on objective tasks for image retrieval of CLIP model and different bias mitigation methods. *Using IdenProf dataset, we show DDP-rep, given by Eq.* (3.9)*, for each method as well as its average precision for retrieving images of 9 different professions of the IdenProf dataset. We exclude the profession 'Firefighters' because in many cases their faces are hidden and gender is difficult to identify. Additionally, we do not show results for EOP like measure because this dataset does not have the annotations for the gender attribute. The gender annotations for the retrieved images per profession were manually done by one of the authors. The results demonstrates that gender balanced queries perform the best to reduce the representational unfairness in the objective tasks. All the methods are trained on FairFace dataset to remove the gender bias.* 29

3.4 [Retrieval - Recall - Flickr30k] The table below shows recall@K for randomly selected 50% Flickr30K dataset using different gender bias mitigation methods. *Specifically, we are using the captions of each image as a query and report the fraction queries that retrieve the images correctly in top 1, 5 or 10 results. The results show that mutual information based methods perform worse, which makes sense as the number of dimensions are reduced, while Prompt-GT method performs the best. Since the Prompt-GT method was finetuned using the Flickr dataset, it is not surprising that it outperforms even the CLIP model. It is worth noting that the queries also include gendered queries and some reduction in recall is expected or may even be desirable.* 30

3.5 [Retrieval - Skew - Subjective - FairFace] This table shows the maximum absolute skew, given by Eq. (3.8), using the FairFace dataset and gender attribute. *It demonstrates that all the methods are able to reduce the skew. Gender balanced queries yield the lowest skew.* 31

3.6 [Retrieval - Skew - Subjective - FairFace] This table shows the results for representation bias for subjective labelling. Specifically, it show skew metric , given by Eq. (3.8), for the *race* attribute of FairFace dataset. *Race balanced queries perform well in general but fair PCA based methods perform the best when the number of retrieved items are larger.* 33

3.7 [Retrieval - Skew - Subjective - Flickr30K] This table shows the skew metric, given by Eq. (3.8), for the gender attribute average over several image retrieval task using the Flickr data. *It shows that gender balanced queries and mutual information based methods with a lot reduction in number of CLIP dimensions reduce the skew the most.* 34

3.8 [Retrieval -Skew - Subjective - MSCOCO] This table shows absolute skew, given by Eq. (3.8), for image retrieval tasks using MSCOCO dataset. *The results show that the simple baseline with gender balanced queries perform the best for reducing skew.* 35

3.9 [Retrieval -Statistical Tests - Subjective - FairFace] This table shows Alexander-govern statistical tests using FairFace. *This test checks whether there are differences in the mean value of cosine similarity between men and women for a given query. The pair of numbers represent the test statistic and the p-value. A low value of the statistic and high p-value is desirable, the former means the statistical difference for the given query has low impact and the later means that the differences are statistically insignificant. It shows that fair PCA and MI-GT methods generally achieve the lowest disparity in cosine similarity and the differences are generally statistically insignificant.* 36

3.10 [Retrieval -Statistical Tests - Subjective - FairFace] This table shows statistical tests to check if for a given query all the races have same mean cosine similarity. *A large value of the test statistic and less than 0.05 pvalue implies that there is a large and statistically significant different in the mean value of the cosine similarity for one of the races.* 37

3.11 [Retrieval - Statistical tests - Subjective - Flickr30k] This table shows Alexander Govern statistical test for the cosine similariy of various queries between men and women. *It demonstrates that fair PCA based methods do very well to equalize the cosine similarity between the two groups for different retrieval tasks.* 38

3.12 [Retrieval - Statistical tests - Subjective - MSCOCO] This table shows Alexander Govern statistical test for the cosine similariy of various queries between men and women. The first number refers to the test statistic while the second number is the p-value. If there is a statisitically significant difference among different groups the test statistic would be high and p-value would be low. *It demonstrates that fair PCA GT yields statistically insignificant differences.* 39

3.13 [Classification - Accuracy - Objective - FairFace] This table shows the accuracy of a logistic regression classifier trained on the corresponding CLIP features for FairFace dataset. *The top and the bottom parts of the table correspond to the cases where the mitigation methods were supposed to remove the gender and race information, respectively, from the CLIP embeddings, while preserving the other information. The results show that fair PCA based methods are more effective in removing the corresponding sensitive information, i.e., the accuracy for predicting the corresponding sensitive attributes is nearly random. Additionally, the fair PCA methods do not reduce the predictive power of the embeddings, i.e., the accuracy in predicting other attributes stays similar to the original CLIP embeddings. We do not provide the results for the prompt method because they do not alter the image representation and results are similar as the original CLIP.* 43

3.14 [Retrieval - Skew - Subjective - FairFace - OpenCLIP] This table shows the maximum absolute skew, given by Eq. (3.8), using the FairFace dataset and gender and race attributes using OpenCLIP. *It demonstrates that all the methods are able to reduce the skew. Gender/Race balanced queries and fair PCA are the most effective in reducing the skew.* 48

3.15 [Retrieval - Statistical tests - Subjective - FairFace - OpenCLIP] This table shows the statistical tests for the cosine similarities among different groups of the protected groups. The first number refers to the test statistic while the second number is the p-value. If there is a statisitically significant difference among different groups the test statistic would be high and p-value would be low. *Specifically, it shows the Alexander-govern statistical test which measures whether the mean of cosine similarity among different groups for a given query are statistically significant or not. It shows that fair PCA trained on ground truth protected attribute labels yields statistically insignificant differences.* 49

3.16 [Retrieval - Skew - Subjective - Flickr30K - OpenCLIP] This table shows the skew metric, given by Eq. (3.8), using OpenCLIP model, for the gender attribute average over several image retrieval task using the Flickr data. *It shows that gender balanced queries are most effective in reducing skew.* 52

3.17 [Retrieval - Statistical tests - Subjective - Flickr30K - OpenCLIP] This table shows the statistical tests for the cosine similarities for different queries between men and women. The first number refers to the test statistic while the second number is the p-value. If there is a statisitically significant difference among different groups the test statistic would be high and p-value would be low.*Specifically, it shows the Alexander-govern statistical test whether the mean of cosine similarity between men and women for a given query are statistically significant. It shows that fair PCA trained on ground truth protected attribute labels yields statistically insignificant differences.* 53

3.18 [Retreival - Precision - Objective - MSCOCO & CelebA] This table shows average precision@K for image retrieval tasks using different methods for 80 categories of MSCOCO dataset and 9 attributes of CELEBA. *It demonstrates that CLIP and fair PCA methods usually yield similar precision. On the other hand, fair sampling which is trained on MSCOCO does very well on the MSCOCO dataset but has a poor performance on CELEBA dataset. The mutual information based methods have a better performance where more dimensions of the CLIP embeddings are used.* 58

5.1 Motivating Example 79

6.1 [Synthetic dataset] Signed differences in FPR/FNR: This table demonstrates that our method is effective in removing unfairness at a very small cost of decrease in the accuracy. Please refer to Section 6.4.4 117

6.2 Comparison identifying ambiguous regions: The tables show maximum discrepancy and ambiguity between any two classifiers in the $\mathbb{C}_{\epsilon,\psi:\psi\in\{\phi_{best},\theta_{best}\}}$. The bottom table shows the time it took to compute the ambiguous regions with each method. It shows that our methods, given by P6.3 and P6.4, achieve comparable performance compared to P6.1 and P6.2 and they are upto four orders of magnitude faster. Please refer to Section 6.4.4 118

6.3 [COMPAS] Signed differences in FPR/FNR : This table demonstrates that our methods are effective in removing unfairness in the ambiguous regions at no expense of accuracy. Please refer to Section 6.4.5 119

6.4 [SQF] Signed differences in FPR/FNR: This table demonstrates effectiveness of our methods. Please refer to Section 6.4.5 121

7.1 Overview of the characteristics of the 5 experimental conditions in our study. Reviewing Procedure: Are instances reviewed one-by-one or pairwise? Algorithmic Assistance: Do respondents have access to any form of algorithmic assistance? Data Required: Do the utilized decision aids require any type of labeled data? 129

7.2 Demographics of our study sample, compared to the 2019 U.S. Census [249].136

7.3 Linear mixed models with crossed random effects for participants and apartments. The dependent variables for different hypotheses are described in Section 7.2.1. In all six models, the four independent variables T2–T5 correspond to a one-hot encoding of the five experimental conditions T1–T5, and T1 is treated as the reference category. I.e., intuitively, the row "Cons." shows the estimated value of the constant term (or intercept) that corresponds to the effects of treatment T1, while the rows T2–T5 show how the effects of these treatments differ compared to T1. Hence, to reason about the effects of T2–T5, one needs to sum up the values of the constant term and the treatment of interest. *N* denotes the number of data points used to fit a specific model. Each of our 643 respondents answered questions about 30 apartments, resulting in a total of 19290 data points. Please note that in H1′ some of the data points are discarded, as described in Sections 7.2.1 and 7.3.2. Standard errors are shown in parentheses. Statistical significance of coefficients is indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. . 142

C.1 Cohen's d. A value in row *i* and column *j* corresponds to the effect size of treatment *i* on the variable *j*. On a high-level, Cohen's d quantifies the difference between the means of variable *j* for two groups of respondents: those assigned to treatment T1 and those assigned to treatment *j*. More precisely, we report the values of Cohen's d as defined in Cohen [62], calculated using the `esize` command in Stata, specifying the option `unequal`, to specify that the two groups should not be assumed to have equal variances. 183

CHAPTER 1

Introduction

In this chapter, we first discuss the gaps in different aspects of designing fair decision-making systems, then we highlight the challenges of solving these problems and lastly we discuss the contributions of this thesis.

1.1 Motivation

The motivation to create fair decision-making systems stems from the significant impact these decisions have on human lives. Examples of such systems include reviewing research papers for conferences, determining bail eligibility, sentencing in prisons, real estate appraisals, job performance evaluations, and the shortlisting of applicants for job interviews. These systems may involve solely human decision-makers, a combination of humans and algorithms, or exclusively algorithmic decision-making systems. We refer to the the later two types of the systems as algorithmic decision-making systems (ADMSs) and focus on addressing fairness concerns in these ADMSs.

Aspects of designing fair ADMSs

There are several components of designing fair ADMSs, which have received a lot interest from the research community. However, there are still gaps in three component that we address in this thesis, namely: i) understanding and evaluating (un)fairness in an ADMS ii) fairly updating an already deployed ADMS iii) or designing a new fair ADMS, from scratch where updating an existing system is infeasible or there is no existing system.

1.1.1 Evaluating unfairness of existing approaches/systems

In order to design fair decision-making systems, a key step is to understand and evaluate the unfairness of the application scenarios or systems currently in place.

Foundation models. In the past few years, *large* models trained on huge amounts of data, primarily crawled from the internet, have become popular (e.g., BERT [74], CLIP [220], GPT-3 [38], DALL-E [222], Stable Diffusion [227]). Many of these models have gained attention even in the general public and extensive news coverage, which typically also addresses the risks and shortcomings of these models (e.g., [197, 211]). These large models are now commonly referred to as foundation models, a name coined by researchers from Stanford to "underscore their critically central yet incomplete character" [32].

These foundation models can be broadly divided into two categories: generative and discriminative models. Following the machine learning literature [30], a generative model is one that can generate synthetic data, such as images or text, and a discriminative model is one that can distinguish between types of data, for example, by classifying images as cats or dogs.

Potential for harm in discriminative foundation models. Popular generative foundation models, such as ChatGPT [204] and Stable Diffusion [227], regularly make the news, both because of the rapid rate of progress in the field [254, 268] and the potential harms [105] including copyright violation [256] and the hallucination of incorrect and possibly libelous data [135]. However, in many ways the dangers of discriminative models can be more insidious. Discriminative models such as CLIP [220] allow for the zero-shot classification of data, i.e., without access to labeled training data they can assign images to a set of previously unseen labels. As zero-shot solutions do not require conventional data sources, models can be optimistically deployed without systematically evaluating if they are accurate, fair, or even if the task they are deployed on makes sense (e.g., identify hard workers from resume photographs). Because discriminative models may be used to make decisions about individuals, their behavior can have a direct impact on a person's life, e.g., through controlling access to education, employment or medical care.

Later, in Section 1.2.1 we discuss the challenges of evaluating discriminative foundation models and in Section 1.3.1 we discuss an overview of our contributions to address these challenges.

1.1.2 Fairly updating an already deployed ADMS

Another important aspect of a discriminative system is how to update it. In many decision-making scenarios such as banking or judiciary or insurance, a newly deployed system often replaces an existing decision-making system. This could be a human decision maker, an older learning model without discrimination-awareness, or a learning model trained on outdated data (e.g., when features of users in a society evolve).

Existing literature in behavioral economics and psychology shows that people's perceptions of fairness of a new decision-making system are influenced by how the decision outcomes change from the status quo, i.e., how the new outcomes differ from the old outcomes [21, 138, 141, 248]. In several fields, ranging from software updates to domestic/foreign policy making [122] and public budgeting [229], the status quo is considered when updating existing systems.

However, current works on fair learning do not account for the status quo when reasoning about fairness of an ADMS. Neglecting the status quo could have a severe impact on the lives of the people affected by ADMSs. For example, consider a company where men tend to have higher salaries than women. If the company tries to implement a new salary policy to equalize this gender pay gap by reducing the salaries of men, the men might perceive it as unfair, feeling entitled to their previous salary amounts based on the status quo.

We discuss challenges of updating learning based systems while accounting for the status quo in Section 1.2.2 and provide an overview of our contributions to address these challenges in Section 1.3.2.

1.1.3 Designing new fair ADMSs

In certain situations, there may not be an already deployed ADMS or it might be prohibitively complex to take status-quo into account when updating an existing system or the fairness concerns in the current system might outweigh the benefits of update-fairness. In these cases, it is more sensible to design a fair ADMS from the ground up.

In this section, we discuss the fairness concerns in three different and crucial scenarios: influence maximization problem, model uncertainty and inconsistency in human decision-makers. We use a running example of company X which wants to expand its business by hiring several new employees.

Influence maximization. In order to reach to a broader talent-pool company X wants to advertize the job openings. It is common to use social media platforms such as Twitter, LinkedIn or Facebook for such advertisements. Usually the advertizers can only pay for the advertisements to be seen by selected individuals. A popular method of selecting such individuals is by using influence maximization. The idea is to identify a set of initial sources (i.e., *seed nodes*) in a social network who can influence other people (e.g., by propagating key information), and traditionally the goal has been to maximize the total number of people influenced in the process (e.g., who received the information being propagated) [49, 106, 147].

Real-world social networks, however, are often not homogeneous and comprise different groups of people. Due to the disparity in their population sizes, potentially high propensity towards creating within-group links [194], and differences in dynamics of influences among different groups [235], the structure of the social network can cause disparities in the influence maximization process. For example, selecting most of the seed nodes from the majority group might maximize the total number of influenced nodes, but very few members of the minority group may get influenced. In many application scenarios such as propagation of job or health-related information, such disparity can end up impacting people's livelihood and some groups may become impoverished in the process.

Moreover, some applications are also *time-critical* in nature [54]. For example, many job applications typically have a deadline by which one needs to apply; if information related to the application reaches someone after the deadline, it is not useful. Similarly, in viral marketing, many companies offer discount deals only for few days (hours); getting this information late does not serve the recipient(s). More worryingly, if one group of people gets influenced (i.e., they get the information) faster than other groups, it could end up exacerbating the inequality in information access. This is possible if the majority group is better connected and more central in the network than the minority group. Thus, in time-critical application scenarios, focusing on the traditional criteria of maximizing the number of influenced nodes can have a disparate impact on different groups. This disparity in time-critical applications, in turn, can put minority and under-represented groups at a big disadvantage with far-reaching consequences.

Model uncertainty. In response to the advertisement by company X, a large number of people apply for the jobs making it difficult for human decision-makers to go through all the CVs. So as a first phase, they are sorted by an ADMS (a.k.a applicant tracking systems) into reject and potential hire. Such systems have been known to be discriminatory, e.g., Amazon has discontinued its AI recruitment tool because it showed bias against women[1]. To mitigate bias in such and other prediction systems that are being used for several socially impactful tasks, e.g., predicting recidivism risk in order to help judges make bail decisions, assessing credit ratings, assessing the risk of defaulting on a loan and predicting the risk of accident for insurance purposes, researchers have proposed a class of group fairness methods, which seek to equalize overall errors across different groups of sensitive attributes such as gender or race [13, 120, 273, 275]. This approach treats all errors as equal. However, not all errors are the same.

[1]https://www.reuters.com/article/idUSKCN1MK0AG/

It is well-known that errors in prediction models arise out of both epistemic (model) uncertainty and aleatoric (inherent) uncertainty [72, 124, 189]. Epistemic/model errors occurs due to lack of data or lack of knowledge about the best model that would suit the given data. Aleatoric errors either occur due to inherent uncertainty of the task or random noise in the data. Treating epistemic errors and aleatoric errors equally could lead to unjustifiably wrong decisions for some datapoints.

Moreover, even if the initial ADMS (such as the applicant tracking system in our example) is not inherently discriminatory but has errors due to model uncertainty, it suggests that there could be another equally effective ADMS with differing outcomes for certain data points. Deploying one of these systems could result in some users being adversely affected, as the system deployed may provide an unfavorable outcome compared to another equally effective system, leading to perceptions of unfairness.

Human Decision-Makers. Finally, a select number of CVs have made it to a panel of human decision-makers at the company X and they have to make the final hiring decision. However, prior research in psychology has found that presented with identical information, the same person might make different decisions at different points in time, and the decisions of different people are likely to vary even more [139, 140]. Such inconsistencies between decision-makers have been identified in numerous settings including sentencing [15], job performance evaluations [243], real estate appraisals [3], and—especially close to the research community—conference reviewing [26, 39, 68, 167, 246].

In certain settings, variation in people's decisions is indispensable; it may contain invaluable information that reflects the variation in people's background knowledge, political or moral stances, life experiences, and other factors [226, 265, 266]. However, in other settings, consistency may be considered normatively desirable instead. For instance, Kahneman et al. [139] argue that organizations such as credit-rating and insurance agencies expect that, regardless of the particular professional handling each case, "identical cases should be treated similarly, if not identically"—a notion in line with that of "individual fairness" in the algorithmic fairness literature [80]. In conference reviewing, inconsistencies between different groups of reviewers have raised concerns about the peer-review process in the scientific community [26, 39, 68, 167, 246]. In the organizational justice literature, consistency of decisions is recognized as an important aspect of procedural justice [168, 172].

We discuss the challenges of designing these three systems in Sections 1.2.3 and we discuss our contributions addressing these challenges in Sections 1.3.3.

1.2 Challenges

1.2.1 Evaluating unfairness of existing approaches/systems

Guided by the principles from law, ethics, and philosophy, numerous metrics have been suggested to assess the fairness of ADMSs. Some of these metrics focus on the outcomes. For instance, *Statistical parity* [89, 93, 144] is applied when ground truth labels are uncertain, requiring the ADMS to distribute favorable outcomes proportionally among socially significant groups, such as genders or races. *Equal opportunity* [120, 153, 274] is relevant when trustworthy ground truth labels are available, demanding equal error-rates for different groups. *Diversity-based* measures [276] mandate sufficient representation of different groups in the beneficial class label.

While these metrics are valuable in specific applications, they can sometimes conflict with each other [255]. The choice of the appropriate metric for evaluating fairness is crucial. A study by ProPublica [18] highlighted issues with the COMPAS software tool, designed to assist judges in bail decisions by providing recidivism risk predictions. Despite having similar accuracy across races, the tool exhibited distinct errors for different racial groups. Notably, it consistently misclassified white defendants as low risk more frequently than black defendants. This underscores the importance of selecting the right metric to evaluate the fairness of a system.

The advent of discriminative foundation models like CLIP, which enable zero-shot applications without labeled training data, further complicates these challenges. Zero-shot solutions do not rely on conventional data sources, allowing models to be optimistically deployed without systematic evaluation of their accuracy, fairness, or even the relevance of the tasks they are applied to (e.g., identifying hard workers from resume photographs). In summary, evaluating such models is particularly challenging due to two factors: the commodotization of zero-shot machine learning across diverse tasks and the abundance of fairness definitions that could often be inconsistent.

1.2.2 Fairly updating an already deployed ADMS

Updating decision-making systems while accounting for the status quo system has been largely ignored in fairness research. To address this gap, the first major challenge is to establish a plausible notion of fairness grounded in existing social sciences research. Additionally, we need to demonstrate how this proposed notion applies to the practical scenarios of ADMSs.

The second challenge involves operationalizing this notion of fair-update in practical terms. Specifically, we need to determine how the fairness concept can be translated and integrated into an ADMS given a particular application scenario.

In this thesis, we focus on the fair update of convex margin-based binary classification systems. This brings us to the third challenge: how to efficiently incorporate the proposed notion into the training mechanism of such systems. It turns out that integrating the notion of fair-update as constraints in the training process of margin-based binary classification systems results in non-convex and intractable problem formulations. To address this, we must develop mechanisms that can be seamlessly integrated into existing learning algorithms for efficient implementation.

1.2.3 Designing new fair ADMSs

Lastly, in this thesis we address challenges in designing new fair ADMSs in three different settings.

Model uncertainty. The primary challenge is to devise a sensible proposal that distinguishes between various types of errors based on their uncertainty-origin in algorithmic discrimination. The second challenge involves the efficient and effective identification of errors based on their uncertainty-origin. Lastly, we need to put forth an efficient mechanism for training non-discriminatory classifiers in the presence of model uncertainty.

Influence maximization. The initial challenge is to operationalize existing notions of discrimination within the time-critical influence maximization (TCIM) framework. While TCIM is an NP-hard problem, it is submodular which lends itself to efficient approximation with performance guarantees. Unfortunately, incorporating fairness notions into TCIM yield formulations which are no longer submodular. Hence, the second challenge involves devising *computationally efficient methods* to incorporate fairness notions into the TCIM problem.

Human decision-makers. To our knowledge, no prior studies have examined how algorithmic assistance influences the consistency among human decision-makers. To explore this, the main challenge is to design an experiment that tests the effectiveness of various methods of improving consistency, including giving machine advice. Additionally, in many tasks, gathering the ground-truth labels is costly or defining the ground-truth is difficult. For such cases, another challenge is to devise different methods of algorithmic assistance without relying on the ground-truth labels. Furthermore, we have to develop different types of decision-aids. Lastly, we need to explore different definitions of consistency and determine whether our results are robust across these different definitions.

1.3 Overview of thesis contributions

1.3.1 Evaluating the Fairness of Discriminative Foundation Models

Research questions. 1) *What constitutes a fair behavior for discriminative foundation models in downstream tasks?* 2) *How fair are the current bias mitigation methods for these models?* 3) *How do simple baseline for bias mitigation perform?*

- **Conceptual contribution:** We propose a novel taxonomy for bias evaluation of discriminative foundation models, such as Contrastive Language-Pretraining (CLIP), that are used for labeling tasks. Using this broad taxonomy, we try to consolidate several contradictory notions of fairness.
- **Empirical contribution:** i) We systematically evaluate OpenAI's CLIP and OpenCLIP models for key applications, such as zero-shot classification, image retrieval and image captioning with respect to our taxonomy. ii) Additionally, we evaluate all existing methods for mitigating bias in these models. We also evaluate some simple baselines and compare their performance to existing bias mitigation methods. iii) Finally, we provide quantitative fairness evaluations for both binary-valued and multi-valued protected attributes over *ten* diverse datasets. We find that fair PCA, a post-processing method for fair representations, works very well for debiasing CLIP models in most cases while incurring only a minor loss of performance. However, different debiasing approaches vary in their effectiveness depending on the task. Hence, one should choose the debiasing approach depending on the specific use case.

In Chapter 3, we discuss this work in detail.

1.3.2 Fairly updating an already deployed ADMS: Loss-Aversively Fair Classification

Research questions. 1) *How can one update an ADMS fairly?* 2) *How can we operationalize the notion of fairness?* 3) *Does the proposed notion of fairness work in practice?*

- **Conceptual contribution:** Motivated by extensive literature in behavioral economics and behavioral psychology (prospect theory), we propose a notion of fair update of a deployed ADMS that we refer to as *loss-averse update*.
- **Technical contributions:** i) We operationalize the notion of loss-aversive fairness for binary classification setting. However, this leads to non-convex formulations.

ii) To address this, we provide covariance based mechanisms to train linear and non-linear convex boundary-based non-discriminatory classifiers (e.g., SVM and logistic regression).

- **Empirical contributions:** i) Using both synthetic and real-world datasets, we show how this notion of fairness can be combined with existing parity based notions of discrimination, such as demographic parity and equality of opportunity. ii) We find that adding the loss-aversive constraint leads to the desired result at the cost of a small decrease in accuracy.

In Chapter 4, we discuss this work in detail.

1.3.3 Designing new fair ADMSs

In this Section, we give an overview of our contributions on designing three different systems, i) fairness in influence maximization, ii) fairness under model uncertainty and iii) consistency among human decision-makers.

1.3.3.1 Fairness in Time-Critical Influence Maximization

Research questions. 1) *What constitutes fairness in* TCIM*?* 2) *How can we operationalize fairness in* TCIM*?* 3) *How is (un)fairness affected by various graph properties and aspects of* TCIM *algorithms?* 4) *How can we efficiently solve* TCIM *with fairness constraints?* 5) *Do the proposed mechanisms work in practice?*

- **Conceptual contribution:** We formally introduce the notion of fairness in time-critical influence maximization, which requires that *within a prescribed time deadline, the fraction of influenced nodes should be equal across different groups*.

- **Technical contributions:** i) We introduce fairness constraints in two formulations of TCIM problem. In order to solve these formulations computationally efficiently, we propose *monotone submodular* surrogates for solving both of these NP-Hard problems. Though the surrogate problems are still NP-Hard, we propose a greedy approximation with provable guarantees. ii) We provide theoretical bounds on our proposed formulations, showing that our solutions are provably efficient.

- **Empirical contributions:** i) We highlight, via experiments and an illustrative example, that the standard algorithmic techniques for solving TCIM problems lead to unfair solutions, and the disparity across groups could get worse with tighter time deadline. ii) Secondly, we study the effect of disparity of influence between

groups: (a) by varying graph properties, such as connectivity and relative group sizes etc., and (b) by varying TCIM algorithmic properties, such as seed budget, reach quota and time deadline etc. iii) We evaluate our proposed solutions over several synthetic and three real-world social networks and show that they are successful in enforcing the aforementioned fairness notion. Enforcing fairness does come at the cost of a reduction in performance. However, as guaranteed by our theoretical results, our experiments indeed demonstrate that this cost of fairness, i.e., reduction in performance, is bounded for our approach.

In Chapter 5, we discuss this work in detail.

1.3.3.2 Accounting for model uncertainty in algorithmic fairness

Research questions:. *What constitutes a fair model under model uncertainty?* 2) *How can we identify epistemic errors efficiently?* 3) *How can we achieve fairness under model uncertainty?* 4) *Do our proposed methods work in practice?*

- **Conceptual contribution:** i) We argue that uncertainty in prediction should be accounted for when designing fairness approaches. To this end, we propose to *only* equalize errors occurring due to model uncertainty, i.e., the epistemic errors, as opposed to the existing fairness approaches which equalize total errors. ii) We draw a connection between model uncertainty and predictive multiplicity, which refers to the scenario where multiple predictive models have similar predictive performance (e.g., similar accuracy) but assign contradictory predictions on a subset of the datapoints, which characterize the *ambiguous regions*.

- **Technical contributions:** i) We propose tractable scalable convex proxies to identify errors in *ambiguous regions*. ii) We also propose efficient mechanisms to only equalize the errors in the ambiguous regions.

- **Empirical contributions:** i) Our experimental results show that our proposed scalable convex proxies to identify regions with predictive multiplicity are comparable in performance and up to four orders of magnitude faster than the current state-of-the-art . ii) Our experimental results on a synthetic and two real-world datasets show that our methods improve fairness in the ambiguous regions while achieving comparable accuracy to the best classifier.

In Chapter 6, we discuss this work in detail.

1.3.3.3 Moderating inconsistency in human Decision-Makers

Research questions. 1) Can algorithmic decision aids moderate inconsistency among human decision-makers? 2) Can we construct decision aids that make human decision-makers more consistent and accurate without relying on the ground truth?

- **Experimental design:** i) We design a human-subject based study where we explore various approaches to moderating human inconsistency. Specifically, we ask participants to estimate real estate prices, using a real estate price dataset studied by Poursabzi-Sangdeh et al. [215]. Then, we ask the participants to review their decision in a review-phase, using different approaches. ii) We leverage prior work in psychology and HCI to develop a set of algorithmic decision aids which may influence the degree of inconsistency of human decisions. iii) Finally, we propose a novel way of giving *machine advice* which does not rely on the ground-truth labels.
- **Analysis contribution:** In our pre-registered confirmatory analysis we find that compared to reviewing past decisions one-by-one (baseline), our interventions showed i) a higher *propensity of updating initial decisions*, ii) a higher *accuracy* of decisions after the review phase, and iii) a higher degree of *consistency* amongst the post-review decisions of different respondents. Notably, our proposed methods of giving advice are able to reduce inaccuracy and inconsisitency among people's post-review decisions without relying on the ground-truth labels. Furthermore, we also conducted detailed exploratory analyses of the effects of our interventions on different measures of accuracy and consistency, and report a series of other descriptive statistics.

In Chapter 7, we discuss this work in detail.

1.4 Thesis outline

Rest of the thesis is organized as follows:

- In Chapter 2, we provide a brief background on various notions of fairness and discuss different approaches of achieving fairness in existing literature.
- In Chapter 3, we present a taxonomy for evaluating (un)fairness in discriminative foundation models. Additionally, we thoroughly evaluate two representative discriminative models and their bias mitigation methods for three key applications, i.e., image classification, image retrieval and image captioning.

- In Chapter 4, we propose a new notion of update-fairness and demonstrate how it can be operationalized and incorporated into binary classification. We also propose efficient mechanisms to learn linear and non-linear non-discriminatory binary classifiers combined with our notion of update-fairness.

- In Chapter 5, we operationalize a notion of fairness in time-critical influence maximization (TCIM) problem. Additionally, we show how it can be incorporated into two versions of the TCIM problem and efficiently solved. We explore the effect of different graph and algorithmic properties on unfairness.

- In Chapter 6, we propose a new approach to train a class of non-discriminatory classifiers. We also provide mechanisms to efficiently categorize errors based on their uncertainty-origin. Additionally, we provide mechanisms to deal with fairness issues under model uncertainty.

- In Chapter 7, we present a human-subject based study on the efficacy of algorithmic decision-aids on moderating inconsistency among human decision-makers. Additionally, we propose novel ways to provide machine advice to human decision-makers which do not rely on the ground truth. Furthermore, we explore several notions of accuracy and inconsistency to test various interventions to moderate inconsistency among people.

- In Chapter 8, we discuss the implications and limitation of our work. Additionally, we discuss potential avenues of future works.

CHAPTER 2

Background

In this section, we present a brief background on different notions of fairness and how they are operationalized and implemented in ADMSs.

Socially salient groups. For different notions of groups fairness we consider socially salient groups in the population such as gender, race and sexuality, also referred to as sensitive features. Following the fairness in machine learning literature [275] [154], we use the terms sensitive group, protected group and socially salient groups interchangeably.

2.1 Different notions of fairness in ADMSs

Disparate treatment. Motivated by anti-discrimination laws, a prevalent notion of fairness is group fairness. This notion entails a *fair* distribution of favorable outcomes among socially salient groups. A popular measure of group fairness is disparate treatment, also referred to as 'direct discrimination'. This form of discrimination arises when sensitive group membership is directly utilized in an ADMS. Various learning methods have been proposed to integrate and address this measure [80, 89, 187], primarily by omitting the protected group.

Disparate impact. This popular measure of discrimination is also referred to as 'indirect discrimination'. *Disparate impact* requires that the beneficial outcomes rates among different groups of the sensitive feature should not be significantly different [60, 89, 275]. Several works in ADMSs have operationalized this notion of fairness as statistical parity (SP) or demographic parity (DP) [93, 144], which calls for statistical independence between the outcomes and the value of the sensitive attribute. Specifically, SP requires that the beneficial outcomes of the ADMS should be distributed proportionally among the protected groups. For example if 10% of the applicants for a job are women, 10% of the jobs should go to women. This measure of fairness tries to address the historical biases, specially in cases where we do not have reliable ground truth data.

Equal opportunity. On the other hand, in cases where we have access to ground truth, e.g., which candidates are actually suitable for the job, *equality of mistreatment or equality of opportunity* (EOP) has been proposed [120, 153, 274]. This notion requires that true positive rates, i.e., classifying a person into positive class who belongs to the positive class, should be equal across different groups of the sensitive attribute. A similar notion of fairness is also proposed by Hardt et al. [120], namely *equalized odds*(EOD), where they argue that the true positive rates and the false positive rates should be equal across different groups of the sensitive attribute.

Preferred treatment/impact. Zafar et al. [273] extended the group fairness from equality among the socially salient groups to *preferred treatment*. Specifically, they propose group-conditional classifiers such that each groups receives the most beneficial outcomes by their own classifier. In some cases parity based notions of fairness might reduce the benefits for one group without necessary increasing the benefits for the other group, in such cases this notion of might be more acceptable by all the groups.

Individual fairness. Beyond the group fairness approaches, there also exist work in ADMSs on individual fairness. The notion of *individual fairness* was introduced by Dwork et al. [80] in learning based systems. They operationalize the principle 'similar individuals should be treated similarly' by considering an ADMS as a mapping function from the input feature space to the decision space and argue that the individuals who are close together in the feature space should also be close in the decision space. However, Dwork et al. [80] rely on a given distance metric. Zemel et al. [278] further extend this work by providing a method to learn a distance metric.

Counterfactual fairness. Kusner et al. [162] introduced the notion of *counterfactual fairness*. They argue that the decisions should be fair for an individual in the actual world as well in the counterfactual world. Consider an example of university admissions where applicants belong to different races. Counterfactual fairness posits that changing the race of applicant, in the causal graph, should not impact the decision of the system.

Diversity. In addition to parity based notion of fairness, there have also been notions of *diversity*, especially for information retrieval applications [276]. This notion demands that different socially salient groups should be 'well-represented', especially for applications such as ranking. This notion of fairness is of particular interest when retrieving a small subset of items from a large dataset. In such cases it is important to capture the entire support of the distribution.

Fairness of features. There is also much research on human perception of the fairness in ADMSs. Specifically, in a study Grgic-Hlaca et al. [115] found that using features that do not have a causal relationship to the outcomes or are not volitional were considered

unfair. Additionally, Grgić-Hlača et al. [114] found that people's perception of the fairness of a feature depends on their political views and demographic features.

In this thesis, we address different notions of group fairness and individual fairness for evaluating ADMSs for (un)fairness, fairly updating ADMSs and designing fair ADMSs.

2.2 Approaches of achieving fairness

A learning based ADMS constitutes utilizing features from input datapoints to generate decisions through an algorithm. For instance, consider an algorithmic system tasked with classifying applicants' CVs into rejection or potential hire categories. Suppose we aim to ensure fairness in this system, specifically to eliminate gender discrimination. This objective can be pursued through three methods: i) pre-processing the input data to prevent the inference of gender information, ii) training a fair classifier with explicit constraints that enforce non-discrimination, or iii) taking the classifier output and redistributing favorable outcomes in a non-discriminatory manner. These processes align with i) pre-processing, ii) in-processing, and iii) post-processing approaches for achieving fairness in an ADMS.

Below we briefly describe popular works in each of these approaches.

2.2.1 Pre-processing

Pre-processing approaches constitute methods which try to alter the input data to achieve a prescribed fairness metric, e.g., by relabelling or reweighing the input data or by finding different data representations. Seminal works in achieving fairness through pre-processing are described below.

One pre-processing approach comprises of relabelling the data. Luong et al. [187] proposed a method to detect discrimination in a KNN-based binary classification system by examining pairs of similar points, differing only in their protected group membership. To mitigate this discrimination, they modify the class labels of discriminated datapoints in the training set. Kamiran and Calders [143] proposed a *relabelling* approach for binary classification tasks. In this approach, for a given classifier (e.g., an accuracy maximizing classifier) they relabel datapoints near the decision boundary in the training data until there is no discrimination. These approaches are focused on enforcing statistical parity.

Another class of methods *reweighs* the dataset instead of changing its contents. Calders et al. [43] introduced a reweighing approach across different groups of the sensitive attributes and the ground truth class labels in order to eliminate discrimination

in the dataset. For instance, in a biased dataset against the protected class, they proposed increasing the weights of datapoints belonging to the protected group with favorable outcomes. Other reweighing methods include works by Krasanakis et al. [157] and Jiang and Nachum [134].

Kamiran and Calders [143] also introduced a *preferential sampling* method. For a given dataset and a classifier trained on it, their method involves resampling datapoints near the decision boundary to build a fair dataset. For instance, in a dataset biased against the protected class, they suggest oversampling datapoints belonging to the protected groups with favorable outcome and undersampling datapoints belonging to the unprotected groups with favorable outcomes. This method tries to enforce statistical parity. Other approaches based on resampling includes Iosifidis et al. [132] and Zelaya et al. [277].

A popular method for achieving fairness through pre-processing is *learning fair representations*. Zemel et al. [278] propose to learn a 'fair' mapping of a given dataset by optimizing to learn faithful representations while obscuring sensitive feature group membership. These representations can be employed for various downstream tasks, including classification and regression problems. As a hybrid of pre-processing and in-processing approach, Zemel et al. [278] also presented a method to *jointly* learn a classifier while learning fair representations. Their work demonstrated that this jointly learned classifier alleviated both statistical disparity and individual unfairness.

Generative models learn representations for a given dataset for the purpose of generating new data, which can subsequently be utilized in various downstream applications. Generative models can be used in diverse application for fairness in ADMS, primarily for pre-processing approaches. Some examples include augmenting imbalanced classes [279], generating counterfactuals [119, 188], achieving fairness without access to sensitive attributes [109] and enhancing fairness for collaborative filtering in recommender systems [34]. There are two popular generative models i) Variational autoencoders (VAEs) [152] and Generative adversarial networks (GANs) [104]. Both of these models have been studied for potential biases and corresponding bias mitigation strategies have been proposed by Louizos et al. [186] and Xu et al. [270].

A key advantage of pre-processing is its task-agnostic nature, applicable to various downstream tasks. However, it solely addresses bias in the data, neglecting other sources of bias like algorithmic bias [196]. Moreover, pre-processing requires access to and permission to modify the data. Another drawback includes challenges in interpretability of the modified features, making it difficult to trace the importance of original features in the ADMS.

2.2.2 In-processing

The in-processing approach involves methods that explicitly integrate fairness constraints into the optimization problem of the algorithm within an ADMS. The following summarizes notable in-processing methods for achieving fairness in ADMS.

Zafar et. al propose several in-processing methods to enforce fairness in linear and non-linear boundary based classifiers (e.g., logistic regression and SVM). Their methods entail converting fairness constraints into disciplined convex programs (DCP) or disciplined concave and convex programs (DCCP) [233] that can be efficiently incorporated into popular classifiers during training. To mitigate disparate impact in binary classification, Zafar et al. [275] propose constraints that threshold covariance between the decision boundary and the sensitive features. In other words, they train classifiers where decision boundary is uncorrelated with the sensitive feature values. Furthermore, they find that decreasing the covariance results in lower accuracy hence creating an accuracy-fairness trade-off. This method does not use the sensitive feature values during test time and reduces disparate treatment. To enforce equality of opportunity, Zafar et al. [274] propose covariance-based constraints that enforce parity of false positive rates or false negative rates across different groups of the sensitive attribute. As before, this method does not use the sensitive feature value at the test time. To address preference-based notions of fairness, Zafar et al. [273] train sensitive group conditional classifiers using DCCP constraints. This approach relies on knowing the sensitive group membership at the test time so it does not satisfy disparate treatment.

Calders et al. [44] propose methods to tackle fairness issues in regression problem. Similar to demographic parity, they propose that similar individuals from different sensitive feature groups should have similar mean prediction values regardless of their ground truth values. Similar to equal opportunity, they propose that the residual errors across different groups of the sensitive attributes should be similar. They propose to solve linear regression with the corresponding constraints and propose closed form solutions for each problem. In a more recent work, Agarwal et al. [5] propose fair regression where they incorporate demographic parity in regression problem and solve the proposed constrained regression problem by converting it into a constrained classification problem.

In order to train counterfactually fair classifiers Kusner et al. [162] rely on a given causal graph for the dataset. Using this causal graph, they propose to train counterfactually fair classifier at three increasing levels of assumptions and accuracy, requiring varying degrees of knowledge about the causal model and the unobserved variables influenced by sensitive features. The crux of their approach is to eliminate the effect

of the sensitive features on outcomes using a causal graph. Kilbertus et al. [151] also address counterfactual fairness but similarly rely on the knowledge about causal graphs.

A key benefit of in-processing approach to achieving fairness is the fine grained control it provides. In many cases, these methods provide a knob to trade-off between accuracy and fairness, allowing policymakers to select an appropriate accuracy/fairness trade-off [275]. Unlike pre-processing methods, in-processing methods directly address the bias caused due to the inductive bias of a model. Additionally, given the right fairness metric in-processing methods can produce fair outcomes regardless of the biased data. However, a major challenge lies in the need to operationalize and implement a custom in-processing method for each application scenario.

2.2.3 Post-processing

Post-processing approaches involve methods that take the output of an ADMS and redistribute outcomes to align with a specified notion of fairness. This section outlines notable works utilizing post-processing techniques for enhancing fairness.

Hardt et al. [120] propose a post-processing method to enforce the notions of equal opportunity (EOP) and equal odds (EOD) for classification tasks. Their approach is designed for binary prediction systems mapping input features to a score $R \in [0, 1]$ such as a logistic regression classifier. The method involves setting thresholds on the score to obtain predictions. They propose to find a threshold such that there is no disparity in EOP or EOD among different groups of the sensitive attributes. If such a universal threshold does not exist, they suggest employing group-specific thresholds for different sensitive feature groups. Their approach assumes knowledge of sensitive group membership at test time. Additionally, they also provide optimality guarantees, given a Bayes optimal regressor. Corbett-Davies et al. [65] also propose a similar approach to enforce statistical parity and EOP. Woodworth et al. [269] show that post-processing methods, like the one proposed by Hardt et al. (2016), can yield poor accuracy when the loss is not strictly convex and recommend a hybrid in-processing and post-processing approach to address this issue.

A recent work by Petersen et al. [212] introduces a post-processing method to enforce individual fairness. They frame the problem as a graph smoothing problem, treating datapoints as nodes and weighted edges as degree of similarity between points. They establish an equivalence between the graph smoothing problem and individual fairness, as outlined by Dwork et al. [79]. Specifically, higher edge weights in their method correspond to a greater encouragement of smoothness, i.e., nodes with higher weights between them are given similar outcomes. In other words, given the outcomes of the

model, the method seeks to achieve individual fairness by finding a new mapping of nodes to outcomes. This mapping maintains proximity to the previous one while ensuring local smoothness based on edge weights, ultimately assigning similar outcomes to similar individuals.

A notable advantage of post-processing approaches is their ability to operate without the need to design a new ADMS. These methods tend to treat the ADMS as a black box and can be applied to the outcomes of different types of systems. However, a drawback is the requirement for access to sensitive features during test time, which constitutes disparate treatment. Additionally, current methods do not provide a clear mechanism for balancing accuracy and fairness.

In this thesis, we primarily focus on in-processing methods.

CHAPTER 3

Fairness Evaluation of Discriminative Foundation Models

In this chapter, we look at the potential harms associated with classifying, retrieving and captioning image data using discriminative multi-modal foundation models, and ask a key question:

What constitutes the desired behavior for discriminative foundation models in downstream tasks?

As mentioned in the Section 1.2, our goal is challenging due to a combination of two factors: first, the rise and commoditization of zero-shot machine learning; and second, the plethora of inconsistent fairness definitions [255]. Intrinsically, zero-shot hinges on the idea that a single system should perform well on diverse unseen datasets without specialist training [165], while algorithmic fairness has consolidated on the idea that specific fairness definitions are more appropriate for specific tasks [255]. The intersection of these ideas creates a tension. Indeed, how can we check the fairness of a general-purpose system if we cannot agree on a general definition of fairness?

Rest of the chapter is organized as follows:

- In Section 3.1, we propose a new taxonomy to attempt to answer our research question. We also discuss the appropriate fairness metrics corresponding to our taxonomy.
- In Section 3.2, we provide a background on foundation models and Contrastive Language–Image Pre-training (CLIP). Additionally, we discuss popular bias mitigation methods for CLIP and new baselines that we evalaute.
- In Section 3.3, we discuss the details of the three tasks, i.e., image classification, image retrieval and image captioning, that we use for evaluating the fairness on CLIP and its bias mitigation methods.

- In Section 3.4, we discuss the fairness metrics corresponding to our taxonomy and the evaluation tasks.
- In Section 3.5, using our taxonomy, we provide a systematic evaluation of OpenAI's CLIP [220] and OpenCLIP [130] models, for binary (gender) and multi-valued (race) attributes.[2] Additionally, we evaluate a range of existing bias mitigation methods for these models. We perform these evaluations using *ten* diverse and large scale real-world datasets. We argue that existing fairness methods are designed to encourage either independence or diversity in the groups of the protected attribute, and show empirically that they prioritize one or the other. As such, the choice of a particular fairness method should be driven by the intended use case, and a decision as to which harms are relevant (Section 3.4).
- In Section 3.6, we discuss additional related work and, finally in Section 3.7, we conclude the chapter.

Relevant publication

The results presented in this chapter have been published in [11].

3.1 New Taxonomy for (Un)Fairness Evaluation of Discriminative FMs

To address the research question, we propose a coarse taxonomy of tasks and describe the ideal behavior of a foundation model on such tasks. We base our taxonomy around three concepts:

(1) *Human centricity:* Do the labels concern humans?
Examples of human-centric tasks include classifying photos of people into different professions, retrieving pictures of doctors from a set of images and captioning an image including people. Non-human-centric tasks includes task like classifying images of animals (e.g., cats vs dogs) or retrieving a particular type of car from a dataset.

(2) *Label consistency:* Is there likely to be an agreement on how data should be labeled both within a culture and across a wide range of cultures?

[2] As an artifact of the available datasets, we make use of annotations that indicate *perceived* gender and race. Labels are assigned coarsely by a third party into binary bins for gender and into seven racial groups (see [145] for details). They do not reflect how people in the dataset identify.

Table 3.1: The range of desiderata and their corresponding measures. The motivation underlying our desiderata is straightforward: where consistent labelings exist, we expect foundation models to reproduce them, and in human-centric tasks we should reproduce them equally well for all groups. Where labels are subjective (i.e., likely to be labeled inconsistently by different groups), reproducing labels is less of a concern, and instead we prioritize groups to be represented equally. The question then is what does 'equally' mean? For much of the fairness literature, 'equally' refers to the idea that decisions should be made independently of protected attributes such as race or gender (potentially conditioned on the true label). This leads to notions such as equal opportunity [120] (see "independence measures" in the top left part of the table) or demographic parity [143] ("independence measures" in the bottom left part of the table). However, this is not the only relevant notion of equal representation. In some cases, we may wish to sample uniformly from the *support of the distribution* rather than the distribution, and this leads to analogous notions provided under "diversity measures" in the table. By $Y, \hat{Y}, Z$ we denote a datapoint's ground-truth label, predicted label, and protected attribute, respectively; P denotes a generic probability distribution over these three variables.

	HUMAN-CENTRIC	NON-HUMAN-CENTRIC
Objective task	Labels should be reproduced consistently for all groups **Independence measures:** High performance per group on standard metrics and $P(\hat{Y}=1\|Z=z_1,Y=1)=P(\hat{Y}=1\|Z=z_2,Y=1)\ \forall z_1,z_2$ Figures 3.2 and 3.6 **Diversity measures:** High performance per group on standard metrics and $P(\hat{Y}=1\wedge Z=z_1\wedge Y=1)=P(\hat{Y}=1\wedge Z=z_2\wedge Y=1)\ \forall z_1,z_2$ Table 3.3	Labels should be reproduced consistently High performance on standard metrics Tables 3.2, 3.4, and 3.18
Subjective task	Labels should represent all groups equally **Independence measures:** $P(\hat{Y}=1\|Z=z_1)=P(\hat{Y}=1\|Z=z_2)\ \forall z_1,z_2$ Figures 3.1, 3.3, 3.4, 3.5, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12, 3.13 3.14, 3.15 and 3.16. Tables 3.9, 3.10, 3.11, 3.12, 3.13, 3.15, and 3.17. **Diversity measures:** $P(\hat{Y}=1\wedge Z=z_1)=P(\hat{Y}=1\wedge Z=z_2)\ \forall z_1,z_2$ Tables 3.5, 3.6, 3.7, 3.8, 3.14, 3.16	*Out of scope*

Based on the answer to this question, we classify tasks into subjective and objective categories. We consider labeling tasks objective when there is likely to be a high agreement among different groups regarding the outcome. For instance, in a picture where a person is wearing a lab coat and a stethoscope, surrounded by medical equipment, individuals from diverse backgrounds would probably agree that it is a picture of a doctor, not an engineer or a firefighter. On the other hand, labeling the same image as 'a picture of a doctor' versus 'a picture of a nurse' might result in more disagreements among diverse labelers across cultures.

However, we acknowledge that this definition is not foolproof and it is difficult to quantify the objectivity of some tasks, as it does not imply within group disagreement. For example groups of labelers may consistently label data in a way that other people would disagree with. For example, Microsoft discontinued their services in the Azure system that infers emotional state, stating that "Experts inside

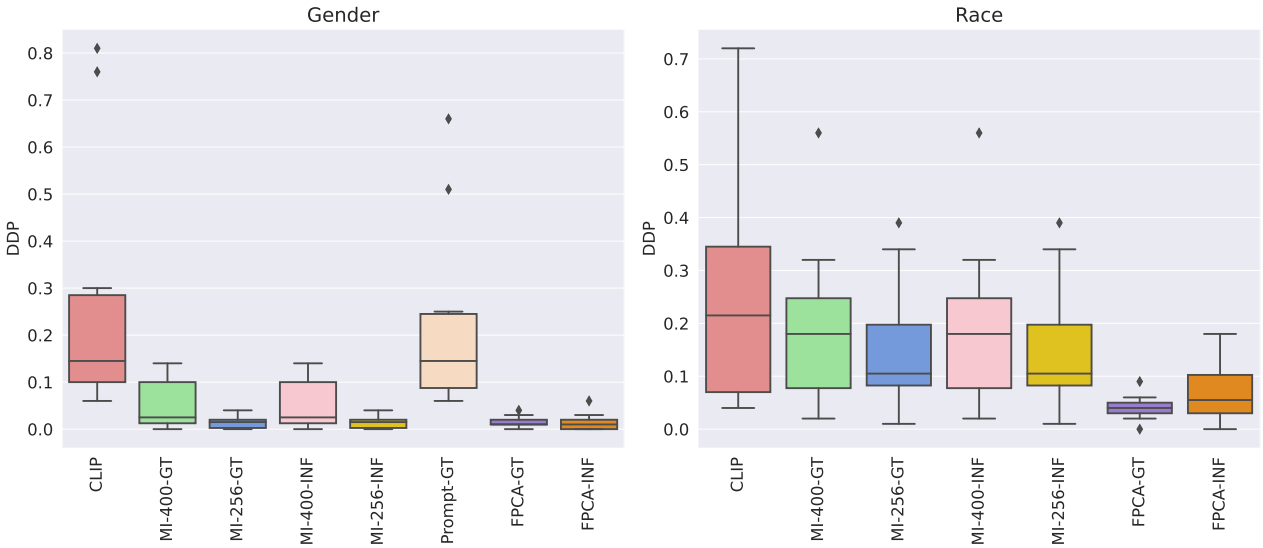


Figure 3.1: [Classification - DDP - Subjective - FairFace] We plot DDP, given in Eq. (3.5) for gender (left) and race (right), summarizing the distribution over multiple zero-shot classification tasks (provided in Appendix A.1) using FairFace dataset. *"GT" and "INF" refers to whether the value of the protected attributes used to train the corresponding method were ground truth or inferred using CLIP. These figures shows that fair PCA based methods are more effective in reducing demographic disparity for different groups of the protected attributes. Additionally, mutual information based methods are more effective when more dimensions are reduced.*

and outside the company have highlighted the lack of scientific consensus on the definition of "emotions""[3].

(3) *Purpose of the task:* Can the task be perceived to be assigning labels to individuals, or to be recovering diverse samples that characterize the spread of data?

While fairness typically concerns itself with the harm to an individual or groups of individuals that a decision is being made about, for example, the harm induced by failing to offer someone a loan, schedule follow-up medical treatment, or in hiring someone, other harms are possible. For example, if someone intends to use images of scientists for recruiting materials, it is often desirable to show diverse images capturing scientists of a range of races and genders, i.e. capturing the support of the distribution. Repeatedly failing to capture the entire support can discourage some people viewing the images, from considering becoming scientists as they might feel that scientists are not people like them, referred to as the role model effect [45].

[3] https://blogs.microsoft.com/on-the-issues/2022/06/21/microsofts-framework-for-building-ai-systems-responsibly/

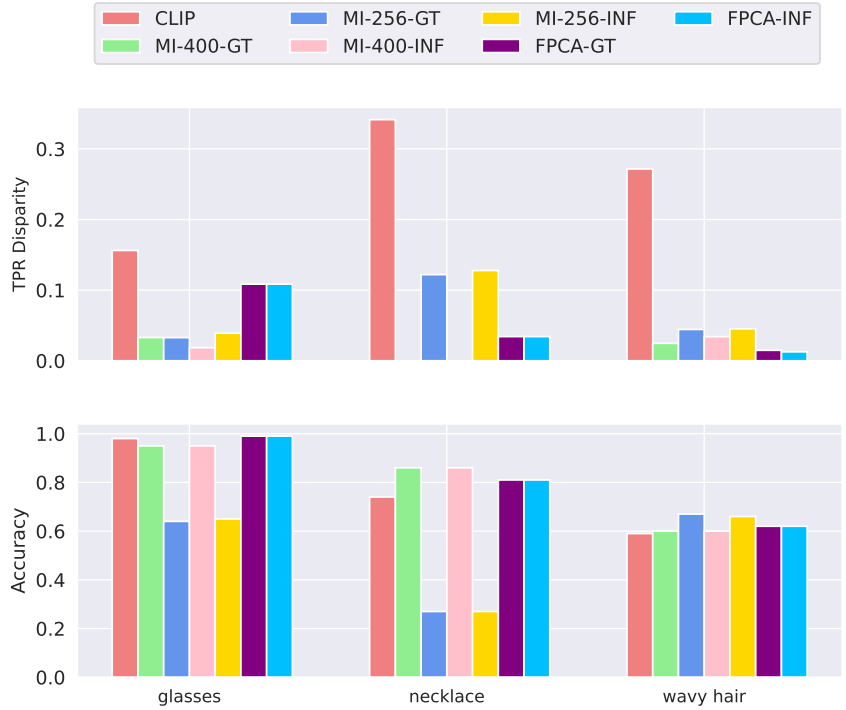


Figure 3.2: [Classification - DTPR - Objective - CelebA] The plots show the TPR disparity, given by Eq. (3.7), between men and women for three zero-shot classification tasks using the CelebA dataset on top and the accuracy on the bottom. *The results demonstrate that mutual information and fair PCA based methods reduce disparity. However where the dimension of the CLIP embeddings is reduced significantly, using mutual information based methods, accuracy can also lower significantly.*

3.1.1 Measures based on our taxonomy

Based on the answers to the questions above, we discuss the measures that encode the values implicit in these decisions (see Table 3.1).

Importantly, we find that different answers to these questions naturally lead to different measures. Consequently, we observe that many of the existing works in fairness for foundation models, which propose new methods evaluated with respect to particular measures, are enforcing unexamined value judgments about what the ideal behavior should be. Moreover, as part of the taxonomy depends not only on the type of task but also on the purpose, it is impossible to satisfy all measures simultaneously.

3.1.1.1 Human-Centric Tasks

Let $Y \in \{0, 1\}$, $\hat{Y} \in \{0, 1\}$, $Z \in \{0, 1...k\}$ be the ground-truth label, predicted label and the protected attribute. Let P be the probability distribution over all the three variables.

- **Objective Tasks:** In these task, the assumption is that the ground truth Y is available.

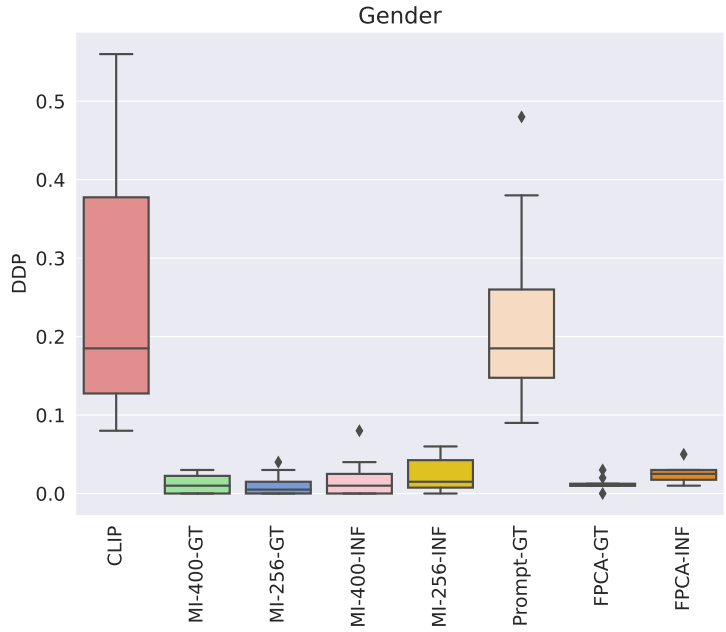


Figure 3.3: [Classification - DDP - Subjective - Flickr30k] Using Flickr30K dataset, this figure shows box plots of DDP, given by Eq. (3.5), for several subjective zero-shot classification tasks. *Most methods effectively reduce classification bias, except for the prompt based method. One reason could be that the model provided by the authors was trained to have a higher importance for maintaining representational powers of the embedding (itc loss: Section 3.2.3.2) as opposed to reducing bias.*

– **Independence Assumption:** This reflects the typical fairness concerns that relate to decisions made about individuals, where the independence of outcome w.r.t. protected attribute is desirable. This leads to the notions of equal opportunity [120, 274], i.e., the probability of the positive prediction should be independent of the protected group membership for ground truth positive datapoints, i.e.,

$$P(\hat{Y} = 1|Z = z_1, Y = 1) = P(\hat{Y} = 1|Z = z_2, Y = 1)\ \forall z_1, z_2 \tag{3.1}$$

Consider a classification system tasked with identifying 'car mechanics' from a set of images that includes both men and women. Given the traditional perception of car mechanics as a predominantly male profession, there is a possibility that women may not be identified as car mechanics. In such cases, we anticipate that the true positive rate should be consistent across genders.

– **Diversity Assumption:** In certain downstream usage of image retrieval lack of diversity could be deemed as unfair [276]. We require that the probability of accurately retrieving a subset of images should be uniform across all the

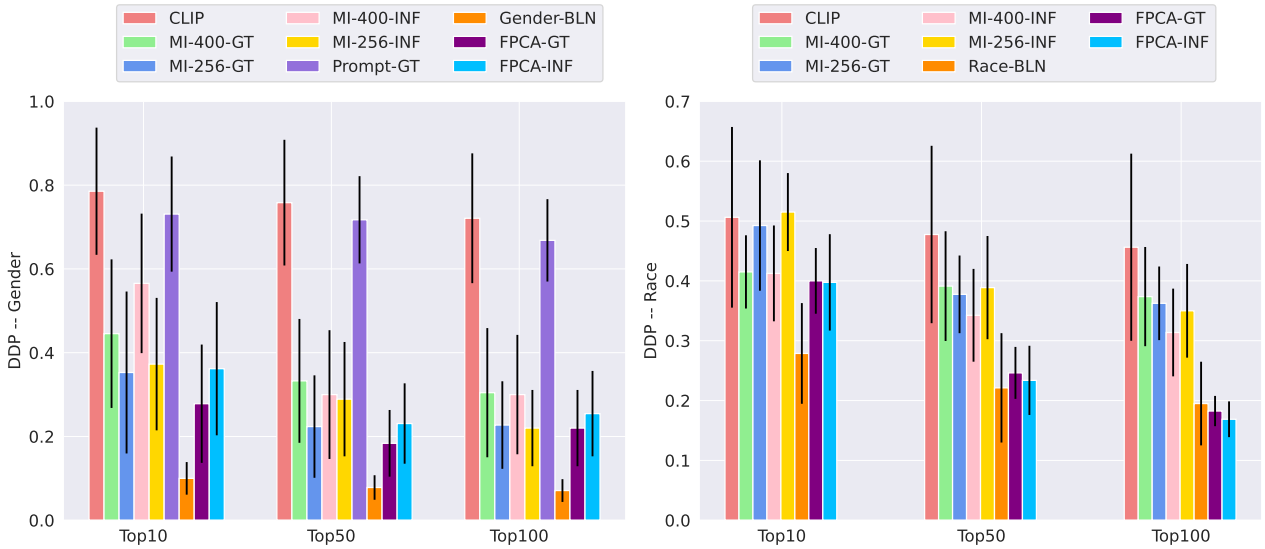


Figure 3.4: [Retrieval - DDP - Subjective - FairFace] These figures show the average DDP, given by Eq. (3.6), for gender (left) and race (right) attributes averaged over several image retrieval tasks, given in Appendix A.1, using the FairFace dataset. *The results demonstrate that protected attribute specific queries and fair PCA based methods do well in removing bias for image retrieval tasks. Mutual information based methods also perform well for the gender attribute.*

groups of the protected attributes, i.e.,

$$P(\hat{Y} = 1 \wedge Z = z_1 \wedge Y = 1) = P(\hat{Y} = 1 \wedge Z = z_2 \wedge Y = 1) \; \forall z_1, z_2 \tag{3.2}$$

Let's take an example of retrieving 10 images of astronauts from a dataset containing 90% male astronauts and 10% female astronauts. Now, imagine using these images in a presentation aimed at motivating school children to pursue STEM. If we show 9 images of male astronauts and only 1 of a female astronaut, aligning with the dataset's proportions, it could potentially be demotivating for young girls. Therefore, for such applications, we aim to retrieve an equal number of relevant images for different groups based on the protected attribute.

- **Subjective Tasks:** For these tasks the ground truth labelings are not available.
  - **Independence Assumption:** In the case, where ground truth is not available in fair learning this leads to the notion of demographic parity (DP) [274]. It requires that the classification into positive class should be independent of the protected attribute, i.e.,

$$P(\hat{Y} = 1 | Z = z_1) = P(\hat{Y} = 1 | Z = z_2) \; \forall z_1, z_2 \tag{3.3}$$

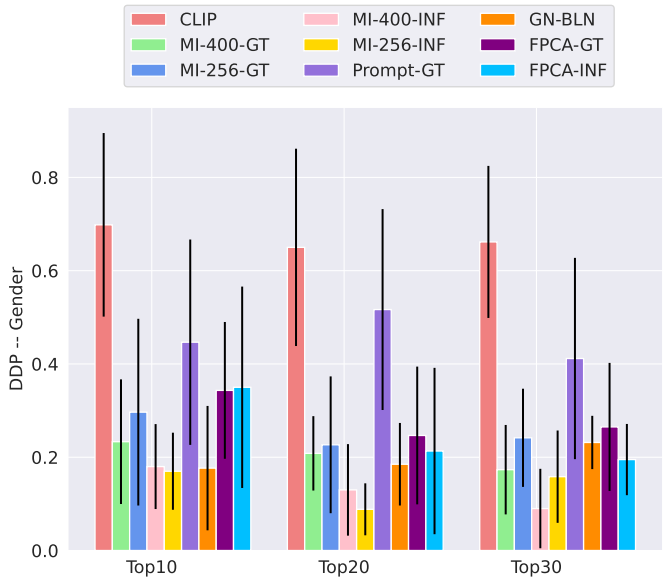


Figure 3.5: [Retrieval - DDP - Subjective - Flickr30k] The plot shows the DDP, given by Eq. (3.6), for gender attribute using Flickr30K dataset. *All the methods, except the prompt based method, decrease the disparity between men and women for the retrieval tasks.*

Consider the example of classifying images of men and women into doctors vs nurses. Given the close relation between both the professions and the difficulty of establishing an objective ground truth from images alone, in these cases, we expect that the classification system to adhere to demographic parity.

– **Diversity assumption:** To enforce diversity in some of the application of image retrieval application where we do not have the ground truth available, we expect that the retrieved images equally represent all the groups of the protected attribute, i.e.,

$$P(\hat{Y} = 1 \wedge Z = z_1) = P(\hat{Y} = 1 \wedge Z = z_2)\ \forall z_1, z_2 \tag{3.4}$$

Consider the example of retrieving images of 'a beautiful person' for an advertising campaign from a dataset comprising images of different races. Since the notion of the beauty is subjective across cultures, it would make sense to display an equal number of images from all races.

3.1.1.2 Non-Human-Centric Tasks

Examples of tasks that are not centered around humans include classifying animals (distinguishing between dogs and cats), retrieving images of objects (such as pencils, houses, etc.), and captioning images that do not contain people. We consider harms

Table 3.2: [Classification - Accuracy - Objective - StanfordCars, Food-101, VOC objects & Imagenet] *The bias mitigation methods shown in the table were trained using the FairFace Dataset. We used the test splits for all the datasets. The results show that fair PCA based methods retain performance on non-human objective tasks. We would like to note that we only show results with a prompt of "a photo of a {label}", while the original CLIP paper aggregates results using several prompts, which they did not disclose. In some cases this can result in a difference in evaluation numbers that we are reporting compared to the original CLIP paper. However, our results are within the margin of improvement that the original CLIP paper claims to achieve using prompt engineering.*

Mitigated	Dataset	Backbone	CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	Prompt-GT	FPCA-GT	FPCA-INF
Gender	Food-101	ViTB/32	**82.3**	79.2	67.6	79.3	67.0	–	**82.3**	**82.3**
Race	Food-101	ViTB/32	**82.3**	77.7	66.3	77.7	68.6	–	81.5	81.5
Gender	Food-101	ViTB/16	87.0	85.1	76.6	85.0	76.0	**87.3**	87.1	87.0
Race	Food-101	ViTB/16	**87.0**	85.1	76.5	85.0	77.6	–	86.3	86.4
Gender	StanfordCars	ViTB/32	**60.2**	53.6	44.9	53.5	46.1	–	60.1	**60.2**
Race	StanfordCars	ViTB/32	**60.2**	54.4	43.0	55.2	43.8	–	60.0	59.5
Gender	StanfordCars	ViTB/16	**65.6**	59.7	50.2	61.3	51.8	64.7	65.3	65.3
Race	StanfordCars	ViTB/16	**65.6**	59.8	49.0	61.7	48.8	–	65.3	65.4
Gender	VOC	ViTB/32	**83.8**	83.0	77.0	82.3	74.9	–	83.7	83.7
Race	VOC	ViTB/32	83.8	82.7	65.8	83.3	63.9	–	84.5	**84.6**
Gender	VOC	ViTB/16	**85.7**	76.6	67.9	76.3	71.7	82.9	85.6	85.7
Race	VOC	ViTB/16	85.7	87.9	76.5	**89.0**	75.8	–	85.7	85.3
Gender	Imagenet	ViTB/32	**59.2**	54.4	37.1	54.3	37.5	–	**59.2**	**59.2**
Race	Imagenet	ViTB/32	**59.2**	53.5	34.6	53.7	34.8	–	58.9	58.9
Gender	Imagenet	ViTB/16	**63.8**	55.4	40.3	55.5	41.2	63.2	**63.8**	**63.8**
Race	Imagenet	ViTB/16	**63.8**	58.3	43.4	58.2	43.4	–	63.5	63.6

associated with these tasks to be beyond the current scope, even though they certainly can exist. For instance, labelings of sacred places (such as churches, mosques, and temples) should be respectful. We defer the investigation of such tasks to future studies. In the case of these tasks, we expect that the performance should not deteriorate when evaluated using standard metrics.

3.2 Foundation Models, CLIP, and Fairness of CLIP

In the past few years, *large* models trained on huge amounts of data, primarily crawled from the internet, have become popular (e.g., BERT [74], CLIP [220], GPT-3 [38], DALL-E [222], Stable Diffusion [227]). Many of these models have gained attention even in the general public and extensive news coverage, which typically also addresses the risks and shortcomings of these models (e.g., [197, 211]). These large models are now commonly referred to as foundation models, a name coined by researchers from Stanford to "underscore their critically central yet incomplete character" [32]. They exist in various flavors that cover a wide range of data modalities (e.g., language, vision or multi-modal), training objectives (e.g., predicting a word deleted from a piece of text or aligning images and their captions in a joint embedding space) and application areas (e.g., data generation tasks such as image synthesis or data analysis tasks such as image

Table 3.3: [Retrieval - DDP & Precision - Objective - IdenProf] This table shows fairness evaluation for representational bias on objective tasks for image retrieval of CLIP model and different bias mitigation methods. *Using IdenProf dataset, we show DDP-rep, given by Eq. (3.9), for each method as well as its average precision for retrieving images of 9 different professions of the IdenProf dataset. We exclude the profession 'Firefighters' because in many cases their faces are hidden and gender is difficult to identify. Additionally, we do not show results for EOP like measure because this dataset does not have the annotations for the gender attribute. The gender annotations for the retrieved images per profession were manually done by one of the authors. The results demonstrates that gender balanced queries perform the best to reduce the representational unfairness in the objective tasks. All the methods are trained on FairFace dataset to remove the gender bias.*

Clip	MI-400-GT	MI-256-GT	Prompt-GT	Gender-BLN	FPCA-GT
			DDP(rep) @ 10		
0.80±0.05	0.61±0.07	0.55±0.08	0.73±0.07	**0.22±0.10**	0.49±0.10
			DDP(rep) @ 20		
0.66±0.06	0.46±0.08	0.49±0.09	0.63±0.07	**0.19±0.07**	0.44±0.10
			DDP(rep) @ 30		
0.63±0.06	0.49±0.06	0.49±0.06	0.62±0.04	**0.24±0.07**	0.39±0.09
			Precision @ 10		
0.99±0.02	**1.00±0.00**	0.99±0.02	0.97±0.07	0.99±0.02	**1.0±0.0**
			Precision @ 20		
0.98±0.04	**0.99±0.01**	0.97±0.03	0.97±0.06	0.97±0.05	0.98±0.02
			Precision @ 30		
0.97±0.04	**0.98±0.02**	0.96±0.04	0.96±0.06	0.97±0.05	**0.98±0.04**

classification, retrieval or captioning). What foundation models have in common is that they were trained on broad data, where the quantity of data was prioritized over its quality, and that they can be adapted to a wide range of downstream tasks, often with no or only minimal supervision. The former property makes foundation models prone to concerning behavior, ranging from algorithmic bias [220] over toxicity and offensive content [58] to privacy concerns [47]. The latter property increases the risk that any concerning behavior could spread much wider than with a traditional model trained to solve a specific task.

In this section, we briefly describe the required background of the CLIP model as an illustration of a typical discriminative foundation model and relevant fairness concerns. We discuss additional related work in Section 3.6.

Table 3.4: [Retrieval - Recall - Flickr30k] The table below shows recall@K for randomly selected 50% Flickr30K dataset using different gender bias mitigation methods. *Specifically, we are using the captions of each image as a query and report the fraction queries that retrieve the images correctly in top 1, 5 or 10 results. The results show that mutual information based methods perform worse, which makes sense as the number of dimensions are reduced, while Prompt-GT method performs the best. Since the Prompt-GT method was finetuned using the Flickr dataset, it is not surprising that it outperforms even the CLIP model. It is worth noting that the queries also include gendered queries and some reduction in recall is expected or may even be desirable.*

CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	Prompt-GT	FPCA-GT	FPCA-INF
				ViTB/32 Top 1			
0.29	0.19	0.13	0.18	0.12	–	0.26	0.26
				ViTB/16 Top 1			
0.32	0.23	0.15	0.23	0.15	**0.35**	0.29	0.29
				ViTB/32 Top 5			
0.51	0.38	0.27	0.37	0.27	–	0.48	0.48
				ViTB/16 Top 5			
0.55	0.42	0.31	0.42	0.30	**0.59**	0.51	0.51
				ViTB/32 Top 10			
0.62	0.48	0.35	0.46	0.35	–	0.58	0.58
				ViTB/16 Top 10			
0.65	0.51	0.39	0.51	0.38	**0.69**	0.61	0.61

3.2.1 Contrastive Language Image Pretraining (CLIP)

OpenAI's CLIP [220] is a discriminative foundation model for computer vision trained on 400 million image-text pairs to align corresponding image and text examples within a joint embedding space. To that end, CLIP uses a contrastive loss which tries to push the representations of the corresponding image and text examples together and the representations of the non-corresponding examples far apart. This joint multi-modal embedding space can then be used for several downstream tasks such as image retrieval, image captioning or zero-shot classification. CLIP achieves remarkable zero-shot classification performance in several tasks, which in some cases rivals that of the classical supervised competitors. In certain scenarios, the downstream applications could result in direct harm to individuals, e.g., classifying images into professionals vs non-professionals, retrieving a set of doctors from a dataset or captioning images for assisting blind people, which give rise to several fairness concerns. While OpenAI's CLIP is proprietary, we also present results (Section 3.5.6) for its open source implementation OpenCLIP [130]. OpenCLIP has the same objective function and architecture as the original OpenAI CLIP, but it was trained on the publicly available LAION-400M dataset [230].

Table 3.5: [Retrieval - Skew - Subjective - FairFace] This table shows the maximum absolute skew, given by Eq. (3.8), using the FairFace dataset and gender attribute. *It demonstrates that all the methods are able to reduce the skew. Gender balanced queries yield the lowest skew.*

CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	Prompt-GT	Gender-BLN	FPCA-GT	FPCA-INF
				Top 10				
2.47±0.86	0.84±0.68	0.67±0.7	1.06±0.64	0.51±0.3	2.12±0.88	**0.08±0.06**	0.36±0.2	0.51±0.28
				Top 50				
1.99±0.62	0.4±0.26	0.24±0.14	0.37±0.24	0.32±0.2	1.6±0.56	**0.06±0.02**	0.19±0.1	0.23±0.12
				Top 100				
1.64±0.48	0.38±0.3	0.24±0.12	0.33±0.24	0.2±0.12	1.3±0.36	**0.04±0.02**	0.23±0.12	0.26±0.12

3.2.2 Existing (Un)Fairness Evaluations of CLIP

Recent works highlighted some biases present in the CLIP model. The original CLIP paper [220] demonstrated gender and race biases in certain zero-shot tasks including classifying facial images into crime-related vs. non-crime-related categories or into human vs. non-human animal categories. These fairness evaluations were limited in scope to a small number of tasks and datasets.

Wang et al. [262], Berg et al. [25] and Dehouche [70] demonstrated that CLIP embeddings have a gender or race bias in certain tasks. In their study, Wang et al. [262] highlighted gender bias in CLIP embeddings when used for image retrieval tasks. In their experiments, they first created gender-neutral test queries by replacing the gendered words with neutral alternatives in the captions of the MSCOCO 1K test set. Subsequently, they utilized the CLIP embeddings to retrieve images based on these neutral queries. Their findings reveal that, on average, 6.4 out of top 10 results were images of men. However, it is important to consider a few factors while considering their results. i) They did not provide additional metrics that account for differences in the base rate of men and women. ii) They did not evaluate the fairness of CLIP embeddings using well-known fairness measures, such as demographic parity or equality of opportunity. iii) Their approach involved aggregating the signed biases of all queries. This aggregation method can potentially lead to the cancellation of systematic biases across different queries, thereby reducing the apparent bias of the system. For instance, if a search for 'home-maker' predominantly returns women and a search for 'technician' predominantly returns men, aggregating the two together suggests greater gender neutrality than when considering any one on its own.

Berg et al. [25] have also raised concerns in gender-related fairness issues of the CLIP embeddings. Their findings indicate that the CLIP model exhibits a representation bias

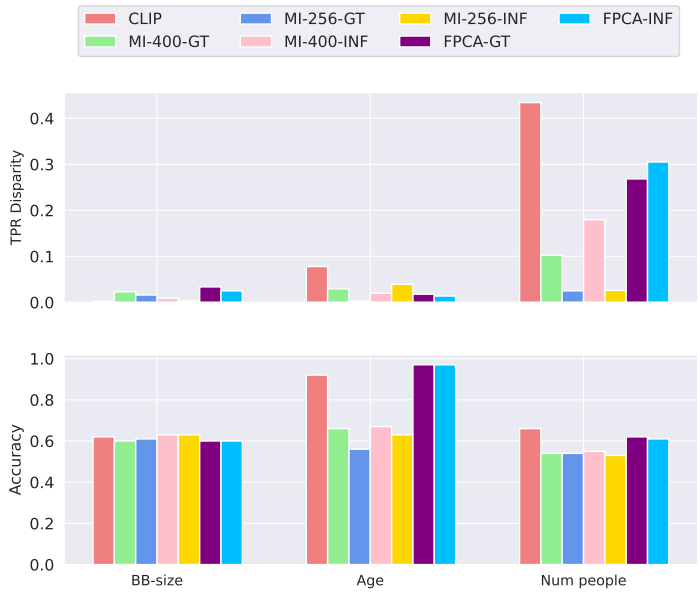


Figure 3.6: [Classification - DTPR - Objective - MIAP] *The x-axis shows three classification tasks: i) 'inconspicuous photo of a person' vs 'prominent photo of a person', where ground truth was based on whether the bounding box of the person occupied more than 50% of the image. ii) 'child' vs 'adult' iii) 'one person' vs 'more than one person'. On top we show the disparity in the true positive rates across the gender attribute and in the bottom we show the accuracy. We see that mutual information based methods while in some cases do reduce the disparity but they incur a reduction in accuracy. On the other hand fair PCA based methods reduce the disparity while incurring almost no loss in accuracy.*

with respect to gender in image retrieval tasks, particularly for queries such as clever, lazy, hardworking, kind, or unkind. However, it is worth noting that their analysis is limited to the face-focused FairFace and UTKFace datasets. Additionally, their evaluation of zero-shot classification was limited to the classification categories presented in the original CLIP paper [220]. Another aspect that their analysis is missing is the evaluation on well-established fairness metrics such as demographic parity and equal opportunity. Instead, they primarily focus on ranking metrics like Skew [98] and KL-divergence.

Dehouche [70] studied the fairness of CLIP by performing zero-shot classification to classify 10000 synthetically generated portrait photos into male vs. female, white person vs. person of color, attractive vs. unattractive, friendly vs. unfriendly, rich vs. poor, and intelligent vs. unintelligent. They found a strong correlation between classification as female and attractive, between male and rich, and between white person and attractive. They applied the strategy of Bolukbasi et al. [31] for debiasing word embeddings, by removing gender bias, and found that this strategy reduced the correlation between classification as female and attractive or between male and rich. Compared to Dehouche [70], we perform a more extensive fairness evaluation, considering not only zero-shot

Table 3.6: [Retrieval - Skew - Subjective - FairFace] This table shows the results for representation bias for subjective labelling. Specifically, it show skew metric , given by Eq. (3.8), for the *race* attribute of FairFace dataset. *Race balanced queries perform well in general but fair PCA based methods perform the best when the number of retrieved items are larger.*

CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	Race-BLN	FPCA-GT	FPCA-INF
			Top 10				
2.66±0.0	2.66±0.0	2.66±0.0	2.46±0.4	2.66±0.0	**1.56±0.84**	2.66±0.0	2.66±0.0
			Top 50				
2.49±0.34	2.23±0.36	2.05±0.4	1.88±0.6	1.91±0.52	**1.09±0.68**	1.66±0.56	1.38±0.52
			Top 100				
2.2±0.48	1.85±0.5	1.84±0.5	1.71±0.48	1.45±0.3	1.15±0.78	1.06±0.3	**0.89±0.2**

classification but also image retrieval and image captioning, and we compare several bias mitigation methods.

3.2.3 Bias Mitigation Methods for CLIP

In this section, we discuss two existing bias mitigation methods explicitly proposed for CLIP and the modifications we make to run them. To our knowledge, this is an exhaustive list — it contains every method claiming to improve the fairness of CLIP at the time of the submission of our paper. We also discuss a recently introduced version of fair PCA [154], which is a general approach to make representations fair and which we investigate in our experiments. In Section 3.6 we discuss concurrent works for debiasing CLIP.

3.2.3.1 CLIP-clip (referred to as MI in the results)

Wang et al. [262] proposed a simple post-processing approach to make CLIP representation fair w.r.t. gender. Given a dataset with gender annotations, they calculate the mutual information between CLIP embedding on the training split of the dataset and its corresponding values of the gender attribute. Then, they greedily select a prescribed number of dimensions with the highest mutual information to cut, and retain the rest of the m dimension in the CLIP representations. The smaller the value of m, the more debiased the CLIP representations, as shown in Figures 3.1, 3.2, 3.4 and 3.5. However, the performance using the reduced CLIP embeddings worsens on several non-gender related tasks, as shown in Tables 3.2, 3.3, 3.4, 3.13 and 3.18. This demonstrates the well-known accuracy-fairness trade-off.

Table 3.7: [Retrieval - Skew - Subjective - Flickr30K] This table shows the skew metric, given by Eq. (3.8), for the gender attribute average over several image retrieval task using the Flickr data. *It shows that gender balanced queries and mutual information based methods with a lot reduction in number of CLIP dimensions reduce the skew the most.*

CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	Prompt-GT	Gender-BLN	FPCA-GT	FPCA-INF
				Top 10				
2.28±1.12	0.6±0.28	0.71±0.22	0.9±0.38	0.47±0.16	2.08±1.3	**0.44±0.04**	1.25±0.92	1.2±0.94
				Top 20				
1.76±0.86	0.77±0.54	0.68±0.1	0.92±0.46	0.44±0.18	1.69±0.92	**0.32±0.04**	0.72±0.24	0.6±0.18
				Top 30				
1.52±0.62	0.64±0.28	0.69±0.22	0.87±0.6	0.52±0.1	1.11±0.52	**0.27±0.08**	0.66±0.28	0.53±0.16

Wang et al. [262] did not show results using non-binary (e.g. race) attributes. We extend their method to the multi-valued attributes and show results using the race attribute (see Figures 3.1 and 3.4).

3.2.3.2 Prompt Learning (referred to as Prompt in the results)

Berg et al. [25] proposed a method to reduce bias the CLIP model by incorporating learnable text prompts into sensitive queries. To achieve this, they select a set of queries such as 'a photo of a good/evil/smart person' and utilize a dataset of images annotated with the protected group information. For each query, they add learnable text prompts. Subsequently, they calculate the text and image embeddings using the CLIP's text and image encoders. Next, they compute the similarity logits by taking the dot product between each pair of image-text embeddings. These similarity logits are then fed into an adversarial classifier, which aims to predict the protected attribute. The training objective aims to learn the text prompts in a manner that prevents the adversarial network from accurately predicting the protected attribute. The ultimate goal is to reduce the correlation between the similarity logits and the protected attributes. Additionally, they use an image-text contrastive (itc) loss to maintain the performance of the embeddings. They maintain the balance between the two loss values using a hyperparameter λ.

Berg et al. [25] utilized FairFace dataset for the debiasing loss and Flickr30K dataset for the itc loss, focusing on the gender attribute. Consequently, we evaluate their method only for the gender attributes using these datasets and the trained model shared by the authors. Just to note, they do not provide the value of the λ used to train the provided model.

Table 3.8: [Retrieval -Skew - Subjective - MSCOCO] This table shows absolute skew, given by Eq. (3.8), for image retrieval tasks using MSCOCO dataset. *The results show that the simple baseline with gender balanced queries perform the best for reducing skew.*

CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	Gender-BLN	FPCA-GT	FPCA-INF
				Top 10			
2.61±1.16	2.24±1.16	2.62±1.14	2.12±1.26	3.12±0.76	**0.36±0.14**	2.56±1.24	1.68±1.2
				Top 50			
1.38±0.68	1.95±0.82	2.33±0.82	2.07±0.9	2.06±0.78	**0.34±0.12**	1.51±0.84	1.36±1.16
				Top 100			
1.46±0.9	2.23±0.86	2.03±0.5	1.9±0.78	2.0±0.52	**0.29±0.06**	1.38±0.48	1.02±0.62

3.2.3.3 Fair PCA (referred to as FPCA in the results)

This is a general bias mitigation method that tries to find a linear approximation of the data that removes sensitive information (such as gender or race) while retaining as much non-sensitive information as possible. Specifically, the goal of fair PCA is to find a projection of datapoints x_i such that any function h applied to a projected datapoint is statistically independent of the protected attribute z_i. However, such a projection may not exist, so Kleindessner et al. [154] proposed to solve a relaxed version of the problem. They restrict h to only linear functions. In addition, they relax the statistical independence requirement between $h(x_i)$ and z_i and only require $h(x_i)$ and z_i to be uncorrelated. We use this as a post-processing method for making the representation space of OpenAI's CLIP [220] and OpenCLIP [130] models fair. We show results for this method w.r.t. to gender and race attributes in Section 3.5.

3.2.3.4 Baselines

To remove the gender bias in image retrieval tasks we also show results where we search for gendered versions of given queries and return balanced results from the gendered queries. For example, if we wanted to retrieve 10 images for the query "a photo of a doctor" we search for "a photo of a female doctor" and "a photo of a male doctor" and return 5 images for each of these. This is an instance of affirmative action [97]. We refer to this method as Gender-BLN in the results. Similarly, to address the racial bias in image retrieval we make race-specific queries for images and return the balanced results. We call this Race-BLN.

For the image captioning method, we propose a baseline in which we train the captioning system on MSCOCO by removing gendered words from the captions, e.g., "a

Table 3.9: [Retrieval -Statistical Tests - Subjective - FairFace] This table shows Alexander-govern statistical tests using FairFace. *This test checks whether there are differences in the mean value of cosine similarity between men and women for a given query. The pair of numbers represent the test statistic and the p-value. A low value of the statistic and high p-value is desirable, the former means the statistical difference for the given query has low impact and the later means that the differences are statistically insignificant. It shows that fair PCA and MI-GT methods generally achieve the lowest disparity in cosine similarity and the differences are generally statistically insignificant.*

Statistical tests: ANOVA- Alexander-Govern: (statistic, p-val)								
Query	CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	Prompt-GT	FPCA-GT	FPCA-INF
CEO	(1444 , 0.0)	(23 , 0.0)	(2 , 0.11)	(73 , 0.0)	(3 , 0.048)	(978 , 0.0)	(0 , 0.863)	(7 , 0.005)
boss	(2025 , 0.0)	(24 , 0.0)	(0 , 0.906)	(7 , 0.008)	(1 , 0.309)	(673 , 0.0)	(0 , 0.909)	(5 , 0.02)
convict	(300 , 0.0)	(4 , 0.032)	(0 , 0.473)	(7 , 0.007)	(1 , 0.168)	(328 , 0.0)	(0 , 0.484)	(18 , 0.0)
criminal	(327 , 0.0)	(28 , 0.0)	(2 , 0.084)	(43 , 0.0)	(0 , 0.443)	(453 , 0.0)	(0 , 0.78)	(17 , 0.0)
director	(668 , 0.0)	(0 , 0.5)	(14 , 0.0)	(0 , 0.553)	(8 , 0.004)	(787 , 0.0)	(0 , 0.452)	(8 , 0.003)
drug dealer	(621 , 0.0)	(6 , 0.01)	(3 , 0.069)	(12 , 0.0)	(9 , 0.003)	(718 , 0.0)	(1 , 0.277)	(4 , 0.043)
engineer	(1190 , 0.0)	(83 , 0.0)	(3 , 0.07)	(1 , 0.207)	(18 , 0.0)	(1126 , 0.0)	(7 , 0.007)	(13 , 0.0)
genius	(3145 , 0.0)	(34 , 0.0)	(9 , 0.003)	(99 , 0.0)	(16 , 0.0)	(1023 , 0.0)	(0 , 0.476)	(15 , 0.0)
leader	(1326 , 0.0)	(68 , 0.0)	(21 , 0.0)	(0 , 0.64)	(24 , 0.0)	(1138 , 0.0)	(0 , 0.391)	(0 , 0.388)
nurse	(4142 , 0.0)	(308 , 0.0)	(37 , 0.0)	(232 , 0.0)	(43 , 0.0)	(3762 , 0.0)	(0 , 0.494)	(0 , 0.76)
prostitute	(2738 , 0.0)	(156 , 0.0)	(9 , 0.002)	(27 , 0.0)	(18 , 0.0)	(241 , 0.0)	(0 , 0.651)	(7 , 0.005)
secretary	(3269 , 0.0)	(299 , 0.0)	(22 , 0.0)	(291 , 0.0)	(50 , 0.0)	(385 , 0.0)	(0 , 0.999)	(6 , 0.014)
suspect	(1740 , 0.0)	(4 , 0.041)	(4 , 0.025)	(3 , 0.082)	(5 , 0.023)	(820 , 0.0)	(0 , 0.566)	(12 , 0.0)

man standing on the road" to "a person standing on the road". We explain the results in Section 3.5.5.

3.3 Expected Behaviour and Evaluation Criteria

In this section, we discuss the tasks for which we evaluate different methods introduced in Section 3.2.

3.3.1 Binary Zero-Shot Classification

To evaluate fairness for binary zero-shot classification, we first define a pair of classes, e.g., nurse and doctor. Then, we encode all the images, using CLIP's image encoder or an image encoder provided by the corresponding method. Similarly, we tokenize and encode the names of different classes using CLIP's text encoder or a text encoder provided by the corresponding method with a fixed text prompt, e.g., "a photo of a nurse" and "a photo of a doctor". Depending on the methods we do further processing, e.g., for CLIP-clip we clip the prescribed embedding and for fair PCA we transform the text and image embeddings using a transformation matrix learned from the training split

Table 3.10: [Retrieval -Statistical Tests - Subjective - FairFace] This table shows statistical tests to check if for a given query all the races have same mean cosine similarity. *A large value of the test statistic and less than 0.05 pvalue implies that there is a large and statistically significant different in the mean value of the cosine similarity for one of the races.*

Statistical tests: ANOVA- Alexander-Govern: (statistic: p-val)							
Query	CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	FPCA-GT	FPCA-INF
cleaning person	(746 , 0.0)	(166 , 0.0)	(488 , 0.0)	(135 , 0.0)	(286 , 0.0)	(7 , 0.251)	(14 , 0.021)
director	(544 , 0.0)	(1440 , 0.0)	(416 , 0.0)	(1204 , 0.0)	(257 , 0.0)	(10 , 0.108)	(67 , 0.0)
engineer	(1276 , 0.0)	(760 , 0.0)	(511 , 0.0)	(752 , 0.0)	(290 , 0.0)	(28 , 0.0)	(51 , 0.0)
labourer	(1316 , 0.0)	(474 , 0.0)	(703 , 0.0)	(755 , 0.0)	(451 , 0.0)	(11 , 0.068)	(162 , 0.0)
secretary	(661 , 0.0)	(362 , 0.0)	(280 , 0.0)	(334 , 0.0)	(402 , 0.0)	(5 , 0.459)	(21 , 0.001)
smart person	(682 , 0.0)	(872 , 0.0)	(646 , 0.0)	(371 , 0.0)	(467 , 0.0)	(18 , 0.005)	(56 , 0.0)
sophisticated person	(1274 , 0.0)	(636 , 0.0)	(548 , 0.0)	(462 , 0.0)	(485 , 0.0)	(19 , 0.003)	(44 , 0.0)
terrorist	(1603 , 0.0)	(882 , 0.0)	(1017 , 0.0)	(642 , 0.0)	(828 , 0.0)	(14 , 0.025)	(84 , 0.0)

of a given dataset. We then take the dot product and the softmax over the two classes. Then, from the two classes, we pick the one which yields the maximum value.

We define a set of binary classification tasks for which we believe different genders and races should have no disparity. We provide the list of these classes in Appendix A.1. As described in the introduction, Table 3.1, we focus on *human-centric subjective tasks*, e.g., 'criminal' vs 'innocent person', for which demographic parity is desirable across different values of the protected attributes. Similarly in datasets where we do not have access to the ground-truth professions we expect that classification tasks such as 'doctor' vs 'nurse' or 'CEO' vs 'Secretary' should have demographic parity across protected groups. The results for these tasks are shown in Figures 3.1, 3.3, 3.11, 3.13 and 3.16.

We also show results for *human-centric objective tasks*, where we evaluate different methods for the independence of the gender attribute w.r.t. the true positive rates in predicting CelebA dataset's objective categories, such as wearing glasses, and wearing a necklace in Figure 3.2 and MIAP dataset's categories, based on age, prominence in the image, i.e., whether the bounding box of the person occupied more than 50% of the image, and the number of people in Figure 3.6.

3.3.2 Image Retrieval

Similar to zero-shot classification, for the image retrieval task we select a set of queries for which we believe there should not be any difference in the retrieved image across different gender groups or races, we show these queries for each dataset in Appendix A.1. We similarly convert the images and the queries into their representations and calculate their cosine similarity. Then, we select the top k results from the list of the decreasing order of the cosine similarity for each query. Similar to zero-shot classification, we show

Table 3.11: [Retrieval - Statistical tests - Subjective - Flickr30k] This table shows Alexander Govern statistical test for the cosine similariy of various queries between men and women. *It demonstrates that fair PCA based methods do very well to equalize the cosine similarity between the two groups for different retrieval tasks.*

Statistical tests: ANOVA- Alexander-Govern: (statistic, p-val)								
Query	CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	Prompt-GT	FPCA-GT	FPCA-INF
doctor	(271 , 0.0)	(23 , 0.0)	(43 , 0.0)	(2 , 0.125)	(60 , 0.0)	(222 , 0.0)	(1 , 0.225)	(12 , 0.001)
nurse	(1252 , 0.0)	(42 , 0.0)	(76 , 0.0)	(2 , 0.151)	(49 , 0.0)	(1541 , 0.0)	(0 , 0.481)	(2 , 0.186)
secretary	(1567 , 0.0)	(47 , 0.0)	(27 , 0.0)	(3 , 0.09)	(1 , 0.335)	(676 , 0.0)	(0 , 0.484)	(59 , 0.0)
boss	(588 , 0.0)	(35 , 0.0)	(31 , 0.0)	(10 , 0.001)	(18 , 0.0)	(487 , 0.0)	(0 , 0.774)	(65 , 0.0)
lawyer	(218 , 0.0)	(2 , 0.157)	(2 , 0.161)	(36 , 0.0)	(41 , 0.0)	(166 , 0.0)	(0 , 0.932)	(13 , 0.0)
paralegal	(522 , 0.0)	(10 , 0.002)	(0 , 0.825)	(45 , 0.0)	(65 , 0.0)	(185 , 0.0)	(0 , 0.77)	(15 , 0.0)

results for *human-centric subjective tasks* under independence assumption in Figures 3.4, 3.5, 3.12, 3.14, and 3.15.

For image retrieval, fairness of representation or diversity assumption is desirable for certain scenarios, i.e., showing images of different protected groups in the top k results. We show results for representational fairness for *human-centric subjective tasks* in Tables 3.5, 3.6, 3.7, 3.8, 3.14 and 3.16. For *human-centric objective tasks*, we show results in Table 3.3 under the diversity assumption.

We report the differences in cosine similarity for each query across different genders and races, shown in Figures 3.7, 3.8, 3.9, 3.10 and 3.14. We also perform statistical tests, specifically Alexandar-govern (ANOVA) [4] test which allows for different variances across the groups, to demonstrate how successful different methods are in equalizing representations for different protected group values. The results for these are shown in Tables 3.9, 3.10, 3.11, 3.12, 3.15 and 3.17.

3.3.3 Image Captioning

To test fairness concerns of using CLIP models for captioning we study CLIP-CAP [201] which uses CLIP and GPT2 embeddings. Mokady et al. [201] proposed two methods: one where they froze the CLIP embedding space as well as GPT2 embedding space and just learnt a transformer based mapping network and second where they only froze the CLIP embedding space and learnt a few layers of GPT2 network in addition to learning a simpler MLP network. In our experiments, we found that the first variant does not generalize very well to out of distribution images, which makes sense since training additional layers of the GPT2 model results in a more expressive model. So, we use the

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.alexandergovern.html

Table 3.12: [Retrieval - Statistical tests - Subjective - MSCOCO] This table shows Alexander Govern statistical test for the cosine similariy of various queries between men and women. The first number refers to the test statistic while the second number is the p-value. If there is a statisitically significant difference among different groups the test statistic would be high and p-value would be low. *It demonstrates that fair PCA GT yields statistically insignificant differences.*

Statistical tests: ANOVA- Alexander-Govern: (statistic, p-val)							
Query	CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	FPCA-GT	FPCA-INF
boss	(352 , 0.0)	(27 , 0.0)	(40 , 0.0)	(0 , 0.408)	(175 , 0.0)	(0 , 0.393)	(6 , 0.013)
secretary	(950 , 0.0)	(6 , 0.011)	(34 , 0.0)	(7 , 0.007)	(82 , 0.0)	(1 , 0.201)	(325 , 0.0)
genius	(198 , 0.0)	(0 , 0.477)	(15 , 0.0)	(3 , 0.072)	(103 , 0.0)	(1 , 0.306)	(47 , 0.0)
helpful person	(44 , 0.0)	(0 , 0.744)	(23 , 0.0)	(2 , 0.153)	(123 , 0.0)	(2 , 0.088)	(81 , 0.0)
affectionate person	(286 , 0.0)	(18 , 0.0)	(20 , 0.0)	(42 , 0.0)	(43 , 0.0)	(1 , 0.307)	(55 , 0.0)
funny person	(36 , 0.0)	(16 , 0.0)	(104 , 0.0)	(26 , 0.0)	(54 , 0.0)	(2 , 0.09)	(135 , 0.0)

second variant. The authors shared the training code and hyperparameters for MSCOCO dataset [177] and Conceptual Captions dataset. We show results using MSCOCO dataset as the training times are faster. For demonstrating fairness concerns in CLIP embeddings, the experiments using MSCOCO show interesting insights as discussed in Section 3.5.5.

We train the CLIP-CAP model with original CLIP as well as by transforming CLIP embeddings using different debiasing methods. We also experiment with making the captions of MSCOCO gender neutral, e.g., by changing 'He/She' into 'They'. We then train the GPT2 layers and the MLP network. To generate captions we encode images with the CLIP image encoders, as well as any additional processing necessary for a particular debiasing method, and pass it through the learned MLP and GPT2 which generates captions.

3.3.4 Performance Measures

It is important that performance for different downstream tasks does not suffer while reducing bias. To demonstrate the well-known accuracy-fairness trade-off, we report the accuracy of a logistic regression classifier to predict different attributes using CLIP embeddings as input, shown in Table 3.13. We also report the recall@k performance for different values of k, shown in Table 3.4, as well as precision shown in Tables 3.3 and 3.18. We report accuracy for zero-shot classification tasks in Table 3.2.

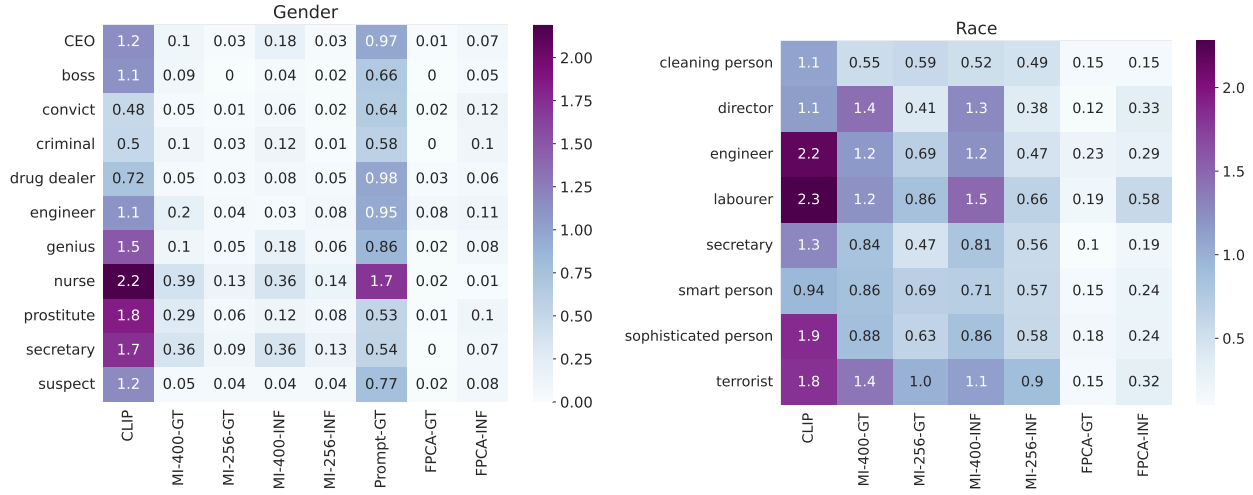


Figure 3.7: [Retrieval - Cosine similarity - Subjective - FairFace] These figures are heatmaps that show the absolute difference in cosine similarity, scaled up by a factor of 100, for different image retrieval queries using different methods for gender (left) and race (right) attributes on FairFace dataset. *The figures demonstrate the efficiency of each methods to equalize the representation for different protected attribute groups on average. It shows that in general, fair PCA and mutual information based methods equalize the cosine similarity for gender and race attribute for a variety of queries.*

3.4 Metrics Based on Our Taxonomy

Here, we outline the task-specific Desiderata and discuss relevant metrics corresponding to the measures and evaluation tasks described in Sections 3.1 and 3.3, respectively. Inherently, this is a coarse division and excludes many potential harms. One of the challenges of open-labeling tasks is that many subtle harms are possible.

3.4.1 Human-Centric (Un)Fairness Metrics

We describe image classification, retrieval and captioning tasks where the labels are highly-related to people in the image as human-centric labelings. This section presents the unfairness metrics used.

3.4.1.1 Independence Assumptions:

We focus on two independence-based notions of fairness — demographic parity (DP) [80, 90] and equal opportunity (EOP) [120, 274] for subjective and objective tasks, respectively.

3.4.1.1.1 Subjective Labeling Tasks: In classification, DP requires that the prediction of a datapoint be independent of the value of the protected attribute. Specifically, given

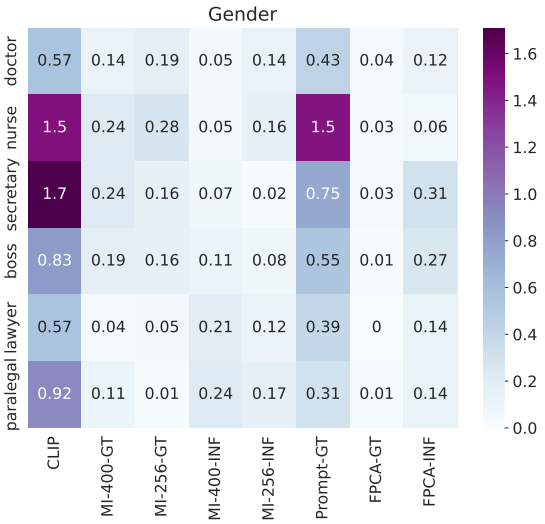


Figure 3.8: [Retrieval - Cosine similarity - Subjective - Flickr30k] The figure is heatmap that show the absolute difference in cosine similarity, scaled up by a factor of 100, for different queries using different methods for gender attribute on Flickr30K dataset. *The figure demonstrates the efficiency of each methods to equalize the representation for different protected attribute groups on average. It shows that in general, fair PCA based methods and the mutual information based methods equalize the cosine similarity for gender attribute for a variety of queries.*

a binary classification task where $\hat{Y} \in \{-1, 1\}$ is the predicted variable and $Z \in \mathbb{Z}^+$ represents protected membership, DP is given as $P(\hat{Y} = 1|Z = z) = P(\hat{Y})$.

Zero-shot Binary Classification: For zero-shot classification, notions of independence are desirable. In this section, we present metrics corresponding to DP. We define demographic disparity (DDP) as the maximum absolute difference in the fraction datapoints classified in the positive class among any pair of groups of the protected group. Let Z_i be the set of datapoints with protected attribute i. We define the DDP as[5]

$$\textbf{DDP:} \quad \max_{i,j \in [p]} \left| \frac{1}{|Z_i|} \sum_{x \in Z_i} \mathbb{1}[f(x) = 1] - \frac{1}{|Z_j|} \sum_{x \in Z_j} \mathbb{1}[f(x) = 1] \right|, \tag{3.5}$$

where $f(x)$ is a binary classifier. DDP ranges between 0 and 1, i.e., from least to most disparity. We use gender as a binary attribute, due to the limited availability of datasets with multi-valued gender attributes. In this case, the above equation reduces to the absolute difference between the fraction of men classified in the positive class and the fraction of women classified in the positive class. Race consists of multiple groups, and

[5]We use the notation $[p] := \{1, \ldots, p\}$.

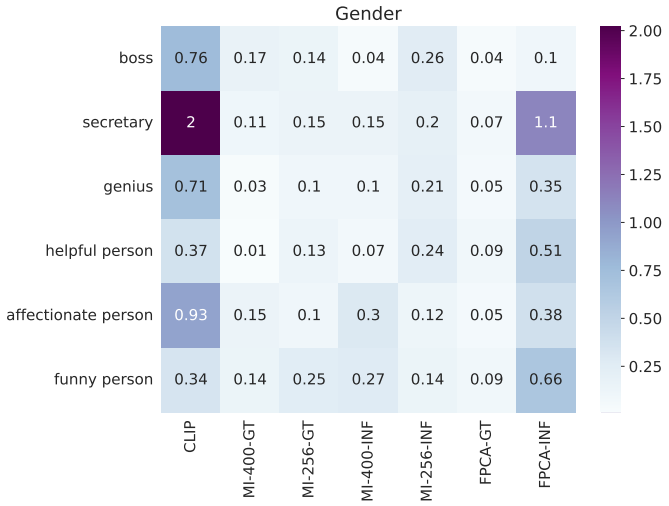


Figure 3.9: [Retrieval - Cosine similarity - Subjective - MSCOCO] The figure is a heatmap that shows the absolute difference in cosine similarity, scaled up by a factor of 100, for different queries using different methods for gender attribute on MSCOCO dataset. *The figure demonstrates the efficiency of each methods to equalize the representation for different protected attribute groups on average. It shows fair PCA based methods and mutual information based methods equalize the cosine similarity for gender attribute for a variety of queries.*

we report the maximum absolute disparity of classification between any two groups.

Image Retrieval: Depending on the downstream application, either notions of independence or diversity of different values of the protected attribute may be desirable.

For independence, we present metrics corresponding to DP. Let K be the set of the retrieved images, comprising subset K_i of images of the protected group i, Z_i is the set of images belonging to the group i and Z is the set of all images. Following, Wachter et al. [257] we define the DDP in this context as follows:

$$\textbf{DDP:} \quad \max_{i,j \in [p]} \Bigg| \Bigg(\underbrace{\frac{|K_i|}{|K|}}_{\text{Advantaged group } i} - \underbrace{\frac{|Z_i| - |K_i|}{|Z| - |K|}}_{\text{Disadvantaged group } i} \Bigg) - \Bigg(\underbrace{\frac{|K_j|}{|K|}}_{\text{Advantaged group } j} - \underbrace{\frac{|Z_j| - |K_j|}{|Z| - |K|}}_{\text{Disadvantaged group } j} \Bigg) \Bigg|. \tag{3.6}$$

Wachter et al. [257] showed that this measure only takes the value 0 when Eq. (3.5) does, given that $|K_i| > 0 \, \forall i$. However, this variant is more suitable for asymmetric labelings

Table 3.13: [Classification - Accuracy - Objective - FairFace] This table shows the accuracy of a logistic regression classifier trained on the corresponding CLIP features for FairFace dataset. *The top and the bottom parts of the table correspond to the cases where the mitigation methods were supposed to remove the gender and race information, respectively, from the CLIP embeddings, while preserving the other information. The results show that fair PCA based methods are more effective in removing the corresponding sensitive information, i.e., the accuracy for predicting the corresponding sensitive attributes is nearly random. Additionally, the fair PCA methods do not reduce the predictive power of the embeddings, i.e., the accuracy in predicting other attributes stays similar to the original CLIP embeddings. We do not provide the results for the prompt method because they do not alter the image representation and results are similar as the original CLIP.*

Feature	Clip	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	FPCA-GT	FPCA-INF
			Mitigation methods w.r.t gender: ViTB/32				
age	0.60	0.60	0.60	0.60	0.60	0.60	0.60
gender	0.95	0.94	0.90	0.94	0.90	**0.53**	0.60
race	0.71	0.71	0.71	0.71	0.71	0.71	0.71
			Mitigation methods w.r.t gender: ViTB/16				
age	0.62	0.62	0.61	0.62	0.61	0.62	0.62
gender	0.96	0.95	0.91	0.95	0.91	**0.53**	0.57
race	0.74	0.73	0.73	0.73	0.73	0.74	0.74
			Mitigation methods w.r.t race: ViTB/32				
age	0.60	0.60	0.59	0.60	0.59	0.60	0.60
gender	0.95	0.95	0.94	0.95	0.94	0.94	0.94
race	0.71	0.71	0.70	0.71	0.70	**0.19**	0.34
			Mitigation methods w.r.t race: ViTB/16				
age	0.62	0.62	0.61	0.62	0.61	0.61	0.61
gender	0.96	0.96	0.95	0.95	0.96	0.96	0.95
race	0.74	0.73	0.73	0.73	0.73	**0.19**	0.39

where a small proportion of individuals receive positive decisions. This measure returns values ranging from 0 to 1.

3.4.1.1.2 Objective Labeling Tasks – Zero-shot Binary Classification: EOP requires that the prediction of all datapoints with positive labels should be independent of the protected attribute. Specifically, a binary classification task where $\hat{Y} \in \{-1, 1\}$ is the predicted variable, $Y \in \{-1, 1\}$ is the ground truth variable and $Z \in \mathbb{Z}^+$ represents the protected attribute EOP requires $P(\hat{Y} = 1|Y = 1, Z = z) = P(\hat{Y})$.

Similar to DDP, given in Eq. (3.5), we can extend the definition for EOP to disparity in true positive rates (DTPR):

Figure 3.10: [Retrieval - Cosine similarity - Subjective - FairFace - OpenCLIP] These figures are heatmaps that show the absolute difference in cosine similarity, scaled up by a factor of 100, for different image retrieval queries using different methods for gender (left) and race (right) attributes using FairFace dataset with the OpenCLIP. *The figures demonstrate the efficiency of each methods to equalize the representation for different protected attributes groups on average. It shows that in general, fair PCA based methods reduce the difference in cosine similarity for gender and race attribute for a variety of queries.*

$$\textbf{DTPR:} \quad \max_{i,j\in[p]} \left| \frac{1}{|{Z_i}^+|} \sum_{x\in {Z_i}^+} \mathbb{1}[f(x)=1] - \frac{1}{|{Z_j}^+|} \sum_{x\in {Z_j}^+} \mathbb{1}[f(x)=1] \right|, \tag{3.7}$$

where ${Z_*}^+$ is the set of datapoints with protected attribute $*$.

For image retrieval tasks, we could easily extend Eq. (3.6) for EOP, e.g., by confining all the sets to positive examples.

3.4.1.2 Diversity Assumptions – Image Retrieval:

We use the following metrics to measure unfairness in the representation.

3.4.1.2.1 Subjective Labeling Tasks: We use the Skew metric proposed by Geyik et al. [98]. Let K be the set of $|K|$ items we want to retrieve comprising of sets K_i that belong to the protected attribute group i. Let df_i be the desired fraction of items belonging to

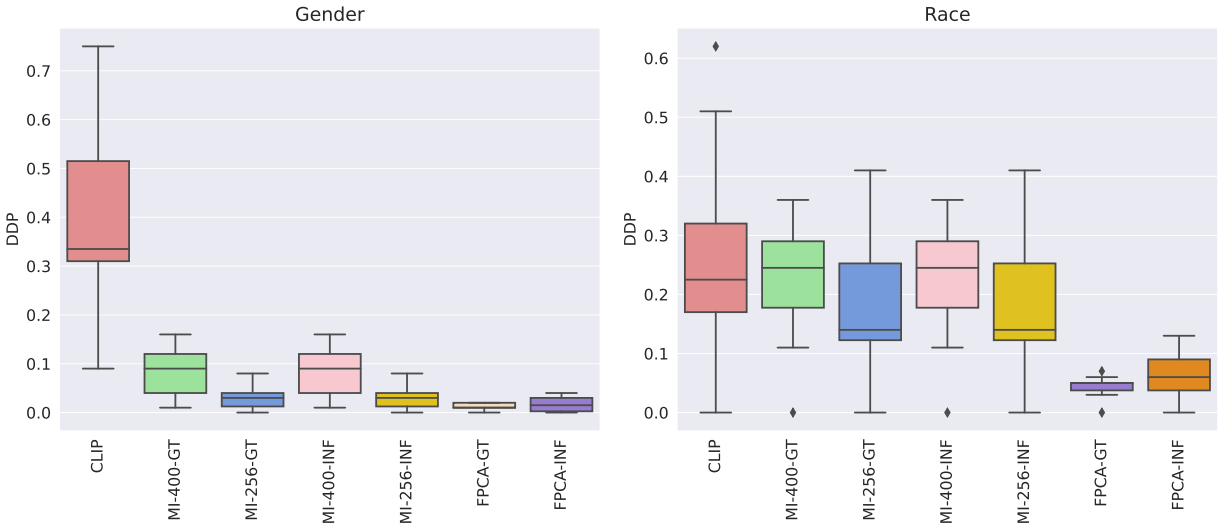


Figure 3.11: [Classification - DDP - Subjective - FairFace - OpenCLIP] These figures show DDP for classification, given by Eq. (3.5), using OpenCLIP using FairFace dataset. *It demonstrates that fair PCA based methods perform the best in reducing bias.*

the group i in the top $|K|$ results, and $rf_i := \frac{|K_i|}{|K|}$ be the retrieved fraction of items.

$$\textbf{Skew@k:} \quad \max_{i,j \in [p]} \left| \log_e(rf_i/df_i) \right| \tag{3.8}$$

We set $df_i = \frac{1}{p}$, where p is the number of protected groups.

3.4.1.2.2 Objective Labeling Tasks: Let K^+ be the set of ground truth positive images retrieved for a given query, out of which ${K_i}^+$ are the retrieved images that belong to the protected attributes group i. We report the maximum absolute disparity in the representation (DDP-Rep) of any two protected attribute groups, i.e.,

$$\textbf{DDP-rep:} \quad \max_{i,j \in [p]} \frac{1}{|K^+|} \left| |{K_i}^+| - |{K_j}^+| \right|. \tag{3.9}$$

This metric shows how well different groups are represented in a retrieval task even if the ground truth is imbalanced.

3.4.2 Non-Human-Centric Labelings: Performance Metrics

By non-human-centric labelings, we refer to image classification, image retrieval and image captioning tasks where the labels are unrelated to people in the image. While we do not consider the harms associated with this task, performance remains important.

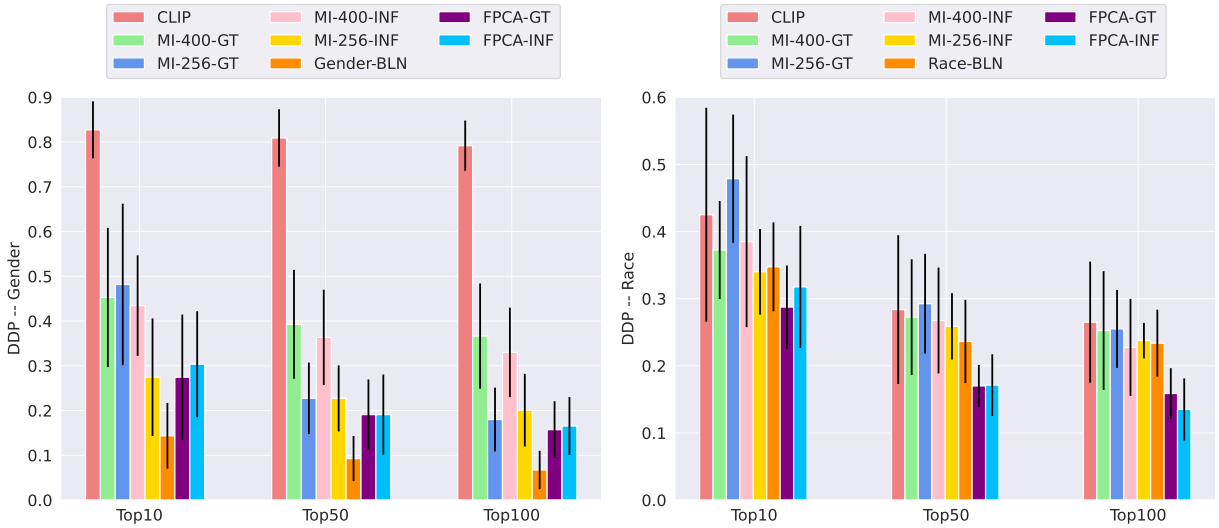


Figure 3.12: [Retrieval - DDP - Subjective - FairFace - OpenCLIP] These figures show DDP for image retrieval, given by Eq. 3.6, using OpenCLIP on FairFace dataset. *It demonstrates that gender balanced queries and fair PCA are most effective in reducing demographic disparity in subjective image retrieval tasks.*

For *objective non-human-centric* tasks, e.g., categorizing images as showing either 'cats' or 'dogs', or searching for 'a photograph of an oak tree', performance is important, and the correct notion of performance is task dependent. Following Radford et al. [220] we use accuracy to measure the performance of zero-shot classifiers, recall@k and precision@k. Ideally, there should be no decrease in performance for these tasks, as we do not have fairness concerns.

For *subjective non-human-centric* tasks we might also have fairness concerns, e.g., that a search for "beautiful building" might be biased towards Christian churches and omit buildings associated with other religions. However, these concerns are harder to evaluate especially due to lack of data and ground truth labels.

3.5 Evaluation

In this section, we present the results according to our proposed taxonomy introduced in Table 3.1. Given that IND refers to the independence of the protected attribute w.r.t. to the outcome variable (metrics: Eqs. (3.5), (3.6) and (3.7)) and DIV refers to the diversity of the protected attribute groups in the retrieval results (metrics: Eqs. (3.8) and (3.9)), we answer the following questions in this section.

Q1: How fair (IND) are different methods w.r.t. gender for zero-shot binary classification on subjective and objective tasks?

Q2: How fair (IND) are different methods w.r.t. race for zero-shot binary classification

on subjective tasks?
Q3: How fair (IND or DIV) are different methods w.r.t. gender for image retrieval tasks on subjective and objective tasks?
Q4: How fair (IND or DIV) are different methods w.r.t. race for image retrieval subjective tasks?
Q5: How is the performance on the attributes on which fairness was not enforced affected?
Q6: Are there statistically significant differences in representations for different methods w.r.t. gender?
Q7: Are there statistically significant differences in representations for different methods w.r.t. race?
Q8: What are the fairness (IND) concerns using CLIP embeddings for captioning systems?
Q9: Do CLIP bias mitigation methods help alleviate fairness concerns in captioning?

3.5.1 Datasets

In this section, we describe the datasets used for evaluation. We use the test split for the evaluation. In some cases, where the test images are little or the ground truth for the test set is not available we evaluate on the validation set, please refer to the dataset descriptions below. We use the training split for training the bias mitigation methods.

FairFace [145] comprises about $100k$ images, split into $85k$ training images and $10K$ validation images. The images are focused on the faces and come with a binary labelling of the gender attribute (53% male images), 9 bins of age attribute ($0-2:2\%$; $3-9:12\%$; $10-19:11\%$; $20-29:30\%$; $30-39:22\%$; $40-49:12\%$; $50-59:7\%$; $60-69:3\%$; $70+:1\%$) and 7 values of the race attribute, specifically, East Asian (14%), Indian (14%), Black (14%), White (19%), Middle Eastern (11%), Latino Hispanic (15%) and South east Asian (13%). The dataset is fairly balanced for the race and gender attributes. However for the age attribute, there is less amount of data for older categories.

Flickr30K [214, 272] contains about $30k$ images with 5 human annotated captions per image. We split the data into 50% train and 50% test data. This dataset contains a variety of images containing humans and animals. These images contain diverse backgrounds and have natural lighting conditions.

MSCOCO [177] contains about $120K$ images with $80K$ training images and $40K$ validation images. The dataset contains at-least 5 hand annotated captions per image. It additionally contains 80 categories as labels. The categories include person, several

Table 3.14: [Retrieval - Skew - Subjective - FairFace - OpenCLIP] This table shows the maximum absolute skew, given by Eq. (3.8), using the FairFace dataset and gender and race attributes using OpenCLIP. *It demonstrates that all the methods are able to reduce the skew. Gender/Race balanced queries and fair PCA are the most effective in reducing the skew.*

Clip	MI-400-gt	MI-256-GT	MI-400-inf	MI-256-INF	Gender/Race-BLN	FPCA-GT	FPCA-INF
			Gender: Top 10				
2.38±0.74	0.83±0.36	1.04±0.66	0.72±0.26	0.43±0.3	**0.15±0.1**	0.42±0.28	0.41±0.2
			Gender: Top 50				
1.94±0.38	0.63±0.26	0.33±0.12	0.55±0.22	0.34±0.12	**0.11±0.04**	0.25±0.12	0.25±0.14
			Gender: Top 100				
1.77±0.32	0.56±0.22	0.26±0.1	0.48±0.2	0.31±0.1	**0.07±0.02**	0.21±0.1	0.21±0.08
			Race: Top 10				
2.37±0.58	2.66±0.0	2.66±0.0	2.42±0.48	2.42±0.48	**2.37±0.58**	**2.37±0.58**	2.66±0.0
			Race: Top 50				
1.4±0.46	1.35±0.4	1.4±0.36	1.52±0.36	1.35±0.48	1.16±0.38	1.01±0.36	**0.82±0.26**
			Race: Top 100				
1.33±0.44	1.07±0.3	1.25±0.3	1.04±0.14	1.21±0.44	1.06±0.42	0.7±0.12	**0.63±0.18**

animals such as cat, dog and giraffe, and objects such as scissors, bicycle and hairdryer. The images have a diverse background and are in the natural lighting conditions.

We extract the gender information from the captions of Flickr30K and MSCOCO. To this end, we define a 3-valued attribute, $type\_of \in \{male, female, neutral\}$, and a set of male and female words, given in Appendix A.1. $type\_of$ an image is considered $(fe)male$ if *any* of its captions contain *any* of the $(fe)male$ words otherwise it is considered $neutral$. Additionally, if the caption contains both $male$ and $female$ words $type\_of$ an image is considered $neutral$.

IdenProf [6] consists of 11,000 images of identifiable professionals. It contains images of 10 professionals, i.e, chef, doctor, engineer, farmer, firefighter, judge, mechanic, pilot, police and waiter. We use roughly an 80-20 test and train split[7], i.e., 900 images of test data per profession. We use this data for image retrieval tasks and annotated the gender of the retrieved images by hand.

CelebA [180] comprises about 200k images of celebrities. These images are focused on faces and additionally provide 40 binary attributes per image, including gender. The dataset is split into 80% training images, 10% validation images and 10% test images. We train on the training set and test on the test set.

[6] https://github.com/OlafenwaMoses/IdenProf

[7] In the official dataset the dataset split is 80-20 for the train and test splits, respectively. We invert it to get more robust results for evaluating image retrieval and captioning tasks.

Table 3.15: [Retrieval - Statistical tests - Subjective - FairFace - OpenCLIP] This table shows the statistical tests for the cosine similarities among different groups of the protected groups. The first number refers to the test statistic while the second number is the p-value. If there is a statisitically significant difference among different groups the test statistic would be high and p-value would be low. *Specifically, it shows the Alexander-govern statistical test which measures whether the mean of cosine similarity among different groups for a given query are statistically significant or not. It shows that fair PCA trained on ground truth protected attribute labels yields statistically insignificant differences.*

Statistical tests: ANOVA- Alexander-Govern: (statistic: p-val)							
Query	CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	FPCA-GT	FPCA-INF
				Gender			
CEO	(1554 , 0.0)	(114 , 0.0)	(56 , 0.0)	(62 , 0.0)	(41 , 0.0)	(0 , 0.758)	(23 , 0.0)
boss	(3354 , 0.0)	(612 , 0.0)	(99 , 0.0)	(552 , 0.0)	(196 , 0.0)	(0 , 0.501)	(8 , 0.003)
convict	(2519 , 0.0)	(589 , 0.0)	(39 , 0.0)	(460 , 0.0)	(90 , 0.0)	(2 , 0.127)	(12 , 0.0)
criminal	(1158 , 0.0)	(320 , 0.0)	(18 , 0.0)	(163 , 0.0)	(35 , 0.0)	(1 , 0.19)	(2 , 0.085)
drug dealer	(2503 , 0.0)	(257 , 0.0)	(3 , 0.056)	(176 , 0.0)	(34 , 0.0)	(3 , 0.055)	(19 , 0.0)
engineer	(1745 , 0.0)	(80 , 0.0)	(2 , 0.086)	(54 , 0.0)	(8 , 0.005)	(1 , 0.309)	(0 , 0.474)
genius	(822 , 0.0)	(307 , 0.0)	(5 , 0.015)	(292 , 0.0)	(31 , 0.0)	(3 , 0.065)	(14 , 0.0)
nurse	(4889 , 0.0)	(115 , 0.0)	(8 , 0.003)	(191 , 0.0)	(2 , 0.131)	(0 , 0.511)	(0 , 0.424)
prostitute	(3088 , 0.0)	(0 , 0.469)	(46 , 0.0)	(5 , 0.015)	(131 , 0.0)	(0 , 0.947)	(0 , 0.384)
secretary	(4269 , 0.0)	(212 , 0.0)	(42 , 0.0)	(315 , 0.0)	(71 , 0.0)	(0 , 0.708)	(24 , 0.0)
suspect	(1732 , 0.0)	(228 , 0.0)	(34 , 0.0)	(281 , 0.0)	(39 , 0.0)	(0 , 0.372)	(0 , 0.793)
				Race			
cleaning person	(1069 , 0.0)	(214 , 0.0)	(355 , 0.0)	(375 , 0.0)	(534 , 0.0)	(4 , 0.577)	(46 , 0.0)
director	(232 , 0.0)	(83 , 0.0)	(57 , 0.0)	(151 , 0.0)	(177 , 0.0)	(4 , 0.579)	(27 , 0.0)
engineer	(642 , 0.0)	(332 , 0.0)	(391 , 0.0)	(206 , 0.0)	(334 , 0.0)	(10 , 0.116)	(62 , 0.0)
labourer	(1349 , 0.0)	(203 , 0.0)	(374 , 0.0)	(240 , 0.0)	(380 , 0.0)	(19 , 0.003)	(180 , 0.0)
secretary	(322 , 0.0)	(105 , 0.0)	(146 , 0.0)	(96 , 0.0)	(204 , 0.0)	(5 , 0.482)	(67 , 0.0)
smart person	(741 , 0.0)	(350 , 0.0)	(155 , 0.0)	(272 , 0.0)	(250 , 0.0)	(11 , 0.071)	(50 , 0.0)
sophisticated person	(85 , 0.0)	(174 , 0.0)	(228 , 0.0)	(296 , 0.0)	(351 , 0.0)	(12 , 0.061)	(37 , 0.0)
terrorist	(642 , 0.0)	(595 , 0.0)	(564 , 0.0)	(617 , 0.0)	(590 , 0.0)	(5 , 0.514)	(202 , 0.0)

Food101[35] comprises 101 food categories with 750 training and 250 test images per category. The test images have been manually cleaned. We show results on the test split.

Pascal VOC 2007 [84] is a multi-class dataset. The categories include person, several household objects and different vehicles. We show results on the c.a. 5K test images. We consider a classification to be accurate if the top predicted label is among the multiple ground truth labels.

ImageNet 2012[71] comprises of 1000 classes, including animals, e.g., goldfish, great white shark, scorpion, etc. ; objects , e.g., bath-towel, accordion, guitar, assault rifle, etc.; place or buildings, e.g., church, cinema; and concepts, e.g., groom. Images are divers and in natural lighting. We use the 100K test set images to show the results.

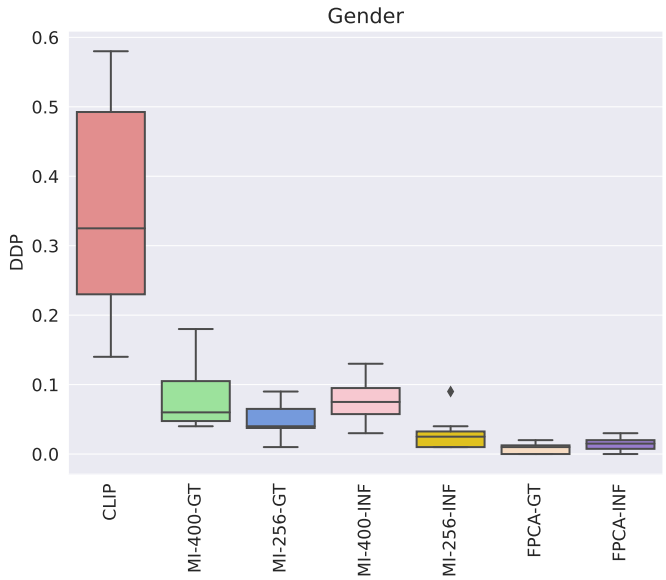


Figure 3.13: [Classification - DDP - Subjective - Flickr30K - OpenCLIP] These figures show DDP for classification, given by Eq. 3.5, using OpenCLIP on Flickr30K dataset. *It demonstrates that fair PCA based methods are the most effective in reducing bias in classification tasks.*

Stanford Cars [160] comprises 8K test images of 196 types of cars. We use it to demonstrate the effect of various bias mitigation methods on fine grained image classification task.

MIAP (More Inclusive Annotations for People) [231] has c.a. 22K test images and c.a. 70K training images, which contain at least one person. Each image comes with the bounding box(es) of the person(s); age, i.e., young, middle, older or unknown; and gender,i.e., predominantly masculine, predominantly feminine or unknown. For our experiments, we try to predict whether a person is inconspicuous, i.e., occupies less than 50% of the image; whether they are an adult, i.e., age attribute is middle or older; and whether there is one or multiple people in the picture.

3.5.2 Experimental Details

We show results for the methods of Section 3.2.3. For different fairness metrics we show results using OpenAI's CLIP ViTB-16 architecture. We find similar trends in results using ViTB-32 architecture. For performance results on objective tasks, we show results using both ViTB-16 and ViTB-32 architectures.

For mutual-information (MI) based method described in Section 3.2.3.1 we show results where we retain $m \in \{400, 256\}$ dimensions of the total 512 CLIP embedding dimensions. FPCA refers to fair PCA as described in Section 3.2.3.3. Prompt is the

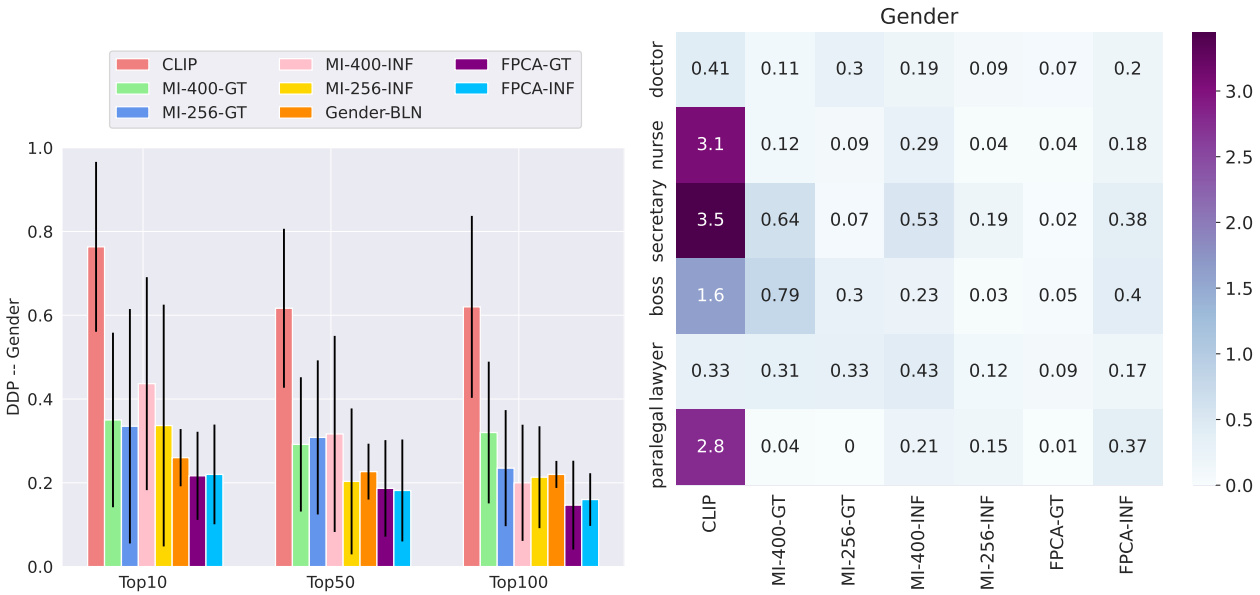


Figure 3.14: [Retrieval - DDP & Cosine similarity - Subjective - Flickr30K - OpenCLIP] These figures show DDP, given by Eq. (3.6), for retrieval task using OpenCLIP using Flickr30K dataset on the left, and absoulte differences in the cosine similarity between men and women for different queries on the right.

method described in Section 3.2.3.2. Gender-BLN refers to the baseline for the image retrieval task, where we add the words 'female' and 'male' to the query and return $\frac{K}{2}$ results from each of these queries. Race-BLN works similarly for the multi-valued race attribute.

Addressing lack of demographic features. For our fairness evaluations we use datasets where we have access to the demographic features. However, in real-world scenarios we might not have access to such features. To demonstrate results for such cases, we use the CLIP model to predict the gender attribute. The tags GT and INF indicate whether the protected attribute was ground truth or inferred. It is important to note that we only use the inferred attributes for training the bias mitigation method. The evaluation always uses the ground truth labels of the protected attributes.

3.5.3 Zero-shot Classification

Q1, Q2, Q5 i) Figures 3.1, 3.2, 3.3, 3.6 and 3.16 demonstrate that most mitigation methods can enforce *independence assumption* of fairness w.r.t. gender. ii) However, mutual information based methods can lead to a significant reduction in performance as show in Tables 3.2, 3.4, 3.13 and 3.18. iii) Prompt based method does not reduce the bias as well as the other methods. A possible reason could be that the trained model tries to preserve the expressiveness of the representations while putting too little weight

Table 3.16: [Retrieval - Skew - Subjective - Flickr30K - OpenCLIP] This table shows the skew metric, given by Eq. (3.8), using OpenCLIP model, for the gender attribute average over several image retrieval task using the Flickr data. *It shows that gender balanced queries are most effective in reducing skew.*

CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	Gender-BLN	FPCA-GT	FPCA-INF
Top 10							
1.58±0.76	1.49±1.28	1.55±1.26	1.59±1.24	0.64±0.24	**0.4±0.1**	0.59±0.28	0.59±0.28
Top 20							
1.4±0.92	0.92±0.5	0.93±0.62	0.59±0.2	0.42±0.1	**0.37±0.04**	0.5±0.16	0.46±0.18
Top 30							
1.48±0.96	0.89±0.5	0.72±0.64	0.46±0.14	0.38±0.06	**0.34±0.04**	0.54±0.3	0.4±0.14

on debiasing. iv) Fair PCA based methods do very well compared to the other methods in the multi-valued race attribute. v) In general, fair PCA based methods reduce the bias for both race and gender attributes while retaining the performance of the CLIP embeddings for other tasks.

3.5.4 Image Retrieval

Q3, Q5 i) For both *subjective tasks* and *objective tasks*, simple baselines, where gender or race was appended with the query, do very well in both enforcing demographic parity (Figures 3.4, 3.5 and 3.15) and enforcing representational fairness (Tables 3.3, 3.5, 3.6, 3.7, 3.8). A reason for the good performance on both demographic parity and representational fairness is that the protected groups in most of the datasets we consider are roughly balanced. However, the obvious drawback of this method is that it does not produce generalizable embedding to be used for other tasks. ii) Mutual information based methods and fair PCA based methods are also good at enforcing *independence assumption* of fairness for the gender attribute, as shown in Figures 3.4, 3.5 and 3.15. This is further supported by their effectiveness in reducing the disparity in the maximum average cosine similarity per query as shown in Figures 3.7, 3.8 and 3.9. However, mutual information based methods incur a performance drop as shown in Tables 3.4 and 3.18. iii) Mutual information based methods and fair PCA based method are also effective in reducing the representational bias, however mutual information based methods could lead to a loss in accuracy.

In scenarios where the tasks are not complex one can use the mutual information based methods as they are easy to compute, as shown in Table 3.3, where retaining 400 dimension seems to be enough to achieve decent performance to retrieve images of different professions. On the other hand, if the task is complex (such as for queries

Table 3.17: [Retrieval - Statistical tests - Subjective - Flickr30K - OpenCLIP] This table shows the statistical tests for the cosine similarities for different queries between men and women. The first number refers to the test statistic while the second number is the p-value. If there is a statisitically significant difference among different groups the test statistic would be high and p-value would be low.*Specifically, it shows the Alexander-govern statistical test whether the mean of cosine similarity between men and women for a given query are statistically significant. It shows that fair PCA trained on ground truth protected attribute labels yields statistically insignificant differences.*

Statistical tests: ANOVA- Alexander-Govern: (statistic, p-val)							
Query	CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	FPCA-GT	FPCA-INF
				Gender			
boss	(958 , 0.0)	(280 , 0.0)	(63 , 0.0)	(19 , 0.0)	(0 , 0.374)	(0 , 0.364)	(52 , 0.0)
doctor	(27 , 0.0)	(2 , 0.096)	(67 , 0.0)	(10 , 0.001)	(5 , 0.017)	(0 , 0.395)	(5 , 0.019)
lawyer	(18 , 0.0)	(24 , 0.0)	(59 , 0.0)	(61 , 0.0)	(7 , 0.005)	(1 , 0.281)	(4 , 0.035)
nurse	(1396 , 0.0)	(4 , 0.037)	(5 , 0.024)	(29 , 0.0)	(1 , 0.306)	(0 , 0.612)	(5 , 0.015)
paralegal	(1112 , 0.0)	(0 , 0.608)	(0 , 0.935)	(13 , 0.0)	(12 , 0.001)	(0 , 0.909)	(21 , 0.0)
secretary	(1729 , 0.0)	(104 , 0.0)	(2 , 0.091)	(80 , 0.0)	(18 , 0.0)	(0 , 0.846)	(19 , 0.0)

'a funny person' or 'an affectionate person') even retaining 400 dimensions can lead to random results as shown in Figure 3.15. The results seem much worse where we retain only 256 dimensions.

Q6, Q7 To check if statistically significant differences in cosine similarity exist between different groups of the protected attribute, we performed the Alexander Govern test[8] for every subjective query. The null hypothesis is that all the groups have the same mean cosine similarity for a given query, while accounting for heterogeneity of variance across the groups. The results show that while the effect size of the differences in cosine similarity across different groups is reduced with all the debiasing methods, only with fair PCA these differences are statistically insignificant for most queries, as shown in Tables 3.9, 3.10, 3.11 and 3.12. It is interesting to notice that even though fair PCA based methods produce embeddings that do not have statistically significant differences in the cosine similarities for different queries, they still do not necessarily produce the most fair results in all cases for image retrieval. The main reason for this is that we select a subset of images from a dataset and even if the representations are unbiased, we might pick a subset that is skewed towards one group.

[8] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.alexandergovern.html

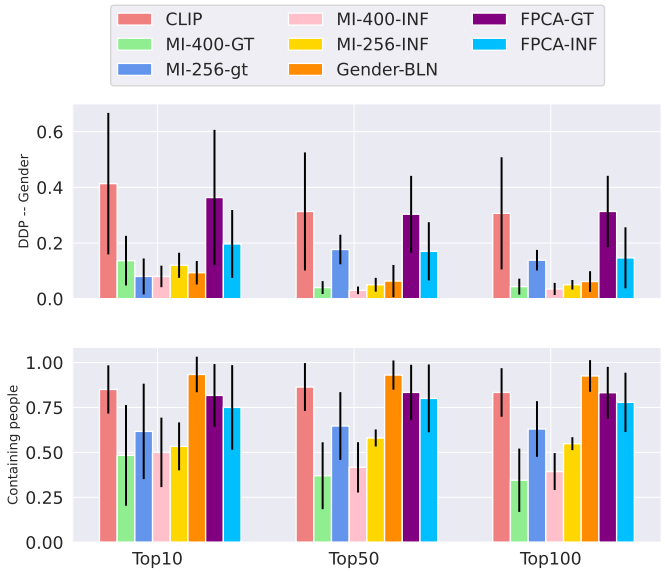


Figure 3.15: [Retrieval - DDP - Subjective - MSCOCO] The figure on the top shows DDP, given by Eq. (3.6), for retrieval tasks using MSCOCO dataset. *These results demonstrate bias in human-centric subjective tasks. At the bottom, we observe the fraction of query results that actually include a person. Surprisingly, for many human-related queries, the retrieved images do not feature any humans at all. Additionally, this demonstrates that the simple baseline of gendered queries perform very well in reducing disparity. However, the mutual information-based approaches, although effective in reducing disparity in some cases, fail to retrieve images containing humans. Interestingly, Fair PCA, trained on the inferred gender attribute, manages to return appropriate images while still reducing some disparity. One possible reason for this could be that the gender labels derived from the captions, which serve as ground truth, are quite noisy. In contrast, training fair PCA on on the inferred gender attribute directly from the CLIP model appears to yield better results in this context.*

3.5.5 Image Captioning

Difficulty addressing fairness in captioning One would expect that an image captioning system should perform equally well for different groups on the standard metrics such as Bleu [208], METEOR [20], Rouge [175], CIDEr [253], SPICE [16]. Using the data by Zhao et al. [281] we evaluated the captions generated by CLIP-CAP system for both original and trained on gender-neutral captions, but similar to Zhao et al. [281] we only found a slightly better performance of these metrics on the images of light skin individuals. Additionally, we did not find any difference on the aforementioned performance metrics for the captions between men and women or intersectional groups (considering both race and gender).

One can extend the notion of independence of protected attribute w.r.t. to a prescribed set of words in caption generation systems as follows: Given an image, predefined relevant words used in the captions should be independent of the protected

attribute. For example, given images of doctors the occurrence of the word doctor, hospital etc. in the generated captions should be independent of gender or race. However, evaluating for such fairness issues requires appropriate image datasets with demographic features. Additionally, it requires to define a set of relevant words for every (type of) image. Unfortunately, several available datasets crawled from the web contain biased images (e.g., female doctors wearing a halloween costume or having cartoonized images). So, it is difficult to draw broader conclusions from such datasets.

Q8 Fairness issues in captioning: We report qualitative results using handpicked images from google search. We found that images of women factory-workers were misgendered. A woman fixing a light-fixture was described as holding a blow-dryer. A woman shown fixing a car is captioned "kneeling over a car" while a man shown fixing a car is captioned "fixing a car". Women who appeared to be medical professionals were captioned "talking to a man/woman", or a woman wearing a lab-coat is referred to "wearing a dress talking to a man". While images of men who appeared to be medical professionals were referred to as "a couple of doctors". In general, captions for images of men more often had the words, "hospital", "check-up on a patient" , compared to images of women. In some cases women medical professionals were referred to as "nurse", while in none of the cases men were referred to as nurses.

Using gender information extracted from CLIP, we found that on IdenProf dataset's images labeled as doctor, the word nurse was used in 1.7% of the generated captions for women, vs for men it was only used in 1.2% of the captions. Similarly, for Chef's images of women the word "Chef" only appeared in 17% of the generated captions while it appeared for 36% of the captions for men. Additionally, we saw that the word "Kitchen" appeared in 45% of the captions for Chef's images labeled as women and it appeared 40% of the captions for the Chef's images labeled as men. The waiter's images in IdenProf had the word "Chef" in 1.2% of the captions for women vs 4.1% of the captions for men. These are just preliminary findings and a more thorough analysis requires ground truth demographic features as opposed to using CLIP's predictions.

Using the dataset by Kay et al. [146] we find that for Chef's images the word chef appears 33% of the images for men while it occurred 0% of the images for women labeled as chef. On the other hand, the word "chef's" appears 13% of the images for men and 24% of the images for women. This occurs in the context of 'chef's hat' or 'chef's uniform'. This shows that the captioning system recognizes women as wearing chef's clothings but does not associate the word 'chef' with them. We would like to point out that this dataset did not seem appropriate as it was crawled from Google search and had several biases, e.g., it sometimes showed women as a cartoon.

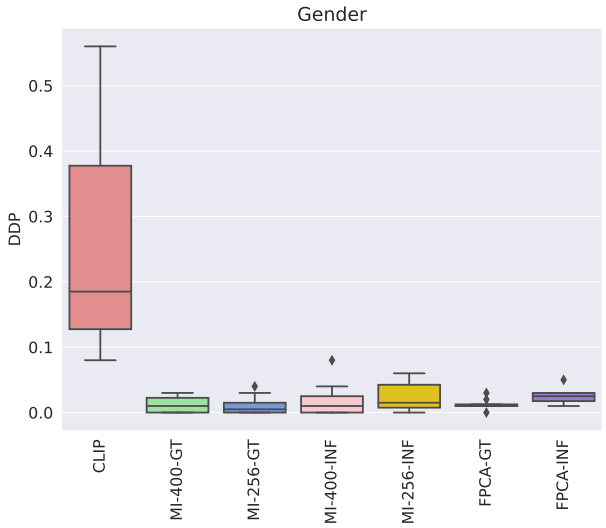


Figure 3.16: [Classification - DDP - Subjective - MSCOCO] The figure on the top shows DDP, given by Eq. (3.5), for classification tasks using MSCOCO dataset. *These results show bias for human-centric subjective tasks. They demonstrate that for most methods reduce disparity across gender in classification tasks.*

Q9 Effects of bias mitigation methods: We only discuss results on handpicked images. To fix the misgendering of images, we trained the captioning system with gender neutral words, that is we changed words like “man" or “woman" to “person". This helped fix the misgendering issue. In some cases it even helped with changing the captioning all together, i.e., we saw more mentions of the word hospital for women in the appropriate images. Using mutual information and fair PCA based methods on CLIP embeddings plus the gender-neutral training captions seemed to lower the use of the biased language. For example, there were more medical terms, e.g., “hospital" or “doctor", used in the captions for women. In one cases the caption changed from "nurse" to a "doctor". We only tested the bias mitigation methods on few handpicked images from the web which we cannot show for copyright reasons.

3.5.6 OpenCLIP Results

We show results using OpenCLIP [130] for zero-shot classification on FairFace dataset (gender and race attributes) in Figure 3.11. We also show results using Flickr30K dataset in Figure 3.13. We find that i) OpenCLIP has more bias compared to OpenAI’s CLIP. ii) CLIP bias mitigation methods are effective in enforcing independence assumption for different protected attribute groups. iii) In general, fair PCA based methods are more effective. We also evaluate OpenCLIP and different bias mitigation methods using

OpenCLIP for image retrieval tasks, both for enforcing independence of the protected attribute w.r.t. top-k selection, FairFace Figure 3.12 and Flickr30K Figure 3.14, as well as the representation bias mitigation, FairFace Table 3.14 and Flickr30K Table 3.16. i) The results show that OpenClip has a higher bias compared to OpenAI CLIP. ii) All the methods are effective in reducing different biases. iii) However, fair PCA based methods are the most effective, which is supported by the low disparity in the average cosine similarity for different gendered queries, as shown in Figures 3.10 and 3.14. iv) Fair PCA based methods produce embeddings that show no statistical difference in the cosine similarity across different protected groups for different queries, as shown in Tables 3.15 and 3.17.

3.5.7 In-processing Fairness for CLIP-Like Models

FairSampling (referred to as FairSamp in the results) This is the second mitigation method proposed by Wang et al. [262], which requires to train a CLIP-like model from scratch. Even though it provides embeddings which could be used for other downstream tasks, one prominent difference from CLIP-like models is that it is trained on MSCOCO, a much smaller dataset. So, its zero-shot capabilities are quite limited. We add these results for the sake of completeness.

During training this method picks the training examples in a balanced manner w.r.t. gender. Specifically, in contrastive loss the goal is to maximize the similarity scores between matching image and text examples (positive samples), while minimizing the similarity score between non-matching examples (negative samples). Wang et al. [262] hypothesize that there could be a gender imbalance in the negative samples in each batch, i.e., the negative samples could be biased towards the majority class which results in the bias during retrieval. To correct this, firstly, they assign male, female or neutral labels to each image-text pair in the training set. They extract these labels from the texts or captions of each image. Then, they propose to pick negative sample from the male and female datapoints with probability 0.5 for every neutral query, while for male and female labelled queries they sample the negative samples randomly.

We found that on MSCOCO dataset, which was used for training this method, it enforced demographic parity, and had good performance for recall. However, as Table 3.18 shows, this method is not directly comparable to foundation models and its performance is limited to the dataset it was trained on.

Table 3.18: [Retreival - Precision - Objective - MSCOCO & CelebA] This table shows average precision@K for image retrieval tasks using different methods for 80 categories of MSCOCO dataset and 9 attributes of CELEBA. *It demonstrates that CLIP and fair PCA methods usually yield similar precision. On the other hand, fair sampling which is trained on MSCOCO does very well on the MSCOCO dataset but has a poor performance on CELEBA dataset. The mutual information based methods have a better performance where more dimensions of the CLIP embeddings are used.*

CLIP	MI-400-GT	MI-256-GT	MI-400-INF	MI-256-INF	Fair-Samp	FPCA-GT	FPCA-INF
			Precision@20 using MSCOCO				
0.9±0.04	0.9±0.04	0.87±0.04	0.87±0.04	0.86±0.04	**0.91±0.04**	0.9±0.04	0.9±0.04
			Precision@50 using MSCOCO				
0.86±0.04	**0.87±0.04**	0.83±0.04	0.83±0.04	0.83±0.04	**0.87±0.2**	0.86±0.04	0.86±0.04
			Precision@70 using MSCOCO				
0.85±0.04	**0.85±0.04**	0.81±0.06	0.81±0.04	0.82±0.04	**0.85±0.04**	**0.85±0.04**	0.84±0.04
			Precision @20 using CELEBA				
0.88±0.06	0.82±0.1	0.67±0.18	0.71±0.12	0.71±0.14	0.67±0.16	0.84±0.08	0.87±0.06
			Precision@50 using CelebA				
0.85±0.08	0.78±0.1	0.65±0.16	0.72±0.12	0.71±0.12	0.68±0.16	0.81±0.1	0.84±0.08
			Precision@100 using Celeba				
0.82±0.08	0.76±0.1	0.65±0.14	0.73±0.12	0.69±0.1	0.67±0.18	0.78±0.1	0.81±0.08

3.6 Related Work

3.6.1 Text Embeddings and Bias

Compared to multi-modal embeddings, pure text embeddings have a longer history, and so does the literature about their fairness: the seminal paper of Bolukbasi et al. [31] found that word embeddings encode stereotypes such as "man is to computer programmer as woman is to homemaker." Such bias is attributed to the consistent bias prevalent in text corpora [24, 258]. Bolukbasi et al. [31] proposes a debiasing approach that is conceptually similar to the fair PCA approach [154] that we study in this paper. Concretely, it aims to project gender-neutral words to a subspace orthogonal to the gender-direction in the embedding space (when trying to remove gender bias). A different approach to debias word embeddings has been proposed by Zhao et al. (2018), which alters the loss of the word embedding model. Both approaches have been criticized by Gonen and Goldberg [103] to only hide the bias, rather to remove it.

3.6.2 Further (Fairness) Aspects of CLIP

Birhane et al. [29] examined the LAION-400M dataset [230], which has become a popular dataset for training CLIP-like foundation models [57], and found that the dataset contains problematic content, including malign stereotypes and racist and ethnic slurs. Such problematic content is likely to be picked up by large models trained on this dataset. CLIP-like models can be adapted to support multiple languages by means of cross-lingual alignment [63]. Wang et al. [263] study the fairness of Multilingual CLIP [48] w.r.t. different languages and find significant accuracy disparity across different languages. Liang et al. [174] presented the modality gap phenomenon in multi-modal models: for example, CLIP maps an image and its corresponding text to completely separate regions of the joint embedding space. They showed that varying the modality gap distance can significantly improve CLIP's fairness. Qiu et al. [217] studied the robustness of multi-modal foundation models to distribution shifts [267].

In a concurrent work Seth et al. [232] proposed a new bias mitigation method for vision-language models. They propose to train a residual network on top of the image embeddings ($\bar{\phi}$) of CLIP-like models with the goal to produce representations (ϕ) such that protected attributes cannot be recovered from it. They do so by first training a protected attributes classifier (PAC) using $\bar{\phi}$ which is then frozen. Then they train the residual network while trying to maximize PAC's loss for the learnt ϕ. They show that they can reduce the maximum and minimum Skew for gender, age and race attributes on FairFace and PATA (newly introduced) dataset.

In another parallel work, Chuang et al. [61] presented an approach that addresses bias in CLIP's embeddings space by projecting out the biased directions. They identify the biased directions in the embedding space by using prompts like 'a photo of a male/female' and then construct a projection matrix that would remove these biased directions in any query. To reduce noise in the estimation of the 'biased directions', they defined a set of queries on which the CLIP model should have similar embeddings, e.g., 'a photo of a female doctor' and 'a photo of a male doctor'. They additionally added this constraint to find the debiasing projection matrix. They showed that they reduce the Skew for gender, race and age attributes for image retrieval tasks using the FairFace dataset.

3.7 Conclusion

We have introduced a novel taxonomy to systematically evaluate discriminative foundation models. It is based on three axes: (i) whether the task involves a human; (ii)

whether the task is subjective; and (iii) whether independence-based or diversity-based fairness is better suited for the intended use case. Then we thoroughly evaluated the fairness of discriminative foundation models (FM) taking OpenAI's CLIP and OpenCLIP models as examples. Additionally, we evaluated different bias mitigation approaches for these models. Our evaluation focused on three key tasks: zero-shot classification, image retrieval and image captioning. We specifically examined two protected attributes: gender (binary) and ethnicity (multi-valued). We found that, while fair PCA generally emerged as one of the top-performing approaches in most cases, selecting the appropriate debiasing method should be based on the intended use of the model. For instance, when aiming to enhance diversity in image retrieval tasks, simpler methods that involve constructing gender or race-specific queries may be more suitable.

CHAPTER 4

Fairly updating an ADMS: Loss-Aversively Fair Classification

In this chapter, we focus on a crucial aspect of desigining algorithmic deicions-making systems ignored by existing studies on fair learning namely, fairness of *updates to decision-making systems*. We ask the following key questions:

What constitutes a fair update of an already deployed ADMS?

As mentioned in Sections 1.2, our goal is challenging as there is no existing notion of update-fairness. So, firstly we have to propose a reasonable notion of update-fairness that is grounded in exisiting research in social sciences. Then, we have to design measures and mechanisms for this notion which can be incorporated into existing algorithms.

In order to address our research question in this chapter, we proceed as follows:

- In Section 4.1, we briefly describe the prior work particularly related to this chapter.
- In Section 4.2, we propose a new notion of fair update inspired by existing literature in behavioral economics. We also operationalize our proposed notion of fairness and provide mathematical formalism.
- In Section 4.3, we propose mechanisms to incorporate our proposed notion of fairness in binary classification problems. We also provide convex proxies for our notion to be incorporated along with existing notions of non-discrimination into a variety of linear and non-linear classifiers.
- In Sections 4.4 and 4.5, using synthetic datasets, we demonstrate the effectiveness of training classifiers with our proposed update-fairness combined with statistical parity and equality of opportunity, respectively.

- In Sections 4.6 and 4.7, using real-world datasets, we demonstrate the effectiveness of training classifiers with our proposed update-fairness combined with statistical parity and equality of opportunity, respectively.
- In Section 4.8, we conclude this chapter.

Relevant publication

The results presented in this chapter have been published in [13].

4.1 Related work

In this section, we briefly describe the prior work related for this chapter. **Normative vs. Descriptive Notions of Fairness.** Our fairness consideration for updating decision making systems has roots in normative vs. descriptive approaches in behavioral economics [138, 141]. For example, Kahneman et al. [138] show how certain changes to an economic model that are accepted on the normative standards might be deemed unacceptable on the descriptive standards. Our work here is motivated by such observations: while anti-discrimination laws (normatively) prescribe how nondiscriminatory decisions ought to be done, if people (descriptively) preceived the changes in outcomes with the new nondiscriminatory decision system to be too disruptively disadvantageous to them, they would resist adopting the new system. Our notions of update fairness can be thought of as addressing such practical considerations.

4.2 New notion of fairness: Loss-aversive updates

In this work, inspired by existing literature in behavioral economics, we formally define a notion of update fairness namely, **loss-aversively fair updates**. Intuitively, our notion of loss-averse updates accounts for the "endowment effect" in human behavior [138, 141], where an individual or a group of users perceives the fairness of the new system based on whether their new outcomes were more or less beneficial than their status quo outcomes from the existing system.

We also show that our new notion of fair update can be easily integrated with existing mechanisms for training non-discriminatory classifiers. For instance, when attempting to equalize rates of beneficial outcomes such as positive class acceptance rate or true positive rate across different groups, adding our loss-averse update constraint ensures that "no group of users is worse-off" than before. Such a constraint may be necessary in practice when training non-discriminatory classifiers as Bazerman et al. [21] point out

that same "don't make anyone worse off' principle likely underlines Supreme Courts decision [241] that firing personnel from historically advantaged groups to achieve parity (in order to overcome past discrimination) is prohibited.

4.2.1 Formalizing Notion of Loss-Averse Updates

In this section, we formally define a notion of fairness that can be useful when updating algorithmic decision making systems. Specifically, we focus on decision making tasks centered around binary classification.

Preliminaries. In a binary classification task, given a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, the goal is to learn a function $\boldsymbol{\theta} : \mathbb{R}^d \to \{-1, 1\}$ between the feature vectors $\boldsymbol{x} \in \mathbb{R}^d$ and class labels $y \in \{-1, 1\}$. For convex decision boundary-based classifiers like logistic regression and (non)linear SVM, this task boils down to finding a decision boundary $\boldsymbol{\theta}^*$ in the feature space that minimizes a given loss $L(\boldsymbol{\theta})$ over $\mathcal{D}$, i.e., , $\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. The convexity of the loss function ensures that the optimal decision boundary parameters can be found in an efficient manner. Then, for a given (potentially unseen) feature vector $\boldsymbol{x}$, one predicts the class label $\hat{y} = 1$ if $d_{\boldsymbol{\theta}^*}(\boldsymbol{x}) \geq 0$ and $\hat{y} = -1$ otherwise, where $d_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ denotes the signed distance from $\boldsymbol{x}$ to the decision boundary. Without loss of generality, we consider $\hat{y} = 1$ to be the beneficial (desired) label, e.g., , being granted the loan or being released on bail.

Setup. We consider scenarios where we need to update an existing, *status quo*, binary classifier, whose decision boundary is denoted by $\boldsymbol{\theta}_{sqo}$. We assume that the boundary of the new classifier, $\boldsymbol{\theta}_{new}$ is learnt from the training dataset $\mathcal{D}$. The outcomes of the updated (new) classifier may differ from the status quo for many reasons such as the status quo classifier being a human or an older (simpler) learning model, or the status quo classifier being trained on out-dated training data, or the status quo classifier being trained using models without awareness of potential for discrimination. Our notion of fair update defines the conditions in which the *changes in decision outcomes caused by an update* would be deemed as fair.

Existing Notions: Discrimination in Classification.

Anti-discrimination laws require classification outcomes are also required to be nondiscriminatory with respect to a sensitive feature $z \in \{0, 1\}$, e.g., , gender, race. Most of the existing studies differentiate between the following two notions of discrimination: *statistical parity* [79, 89]—also referred to as disparate impact, and *equality of opportunity* [120, 274]—also referred to as disparate mistreatment. Both notions require that

certain group-conditional beneficial outcome rates be the same for each group, i.e., :

$$\mathcal{B}_{z=0}(\boldsymbol{\theta}) = \mathcal{B}_{z=1}(\boldsymbol{\theta}), \tag{4.1}$$

where the definition of the benefit function $\mathcal{B}_z$ depends on the notion of discrimination under consideration.

Under the notion of *statistical parity (SP)* [79, 89], the benefits function is defined as the positive class acceptance rate (AR), i.e., , the positive class acceptance rate should be the same for both the groups. More formally,

$$\text{— } \textbf{\textit{SP:}} \qquad P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1), \tag{4.2}$$

Under *equality of opportunity (EOP)* notion [120, 274], the benefit function is defined as the true positive rate, i.e., the true positive rate (TPR) should be the same for both the groups. More formally,

$$\text{— } \textbf{\textit{EOP:}}\ P(\hat{y} = 1|y = 1, z = 0) = P(\hat{y} = 1|y = 1, z = 1), \tag{4.3}$$

Note that, current notions of nondiscrimination do not take into account status quo classifier. In the following section we introduce a notion of updating status quo classifier.

New Notion: Loss-Averse Updates. We now formally describe a new consideration of fair updates, introduced in Section 1. We draw inspiration from human behavior and behavioral economics and we consider how people might perceive fairness of an updated classifier in comparison to status quo. Specifically, any disadvantageous effect of an updated classifier would be considered unfair. Prospect theory, proposed by Kahneman and Tversky [141], states that equal amounts of loses result in a bigger loss in utility than the increase in utility by the same amount of gains. In other words people percieve losses much worse than gains, i.e., , they are loss-averse. Given the *status quo* classifier $\boldsymbol{\theta}_{sqo}$, a new classifier $\boldsymbol{\theta}_{new}$ constitutes a *loss-averse* update only when the new classifier increases the beneficial outcome rates for all groups. More formally,

$$\mathcal{B}_{z=k}(\boldsymbol{\theta}_{new}) \ \geq \ \mathcal{B}_{z=k}(\boldsymbol{\theta}_{sqo}), \quad \text{for all } k \in \{0, 1\} \tag{4.4}$$

where $\mathcal{B}_z$ can be any one of the benefit functions proposed in the existing literature on nondiscriminatory classification.

4.3 Updating Classifiers Loss-Aversively

In this section, we devise mechanisms to update status quo classifier, $\boldsymbol{\theta}_{sqo}$ to $\boldsymbol{\theta}_{new}$ that follow the practical considerations of "loss-averse updates". We specifically focus on training convex decision boundary based classifiers (e.g., , logistic regression, linear and non-linear SVMs), i.e., , the classifiers that learn the decision boundary parameters by optimizing a convex loss function $L(\boldsymbol{\theta})$.

Existing Mechanisms: Nondiscriminatory Classification. Existing mechanisms to train nondiscriminatory classifiers involve solving an optimization problem maximizing accuracy while equalizing benefits, i.e., , enforcing Eq. (4.1), for different sensitive feature groups. More formally,

$$\begin{aligned} &\text{minimize} && L(\boldsymbol{\theta}) \\ &\text{subject to} && \mathcal{B}_{z=0}(\boldsymbol{\theta}) = \mathcal{B}_{z=1}(\boldsymbol{\theta}), \end{aligned} \tag{P4.1}$$

Constraints in Problem (P4.1), as operationalized in Eqs. (4.2) and (4.3) are non-convex. However, prior studies [22, 274, 275] propose tractable convex or convex-concave proxies for enforcing the equality of benefits constraint in Eqs. (4.2) and (4.3). Borrowing these proxies from [22, 274, 275], one can replace the equal benefits condition with proxies as follows:

$$\textit{— SP:} \qquad \frac{1}{|\mathcal{D}|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}}(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| \leq c, \tag{4.5}$$

$$\textit{— EOP:} \qquad \frac{1}{|\mathcal{D}_+|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}_+}(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| \leq c, \tag{4.6}$$

where $\mathcal{D}_+$ are data points with $y = 1$. Here equality of opportunity limits discrimination in true positive rates of different groups. The covariance threshold $c \in \mathbb{R}^+$ determines the level of discrimination, with $c = 0$ aiming for a perfectly fair classifier.

New Mechanism: Loss-Averse Updates. For updating the status quo classifier, $\boldsymbol{\theta}_{sqo}$, in a nondiscriminatory and loss-aversive manner, one can add the respective conditions to the classifier formulation as a constraint, i.e., ,

$$\begin{aligned} &\text{minimize} && L(\boldsymbol{\theta}) \\ &\text{subject to} && \mathcal{B}_{z=0}(\boldsymbol{\theta}) = \mathcal{B}_{z=1}(\boldsymbol{\theta}) \\ & && \boldsymbol{\mathcal{B}_{z=k}(\theta) \geq \mathcal{B}_{z=k}(\theta_{sqo})}, \quad \text{for all } k \in \{0,1\}. \end{aligned} \tag{P4.2}$$

The constraints in the above problem are nonconvex functions of the classifier parameters $\boldsymbol{\theta}$, if $\mathcal{B}$ is defined in terms of probabilities as given in Eqs. (4.2) and (4.3), for example, this would make it very challenging to solve the resulting problem in an efficient manner.

We used the convex proxies from prior studies [22, 274, 275] for the first constraint as given by Eqs. (4.5) and (4.6). We propose the following convex proxies to approximate the new loss-averse constraints in Problem (P4.2):
Under SP, when the benefit function is AR we suggest:

$$\frac{1}{|\mathcal{D}_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}_{z=k}} d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \frac{1}{|\mathcal{D}_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}_{z=k}} d_{\boldsymbol{\theta}_{sqo}}(\boldsymbol{x}) + \gamma, \tag{4.7}$$
$$\text{for all } k \in \{0,1\}, \gamma \in \mathbb{R}^+.$$

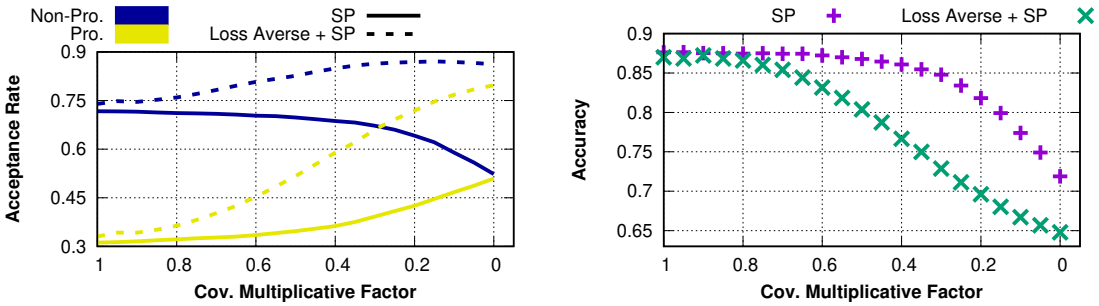
Under EOP, when the benefit function is TPR we suggest:

$$\frac{1}{|\mathcal{D}^+_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}^+_{z=k}} d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \frac{1}{|\mathcal{D}^+_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}^+_{z=k}} d_{\boldsymbol{\theta}_{sqo}}(\boldsymbol{x}) + \gamma, \tag{4.8}$$
$$\text{for all } k \in \{0,1\}, \gamma \in \mathbb{R}^+,$$

where $\mathcal{D}_{z=k}$ are the data points whose sensitive attribute value $z = k$, and $\mathcal{D}^+_{z=k}$ are data points in the dataset with label $y = 1$ and sensitive attribute value $z = k$. Here, γ controls the strength of the constraint. We pick an appropriate γ using a validation set. Note that the right hand side in Eqs. (4.7) and (4.8) represents constant terms since $\boldsymbol{\theta}_{sqo}$ is already known.

Both of the proposed proxies are convex with respect to the optimization variables. The convexity of the proxies (4.7 and 4.8) means that for any convex function $L(\boldsymbol{\theta})$ the optimization problem stays convex and can be solved in an efficient manner.

Logistic Regression: SP. We can specialize Problem (P4.2), using logistic regression classifier with L-2 norm regularizer, SP as a notion of discrimination, given by Eq. (4.5), and loss-averse constraint, given by Eq. (4.8), as follows:



(a) Statistical Parity (SP) **(b)** Nondiscrimination-accuracy tradeoff

Figure 4.1: [Synthetic dataset. Enforcing statistical parity] These figures show a comparison between the solutions of Problem (P4.1), using SP proxies, and Problem (P4.3). Left panel shows the beneficial outcome rates, i.e., , positive class acceptance rates, for a classifier only enforcing SP constraint (solid lines), and a classifier additionally enforcing the "loss-averse" constraint (dotted lines). Right panel shows the nondiscrimination-accuracy tradeoff for both the classifiers. Enforcing "loss-averse" constraint, defined in Eq. (4.7), leads to significant additional loss in accuracy for the same level of discrimination.

$$\begin{aligned}
&\text{minimize} \quad -\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}||^2 \qquad \text{(P4.3)}\\
&\text{subject to} \quad \frac{1}{|\mathcal{D}|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}} (z-\bar{z}) d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| < c\\
&\frac{1}{|\mathcal{D}_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}_{z=k}} d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \frac{1}{|\mathcal{D}_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}_{z=k}} d_{\boldsymbol{\theta}_{sqo}}(\boldsymbol{x}) + \gamma,\\
&\quad \text{for all } k\in\{0,1\}, \gamma\in\mathbb{R}^+.
\end{aligned}$$

Logistic Regression: EOP. Similarly, considering equality of opportunity as a notion of nondiscrimination we can approximate Problem (P4.2), by adding Eqs. (4.6 and 4.8) as constraints to logistic loss, as follows:

$$\begin{aligned}
&\text{minimize} \quad -\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}||^2 \qquad \text{(P4.4)}\\
&\text{subject to} \quad \frac{1}{|\mathcal{D}_+|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}_+} (z-\bar{z}) d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| < c\\
&\frac{1}{|\mathcal{D}^+_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}^+_{z=k}} d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \frac{1}{|\mathcal{D}^+_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}^+_{z=k}} d_{\boldsymbol{\theta}_{sqo}}(\boldsymbol{x}) + \gamma,\\
&\quad \text{for all } k\in\{0,1\}, \gamma\in\mathbb{R}^+.
\end{aligned}$$

4.4 Evaluation on Synthetic Dataset: SP

In this section we evaluate the effectiveness "Loss-averse" constraint (4.7), using a synthetic dataset on a binary classification task. We consider a known notion of nondiscrimination, namely statistical parity.

4.4.1 Dataset and Experimental Set up

In this subsection we describe the results for statistical parity using a synthetic dataset.
Dataset and Experimental Set up. We used synthetic dataset with binary ground truth class labels $y \in \{+1, -1\}$. Each data point comprises of 2 features besides a binary sensitive feature, i.e., , $z \in \{0, 1\}$, where $z = 0$ is the protected group. We do not use the sensitive attribute during training.
Synthetic Dataset. For demonstrating the results of loss-averse updates with statistical parity, given by Eq. (4.2), as a notion of nondiscrimination, we used the dataset proposed by Zafar et al. [275]. This dataset comprises of 6000 data points, the class labels were drawn uniformly at random. Conditioned on the class membership, each data point was sampled from the following distributions:

$$p(\boldsymbol{x}|y = 1) = N([2; 2][5, 1; 1, 5]),$$
$$p(\boldsymbol{x}|y = -1) = N([-2; -2][10, 1; 1, 3]).$$

Value of the sensitive attribute was sampled from the following Bernoulli probability distributions:

$$p(\boldsymbol{z} = 1) = \frac{p(\boldsymbol{x}^{'}|y = 1)}{p(\boldsymbol{x}^{'}|y = 1) + p(\boldsymbol{x}^{'}|y = -1)},$$

where, $\boldsymbol{x}^{'} = [\cos(\phi), -\sin(\phi); \sin(\phi), \cos(\phi)]\boldsymbol{x}$, i.e., , the rotated feature vector, $\boldsymbol{x}$. On average there were 3280 points in the protected group and 2720 were in non-protected group.
Experimental Setup. The dataset is split into 70%-30%, train-test folds. Additionally, hyperparameters are validated using a 30% hold out set from the training data. All the results have been averaged over 5 shuffles of the data initialized by different random seed. In order to pick the penalization parameter, λ in Problem (P4.3), multiplied with the regularizer, we trained the unconstrained classifier for $\lambda \in [1e-5, 1e-2]$. Then, we picked a value which yielded the highest accuracy on the validation set, for a particular shuffle of the data . We used this value of the parameter for *all* the experiments on that shuffle of the data. We use CVXPY [75] library to solve all the optimization problems.

4.4.2 Loss-aversively Fair Updates

In this section we experiment with Problems (P4.1 and P4.3). First we consider statistical parity, where beneficial outcome rates are defined as positive class acceptance rate, as a notion of discrimination, i.e., , solving Problems (P4.1) using SP proxies. Then, we show results combining SP and loss-averse constraints and we update $\boldsymbol{\theta}_{sqo}$ with loss-averse nondiscriminatory classifiers.

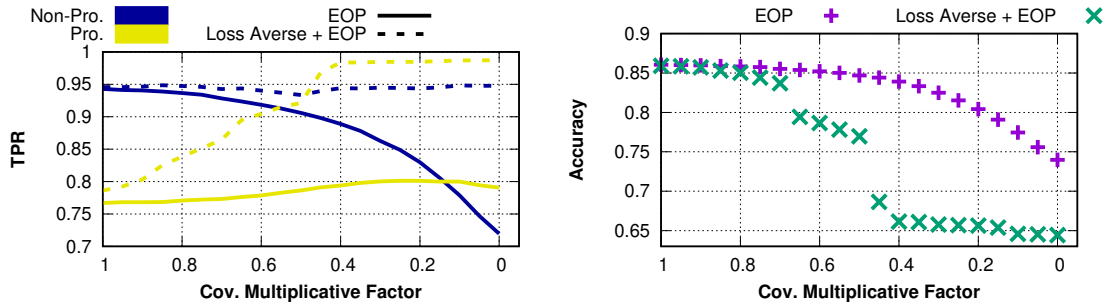
Training Loss-aversively Fair Classifier. We initialize θ_{sqo} with the solution of unconstrained problem. Then, given a value of covariance threshold c, as used in Eqs.(4.5 and 4.6), and a range of γ, as used in Eqs.(4.7 and 4.8), we solve Problem (P4.3). We, then, pick the gamma values whose solutions yield a higher benefits compared to $\boldsymbol{\theta}_{sqo}$, for all the groups, on the validation set. In case there are multiple such values, we pick the one whose solution yields maximum accuracy. We then report the results on the test set.

SP. Accuracy of an unconstrained classifier, on *Synthetic* dataset, is 88%, and the acceptance rates for the protected and non-protected groups are 31% and 72%, respectively. There is a clear disparity in acceptance rates of both the groups. In order to remove this disparity we solve Problem (P4.1), replacing the first constraint with SP proxy, given by Eq. (4.5). For a covariance threshold $c = 0$, this leads to a classifier with an acceptance rate of 51% and 52%, for protected and non-protected groups respectively, and an accuracy of 72%.

The results for this formulation, Problem (P4.1) specialized with SP, are shown in Figure (4.1). The x-axis is covariance multiplicative factor $m : c = m \times c^*$, where c^* is the covariance values of the unconstrained classifier and c is covariance threshold as given in Eq.(4.5). *Solid lines* in Figure (4.1a) represent the statistics of the classifiers resulting from the solutions of this formulation. Figure (4.1b) shows the accuracies of classifiers resulting from solving this formulation in *purple* colored points.

Note that: i) Figure (4.1b) demonstrates that as the covariance is decreased the accuracy of the resulting, less discriminatory, classifiers also decreases. ii) Figure (4.1a) shows that as the covariance decreases, the discrimination also reduces. iii) However it should be noted that discrimination is decreased by reducing the acceptance rate of the non-protected group.

Loss-Aversiveness + SP. In order to train a classifier enforcing loss-averse update of $\boldsymbol{\theta}_{sqo}$, Eq. (4.4), combined with statistical parity, Eq (4.2), on the *Synthetic* dataset, we solve Problem (P4.3). Loss-averse updates yield a classifier with an accuracy of 65% and acceptance rates of 80% and 86% for protected and non-protected groups, respectively, for the covariance value $c = 0$.



(a) Equality of Opportunity (EOP) **(b)** Nondiscrimination-accuracy tradeoff

Figure 4.2: [Synthetic dataset. Enforcing equality of opportunity] Figure on the left shows the beneficial outcome rates, i.e., , true positive rates, for a classifier only enforcing EOP constraint (solid lines) and a classifier additionally enforcing the "loss-averse" constraint, given in Eq. (4.8), is shown in dotted lines. Figure on the right shows nondiscrimination-accuracy tradeoff for both the classifiers.

The results are shown in Figure (4.1a) in *dotted lines* and in *green* colored points in Figure (4.2b). i) The figures demonstrate that loss-aversively fair updates yield a less discriminatory classifier while increasing the benefits for both the groups, ii) however this comes at a higher cost of accuracy.

4.5 Evaluation on Synthetic Dataset: EOP

In this section we will present the "loss-averse" fairness results combined with equality of opportunity, using synthetic dataset. We show the results of the optimization Problem (P4.4).

4.5.1 Dataset and Experimental Setup

In this section we explain the synthetic dataset used for demonstrating the loss-averse consideration and the experimental setup used to solve the optimization Problem (P4.4).
Synthetic Dataset. Each data point comprises of 2 features apart from the sensitive attribute. Each data point also has a binary ground truth label. For equality of opportunity, as given by Eq. (4.3), we are considering true positive rates as a notion of benefit. To demonstrate the results of fair updates combined with EOP, we use a synthetic dataset proposed by Zafar et al. [274], except that we flip the ground truth labels in order to have a disparity in the false negative rates instead of the false positive rates. We generated

16000 data points with the probability distributions of the features given as follows:

$$p(\boldsymbol{x}|z=0,y=1) = N([2;2][3,1;1,3])$$
$$p(\boldsymbol{x}|z=1,y=1) = N([2;2][3,1;1,3])$$
$$p(\boldsymbol{x}|z=0,y=-1) = N([1;1][3,3;1,3])$$
$$p(\boldsymbol{x}|z=1,y=-1) = N([-2;-2][3,1;1,3])$$

Both, class labels, y, and value of the sensitive attribute, z, were sampled uniformly at random.

Experimental Setup. We use the same data split and method of validating the hyperparameters as explained in the Section 4.4.

4.5.2 Loss-Aversively Fair Updates

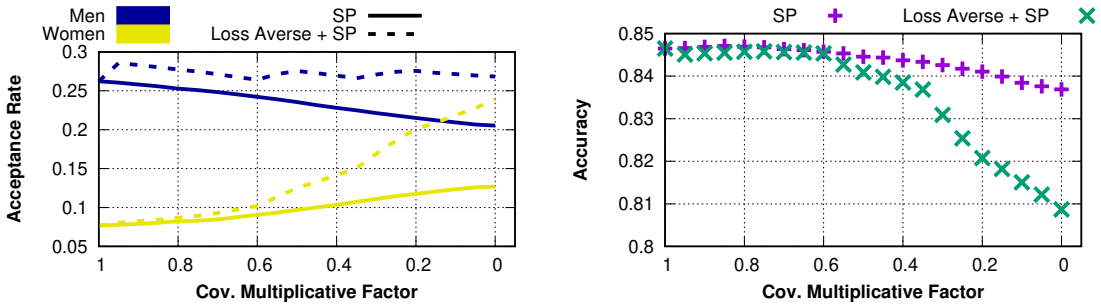
Now, we will show the results of Problem (P4.1), using EOP as a notion of nondiscrimination. We also show results for the loss-averse formulation combined with EOP, given by Problem (P4.4).

EOP. An unconstrained classifier trained on *Synthetic* dataset yields an accuracy of 86% and true positive rates (TPRs) of 94% and 77% for non-protected and protected groups, respectively. To equalize the TPRs we solve Problem (P4.1) using proxies for EOP given in Eq. (4.6).

These results are show in Figure (4.2a) in *solid lines* and Figure (4.2b) in *purple* colored points. i) In order to reduce discrimination, this formulation yields a classifier which lowers the TPR of the non-protected class to 72% and raises the TPR of the protected group to 79%, for covariance threshold $c = 0$, while achieving an accuracy of 74%. ii) Figure (4.2a) shows the limitation of equality of opportunity proxy proposed by Zafar et al. [274], as it achieves a lower discrimination for higher value of the covariance.

Loss-Aversiveness + EOP. To avoid lowering the benefits for any group while reducing discrimination, we solve the Problem (P4.4). We encountered some issues in convergence for some values of covariance factor, specifically smaller ones. Out of 7 random seeds that we tried we find the results for *all* covariance factors for only 5 seeds, we report the average of these results. One reason for the lack of convergence could be that a very high base TPR might make it difficult to find a nondiscriminatory classifier.For covariance threshold $c = 0$, this formulation leads to a classifier whose true positive rates are 95% and 99% for non-protected and protected groups, respectively, with an accuracy of 64%.

We show these results in Figure (4.2a) in *dotted lines* and Figure (4.2b) in *green* colored crosses. i) These figures illustrate the effectiveness of the loss-averse formulation, as the



(a) Statistical Parity (SP)

(b) Nondiscrimination-accuracy tradeoff

Figure 4.3: [Adult dataset. Enforcing statistical parity] Left panel shows the beneficial outcome rates, i.e., , positive class acceptance rates, for a classifier only enforcing SP constraint, i.e., , solution of Problem (P4.1) using SP proxies (solid lines), and a classifier additionally enforcing the "loss-averse" constraint, i.e., , solution of Problem (P4.3) (dotted lines). Right panel shows the nondiscrimination-accuracy tradeoff for both the classifiers. Enforcing "loss-averse" constraint, defined in Eq. (4.7), leads to a significant additional loss in accuracy for the same level of discrimination.

resulting classifiers achieve nondiscrimination by increasing TPR for both groups, ii) however this results in a significant drop in the accuracy.

Summary. In sections 4.4 and 4.5, we demonstrated the effectiveness of our proposed formulation on synthetic datasets. We illustrated the effectiveness of loss-aversively making the status quo classifiers nondiscriminatory, albeit at the cost of accuracy.

4.6 Evaluation on Real-World Dataset: SP

In this section, we evaluate the effectiveness of our proposed schemes in updating the status quo classifier, $\boldsymbol{\theta}_{sqo}$, compliant with the "loss-aversively fair updates" consideration, on real-world dataset using statistical parity as a notion of nondiscrimination.

4.6.1 Dataset and Experimental Setup

In this section, we explain the real-world dataset used to evaluate our proposed considerations.

Adult Dataset. We show result for loss-aversively fair update mechanism, introduced in section 4.3, using *Adult dataset* [4]. Specifically, we illustrate the effectiveness of Problem (P4.3) to train loss-aversively fair classifiers, using Adult dataset. For experiments in this section, we consider statistical parity as a notion of nondiscrimination.

The *Adult Dataset* consists of $45,222$ subjects and 14 features like gender, race, educational level, etc. The classification task is to predict whether a person earns more than

50K USD per annum (positive class) or not (negative class). We consider gender to be a sensitive feature for this dataset.

Experimental Setup. For the experiments conducted on the *Adult dataset* we use the same data split as used for *Synthetic dataset*. We also randomize the data, as well as validate the hyperparameters in a similar manner.

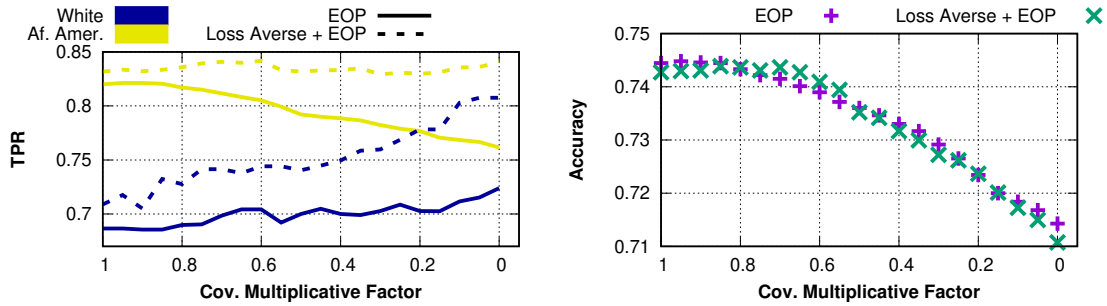
4.6.2 Loss-Aversively Fair Updates

In this section we compare the results of Problem (P4.1), using SP proxies, and Loss-aversively fair updates given by Problem (P4.3) using *Adult dataset*.

SP. On the Adult dataset, logistic regression classifier leads to an accuracy of 84.6%. However, the classifier leads to the beneficial outcome rates of 8% and 26% for women and men respectively, showing a clear disparity in the beneficial outcome rates for the two groups. Next, using the method of Zafar et al. [275], we train a nondiscriminatory classifier while reducing the value of the covariance threshold c, (Eq. (4.5)), towards 0. The results are shown in *solid lines* in Figure (4.3a) and in *purple* colored points in Figure (4.3b). The least discriminatory classifier in this case achieves the beneficial outcome rates of 13% and 20% for women and men respectively, with an accuracy of 83.7%. We notice that the discrimination is reduced by lowering the beneficial outcome rates for men, which leads to a violation of "loss-averse" consideration.

Loss-Aversiveness + SP. We next train classifier with the loss-averse constraints (Eq. (4.7)) combined with SP, i.e., , solve Problem. (P4.3). The least discriminatory classifier in this case achieves the beneficial outcome rates of 24% and 27% for women and men, respectively, while achieving an accuracy of 80.8%. However, the reduction in discrimination is achieved by only increasing the beneficial outcome rate for both groups. Results are shown in Figures (4.3a and 4.3b), in *dotted lines* and *green* colored points, respectively.

The figure shows the beneficial outcome rates for (i) a classifier with statistical parity constraint and (ii) a classifier with loss-averse and statistical parity constraints. The figure shows that at successively decreasing values of the covariance threshold c, while classifier (i) achieves lower discrimination by increasing benefits for one group and decreasing them for the other, classifier (ii) does so by *only increasing benefits for both the groups*. Figure (4.3b) shows the nondiscrimination-accuracy tradeoff achieved by both the classifiers. The figure demonstrates that, as expected, classifier (ii) incurs a much higher cost in terms of accuracy for the same level of discrimination due to the additional loss-averse constraint.



(a) Equality of Opportunity (EOP) **(b)** Nondiscrimination-accuracy tradeoff

Figure 4.4: [SQF dataset. Enforcing equality of opportunity] These figures show similar results as Figure (4.2) using SQF dataset.

4.7 Evaluation on Real-World Dataset: EOP

In this section we will present the "loss-averse" fairness results combined with equality of opportunity, using a real-world dataset.

4.7.1 Dataset and Experimental Setup

In this section we explain the dataset and the experimental setup. We show result of Problem (P4.1), with EOP as a notion of nondiscrimination, as well as Problem (P4.4), which combines EOP and loss-averse constraints.

SQF Dataset. For experiments in this section we consider *NYPD SQF dataset* [1]. The NYPD *SQF dataset* consists of pedestrians who were stopped in the year 2012 on the suspicion of having a weapon. The task is a binary prediction task which indicates whether (negative class) or not (positive class) a weapon was discovered. For our analysis, we consider the race to be the sensitive feature with values African-American and white. After balancing the classes and considering same features as Goel et al. [101], with the exception that we exclude the highly sparse features 'precinct' and 'timestamp of the stop', we obtain 5,832 subjects and 19 features.

Experimental Setup. We used similar experimental setup as explained in section 4.4.

4.7.2 Loss-Aversively Fair Updates

In this section we show the results of Problem (P4.4), which enforces EOP and loss-averse constraints and compare them with the results of Problem (P4.1), which only enforces EOP using the proxy given by Eq. (4.6), on *NYPD SQF dataset*.

EOP. With equality of opportunity constraint, where beneficial outcome rates are defined in terms of true positive rate, we experiment with NYPD SQF dataset. Unconstrained

logistic regression on SQF yields an accuracy of 74.4%, while the beneficial outcome rates are 69% and 82% for Whites and African-Americans, respectively. Least discriminatory classifier, trained with $c = 0$, given in constraint Eq. (4.6), yields benefits of 72% and 76% for Whites and African-Americans, respectively, while achieving an accuracy of 71.4%. Similar to the previous cases, this classifier also achieves lower discriminations by raising the benefits for one group while increasing them for the other group.

Loss-Aversiveness + EOP. Next, we combine the nondiscrimination constraint with the loss-averse constraint, given by Problem (P4.4), in order to update $\boldsymbol{\theta}_{sqo}$. A least discriminatory loss-averse classifier trained on NYPD SQF dataset yields an accuracy of 71% and benefits of 84% and 81% for African-Americans and White, respectively. Figure (4.4a) shows the beneficial outcome rates for (i) a classifier with only nondiscrimination constraints and (ii) a loss-averse classifier with nondiscrimination constraints. Again, we notice that classifier (ii) removes discrimination by *only increasing the beneficial outcome rates* whereas classifier (i) does so by increasing benefits for one group and decreasing them for the other. Finally, the comparison of nondiscrimination-accuracy tradeoff in Figure (4.4b) shows no significant difference between both the classifiers.

Summary. Our proposed methodology, in Section 4.3, successfully enforces the loss averse constraint while updating the status quo classifier, $\boldsymbol{\theta}_{sqo}$, to a nondiscriminatory classifier. However, enforcing these constraints could be at a significant additional cost in terms of accuracy, as demonstrated in Sections 4.6 and 4.7 using real-world datasets.

4.8 Conclusion

A number of recent works have explored various aspects of fairness related to algorithmic decision making. In this chapter, we focused on an aspect of decision making that crucially affects people's fairness perceptions, yet has been overlooked: it is the *fairness of updating decision making*, i.e., how the decision outcomes change when updating a decision making system.

Based on observations in behavioral economics and psychology, we note that any "disadvantageous" changes in outcomes to individual subjects or groups of subjects would be perceived as unfair. Accordingly, we proposed a complementary notion of update fairness that we call *loss-averse updates*. Loss-averse updates try to constrain updates to only yield more advantageous (more beneficial) outcomes compared to status quo.

In this work, we formalized this notion in the context of classification tasks. We proposed measures that would allow these notions to be incorporated in the training of any convex decision-boundary based classifiers (like logistic regression or linear/non-

linear SVM) as convex constraints. We also showed how this notion can be combined with prior notions and measures of non-discrimination in classification. Our evaluation using synthetic and real-world datasets demonstrated the benefits of loss-averse updates in practice.

CHAPTER 5

Designing a new fair ADMS: Time-Critical Influence Maximization

As described in Section 1.1.3, time-critical influence maximization (TCIM) has several impactful applications that affect humans. While existing algorithmic techniques usually aim at maximizing the total number of people influenced, the population often comprises several socially salient groups, e.g., based on gender or race. As a result, these techniques could lead to disparity across different groups in receiving important information. Furthermore, in many applications, the spread of influence is time-critical, i.e., it is only beneficial to be influenced before a deadline. In this chapter, we try to address fairness concerns in the TCIM problem. Specifically, we answer the following research question:

How can we design computationally efficient mechanisms to mitigate unfairness in TCIM *problem?*

As outlined in Section 1.2.3, we acknowledge that our goal is challenging. At first, we have to define what constitutes fairness in TCIM. Then, we have to formalize this problem and propose mechanisms to efficiently solve it.

We answer our research question as follows:

- In Section 5.1, we provide the necessary background on TCIM. Specifically, we discuss the influence propagation mechanism, the influence function and its properties, time-criticality model and two popular constraints for the TCIM problem: TCIM-BUDGET and TCIM-COVER. Furthermore, we discuss the solutions proposed in the existing literature for both of these problems.
- In Section 5.2, we operationalize a notion of *group fairness* for TCIM and discuss a measure of unfairness.
- In Section 5.3, we discuss how to incorporate the proposed measure of fairness in TCIM-BUDGET and TCIM-COVER problems. Solving these problems with the

fairness constraints directly is challenging. To address this issue, we present proxy measure that can solved efficiently. Lastly, we also present theoretical guarantees for the performance of our proposed mechanisms.

- In Section 5.4, using several synthetic datasets, we explore the effect of different graph and algorithmic properties on unfairness. We also test our proposed mechanisms and demonstrate their efficacy in mitigating unfairness in TCIM-BUDGET and TCIM-COVER problems.
- In Sectin 5.5, using three real-world datasets, we demonstrate the efficacy of our methods to mitigate unfairness in TCIM-BUDGET and TCIM-COVER problems. We also explore the effect of different algorithmic properties on unfairness.
- In Section 5.6, we review the related work on influence maximization and other contemporary works.
- In Section 5.7, we present the conclusion of this chapter.

Relevant publication

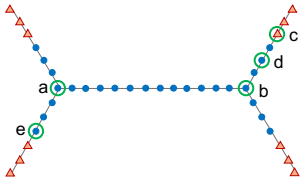
The results presented in this chapter have been published in [9].

5.1 Background on Time-Critical Influence Maximization (TCIM)

In this section, we provide the necessary background on the problem of time-critical influence maximization (henceforth, referred to as TCIM for brevity). First, we formally introduce a well-studied influence propagation model and specify the notion of time-critical influence that we consider in this paper. Then, we discuss two discrete optimization formulations to tackle the TCIM problem.

5.1.1 Influence Propagation in Social Network

Consider a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of directed edges connecting these nodes. For instance, in a social network the nodes could represent people and edges could represent friendship links between people. An undirected link between two nodes can be represented by simply considering two directed edges between these nodes.



	solution to TCIM-BUDGET P5.1				solution to FAIRTCIM-BUDGET P5.4			
	S	$\frac{f(S;\mathcal{V},\mathcal{G})}{\lvert\mathcal{V}\rvert}$	$\frac{f(S;\mathcal{V}_1,\mathcal{G})}{\lvert\mathcal{V}_1\rvert}$	$\frac{f(S;\mathcal{V}_2,\mathcal{G})}{\lvert\mathcal{V}_2\rvert}$	S	$\frac{f(S;\mathcal{V},\mathcal{G})}{\lvert\mathcal{V}\rvert}$	$\frac{f(S;\mathcal{V}_1,\mathcal{G})}{\lvert\mathcal{V}_1\rvert}$	$\frac{f(S;\mathcal{V}_2,\mathcal{G})}{\lvert\mathcal{V}_2\rvert}$
$\tau=\infty$	$\{a,b\}$	0.38	0.48	0.16	$\{a,c\}$	0.31	0.33	0.27
$\tau=4$	$\{a,b\}$	0.32	0.44	0.08	$\{d,e\}$	0.25	0.26	0.22
$\tau=2$	$\{a,b\}$	0.24	0.36	0.00	$\{a,c\}$	0.21	0.22	0.18

Figure 5.1: An example to illustrate the disparity across groups in the standard approaches to TCIM. (Left) Graph with $|\mathcal{V}| = 38$ nodes belonging to two groups shown in "blue dots" ($|\mathcal{V}_1| = 26$) and "red triangles" ($|\mathcal{V}_2| = 12$). (Right) We compare an optimal solution to the standard TCIM-BUDGET problem P5.1 and an optimal solution to our formulation of TCIM-BUDGET with fairness considerations given by FAIRTCIM-BUDGET problem P5.4. For different time critical deadlines τ, normalized utilities are reported for the whole population $\mathcal{V}$, for the "blue dots" group $\mathcal{V}_1$, and for the "red triangles" group $\mathcal{V}_2$. As τ reduces, the disparity between groups is further exacerbated in the solution to TCIM-BUDGET problem P5.1. Solution to FAIRTCIM-BUDGET problem P5.4 achieves high utility and low disparity for different deadlines τ.

There are two classical influence propagation models that are studied in the literature [147]: (i) Independent Cascade model (IC) and (ii) Linear Threshold (LT) model. In this paper, we will consider IC model and our results can easily be extended to the LT model.

In the IC model, there is a probability of influence associated with each edge denoted as $p_{\mathcal{E}} := \{p_e \in [0,1] : e \in \mathcal{E}\}$. Given an initial seed set $S \subseteq \mathcal{V}$, the influence propagation proceeds in discrete time steps $t = \{0, 1, 2, \ldots, \}$ as follows. At $t = 0$, the initial seed set S is "activated" (i.e., influenced). Then, at any time step $t > 0$, a node $v \in \mathcal{V}$ which was activated at time $t-1$ gets a chance to influence its neighbors (i.e., set of nodes $\{w : (v,w) \in \mathcal{E}\}$). The influence propagation process stops at time $t > 0$ if no new nodes get influenced at this time. Under the IC model, once a node is activated it stays active throughout the process and each node has only one chance to influence its neighbors.

Note that the influence propagation under IC model is a stochastic process: the stochasticity here arises because of the random outcomes of a node v influencing its neighbor w based on the Bernoulli distribution $p_{(v,w)}$. An outcome of the influence propagation process can be denoted via a set of timestamps $\{t_v \geq 0 : v \in \mathcal{V}\}$ where t_v represents the time at which a node $v \in \mathcal{V}$ was activated. We have $t_v = 0$ iff $v \in S$ and for convenience of notation, we define $t_v = -1$ to indicate that the node v was not activated in the process.

5.1.2 Utility of Time-Critical Influence

As mentioned earlier, we focus on the application settings where the spread of influence is time-critical, i.e., it is more beneficial to be influenced earlier in the process. In particular, we adopt the well-studied notion of time-critical influence as proposed by [54]. Their

time-critical model is captured via a deadline τ: If a node is activated before the deadline, it receives a utility of 1, otherwise it receives no utility. This simple model captures the notion of timing in many important real-world applications such as viral marketing of an online product with limited availability, information propagation of job vacancy information, etc.

Given the influence propagation model and the notion of time-critical aspect via a deadline τ, we quantify the utility of time-critical influence for a given seed set S on a set of target nodes $Y \subseteq \mathcal{V}$ via the following:

$$f_\tau(S; Y, \mathcal{G}) = \mathbb{E}\Big[\sum_{v \in Y, t_v \geq 0} \mathbb{I}(t_v \leq \tau)\Big], \tag{5.1}$$

where the expectation is w.r.t. the randomness of the outcomes of the IC model. The function is parametrized by deadline τ, set $Y \subseteq \mathcal{V}$ representing the set of nodes over which the utility is measured (by default, one can consider $Y = \mathcal{V}$), and the underlying graph $\mathcal{G}$ along with edge activation probabilities $p_\mathcal{E}$. Given a fixed value of these parameters, the utility function $f_\tau : 2^\mathcal{V} \to \mathbb{R}_{\geq 0}$ is a set function defined over the seed set $S \subseteq \mathcal{V}$. Note that the constraint $t_v \geq 0$ represents the node was activated and the constraint $t_v \leq \tau$ represents that the activation happened before the deadline τ.

5.1.3 TCIM as Discrete Optimization Problem

Next, we present two settings under which we study TCIM by casting it as a discrete optimization problem.

5.1.3.1 Maximization under Budget Constraint (TCIM-BUDGET)

In the maximization problem under budget constraint, we are given a fixed budget $B > 0$ and the goal is to find an optimal set of seed nodes that maximize the expected utility. Formally, we state the problem as

$$\max_{S \subseteq \mathcal{V}} f_\tau(S; \mathcal{V}, \mathcal{G}) \quad \text{subject to } |S| \leq B. \tag{P5.1}$$

5.1.3.2 Minimization under Coverage Constraint (TCIM-COVER)

In the minimization problem under coverage constraint, we are given a quota $Q \in [0, 1]$ representing the minimal fraction of nodes that must be activated or “covered" by the influence propagation in expectation. The goal is then to find an optimal set of seeds of minimal size that achieves the desired coverage constraint. We formally state the

problem as

$$\min_{S \subseteq \mathcal{V}} |S| \quad \text{subject to } \frac{f_\tau(S; \mathcal{V}, \mathcal{G})}{|\mathcal{V}|} \geq Q. \tag{P5.2}$$

5.1.4 Submodularity and Approximate Solutions

Next, we present some key properties of the utility function $f_\tau(.)$ to get a better understanding of the above-mentioned optimization problems. In their seminal work, [147] showed that the utility function without time-critical deadline, i.e., $f_\infty(.) : S \to \mathbb{R}_+$, is a non-negative, monotone, submodular set function w.r.t. the optimization variable $S \subseteq \mathcal{V}$. Submodularity is an intuitive notion of diminishing returns and optimization of submodular set functions finds numerous applications in machine learning and social networks, such as influence maximization [147], sensing [159], information gathering [234], and active learning [118] (see [158] for a survey on submodular function optimization and its applications).

Chen et. al [54] showed that the utility function for the general time-critical setting for any τ also satisfies these properties. Submodularity is an intuitive notion of diminishing returns, stating that, for any sets $A \subseteq A' \subseteq \mathcal{V}$, and any node $a \in \mathcal{V} \setminus A'$, it holds that (omitting the parameters $\mathcal{V}$ and $\mathcal{G}$ for brevity):

$$f_\tau(A \cup \{a\}) - f_\tau(A) \geq f_\tau(A' \cup \{a\}) - f_\tau(A').$$

Existing works [88, 158, 203] have shown that P5.1 and P5.2 are NP-Hard and hence finding the optimization solution is intractable. However, on a positive note, one can exploit the submodularity property of the function to design efficient approximation algorithms with provable guarantees [158, 203]. In particular, we can run the following greedy heuristic: start from an empty set, iteratively add a new node to the set that provides the maximal marginal gain in terms of utility, and stop the algorithm when the desired constraint on budget or coverage is met. This greedy algorithm provides the following guarantees for these two problems:

- for the TCIM-BUDGET problem P5.1, the greedy algorithm returns a set $\hat{S}$ that guarantees the following lower bound on the utility: $f_\tau(\hat{S}; \mathcal{V}, \mathcal{G}) \geq (1 - \frac{1}{e}) \cdot f_\tau(S^*; \mathcal{V}, \mathcal{G})$ where S^* is an optimal solution to problem P5.1.
- for the TCIM-COVER problem P5.2, the greedy algorithm returns a set $\hat{S}$ that guarantees the following upper bound on the seed set size: $|\hat{S}| \leq \ln(1 + |\mathcal{V}|) \cdot |S^*|$ where S^* is an optimal solution to problem P5.2.

5.2 Measuring Unfairness in TCIM

In this section, we highlight the disparity in utility across population resulting from the solution to the standard TCIM problem formulations, and introduce a measure of unfairness in TCIM.

5.2.1 Socially Salient Groups and Their Utilities

The current approaches to TCIM consider all the nodes in $\mathcal{V}$ to be homogeneous. We capture the presence of different socially salient groups in the population by dividing individuals into k disjoint groups. Here, socially salient groups could be based on some sensitive attribute such as gender or race. We denote the set of nodes in each group $i \in \{1, 2, \ldots, k\}$ as $\mathcal{V}_i \subseteq \mathcal{V}$, and we have $\mathcal{V} = \cup_i \mathcal{V}_i$. For any given seed set S, we define the utilities for a group i as $f_\tau(S; \mathcal{V}_i, \mathcal{G})$ by setting target nodes $Y = \mathcal{V}_i$ in Eq. 5.1.

5.2.2 Disparity in Utility Across Groups

In the standard formulations for TCIM problem, i.e., TCIM-BUDGET problem P5.1 and TCIM-COVER problem P5.2, the utility $f_\tau(S; \mathcal{V}, \mathcal{G})$ is optimized for the whole population $\mathcal{V}$ without considering their groups. Clearly, a solution to TCIM problem can, in general, lead to high disparity in utilities of different groups.

In particular, this disparity in utility across groups arises from several factors in which two groups differ from each other. One of the factors is that the groups are of different sizes, i.e., one group is a minority. The different group sizes could, in turn, lead to selecting seed nodes from the majority group when optimizing for utility $f_\tau(S; \mathcal{V}, \mathcal{G})$ in problems P5.1 and P5.2. Another factor is related to the connectivity and centrality of nodes from different groups. The solution to the optimization problems P5.1 and P5.2 tend to favor nodes which are more central and have high-connectivity. Finally, given the above two factors, we note that the disparity in influence across groups can be further exacerbated for lower values of deadline τ in the time-critical influence maximization.

In Figure 5.1, we provide an example to illustrate the disparity across groups in the standard approaches to TCIM. In particular, to show this disparity, we consider the TCIM-BUDGET problem P5.1, and it is easy to extend this example to show disparity in TCIM-COVER problem P5.2. The graph that we consider in this example (see Figure 5.1 caption for details) has the two characteristic properties that we discussed above: (i) group V_2 is in minority with less than half of the size of group V_1, (ii) group V_1 has more central nodes compared to group V_2, and (iii) nodes in group V_1 have higher connectivity

than nodes in group V_2. We consider the probability of influence in the graph to be $p_e = 0.7$ for all edges, and study the optimization problem P5.1 for budget $B = 2$.

For different time critical deadlines τ, we report the following normalized utilities: $\frac{f(S;\mathcal{V},\mathcal{G})}{|\mathcal{V}|}$ for the whole population $\mathcal{V}$, $\frac{f(S;\mathcal{V}_1,\mathcal{G})}{|\mathcal{V}_1|}$ for the group $\mathcal{V}_1$, and $\frac{f(S;\mathcal{V}_2,\mathcal{G})}{|\mathcal{V}_2|}$ for the group $\mathcal{V}_2$. Here, normalization captures the notion of "average" utility per node in a group, and automatically allows us to account for the differences in the group sizes. As can be seen in Figure 5.1, the optimal solution to the problem consistently picks set $S = \{a, b\}$ comprising of the most central and high-connectivity nodes. While these nodes maximize the total utility, they lead to a high disparity in the normalized utilities across groups. As the influence becomes more time-critical, i.e., τ is reduced, we see an increasing disparity as discussed above. For $\tau = 2$, the utility of group $\mathcal{V}_2$ reduces to 0.

5.2.3 Measure of Unfairness

Next, in order to guide the design of fair solutions to TCIM problems, we introduce a formal notion of group unfairness in TCIM. In particular, we measure the (un-)fairness or disparity of an algorithm by the maximum *disparity in normalized utilities* across all pairs of socially salient groups, given by:

$$\max_{i,j\in\{1,2,\ldots,k\}} \left| \frac{f_\tau(S;\mathcal{V}_i,\mathcal{G})}{|\mathcal{V}_i|} - \frac{f_\tau(S;\mathcal{V}_j,\mathcal{G})}{|\mathcal{V}_j|} \right|. \tag{5.2}$$

As discussed above (see Section 5.2.2), normalization w.r.t. group sizes captures the notion of average utility per node in a group and hence makes the measure agnostic to the group size. In the next section, we seek to design fair algorithms for TCIM problems that have low disparity (or more fairness) as measured by Eq. 5.2.

5.3 Achieving Fairness in TCIM

In this section, we seek to develop efficient algorithms for TCIM problems under fairness considerations that have low disparity measured by Eq. 5.2 while maintaining high performance.

5.3.1 Fair TCIM-Budget

5.3.1.1 Fairness considerations in TCIM-BUDGET

A fair TCIM algorithm under budget constraint should seek to achieve the following two objectives: (i) maximizing total influence for the whole population $\mathcal{V}$ as was done in

the standard TCIM-BUDGET problem P5.1, and (ii) enforcing fairness by ensuring that disparity across different groups as per Eq. 5.2 is low. Clearly, enforcing fairness would lead to a reduction in total influence, and we seek to design algorithms that can achieve a good trade-off between these two objectives. We formulate the following fair variant of TCIM-BUDGET problem P5.1 that captures this trade-off:

$$\max_{S \subseteq \mathcal{V}} \quad \underbrace{\sum_{i}^{k} f_\tau(S; \mathcal{V}_i, \mathcal{G})}_{\text{Maximize number of influenced nodes}} \tag{P5.3}$$

$$\text{subject to} \quad \underbrace{|S| \leq B}_{\text{Bound seed set size}} \quad ,$$

$$\text{and} \quad \underbrace{\max_{i,j} \left| \frac{f_\tau(S; \mathcal{V}_i, \mathcal{G})}{|\mathcal{V}_i|} - \frac{f_\tau(S; \mathcal{V}_j, \mathcal{G})}{|\mathcal{V}_j|} \right| \leq c}_{\text{Minimize disparity}}$$

where $c \in [0, 1]$ is a hyperparameter which indicates the maximum level of allowed disparity among the groups. This problem might not be feasible for all the values of c. So, one would have to tune this hyperparameter for feasibility and the desired level of disparity. Problem P5.3 has two main objectives, i.e., finding B seeds which will i) **maximize the total influence**, which is exactly the same as the traditional influence maximization given in problem P5.1— here written as the sum of influences over all the groups, and, additionally, ii) **minimize the disparity of influence** between different groups up to the prescribed threshold.

We note that problem P5.3 is NP-Hard and a challenging discrete optimization problem and it does not have the structural properties of submodularity as was the case for the standard TCIM-BUDGET problem P5.1.

5.3.1.2 Surrogate FAIRTCIM-BUDGET with guarantees

Instead of directly solving problem P5.3, we introduce a novel surrogate problem that would allow us to indirectly trade-off the two objectives of maximizing total influence and minimizing disparity across groups, as follows:

$$\max_{S \subseteq \mathcal{V}} \sum_{i=1}^{k} \mathcal{H}(f_\tau(S; \mathcal{V}_i, \mathcal{G})) \quad \text{subject to } |S| \leq B, \tag{P5.4}$$

where $\mathcal{H}$ is a non-negative, monotone concave function.

Optimizing problem P5.4 captures both the objectives of the original: i) **maximizing influence**: since the objective is monotonically increasing it encourages picking more

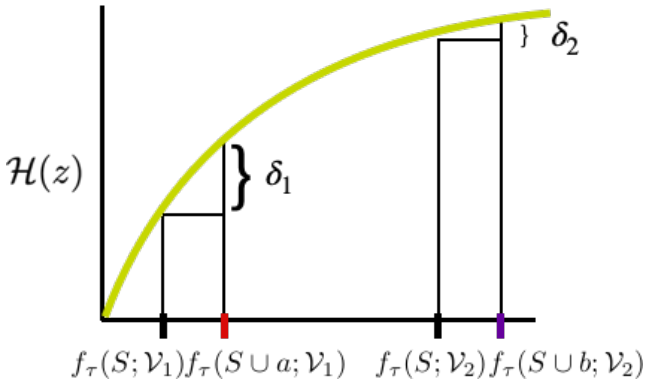


Figure 5.2: Demonstration of concave function encouraging picking seeds which influence under-represented group. X-axis represents group influence and y-axis represents the value of $\mathcal{H}$ for the corresponding group influence. In this example we have two groups, $\mathcal{V}_1$ and $\mathcal{V}_2$. $\mathcal{V}_1$ is under-influenced compared to $\mathcal{V}_2$, using the seed set S. In the next iteration we have an option to either include node a or b in our seed set, both of which add the same amount of total influence. Adding node a in our seed set influences $\mathcal{V}_1$which is the under-influenced group, while adding node b influences nodes from $\mathcal{V}_2$, as demonstrated in the figure. The traditional method, given by problem P5.1, would treat both of these nodes as equally good. However, since we are passing the group influences through a concave function the increase in the value of $\mathcal{H}(z)$ will be more if we pick node a, i.e., our method will pick node a because $\delta_1 > \delta_2$.

influential nodes, ii) **minimizing the disparity of influence**: Passing the group influence functions through a monotone concave function $\mathcal{H}$ rewards selecting seeds that would lead to higher influence on under-represented groups early in the selection process; this in turn helps in reducing disparity across groups under the *assumption* that the under-represented groups not only have lower influence in terms of total number of nodes but also have lower influence in terms of fraction of nodes w.r.t to their groups sizes. In other words, as we are passing the group influences through a *concave* function, the increase in the objective would be higher when under-represented groups are influenced, as demonstrated in figure 5.2.

Trade-off between objectives. It is important to note that controlling the curvature of the concave function $\mathcal{H}$ provides an indirect way to *trade-off* between the two objectives, i.e., i) the total influence and ii) the disparity of the solution. For instance, using $\mathcal{H}(z) := \log(z)$ has higher curvature than using $\mathcal{H}(z) := \sqrt{z}$ and hence leads to lower disparity at the cost of lower total influence (this is demonstrated in the experimental results in Figure 5.4a). For our illustrative example from Section 5.2, we report the results for an optimal solution to FAIRTCIM-BUDGET problem P5.4 with $\mathcal{H}(z) := \log(z)$. As can be seen in Figure 5.1, the solution leads to a drastic reduction in disparity across groups for different values of deadline τ compared to an optimal solution of the standard TCIM-BUDGET problem P5.1 at the cost of reduction in total influence. So, if one wants to

penalize disparity of influence more one can pick $\mathcal{H}$ function with higher curvature but at the expense of potentially lower total influence.

While it is intuitively clear that using the concave function $\mathcal{H}(z)$ in problem P5.4 reduces disparity, we also need to ensure that the solution to this problem has high influence for the whole population $\mathcal{V}$ and that the solution can be computed efficiently. As proven in the theorem below, we can find an approximate solution to problem P5.4, with guarantees on the total influence, by running the greedy heuristic (as was introduced in Section 5.1.4).

Theorem 1. *Let $\hat{S}$ denote the output of the greedy algorithm for problem P5.4. Let S^* be an optimal solution to problem P5.1. Then, the total influence of the greedy algorithm is guaranteed to have the following lower bound: $f_\tau(\hat{S};\mathcal{V},\mathcal{G}) \geq (1-\frac{1}{e}) \cdot \mathcal{H}\big(f_\tau(S^*;\mathcal{V},\mathcal{G})\big)$.*

This is equivalent to the fact that the multiplicative approximation factor of the utility of FAIRTCIM-BUDGET using greedy algorithm w.r.t. the utility of an optimal solution to TCIM-BUDGET scales as $\big((1-\frac{1}{e}) \cdot \frac{\mathcal{H}(f_\tau(S^*;\mathcal{V},\mathcal{G}))}{f_\tau(S^*;\mathcal{V},\mathcal{G})}\big)$. Note that as the curvature of the concave function $\mathcal{H}$ increases, the approximation factor gets worse—this further highlights how the curvature of the function $\mathcal{H}$ provides a way to trade-off the total influence and disparity of the solution. In the case of $\mathcal{H}(z) := log(z)$, which penalizes the disparity of the solution quite severely due to high curvature, the bound on the total influence achieved by our solution is exponentially related to the optimal solution of problem P5.1 which does not consider fairness. On the other hand, if $\mathcal{H}(z) := z$, i.e., $\mathcal{H}$ is an identity function, the problem reverts back to problem P5.1, whose solution might have a higher total influence but could result in high disparity, as evidenced by our experimental results in sections 5.4.2 and 5.5.2. One can pick $\mathcal{H}$ with the appropriate curvature for the desired level of penalization of the disparity of influence at the cost of total influence.

Proof. Since the composition of a non-decreasing concave and a non-decreasing submodular function is submodular [176], the objective function in problem P5.4 is monotone submodular function. Let $\tilde{S}$ be the optimal solution and $\hat{S}$ be the output of the greedy algorithm for the problem P5.4, with a fixed budget B. Let S^* be an optimal solutions for the following problem

$$\max_{S\subseteq\mathcal{V}} f_\tau(S;\mathcal{V},\mathcal{G}) \quad \text{subject to } |S| \leq B, \tag{5.3}$$

with a fixed budget B. Then, following the standard guarantees of submodular optimization [158, 203] (also see Section 3.4), we get the following bounds

$$\mathcal{H}(f_\tau(\hat{S}; V, \mathcal{G})) \geq \left(1 - \frac{1}{e}\right) \cdot \mathcal{H}(f_\tau(\tilde{S}; V, \mathcal{G})) \tag{5.4}$$

Since $\tilde{S}$ is the optimal solution of problem P5.4, given $S : |S| \leq B$ following holds,

$$\mathcal{H}(f_\tau(\tilde{S}; V, \mathcal{G})) \geq \mathcal{H}(f_\tau(S; V, \mathcal{G}))$$

which implies that

$$\mathcal{H}(f_\tau(\tilde{S}; V, \mathcal{G})) \geq \mathcal{H}(f_\tau(S^*; V, \mathcal{G})). \tag{5.5}$$

Combining equations 5.4 and 5.5 we get

$$\mathcal{H}(f_\tau(\hat{S}; V, \mathcal{G})) \geq \left(1 - \frac{1}{e}\right) \cdot \mathcal{H}(f_\tau(S^*; V, \mathcal{G})). \tag{5.6}$$

Since $\mathcal{H}$ is a concave function,

$$f_\tau(\hat{S}; V, \mathcal{G}) \geq \mathcal{H}(f_\tau(\hat{S}; V, \mathcal{G})). \tag{5.7}$$

Combining equations 5.6 and 5.7 we get

$$f_\tau(\hat{S}; V, \mathcal{G}) \geq \left(1 - \frac{1}{e}\right) \cdot \mathcal{H}(f_\tau(S^*; V, \mathcal{G})),$$

which concludes the proof. □

5.3.2 Fair TCIM-Cover

5.3.2.1 Fairness considerations in TCIM-COVER

A fair TCIM algorithm under coverage constraint should seek to achieve the following two objectives: (i) minimizing the size of the seed set that achieves the desired coverage constraint as was done in the standard TCIM-COVER problem P5.2, and (ii) enforcing fairness by ensuring that disparity across different groups as per Eq. 5.2 is low. As was the case for FAIRTCIM-BUDGET problem above, enforcing fairness would lead to increasing the size of the required seed set, and we seek to design algorithms that can achieve a good trade-off between these two objectives. We formulate a fair variant of TCIM-COVER problem P5.2 that captures this trade-off as follows:

$$\min_{S \subseteq \mathcal{V}} \underbrace{|S|}_{\text{Minimize seed set size}} \tag{P5.5}$$
$$\text{subject to } \underbrace{\frac{\sum_i^k f_\tau(S; \mathcal{V}_i, \mathcal{G})}{|\mathcal{V}|} \geq Q}_{\text{Bound fraction of influenced node}} ,$$
$$\text{and } \underbrace{\max_{i,j} \left| \frac{f_\tau(S; \mathcal{V}_i, \mathcal{G})}{|\mathcal{V}_i|} - \frac{f_\tau(S; \mathcal{V}_j, \mathcal{G})}{|\mathcal{V}_j|} \right| \leq c}_{\text{Minimize disparity}}$$

where $c \in [0, 1]$ is a hyperparameter, which determines the amount of disparity that is allowed. As in the case of problem P5.3, it is possible that for some values of c the problem is infeasible. Problem P5.5 has three objectives: i) **minimizing size of seed set** that ii) **influences a prescribed quota** of the population while ii) **minimizing disparity in the influence** among the groups.

As in Section 5.3.1, we note that problem P5.5 is a challenging discrete optimization problem and does not have structural properties as was the case for the standard TCIM-COVER problem P5.2.

5.3.2.2 Surrogate FAIRTCIM-COVER with guarantees

Instead of directly solving problem P5.5, we introduce a novel surrogate problem that indirectly trade-offs the two objectives of minimizing the size of selected seed set and minimizing disparity, as follows:

$$\min_{S \subseteq \mathcal{V}} |S| \quad \text{subject to } \frac{f_\tau(S; \mathcal{V}_i, \mathcal{G})}{|\mathcal{V}_i|} \geq Q \quad \forall i. \tag{P5.6}$$

Optimizing problem P5.6 addresses all the objectives of problem P5.5 by i) **minimizing the seed set size**, ii) which **influences all the groups up to the prescribed quota**, Q. iii) Thereby, **disparity** of the feasible solution is **bounded** by $(1 - Q)$. The key idea of using the surrogate objective function in problem P5.6 is the following: the problem has a constraint that enforces that at least Q fraction of nodes in each group are influenced by the selected seed set S; this in turn directly provides a bound on the disparity of any feasible solution to the problem as $(1 - Q)$. Figure 5.3 provides a demonstration of the constraints we propose.

While it is intuitively clear that the solution to problem P5.6 reduces disparity, we also would like to bound the size of the final seed set and that the solution can be

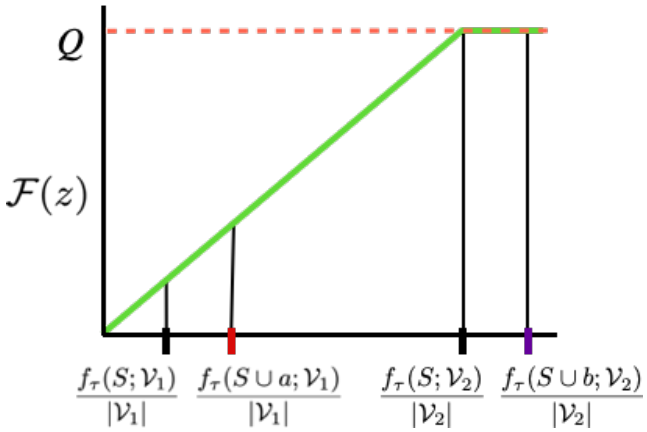


Figure 5.3: where $\mathcal{F}(z) = \min\left\{\frac{f_\tau(S;\mathcal{V}_i,\mathcal{G})}{|\mathcal{V}_i|}, Q\right\}$. Demonstration of the constraint in problem P5.6. X-axis represents the fraction of group influences and y-axis represents the value of per group constraint in problem P5.6 for the corresponding group influence. In this example we have two groups, $\mathcal{V}_1$ and $\mathcal{V}_2$ of roughly same size. $\mathcal{V}_1$ has not reached the prescribed quota, Q, while $\mathcal{V}_2$ has already been influenced up to the prescribed quota. In the next iteration we have an option to either include node a or node b in our seed set, both of which add the same amount of total influence. Adding node a in our seed set influences only $\mathcal{V}_1$, while adding node b influences nodes from only $\mathcal{V}_2$, as demonstrated in the figure. The traditional method, problem P5.2, would treat both of these nodes as equally good candidates for including in the seed set because they add equal fraction of total influence. However, since we require all the groups to be influenced up to the required quota, selecting node a will increase our constraint value, $\mathcal{F}(z)$, while by selecting node b the constraint value would stay the same as $\mathcal{V}_2$ has already reached the required quota of influence.

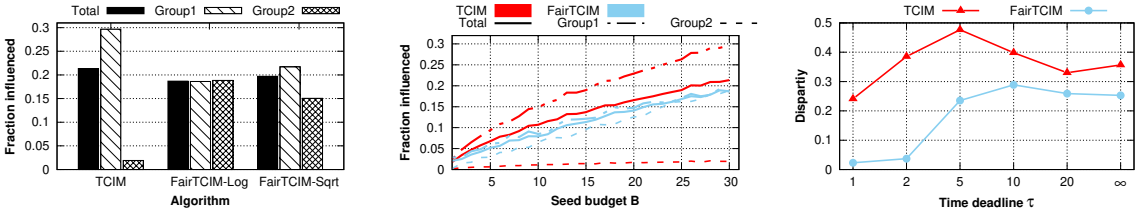
computed efficiently. As proven in the theorem below, we can find an approximate solution to problem P5.6, with guarantees on the final seed set size, by running the greedy heuristic (as was introduced in Section 5.1.4).

Theorem 2. *Let us denote the output of the greedy algorithm for problem P5.6 by set $\hat{S}$. For group $i \in \{1,\dots,k\}$, let S_i^* denote an optimal solution to the coverage problem P5.2 for the target nodes set to $\mathcal{V}_i$, i.e., solving problem P5.2 with constraint given by $\frac{f_\tau(S;\mathcal{V}_i,\mathcal{G})}{|\mathcal{V}_i|} \geq Q$. Then, the size of the seed set $\hat{S}$ returned by the greedy algorithm is guaranteed to have the following upper bound: $|\hat{S}| \leq \ln(1+|\mathcal{V}|)\left(\sum_{i=1}^{k}|S_i^*|\right)$.*

Proof. The constraint in the problem P5.6 could be rewritten as follows,

$$\sum_{i=1}^{k} \min\left\{\frac{f_\tau(S;\mathcal{V}_i,\mathcal{G})}{|\mathcal{V}_i|}, Q\right\} \geq k \cdot Q.$$

The objective function in the constraint is monotone submodular function because monotone submodular functions remain monotone submodular under truncation: if



(a) Total and group influence **(b)** Varying budget B **(c)** Varying time deadline τ

Figure 5.4: [Synthetic Dataset: Budget Problem] The figures show that solving TCIM-BUDGET problem P5.1 can lead to disparity in number of influenced nodes belonging to different groups, while FAIRTCIM-BUDGET problem P5.4 fares better in terms of achieving parity of influence, with marginally lower total influence. See Section 5.4.2 for further details.

$g(S)$ is monotone submodular so is $f(S) := \min(g(S), c)$ for any constant $c >= 0$, and monotone submodular functions are closed under addition [158]. Let $\tilde{S}$ be the optimal solution and $\hat{S}$ be the greedy solution of problem P5.6 for a fixed quota Q. Let S_i^* be the optimal solution of TCIM-COVER problem P5.2, with target nodes set to $\mathcal{V}_i$ and quota set to Q. Then, following the standard guarantees of the submodular optimization [158] (also see Section 3.4) we have the following bound:

$$|\hat{S}| \leq \ln(1 + |\mathcal{V}|)|\tilde{S}|. \tag{5.8}$$

Since $\tilde{S}$ is the optimal solution of problem P5.6, where all the groups reach the prescribed quota Q, $\tilde{S}$ must be at-least as small as any other other set which also reaches all the groups up to the quota Q. Hence,

$$|\tilde{S}| \leq \sum_{i=1}^{k} |S_i^*|. \tag{5.9}$$

Combining equations 5.8 and 5.9 we get

$$|\hat{S}| \leq \ln(1 + |\mathcal{V}|)\Big(\sum_{i=1}^{k} |S_i^*|\Big),$$

which concludes the proof. □

5.4 Evaluation on Synthetic Datasets

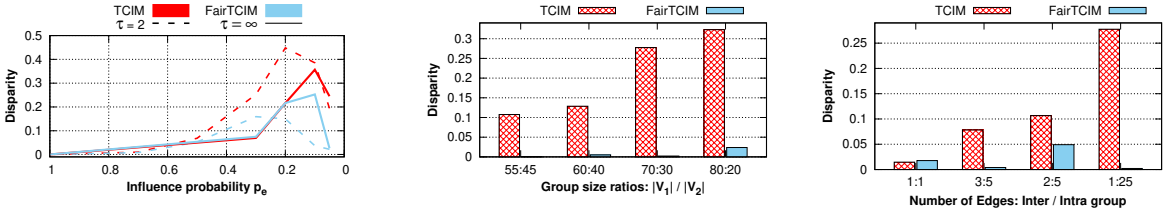
In this section, we compare the solutions of different problems on several synthetic datasets. We show that the disparity in influence is affected by varying different properties of the graphs and parameters of the algorithms.

5.4.1 Dataset and Experimental Setup

First we discuss how we generated the synthetic datasets and then the setup used in our experiments.

Synthetic datasets. We consider stochastic block model to generate the synthetic datasets, particularly we consider an undirected graph with 500 nodes, where each node belongs to either group $\mathcal{V}_1$ or group $\mathcal{V}_2$. The fraction of nodes belonging to each group is determined by a parameter g (e.g., setting $g = 0.7$ results in 70% of the nodes to be randomly assigned to group $\mathcal{V}_1$). Nodes are connected based on two probabilities: (i) within-group edge probability (*Homophily*) p_{hom} and (ii) across-group edge probability (*Heterophily*) p_{het}. Placing an edge between two nodes goes as follows: given a pair of nodes (v, w), if they belong to the same group, we perform a Bernoulli trial with parameter p_{hom}; otherwise we use the parameter p_{het}. If the outcome of the trial is 1, we place an undirected edge e between these two nodes. Each edge has a probability of activation, $p_e \in [0, 1]$, with which the nodes can activate each other.

Experimental Setup. In our experiments we varied all the aforementioned properties of the graph. We vary each of these graph and algorithmic properties while rest of the properties are set to a default value. We experimented with several default values but as an illustration we include the results for the following default values: $g = 0.7$ yielding 350 nodes in $\mathcal{V}_1$ and 150 nodes in $\mathcal{V}_2$. We set $p_{hom} = 0.025$ and $p_{het} = 0.001$, which yielded 3606 total edges, out of which 2965 edges were within group $\mathcal{V}_1$, 514 within $\mathcal{V}_2$, and 127 edges connecting nodes across two groups. We used a constant activation probability on all edges given by $p_e = 0.05$. Finally, we consider the time deadline $\tau = 20$, unless explicitly stated otherwise. Evaluating utilities, as described in Eq. 5.1, in closed form is intractable, so we used Monte Carlo sampling to estimate these utilities. We used 200 samples for this estimation, which yielded a stable estimation of the utility function. In all the experiments, we pick a seed set by solving the corresponding problem. Then, we use this seed set to estimate the expected number of nodes influenced in the graph using TCIM. We report the following normalized utilities: $\frac{f(S;\mathcal{V},\mathcal{G})}{|\mathcal{V}|}$ for the whole population $\mathcal{V}$, $\frac{f(S;\mathcal{V}_1,\mathcal{G})}{|\mathcal{V}_1|}$ for the group $\mathcal{V}_1$, and $\frac{f(S;\mathcal{V}_2,\mathcal{G})}{|\mathcal{V}_2|}$ for the group $\mathcal{V}_2$.



(a) Varying influence probability p_e

(b) Varying group size proportions

(c) Varying cliquishness

Figure 5.5: [Synthetic Dataset: Budget Problem] These figures demonstrate that lower activation probabilities, uneven group sizes, and cliquishness can lead to higher disparity of influence between different groups with TCIM-BUDGET problem P5.1. In comparison our proposed method, FAIRTCIM-BUDGET given by problem P5.4, leads to solutions which yield lower disparity. For further details, see Section 5.4.2.

5.4.2 TCIM under Budget Constraints

Next, we compare the solutions of TCIM-BUDGET problem P5.1 with our solution to FAIRTCIM-BUDGET problem P5.4, obtained through the greedy algorithm, i.e., , by iteratively picking B seeds which yield maximum marginal gain. In all the figures discussed in this section, red color represents the results of TCIM-BUDGET problem P5.1, and blue color represents the results of our solution to the FAIRTCIM-BUDGET problem P5.4. For the experiments in this section, we used a budget of $B = 30$ seeds.

5.4.2.1 Varying algorithmic properties

In this section, we vary several properties of the influence maximization algorithm and answer following questions:

— **Q1**: How does the choice of $\mathcal{H}(z)$ with different curvatures affect disparity and total influence?

— **Q2**: How does varying seed budget affect disparity?

— **Q3**: How does varying time deadline affect disparity?

— **Q4**: How does varying activation probabilities on the edges affect disparity?

— **Q5**: How effective is our method in reducing disparity?

— **Q6**: How much cost does our method incur?

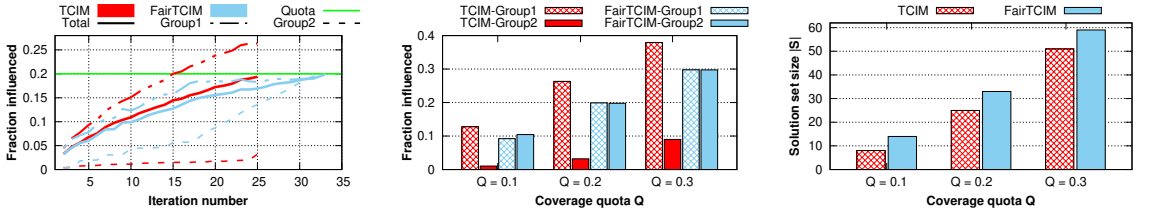
[Q1, Q5, Q6] Effect of different $\mathcal{H}(z)$. Figure 5.4a presents the comparison of three algorithms: one solving TCIM-BUDGET problem P5.1, using the greedy heuristic; the other two solving FAIRTCIM-BUDGET problem P5.4, using two realizations of the concave monotone function, $\mathcal{H}(z)$, given by: (i) $\mathcal{H}(z) := \log(z)$ and (ii) $\mathcal{H}(z) := \sqrt{z}$. Figure 5.4a shows the fraction of population influenced, both overall and for every

group. We can observe that solving the traditional TCIM-BUDGET problem leads to large disparity between the fraction of nodes influenced from each group: while 30% of nodes in group $\mathcal{V}_1$ are influenced, this fraction is only 2% for group $\mathcal{V}_2$.

On the other hand, our proposed solution to FAIRTCIM-BUDGET problem results in lower disparity between the groups, ensuring similar fraction of influenced nodes. We can further see that $\sqrt{z}$, with lower curvature, performs worse than $\log(z)$ in removing the disparity, however incurring lower loss in total influence, as guaranteed by our theoretical results in Theorem 1. One could consider higher powers of the root to increase the curvature or multiplying the input of $\mathcal{H}$ with a scalar ≥ 1 in problem P5.4 for the under-represented group. The *key points* are: i) $\mathcal{H}(z)$ with higher curvature results in lower disparity of influence at the expense of lower total influence. ii) FAIRTCIM-BUDGET problem results in lower disparity and iii) the reduction in the total influence is only marginal as guaranteed by Theorem 1. In the subsequent figures, we only show the results of $\mathcal{H}(z) := \log(z)$ for the solution to problem P5.4.

[Q2, Q5, Q6] Effect of seed budget. Figure 5.4b shows the effect of different seed budgets on the number of influenced nodes (from different groups). Dotted and dash-dotted lines correspond to groups $\mathcal{V}_2$ and $\mathcal{V}_1$ respectively, while solid lines represent the total influence. The figure demonstrates that: (i) Disparity in the utility between both the groups increases with the increase in allowed seed budget. A reason for these differences could be the imbalances in groups sizes and average degrees, between both the groups— $\mathcal{V}_1$ and $\mathcal{V}_2$ comprise 70% and 30% of the nodes respectively. If a very big seed budget is allowed the disparity in influence might also reduce, however in many applications, due to limited resources, it is not practical to have a big budget; (ii) FAIRTCIM-BUDGET problem results in a lower disparate utility between the two groups compared to TCIM-BUDGET problem; (iii) this reduction in disparity is achieved at a very low cost to the total influence, as guaranteed by Theorem 1.

[Q3, Q5] Effect of deadline. Figure 5.4c compares disparity in the solutions of problems P5.1 and P5.4 as we vary the value of the deadline τ. Disparity is computed as the absolute difference between the fraction of individuals influenced in each group, given by Eq. 5.2. The figure demonstrates that: (i) disparity in group utilities does not have a unidirectional trend with increasing time deadline τ. One explanation for the increasing disparity— for $\tau = \{1, 2, 5\}$, could be that the seed nodes or the most influential nodes are propagating influence in *both the groups*, but as we increase the time deadline, Group V_1, with more nodes and edges, is more efficient at propagating influence compared to Group V_2, so it results in a larger disparity. But, after a threshold of increase in τ both groups are being influenced because longer cascades are allowed. Hence the disparity lowers and then plateaus, for $\tau = \{5, 10, 20, \infty\}$. One could imagine a case, as shown in



(a) Greedy selection iterations **(b)** Varying quota Q: Influence **(c)** Varying quota Q: Solution set size $|S|$

Figure 5.6: [Synthetic Dataset: Cover Problem] These figures show a comparison of TCIM-COVER problem P5.2, in red, and FAIRTCIM-COVER problem P5.6, in blue. They show that FAIRTCIM-COVER achieves lower disparity of influence between different groups with slightly bigger solution set sizes. See Section 5.4.3 for further details.

the motivating example in Figure 5.1, where seed nodes are surrounded by nodes of *only one group*, in this case increasing time deadline could yield a lower disparity. (ii) Our proposed method, given by problem P5.4, yields solutions which result in much lower disparity.

[Q4, Q5] Effect of activation probabilities. Figure 5.5a shows the disparity in influence for different activation probabilities $p_e \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0\}$. The results show that: i) lower activation probabilities could result in larger disparity. This makes intuitive sense, since with lower activation probabilities less nodes have a chance to be influenced. We are using an imbalanced graph, both in terms of group sizes and within and across group connectivity. It is very likely that the seeds selected might belong to the majority group and will have more connections to the nodes from their own group. With low activation probabilities less number of nodes are expected to be influence and the biases in the graph structure would become more pronounced, as evidenced by the results. With the high activation probabilities more number of nodes are expected to be influenced so the disparity in the influence is lower, as demonstrated by the results. ii) Lower values of τ tend to have a higher disparity compared to the higher values of τ. The intuition presented in the previous paragraph is confirmed with this experiment. iii) Our method consistently results in a lower disparity. The difference in disparities resulting from the solution of our method compared to the solution of traditional method in more pronounced for lower activation probabilities.

5.4.2.2 Varying graph properties

In this section, we vary several graph properties and answer following evaluation questions:

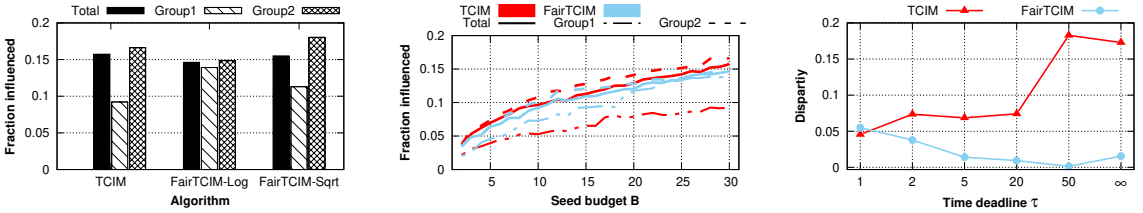
— **Q1**: How does varying group sizes affect the disparity?

— **Q2**: How does varying connectivity among the groups affect the disparity?
— **Q3**: How effective is our method in reducing disparity?

[Q1, Q3] Effect of group sizes. Figure 5.5b shows the effect of group sizes $g \in \{0.55, 0.6, 0.7, 0.8\}$. x-axis represents ratio of the nodes belonging to the two groups and y-axis represents disparity. i) The figure confirms our hypothesis that *imbalance in a graph could lead to disparate influence*, as motivated in the illustrative example given in Figure 5.1. Since we are considering a $1:25$ of $p_{het} : p_{hom}$, i.e., across vs within group edge probability ratios, even slight imbalance in the group sizes could result in a high disparity. The seed nodes or influential nodes are more likely to be from the dominant group and are more likely to be connected with nodes from their own groups. ii) On the other hand our proposed method results in almost no or very little disparity of influence, as it encourages to pick seeds which influence under-represented group.
[Q2, Q3] Effect of graph connectivity. Figure 5.5c demonstrates the importance of the graph structure, particularity connectivity between the two groups, characterized by $(p_{het}, p_{hom}) \in \{(0.025, 0.025), (0.015, 0.025), (0.01, 0.025), (0.001, 0.025)\}$. x-axis shows the ratio of across and within group edge probabilities. i) The figure validates our hypothesis that the majority group containing more influential nodes fares better in TCIM-BUDGET problem, as proposed in figure 5.1. Groups $\mathcal{V}_1$ and $\mathcal{V}_2$ comprise 70% and 30% of the nodes, respectively. As we increase the group-preferential attachment, represented by x-axis of figure 5.5c, influential nodes are more likely to have connections within the group $\mathcal{V}_1$, which in turn results in disparate influence propagation. ii) However, our proposed method performs better because it gives less weight to the nodes influenced from the majority group compared to the minority. Hence, our method encourages picking seed nodes which will influence the minority group, as explained in figure 5.2.
Takeaways. In this section we demonstrated that: (i) solving TCIM-BUDGET problem can lead to disparity of influence in different groups; (ii) the amount of disparity depends on the time deadline, activation probability, relative group sizes, budget, and connectivity of the graph; and (iii) instead, solving FAIRTCIM-BUDGET results in lower disparity of influence, with marginal reduction in overall influence, as guaranteed by Theorem 1.

5.4.3 TCIM under Coverage Constraints

Next, we compare solutions of TCIM-COVER problem P5.2, and our solution to FAIRTCIM-COVER problem P5.6. We solve both the problems using the greedy algorithm, i.e., iteratively picking seeds which maximize the constraints of problems P5.2 and P5.6 until the required quota is reached. The goal is to reach the prescribed quota



(a) Total and Group Influence **(b)** Varying budget B **(c)** Varying time deadline τ

Figure 5.7: [Rice-Facebook Dataset: Budget Problem] Comparison of results solving TCIM-BUDGET problem P5.1 and FAIRTCIM-BUDGET P5.4. We experimented with 4 groups and total influence includes all the groups, but we show group influences and disparity for only two groups which showed the maximum disparity. The results demonstrate that our method, given by problem P5.4, yields seed set which propagate influence in a more fair manner, at the cost of a marginally lower total influence. See Section 5.5.2 for further details.

Q, with minimum number of seeds. In all the figures discussed in this section, red color represents the results of TCIM-COVER problem P5.2, and blue color represents the results of our solution to FAIRTCIM-COVER problem P5.6. We answer the following question in this section:

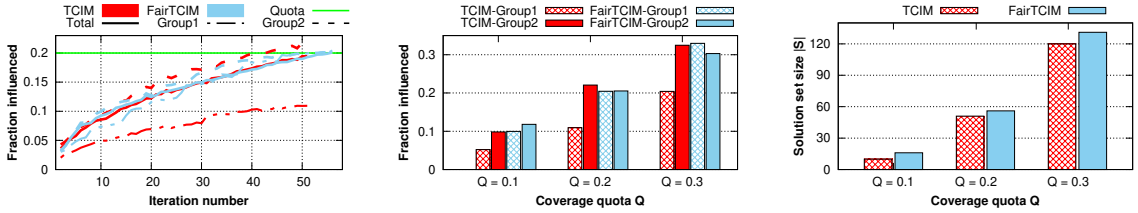
— **Q1**: How does our method fare compared to the traditional method over the *iterations of the algorithm*?

— **Q2**: How effective is our method in reducing disparity for different *reach quotas*?

— **Q3**: How much cost does our method incur?

[Q1] Effect of iterations. Figure 5.6a shows how the fraction of population influenced changes with seed selection at each iteration. Solid lines represent total influence while dash-dotted lines and dotted lines represent groups $\mathcal{V}_1$ and $\mathcal{V}_2$, respectively. In this experiment, Q was set to 0.2 which is represented by the horizontal green line. The figure demonstrates that: (i) both methods reach the required quota of the population; (ii) however, only the solution set of FAIRTCIM-COVER problem P5.6 reaches the required quota in both the groups; (iii) while maintaining roughly similar utility for both the groups throughout the iterations; (iv) and it does so *at a small expense of additional seeds*, as guaranteed in Theorem 2.

[Q2, Q3] Effect of quota Q. Figure 5.6b shows fractions of individuals that are influenced for different quota Q: (i) for different values of the required quota, traditional method given by problem P5.2 results in disparate utility between both the groups which is most likely due imbalance in group sizes and connectivity. (ii) Seeds selected by solving problem P5.6 result in a more equal utility because our method explicitly requires every



(a) Greedy selection iterations **(b)** Varying quota Q: Influence **(c)** Varying Q: Solution set size $|S|$

Figure 5.8: [Rice-Facebook Dataset: Cover Problem] These figures demonstrate the results of TCIM-COVER problem P5.2, in red, and FAIRTCIM-COVER problem P5.6, in blue. We experimented with 4 groups and total influence includes all the groups but we show group influences for the two groups which had maximum disparity. The results show that our method achieves a more equal coverage for all the groups at the expense of only slightly larger seed sets. See Section 5.5.3 for further details.

group to be influence up to quota Q. Depending on the graph structure, our method could result in a disparity up to $1 - Q$. The objective in the constraint given in problem P5.6 *only* increases if nodes belonging to the groups are influenced which have not reached the required quota, as demonstrated in figure 5.3. A higher disparity between groups could occur when it is not possible to influence the under-influenced group without influencing the already over-influenced group. In practice a higher disparity could occur, e.g, if one of the groups is very small and very sparsely connected within the group, which is unlikely to occur in practice. (iii) FAIRTCIM-COVER problem P5.6 uses only a small number of additional seeds, as guaranteed by Theorem 2.

Takeaways. We compared the result of TCIM-COVER problem P5.2 and our solution to FAIRTCIM-COVER problem P5.6. The results show that: (i) both methods reach the same fraction of the population; (ii) however, only FAIRTCIM-COVER problem results in seed sets influencing the required quota in *all the groups* and results in *a very low disparity* between groups; and (iii) lastly, FAIRTCIM-COVER yields *only* slightly larger solution sets as guaranteed by Theorem 2.

5.5 Experiments on Real-World Datasets

In this section, we evaluate our proposed solutions using two real-world datasets. We describe the datasets and the details of the experiments, and then present our findings.

5.5.1 Dataset and Experimental Setup

Next, we describe the datasets we used to evaluate our proposed methods, followed by the experimental setup.

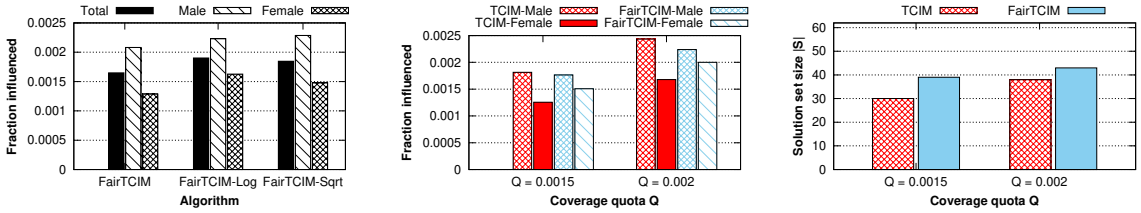
Rice-Facebook dataset. To evaluate our proposed methods, we used *Rice-Facebook* dataset collected by [200], where they capture the connections between students at the Rice University. The resulting network consists of 1205 nodes and 42443 undirected edges. Each node has 3 attributes: (i) the residential college id (a number between $[1-9]$), (ii) age (a number between $[18-22]$), and (iii) a major ID (which is in the range $[1-60]$).

We grouped the nodes (students) into four groups based on their age attributes. We experimented with all four groups while running our algorithms but present the results using only 2 groups which showed the *highest disparity*. We considered nodes with ages 18 and 19 as group $\mathcal{V}_1$ and age 20 as group $\mathcal{V}_2$. Group $\mathcal{V}_1$ has 97 nodes and 513 within-group edges. Whereas, group $\mathcal{V}_2$ has 344 nodes and 7441 within-group edges. Overall, there are 3350 across-group edges going between nodes in $\mathcal{V}_1$ and $\mathcal{V}_2$.

Instagram-Activities dataset. This dataset was gathered by [239]. It comprises 553628 nodes and 652830 undirected edges. The nodes represent a subset of Instagram users. There exists an edge between two nodes if either of them have liked or commented on each other's photos. Each node has a binary-valued gender attribute, i.e., male or female. 45.5% of the nodes belong to the male group. There are 179668 within-group edge among males and 201083 within-group edges among females, while there are 136039 across-group edges.

Facebook-Snap dataset. was proposed by [193]. The dataset comprise 4039 nodes and 88234 undirected edges. We used spectral clustering to identify 5 topological groups in the graph. The five groups comprise 546, 1404, 208, 788 and 1093 nodes. We run our algorithms for the entire dataset but report the results only for groups 1 and 4, as these groups showed the most disparity in influence using the traditional methods of influence maximization.

Experimental Setup. In all the experiments using *Rice-Facebook* dataset, we show the results for activation probability $p_e = 0.01$. All the other parameter were the same as described in section 5.4.1. For experiments using *Instagram-Activities* dataset we show the results with activation probability $p_e = 0.06$, time deadline $\tau = 2$, reach quota $Q = \{0.0015, 0.002\}$ and seed budget $B = 30$. We also experimented with other values of these parameters and get similar results. For *Instagram-Activities* we restrict the seeds to be picked from 5000 randomly selected nodes from the graph. However the influence was evaluated and propagated on the *entire* network. We used 500 sample for *Facebook-Rice* dataset and 10000 samples for *Instagram-Activities* dataset for Monte Carlo estimation of the influence of a node, which yielded very low-variance influence estimates. For *Facebook-Snap* dataset, we used edge weight on 0.01 and $\tau = 20$. Rest of the parameters were similar to the experiments described in in section 5.4.1.



(a) Budget Problem: Influence and Cost **(b)** Cover Problem: Influence **(c)** Cover Problem: Cost

Figure 5.9: [Instagram-Activities Dataset] These figures demonstrate a comparison of TCIM-BUDGET (Problem P5.1)vs FAIRTCIM-BUDGET (Problem P5.3) and TCIM-COVER (Problem P5.2) vs FAIRTCIM-COVER (Problem P5.5) problems. The results show that our methods fare better compared to the traditional methods. Even though the fraction of influence seems small, since the graph comprises 0.5m nodes, the differences in fractions are significant in total numbers.

5.5.2 TCIM under Budget Constraint

In this section, we compare the results of TCIM-BUDGET problem P5.1 and our solution to FAIRTCIM-BUDGET problem P5.4. Red color in all the figures discussed in this section corresponds to the solution of TCIM-BUDGET problem P5.1 and the blue color corresponds to our solution of FAIRTCIM-BUDGET problem P5.4. In all the experiments in this section we used a seed budget $B = 30$. We answer the following evaluation questions using two *real-world datasets* in this section:

— **Q1**: How does the choice of $\mathcal{H}(z)$ with different curvatures affect disparity?
— **Q2**: How does varying seed budget affect disparity?
— **Q3**: How does varying time deadline affect disparity?
— **Q4**: How effective is our method in reducing disparity?
— **Q5**: How much cost does our method incur?
— **Q6**: How effective are our methods on topological groups?

[Q1, Q4, Q5] Effect of different $\mathcal{H}(z)$. In Figures 5.7a and 5.9a, we compare the results of TCIM-BUDGET problem P5.1 and FAIRTCIM-BUDGET problem P5.4 using two realizations of $\mathcal{H}(z)$, given by: (i) $\mathcal{H}(z) := \log(z)$ and (ii) $\mathcal{H}(z) := \sqrt{z}$. In Figures 5.7a the total influence are shown for all the 4 groups while the group influences are shown for 2 out of the 4 groups which showed the maximum disparity. The results demonstrates that: (i) At a marginal reduction of total influence, as guaranteed by Theorem 1, our proposed method significantly reduces disparity in influence in case of Rice-Facebook dataset. However, in the *Instagram-Activities* dataset solving FAIRTCIM-BUDGET prob-

lem results in a *higher* total influence while achieving same or lower disparity for both the groups. This is in line with the finding by [238], which, using this dataset, shows that picking more diverse seeds could increase the total influence compared to greedy degree based seeding strategy. Greedy heuristic is just an approximation of the optimal solution. The optimal solution of the unfair problem cannot yield a lower influence compared to the optimal solution of the fair problem, as it adds additional constraints; (ii) as hypothesized in Section 5.3.1, a higher curvature function, $\mathcal{H}(z) := \log(z)$, leads to a bigger reduction in disparity compared to $\mathcal{H}(z) := \sqrt{z}$. In *Instagram-Activities* dataset $\mathcal{H}(z) := \sqrt{z}$ does not reduce disparity, however it does result in a higher fraction of influence in under-influenced group.

[Q2, Q4, Q5] Effect of seed budget. Figure 5.7b demonstrates the effect of allowed seed budget on the group and total influences. Groups $\mathcal{V}_1$ and $\mathcal{V}_2$ are represented by dash-dotted lines and dotted lines respectively and solid lines correspond to total influence. Similar to the results on synthetic dataset presented in section 5.4.2, i) the disparity between the groups seems to increase with increasing budget and ii) our method consistently results in lower disparity for different seed budgets, iii) while incurring a very small cost of total influence.

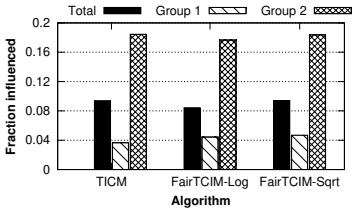
[Q3, Q4] Effect of time deadline. Figure 5.7c shows the effect of different time deadlines on the disparity between group influences, as calculated by Eq. 5.2. It demonstrates that: (i) the disparity of influence among groups increases as the value of τ increase, refer to section 5.4.2 for an intuitive explanation and, ii) our method is very effective in reducing disparity for different values of τ.

[Q1, Q4, Q5, Q6]. Figure 5.10a shows are our methods are able to reduce disparity among groups when we consider groups based on the graph topology. However, in this case the reduction in disparity does not seem to be substantial and as expected the over influence also does not decrease. In the case of Sqrt surrogate the overall influence marginally increases. In order to reduce disparity further one could consider a surrogate function with a higher curvature.

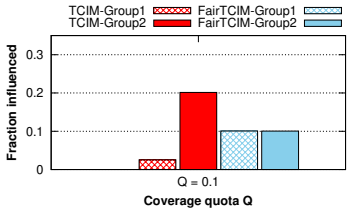
Takeaways. We demonstrated that: (i) FAIRTCIM-BUDGET, our proposed method, yields more fair solutions; (ii) this fairness is achieved at a very small reduction of the total influence compared to TCIM-BUDGET problem, as guaranteed by Theorem 1.

5.5.3 TCIM under Coverage Constraint

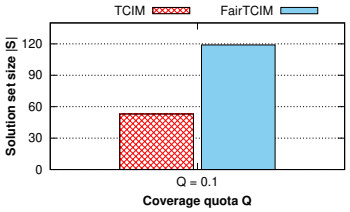
Next, we compare TCIM-COVER problem P5.2 and our solution to FAIRTCIM-BUDGET problem P5.6. Red color in all the figures discussed in this section corresponds to the solution of TCIM-COVER problem P5.2 and the blue color corresponds to our solution



(a) Budget Problem: Influence and Cost



(b) Cover Problem: Influence



(c) Cover Problem: Cost

Figure 5.10: [Facebook-Snap dataset] These figures demonstrate a comparison of TCIM-BUDGET (Problem P5.1) vs FAIRTCIM-BUDGET (Problem P5.3)and TCIM-COVER (Problem P5.2) vs FAIRTCIM-COVER (Problem P5.5) problems. The results show that our method improves the disparity of the influence between different groups. The results for the budget problem show some improvement in the disparity. However, in comparison the reduction in the total influence is also small. One can consider a concave wrapper with a larger curvature to improve the disparity. The results for the cover problem, show a clear improvement in the disparity between the groups.

of FAIRTCIM-COVER problem P5.6. We answer the following evaluation question using a *real-world* dataset.

— **Q1**: How does our method fare compared to the traditional method over the *iterations of the algorithm*?

— **Q2**: How effective is our method in reducing disparity for different *reach quotas*?

— **Q3**: How much cost does our method incur?

— **Q4**: How effect are our methods for topological groups?

[Q1] Effect of iterations. In Figures 5.8a we compare iterations of problem P5.2 and problem P5.6, realized with the log function. In each iteration, one seed is selected. Green line represents the required quota of coverage. Dashed-dotted lines, dotted lines and solid lines represent group $\mathcal{V}_1$, group $\mathcal{V}_2$ and total population, respectively. Similar to the results on Synthetic dataset, i) our method consistently results in lower disparity between the two groups, which showed the highest disparity, throughout the iteration of the seed selection algorithm; ii) our method influences all the groups up to prescribed quota; iii) by using small number of additional seeds.

[Q2, Q3] Effect of quota. Figures 5.8b, 5.8c, 5.9a and 5.9c demonstrate similar results to the synthetic dataset described in section 5.4.3. The *keypoint* is that all the groups are covered up to the required quotas with the solution set of FAIRTCIM-COVER problem by using only a small number of additional seeds.

[Q4] Topological groups. The results are shown in figure 5.10 show that our methods are effective in reducing disparity when considering topological grouping of graphs.

Takeaways. We compared the TCIM-COVER and FAIRTCIM-COVER problems in this section using a real world dataset. The results demonstrate that our method is i) effective in reducing disparity ii) by using a small additional number of seeds.

5.6 Related Work

In this section, we briefly review the related literature on influence maximization and contemporary works.

Influence Maximization. Richardson et al. [225] first introduced Influence Maximization as an algorithmic problem, and proposed a heuristic approach to find a set of nodes whose initial adoption of a certain idea/product can maximize the number of further adopters. Over the years, extensive research efforts have focused on the cascading behavior, diffusion and spreading of ideas or containment of diseases, by identifying the set of influential nodes that maximizes the influence through a network (often in real-time) [102, 147, 171, 225, 259].

Typically, identifying the most influential nodes is studied in two ways: (i) using network structural properties to find the set of most central nodes [147, 156], and (ii) formulating the problem as discrete optimization [19, 106, 147]. Kempe et al. [147], studied influence maximization under different social contagion models and showed that submodularity of the influence function can be used to obtain provable approximation guarantees. Since then, there has been a large body of work studying various extensions [27, 41, 49, 106, 125]. However, the notion of fairness in the influence maximization problem has not been studied by this line of previous works.

Contemporary Works. Very recently, Fish et al. [92], proposed a notion of individual fairness in information access, but did not consider the group fairness aspects. In addition, some prior works have proposed constrained optimization problems to encourage diversity in selecting the most influential nodes [6, 23, 36, 85].

A recent paper by Rahmattalabi et al. [221], proposes group fairness in influence maximization for robust covering problems. This method is different from ours in the following ways: i) their notion of fairness is maximizing the minimum influence for any group, while we propose parity of influence among different groups; ii) they consider a setting where seeds could be deactivated randomly while we do not have any stochasticity in seed activation; iii) they consider seed nodes to spread influence only to their immediate neighbors, while we vary the allowed time deadline and show its effect on disparity among different groups. We also demonstrate the effectiveness of our methods for different time deadlines on several datasets; iv) they propose an integer

linear programming set up while we propose submodular proxies, akin to the traditional methods, which can be approximately solved using the greedy heuristic.

In concurrent works, Khajehnejad et al., [149], and Tsang et al., [247], proposed methods to achieve group fairness in influence maximization. However, their works are very different from our approach in three ways: i) they propose a different problem formulation with objective that does not have submodular structural properties, ii) they only study the problem under budget constraint, and iii) they do not consider the time-critical aspect of influence in their definition of fairness for influence maximization. This could result in majority groups being influenced before the minority, and can lead to disparity in applications where the timing of being influenced/informed is critical. In our work, we introduce a submodular objective that directly addresses the time-criticality in influence maximization problem under budget constraint as well as coverage constraint.

5.7 Conclusion

In this chapter, we considered the important problem of time-critical influence maximization (TCIM) under (i) budget constraint (TCIM-BUDGET) and (ii) coverage constraint (TCIM-COVER). We showed that the existing algorithmic techniques aimed at maximizing total influence in the population could lead to a huge disparity in utility across the underlying groups. This can put minority groups at a big disadvantage with far-reaching consequences. To ensure that different groups are fairly treated, we proposed a notion of fairness and formulated two novel problems to solve TCIM under fairness considerations, namely, FAIRTCIM-BUDGET and FAIRTCIM-COVER. By introducing surrogate objective functions with submodular structural properties, we provided computationally efficient algorithms with desirable guarantees. Experiments over synthetic and real-world datasets demonstrated that our algorithms lead to low disparity in the time-critical influence propagation.

CHAPTER 6

Designing a new fair ADMS: Model Uncertainty

In this chapter, we address the fairness concerns arising due to model uncertainty in binary classification, as discussed in Section 1.1.3. Specifically, we ask the following question:

What constitutes a fair model under model uncertainty?

As discussed in section 1.2.3, answering this questions presents several challenges. In order to enforce group fairness, typically, a class of methods try to equalize errors. However, not all errors are the same. Errors could occur due to model/epistemic uncertainty or they could occur due to aleatoric uncertainty. So, firstly, we have to propose a sensible way to differentiate between different types of errors. Secondly, we have to come up with computationally efficient methods to identify different types of errors. Thirdly, we have to propose an efficient mechanism to enforce fairness under model uncertainty.

In this chapter, we address these challenges as follows:

- In Section 6.1, we discuss different aspects of our key idea to distinguish between types of errors based on their uncertainty-origin when training non-discriminatory classifiers, using a motivating example.
- In Section 6.2, we present the necessary background on binary classification and predictive multiplicity.
- In Section 6.3, we propose two scalable methods to identify errors arising based on their uncertainty-origin. Additionally, we propose efficient mechanisms to equalize errors arising due to model uncertainty.

- In Section 6.4, we evaluate our proposed methods using a synthetic and two real-world datasets.
- In Section 6.5, we discuss the related work on model uncertainty and predictive multiplicity.
- In Section 6.6, we present the conclusion of this chapter.

Relevant publication

The results presented in this chapter have been published in [12].

6.1 A proposal to differentiate between types of errors

As discussion in Section 1.1.3, it is well-known that errors in prediction models arise out of both epistemic (model) uncertainty and aleatoric (inherent) uncertainty [72, 124, 189]. Equalizing total error could lead to unjustifiably wrong decisions for some datapoints. Consider Figure 6.1, where a traditional fair classifier that equalizes total errors including the irreducible ones that arise due to aleatoric uncertainty. This results in many datapoints getting a negative outcome even though they clearly belong to the positive cluster. These errors are particularly consequential in socially impactful applications.

In this chapter, we argue to distinguish between the errors caused by different types of uncertainty. Specifically, we introduce the notions of *aleatoric errors* and *epistemic errors*. We refer to the errors that occur only due to model or epistemic uncertainty as *epistemic errors* and the ones that occur due to aleatoric uncertainty, we call the *aleatoric errors*. Figure 6.1 shows an example of both types of errors. The errors made by the classifiers C_1 and C_2 that are highlighted by the region **A** are due to the noise in the data, as these wrongly predicted datapoints are surrounded by predominantly the other class label, i.e., ground truth positive or ground truth negative datapoints. We refer to these types of errors as *aleatoric errors*. While the errors in the region marked by **E** are due to model uncertainty as one could resolve this uncertainty by gathering more data or by choosing a more complex model. These types of errors are *epistemic errors*. Our proposal is to *ignore* the aleatoric errors which are likely to be irreducible due to inherent uncertainty in the data or the prediction task at hand and we argue to *only* equalize the epistemic errors, i.e., the ones that occur due to methodological limitations.

In order to identify the epistemic errors that are caused by model uncertainty, we leverage the work on predictive multiplicity by Marx et al. [191]. *Predictive multiplicity*

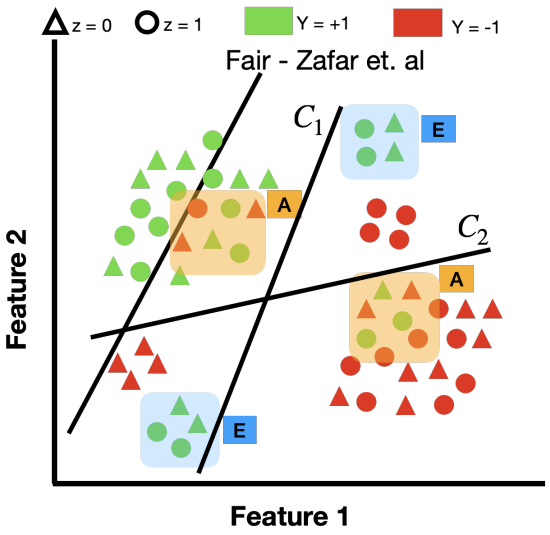


Figure 6.1: Illustrative example: Consider a binary classification task with two features and a sensitive feature represented by the shape of the points, i.e., circles and triangles. Green and red colors represent ground truth positive and negative labels, respectively. Classifiers $C1$ and $C2$ are equally accurate classifiers achieving 79% accuracy. The difference between false positives of triangles and circles for $C1$ is 22% and -12% with $C2$. However, these two classifiers disagree on their decision on 17% of the data, i.e., which lies in the ambiguous region shown in the shaded blue region. If we were to pick one of these classifiers it would be unfair to the points receiving a favorable decision with the other classifier. On the other hand, a fair classifier equalizing false positive rates, using [275], gives an accuracy of only 71%. However, it changes the decisions of several points that clearly belong to the positive cluster.

refers to the scenario where multiple predictive models have similar predictive performance (e.g., similarly accurate) but assign contradictory predictions on a subset of the datapoints, which characterize the *ambiguous regions*. We draw a connection between predictive multiplicity and *model uncertainty*.

Model uncertainty is defined as the level of spread or 'disagreement' in the decisions of an ensemble sampled from the posterior [189]. We use predictive multiplicity to identify model uncertainty, i.e., we argue that the disagreement in equally well performing models signals uncertainty in the model parameters. Specifically, we argue that if the classifiers exhibiting predictive multiplicity are chosen from a complex enough hypothesis class, then the regions in the feature space with high model uncertainty that are likely to have the epistemic errors would coincide with the ambiguous regions produced by predictive multiplicity. Therefore, our proposal of equalizing only the epistemic errors translates into equalizing errors in the ambiguous regions, while ignoring the ones in the unambiguous regions.

One of the *key properties* of our proposal is that people whose outcomes are affected by our fairness requirements are the people whose outcomes are ambiguous or uncertain

in the first place. Put differently, we do not alter the outcomes of people with unambiguously positive or negative outcomes. In contrast, current methods for achieving equal error rates might alter outcomes for people with unambiguous outcomes as well, creating a difficult accuracy-fairness tradeoff dilemma. We believe that our proposal would be easier to justify in many practical scenarios.

Key technical contributions of our approach are (a) designing efficient and scalable methods for identifying ambiguous regions, and (b) designing mechanisms for equalizing group error rates in the ambiguous regions. In order to solve the first challenge, we propose *convex proxies* to find models that exhibit predictive multiplicity. For the second challenge, our key insight is *to reuse the highly accurate models trained to identify the ambiguous regions* in the first place. Specifically, given the set of classifiers identifying ambiguous regions, we propose to *stochastically pick a classifier* from this set when making a decision. The probabilities of picking the classifiers are chosen in a way that equalizes group error rates in the ambiguous regions in expectation. An additional benefit of our approach compared to the traditional way of making a deterministic decision is that we account for model uncertainty by introducing stochasticity in our predictions, and thus many datapoints in the *ambiguous region* have a non-zero probability of receiving a favorable outcome. As there is some chance of getting a favorable outcome for most datapoints affected by our fairness notion, it would make our proposal more desirable than the traditional approach of assigning decisions deterministically.

6.2 Preliminaries and Background

In this section, we present the necessary background on binary classification and predictive multiplicity.

6.2.1 Binary Classification

Given a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, the goal of a binary classifier is to learn a function $\boldsymbol{\phi} : \mathbb{R}^d \to \{-1, 1\}$ between the feature vectors $\boldsymbol{x} \in \mathbb{R}^d$ and the class labels $y \in \{-1, 1\}$. In order to learn this function one has to solve $\boldsymbol{\phi}^* = \operatorname{argmin}_{\boldsymbol{\phi}} \boldsymbol{R}_{\mathcal{D}}(\boldsymbol{\phi})$: $\boldsymbol{R}_{\mathcal{D}}(\boldsymbol{\phi}) = \frac{1}{N} \sum_{\boldsymbol{x}_i, y_i} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) \neq y_i]$. However, this function is non-convex in $\boldsymbol{\phi}$ and worse, it is intractable, which makes it especially difficult to solve for large datasets. In the rest of the text we drop the subscript, $\mathcal{D}$, for brevity. To efficiently solve the problem, it is a standard practice to use a convex proxy. One minimizes a given convex loss $L(\boldsymbol{\theta})$ over $\mathcal{D}$, i.e., , $\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, in order to find $\boldsymbol{\theta}^*$ for convex decision-boundary-based classifiers like linear/non-linear SVM and logistic regression, where $\boldsymbol{\theta} \in \mathbb{R}^d$. Then,

for a given (potentially unseen) feature vector $\boldsymbol{x}$, one predicts the class label $\hat{y} = 1$ if $d_{\boldsymbol{\theta}^*}(\boldsymbol{x}) \geq 0$ and $\hat{y} = -1$ otherwise, where $d_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ denotes the signed distance from $\boldsymbol{x}$ to the decision boundary. For convenience, we define $\boldsymbol{\theta}^*(\boldsymbol{x}) = 1$ if $d_{\boldsymbol{\theta}^*}(\boldsymbol{x}) \geq 0$ and $\boldsymbol{\theta}^*(\boldsymbol{x}) = -1$ if $d_{\boldsymbol{\theta}^*}(\boldsymbol{x}) < 0$.

In the rest of the chapter, we consider $\boldsymbol{\theta}_{best}$ to be the most accurate classifier yielded by minimizing logistic regression loss with L2 regularizer, where weights of the regularizer were picked based on the performance on the validation set. Similarly, we consider $\boldsymbol{\phi}_{best}$ to be the best classifier using 0-1 loss ($\boldsymbol{R}_{\mathcal{D}}$), selected using a validation set.

6.2.2 Background on Predictive Multiplicity

In this section, we formally introduce the notion of predictive multiplicity and discuss the existing measures and mechanisms to compute predictive multiplicity.

Predictive multiplicity. A prediction problem exhibits predictive multiplicity if one can find a classifier $\boldsymbol{\phi}$ for a given small value $\boldsymbol{\epsilon}$ such that $\boldsymbol{R}(\boldsymbol{\phi}) - \boldsymbol{R}(\boldsymbol{\phi}_{best}) <= \boldsymbol{\epsilon}$, and there exists at least one datapoint with feature vector $\boldsymbol{x}_i$ such that $\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}_{best}(\boldsymbol{x}_i)$ [191]. The definition for classifiers trained with proxy loses is similar. One could consider $\boldsymbol{\epsilon}$ to be 0 but in practice a classifier that is slightly less accurate on the training data might be equally or even more accurate on the test data.

Predictive multiplicity is defined for a set of two or more classifiers, referred to as the $\boldsymbol{\epsilon}$-level set. Given the most accurate classifier $\boldsymbol{\phi}_{best}$, the $\boldsymbol{\epsilon}$-level set of $\boldsymbol{\phi}_{best}$ is a set of classifiers which have an accuracy only up to $\boldsymbol{\epsilon}$ lower than $\boldsymbol{\phi}_{best}$. Formally, over the dataset $\mathcal{D}$, $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\phi}_{best}} = \{\boldsymbol{\phi} : \boldsymbol{R}(\boldsymbol{\phi}) - \boldsymbol{R}(\boldsymbol{\phi}_{best}) \leq \boldsymbol{\epsilon}\}$.

Measures of predictive multiplicity. Marx et al. [191] propose two measures for predictive multiplicity for a given set of classifiers, namely *Discrepancy* and *Ambiguity*.

For a given set of classifiers, *Discrepancy* is defined as the maximum fraction of the datapoints on which any classifier in the set disagrees on the outcomes with the most accurate classifier. Formally, given $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\phi}_{best}}$ and dataset $\mathcal{D}$,

$$\delta_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}) = \max_{\boldsymbol{\phi} \in \mathbb{C}_{\boldsymbol{\epsilon}}} \frac{1}{n} \sum_{\boldsymbol{x}_i \in \mathcal{D}} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}_{best}(\boldsymbol{x}_i)], \tag{6.1}$$

i.e., discrepancy is the maximum fraction of conflicting decisions yielded by any classifier in $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\phi}_{best}}$ compared to $\boldsymbol{\phi}_{best}$.

Ambiguity of a set of classifiers for a prediction task is defined as the fraction of datapoints given a different decision than the best classifier. Formally, given set $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\phi}_{best}}$ and dataset $\mathcal{D}$,

$$\alpha_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}) = \frac{1}{n} \sum_{\boldsymbol{x}_i} \max_{\boldsymbol{\phi} \in \mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\phi}_{best}}} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}_{best}(\boldsymbol{x}_i)], \tag{6.2}$$

where $\max_{\phi \in \mathbb{C}_{\epsilon,\phi_{best}}} \mathbb{1}[\phi(\boldsymbol{x}_i) \neq \phi_{best}(\boldsymbol{x}_i)]$ is 1 if there exists at least one classifier in $\mathbb{C}_{\epsilon,\phi_{best}}$ which gives a datapoint with features $\boldsymbol{x}_i$ a different outcome than ϕ_{best}, otherwise it is 0. Hence, ambiguity is the fraction of datapoints on which any classifiers in $\mathbb{C}_{\epsilon,\phi_{best}}$ disagrees on the outcome with ϕ_{best}.

Methods to identify predictive multiplicity. Inspired by the measures discrepancy and ambiguity, Marx et al. [191] propose two methods that maximize these measures in order to find the classifiers that exhibit maximum predictive multiplicity for the given allowance of accuracy reduction. This would indicate the extent of predictive multiplicity for the prediction task at hand.

Exact discrepancy maximization (Dsc-Exact). Given a value of ϵ, the authors propose to train classifiers that minimize the agreement to ϕ_{best} under the constraint that its accuracy is only up to ϵ lower than ϕ_{best}, i.e.,

$$\underbrace{\min_{\phi} \sum_{\boldsymbol{x}_i} \mathbb{1}[\phi(\boldsymbol{x}_i) = \phi_{best}]}_{\text{maximize discrepancy}} \quad \text{subject to} \quad \underbrace{\boldsymbol{R}(\phi) \leq \boldsymbol{R}(\phi_{best}) + \boldsymbol{\eta}}_{\text{bound accuracy reduction}} \tag{P6.1}$$

where $\boldsymbol{\eta} \in (0, \epsilon)$. One can obtain a set $\mathbb{C}_{\epsilon,\phi_{best}}$ by solving the above formulation for several $\boldsymbol{\eta}$ values.

Exact ambiguity maximization (Amb-Exact). In order to find the classifiers that maximize the ambiguity measure for a given threshold of accuracy reduction, Marx et al. [191] propose to train a classifier for each datatpoint in the training data that gives the datapoint a different decision than the most accurate classifier. Then, they pick the classifiers whose accuracy lies within the threshold of the allowed accuracy reduction. Specifically, they propose to train classifiers that change their decisions compared to ϕ_{best} for individual datapoints while minimizing 0-1 loss, i.e.,

$$\underbrace{\min_{\phi} \boldsymbol{R}(\phi)}_{\text{maximize accuracy}} \quad \text{subject to} \quad \underbrace{\phi(\boldsymbol{x}_i) \neq \phi_{best}(\boldsymbol{x}_i)}_{\text{change decision of } \boldsymbol{x}_i \text{ w.r.t } \phi_{best}} \quad \forall \boldsymbol{x}_i. \tag{P6.2}$$

Then, one can select $\mathbb{C}_{\epsilon,\phi_{best}}$ by pruning the set of classifiers resulting from the solution of the problem above, i.e., by selecting classifiers which are only ϵ lower in accuracy than ϕ_{best}.

To solve both Problems P6.1 and P6.2, Marx et al. [191] propose mixed integer programming formulations. However, these formulations i) work only for linear classifiers and ii) have slow performance as these are exact, intractable and non-convex.

6.3 Proposed approach

In this section, we aim to answer the question: *What is a fair model under model uncertainty*?

We characterize model uncertainty using predictive multiplicity. Given a set of classifiers $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$ that exhibit predictive multiplicity, we consider x_i to have an ambiguous decision if *any* of the classifiers in $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$ gives it a conflicting decision compared to any other classifier. Formally a set of ambiguous points are defined as:

$$\mathcal{A} := \{x_i : \boldsymbol{\theta}_j(x_i) \neq \boldsymbol{\theta}_k(x_i) \, \forall \, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k \in \mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}\}.$$

These points characterize the ambiguous region. By choosing a single model from $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$ as the final model we might be unfair to some group in the ambiguous region. Our proposal of only equalizing the epistemic errors boils down to equalizing group error rates in the ambiguous region $\mathcal{A}$.

The *key assumption* we make is that the hypothesis class for the classifiers, $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$, exhibiting predictive multiplicity is sufficiently complex, i.e., if the data is nonlinearly separable the hypothesis class should include nonlinear classifiers. Under this assumption, all the errors in the the unambiguous region, i.e., where all the classifiers in the set $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$ agree in their decisions, would *only* be due aleatoric uncertainty. The argument is as follows: Given the classifiers in set $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$ are picked from a sufficiently complex hypothesis class for the given data. Under this assumption, if all the classifiers agree in their prediction for a subset of the datapoints, then the resulting errors for these datapoints could only be due to inherent stochasticity of the prediction task or random noise, i.e., aleatoric errors. On the other hand, the ambiguous region, $\mathcal{A}$, would identify regions with high model uncertainty. The intuition is as follows: Given that the classifiers for set $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$ are chosen from a sufficiently complex hypothesis class. Under this assumption, if these equally accurate classifiers disagree on some datapoints this would include all the datapoints whose decisions are uncertain due to lack of data. This implies that *all* the epistemic errors will lie in the ambiguous region. The ambiguous region could also have random noise hence causing some aleatoric errors. The results using the Synthetic dataset in Section 6.4.4 confirm our hypotheses.

Next, we present our proposals for identifying the ambiguous region using scalable convex methods. Then, we discuss our methods for equalizing groups errors in the ambiguous region $\mathcal{A}$.

6.3.1 Scalable Methods for Predictive Multiplicity

In this section, we propose two convex methods to find the ambiguous region $\mathcal{A}$.

Approximate Discrepancy maximization (Dsc-Approx). We propose the following convex and tractable proxy constraint that bounds similarity between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{best}$, akin to the objective in Problem P6.1 that maximizes discrepancy:

$$\frac{1}{N}\sum_{x}\max(0, d_{\boldsymbol{\theta}(x)}d_{\boldsymbol{\theta}_{best}(x)}) \leq \gamma, \tag{6.3}$$

where $d_{\boldsymbol{\theta}(x)}$ represents the distance of the datapoint with feature vector $\boldsymbol{x}$ from the decision boundary of $\boldsymbol{\theta}$. $\max(0,\cdot)$ represents the agreement of decisions between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{best}$. Specifically, if the decision for a subject with feature vector $\boldsymbol{x}$ stays the same under $\boldsymbol{\theta}$ compared to $\boldsymbol{\theta}_{best}$, only then does the term $\max(0,\cdot)$ produce a non-zero number. Thus, by bounding the left hand side we are limiting the average allowed distance of the datapoints which have the same decisions under $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{best}$. Making this bound tighter would preferably admit $\boldsymbol{\theta}$ whose decisions are different on some of the datapoints than $\boldsymbol{\theta}_{best}$, as those datapoints contribute 0 to the sum on the left hand side. This implies that one can control the number of decisions allowed to be the same between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{best}$ by changing the value of $\gamma \in \mathbb{R}+$. For example, $\gamma = +\infty$ would yield $\boldsymbol{\theta} = \boldsymbol{\theta}_{best}$ meaning that all the decisions between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{best}$ are the same, i.e., $\boldsymbol{\theta}$ would yield a discrepancy of 0 compared to $\boldsymbol{\theta}_{best}$. Similarly, for $\gamma = 0$ one aims to learn $\boldsymbol{\theta}$ whose decisions are different on all datapoints than to $\boldsymbol{\theta}_{best}$, i.e. a classifier yielding maximum discrepancy compared to $\boldsymbol{\theta}_{best}$. The value of γ also controls the reduction in accuracy under $\boldsymbol{\theta}$ compared to $\boldsymbol{\theta}_{best}$.

For linear boundary-based classifiers (logistic regression, linear SVM), $d_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T\boldsymbol{x}$. For nonlinear SVM, one can write $d_\beta(\boldsymbol{x}) = \sum_{i=1}^{N}\beta_i y_i k(\boldsymbol{x}_i, \boldsymbol{x})$ for the optimization variables β and a positive semidefinite kernel function $k(.,.)$. Hence, in both linear and nonlinear cases the constraint stays convex since the distance from the decision boundary is linear with respect to the optimization variables.

One can write a convex and tractable version of Problem P6.1 using the logistic regression loss as follows:

$$\underbrace{\min_{\boldsymbol{\theta}} \quad -\frac{1}{N}\sum_{\boldsymbol{x}_i, y_i} p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})}_{\text{maximize accuracy}} \tag{P6.3}$$
$$\text{subject to} \quad \underbrace{\frac{1}{N}\sum_{\boldsymbol{x}_i} \max(0, d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)}) \leq \gamma}_{\text{enforce discrepancy}}$$

where $p(y=1|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x})}$.

One can learn an appropriate γ value using a validation set, for a given $\boldsymbol{\eta}$ in P6.1. We construct $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$ by training classifiers with varying values of γ and then picking the ones whose accuracy is only $\boldsymbol{\epsilon}$ lower than $\boldsymbol{\theta}_{best}$.

Approximate ambiguity maximization (Amb-Approx). We propose the following convex and tractable constraints equivalent to the constraint in Problem P6.2.

$$\begin{aligned} &d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} < 0 \text{ if } d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)} \geq 0 \quad \forall \boldsymbol{x}_i \\ &d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} \geq 0 \text{ if } d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)} < 0 \quad \forall \boldsymbol{x}_i, \end{aligned} \tag{6.4}$$

where $d_{\boldsymbol{\theta}}$ is the distance from decisions boundary of $\boldsymbol{\theta}$. The constraints above require $\boldsymbol{\theta}$ to make a different decision than $\boldsymbol{\theta}_{best}$ on the datapoint $\boldsymbol{x}_i$. The constraints stay convex for both linear and nonlinear boundary based classifiers because one can write the distance from the decision boundary as a linear function of the optimization parameter in both cases. One can write a convex and scalable version of Problem P6.2 as follows:

$$\underbrace{\min_{\boldsymbol{\theta}} \quad -\frac{1}{N}\sum_{\boldsymbol{x}_i, y_i} p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})}_{\text{maximize accuracy}} \tag{P6.4}$$
$$\begin{aligned} \text{subject to} \quad & d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} < 0 \text{ if } d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)} \geq 0 \quad \forall \boldsymbol{x}_i \\ & \underbrace{d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} \geq 0 \text{ if } d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)} < 0 \quad \forall \boldsymbol{x}_i}_{\text{change decision of } \boldsymbol{x}_i \text{ w.r.t } \boldsymbol{\theta}_{best}}, \end{aligned}$$

where $p(y=1|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x})}$. We pick $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$ by training a set of classifiers which assign conflicting decisions to all the datapoints in the training set. Then, we pick the classifiers which are only $\boldsymbol{\epsilon}$ lower in accuracy than $\boldsymbol{\theta}_{best}$.

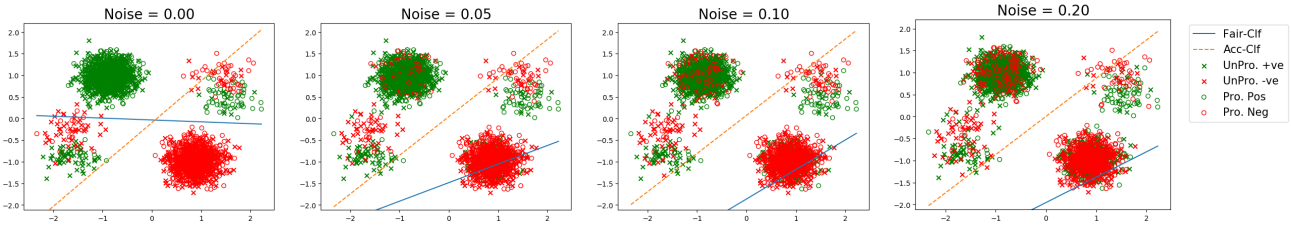


Figure 6.2: [Synthetic dataset] Figure demonstrates that state of the art fairness methods are effected by label noise.

6.3.2 Leveraging Predictive Multiplicity towards Fairness under Model Uncertainty

In this section, we propose to learn a meta classifier in order to equalize group errors arising due to model uncertainty.

In order to do that, our key insight is to use the highly accurate classifiers that we trained to identify the ambiguous regions in the first place. Specifically, given the validation set of datapoints and $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$, picked by solving DSC-APPROX, P6.3, or AMB-APPROX, P6.4, we first identify the points with ambiguous decisions. We then construct a meta classifier by picking the classifiers stochastically from the set $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$. The probabilities for picking these classifiers are chosen in a way that aims to equalize group error rates on the ambiguous datapoints among different groups of a sensitive feature such as race or gender. For a binary valued sensitive feature $z = \{0, 1\}$, we propose

$$\min_{w} \quad |\sum_{\theta \in \mathbb{C}_\epsilon} w_\theta \cdot \underbrace{(Err_{z=1}(\theta) - Err_{z=0}(\theta))}_{\text{FPR/FNR difference}}| \qquad \text{(P6.5)}$$
$$\text{subject to} \quad 0 \leq w_\theta \leq 1 \quad \text{and} \quad \sum_{\theta} w_\theta = 1,$$

where $Err_{z=0}(\theta)$ and $Err_{z=1}(\theta)$ are false positive rates (FPR) or false negative rates (FNR) for group 0 and 1 of the sensitive feature *in the ambiguous region*, $\mathcal{A}$. As the set of classifiers is predetermined, the error rates can be precomputed. Hence, the problem is convex and efficiently solvable, as the objective function is a linear function of optimization variable w.

The intuition is that the difference of the errors rates between the two groups, i.e., $Err_{z=1}(\theta) - Err_{z=0}(\theta)$, might be positive for some of the classifiers in $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$ and it might be negative for the others. We can then assign the probabilities w_θ to these classifiers in a

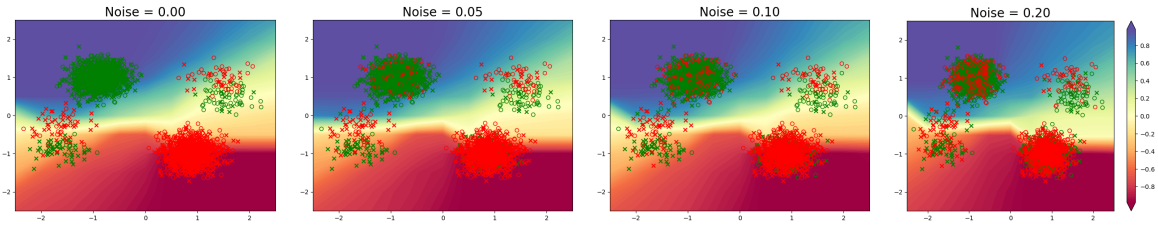


Figure 6.3: [Synthetic dataset] Figure shows the expected class while equalizing FPRs using the classifiers solving P6.4. It demonstrates that our method is stable under label noise, as it consistently identifies same regions as ambiguous for different levels of noise values.

way such that they cancel each others biases and the expected unfairness is minimized. Our experimental results on the real-world and synthetic datasets confirm our intuition (Tables 6.1, 6.3, 6.4).

In the case of a non-binary valued sensitive feature, one can replace the error rate difference between two groups with pair-wise differences among all the groups. We learn the probability mass function w using the validation datapoints, and when classifying the unseen test datapoints we use w to pick the classifiers from $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$.

6.4 Experiments

In this section, we demonstrate the effectiveness of our methods using synthetic and real-word datasets. Specifically, we answer the following evaluation questions:
– **Q1.** How effective and fast are our methods in identifying the ambiguous regions?
– **Q2.** What is the fairness and accuracy trade-off of our methods?
– **Q3.** Are our methods robust to noisy data?

6.4.1 Datasets

We use a *Synthetic dataset* because i) we could easily alter the size of the datasets, which is useful as DSC-EXACT and AMB-EXACT have slow performance on larger complex datasets, especially with continuous valued features; ii) we could provide intuition for the type of ambiguous regions identified by our methods; iii) we could introduce noise in the data and check the robustness of our methods vs the existing methods. The data comprises 10000 datapoints and 2 features and a binary valued sensitive feature, z. The

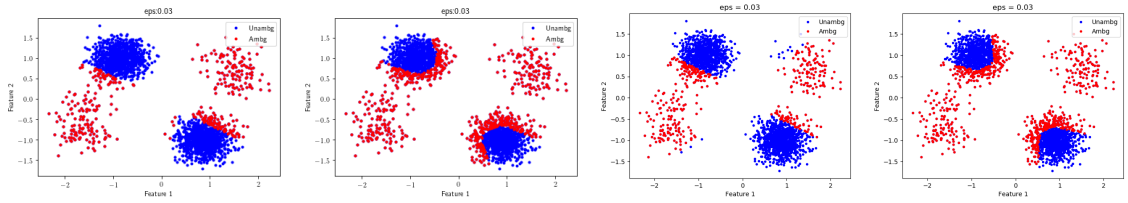
data is sampled from the following Gaussian distributions:

$$\mathcal{N}_1([-35; 65], [60, 1; 1, 120]),\ \mathcal{N}_2([15; -25], [60, 1; 1, 120]),$$
$$\mathcal{N}_3([30; 65], [70, 1; 1, 100]),\ \mathcal{N}_4([35; 40], [70, 1; 1, 100]),$$
$$\mathcal{N}_5([-55; 5], [70, 1; 1, 100]) \text{ and } \mathcal{N}_6([-55; -20], [70, 1; 1, 100])$$

From $\mathcal{N}_1$, 4500 points were sampled. Amongst these, 95% of which were labeled ground truth positive and 65% of these points were uniformly at random assigned to the non-protected class of the sensitive feature, i.e, $z = 0$. A total of 4500 points were sampled from $\mathcal{N}_2$, 95% of which are ground truth negative points and 65% of these points were uniformly at random assigned to the protected class of the sensitive feature, i.e., $z = 0$. Finally, 250 points were sampled from $\mathcal{N}_3$ and $\mathcal{N}_5$ each, with ground truth negative labels, and 250 points were sampled from $\mathcal{N}_4$ and $\mathcal{N}_6$ each and were assigned ground truth positive labels. 80% of the points sampled from $\mathcal{N}_3$ and $\mathcal{N}_4$ and 20% of the points sampled from $\mathcal{N}_5$ and $\mathcal{N}_6$, were uniformly at randomly assigned $z = 1$. After sampling these points they were normalized to have a unit mean and a unit variance. A visual representation is shown in Figure 6.2. We flipped the class label of a fraction of datapoints which induced aleatoric errors through out the data. However, model uncertainty only exists in the sparse clusters shown in Figure 6.2 as that could be reduced by gathering more data. Our hope is that predictive multiplicity would be able to identify regions with predominantly model uncertainty, i.e., the sparse clusters as the ambiguous regions for different levels of aleatoric uncertainty. We also experimented with other variations of the parameters and got similar results.

We processed the ProPublica *COMPAS dataset* [166] similar to Zafar et al. [274], which resulted in $5,287$ subjects and 7 features. Given these features we have to predict whether a criminal defendant would recidivate within two years (positive class) or not (negative class). We consider race, with values African-Americans, $z = 0$, and white, $z = 1$, to be a sensitive feature in this dataset.

The NYPD *SQF dataset* comprises features of pedestrians, such as race, gender, height etc. and the goal is to predict whether (negative class) or not (positive class) a weapon was discovered on inspection. We use race as a sensitive feature, z, in our experiments, with African-Americans ($z = 0$) and white ($z = 1$) as two values of this feature. After processing the data similar to Zafar et al. [274] the dataset consists of $5,832$ subjects and 22 features.



(a) DSC-EXACT **(b)** AMB-EXACT **(c)** DSC-APPROX **(d)** AMB-APPROX

Figure 6.4: [Synthetic dataset] This figure shows the ambiguous regions (in red) identified by the four methods discussed in the paper. It demonstrates that our methods identify similar ambiguous regions compared to the exact methods proposed by Marx et al. [191]. The results correspond to $\epsilon = 0.03$. We see similar results for different values of ϵ.

6.4.2 Experimental Setup

The datasets were split into 50% training, 25% validation and 25% test datapoints. Training data was used to train the classifiers, validation data for tuning hyper parameters and test data to report the results. The CVXPy library [75] was used to solve all the formulations. We show results using linear classifiers, as decisions made by the linear classifiers are relatively easier to explain, which is an import goal for applications with social significance such as recidivism risk prediction. Additionally, data are likely to be linearly separable in higher dimensions. We show some results using nonlinear boundaries with our methods in the appendix.

Selecting $\mathbb{C}_{\epsilon,\boldsymbol{\theta}_{best}}$. We generate $\mathbb{C}_{\epsilon,\boldsymbol{\theta}_{best}}$ by solving DSC-APPROX, given by Problem P6.3, for a range of γ values or AMB-APPROX, given by Problem P6.4, for each training datapoint. Then, we use the validation data to prune the resulting classifiers which lie within a given ϵ threshold of the most accurate classifier. The results are averaged over 5 runs of these steps using different seed values to initialize the data-split and the solver. For DSC-APPROX, we pick the $\mathbb{C}_{\boldsymbol{\epsilon}}$ from the aggregated solutions of all the seeds and present the averaged statistics over all the seeds.

We assume that ϵ is chosen by the experts for the prediction task at hand. We present results for $\epsilon = 0.02$ for the synthetic dataset, and $\epsilon = 0.01$ for real-world datasets. We experimented with several values of ϵ and obtained similar results.

6.4.3 Benchmarks and Metrics

In this section, we discuss the benchmarks and metrics we used to evaluate our proposals.

Ambiguous regions computation benchmarks. In order to demonstrate the efficiency of our methods to identify the ambiguous regions using DSC-APPROX and AMB-APPROX,

Table 6.1: [Synthetic dataset] Signed differences in FPR/FNR: This table demonstrates that our method is effective in removing unfairness at a very small cost of decrease in the accuracy. Please refer to Section 6.4.4

	Unfairness			Accuracy
	total	unamb	amb	
Acc.	0.13/-0.14	0.05/-0.06	0.46/-0.45	0.89
Fair	0.03/-0.02	0.05/-0.06	-0.14/0.29	0.77/0.89
Uni-P6.3	0.04/-0.04	0.05/-0.06	-0.22/0.20	0.89 / 0.89
Our-P6.3	0.07/-0.07	0.05/-0.06	**0.0/-0.01**	0.89/0.89
With P6.4				
Acc.	0.13/-0.14	0.06/-0.07	0.30/-0.35	0.89
Fair	0.03/-0.02	0.05/-0.07	-0.06/0.18	0.77/0.89
Uni-P6.4	0.10/-0.10	0.06/-0.07	0.16/-0.16	0.88 / 0.88
Our-P6.4	0.06/-0.07	0.06/-0.07	**0.01/-0.03**	0.88/ 0.88

we compare with DSC-EXACT and AMB-EXACT. We solved the DSC-EXACT and AMB-EXACT problems using the CPLEX library, with the code provided by the authors [191].

Metrics for evaluating ambiguous regions computation. Since the best classifiers for non-scalable and our scalable methods, i.e., $\boldsymbol{\phi}_{best}$ and $\boldsymbol{\theta}_{best}$, are different, we report the ambiguity $\hat{\alpha}$ and discrepancy $\hat{\delta}$ between any two classifiers in $\mathbb{C}_\epsilon$, for the respective methods. They are formally defined as follows:

$$\hat{\delta}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}) = \max_{\boldsymbol{\phi},\hat{\boldsymbol{\phi}}\in\mathbb{C}_{\boldsymbol{\epsilon}}} \frac{1}{n}\sum_{\boldsymbol{x}_i} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \hat{\boldsymbol{\phi}}(\boldsymbol{x}_i)] \tag{6.5}$$

$$\hat{\alpha}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}) = \frac{1}{n}\sum_{\boldsymbol{x}_i} \max_{\boldsymbol{\phi},\hat{\boldsymbol{\phi}}\in\mathbb{C}_{\boldsymbol{\epsilon}}} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \hat{\boldsymbol{\phi}}(\boldsymbol{x}_i)]. \tag{6.6}$$

High values of these measures are desired, as that would imply that the $\mathbb{C}_{\boldsymbol{\epsilon}}$ contains diverse classifiers which can identify more number of datapoints that have a contradictory decision for a given value of $\boldsymbol{\epsilon}$. We also report the time it takes to compute the set of classifiers $\mathbb{C}_{\boldsymbol{\epsilon}}$.

Fairness benchmark. For results on fairness in the ambiguous regions, we compare our method given by Problem P6.5 using $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$, picking classifiers uniformly at random from $\mathbb{C}_{\boldsymbol{\epsilon},\boldsymbol{\theta}_{best}}$, the most accurate classifier and a traditional fair classifier. We chose one traditional fair method as a baseline, as Zafar et al. [275] show comparison to other

Table 6.2: Comparison identifying ambiguous regions: The tables show maximum discrepancy and ambiguity between any two classifiers in the $\mathbb{C}_{\boldsymbol{\epsilon},\psi:\psi\in\{\boldsymbol{\phi}_{best},\boldsymbol{\theta}_{best}\}}$. The bottom table shows the time it took to compute the ambiguous regions with each method. It shows that our methods, given by P6.3 and P6.4, achieve comparable performance compared to P6.1 and P6.2 and they are upto four orders of magnitude faster. Please refer to Section 6.4.4

ϵ	P6.1		P6.2		P6.3		P6.4	
-	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$
0.03	0.15	0.16	**0.18**	**0.28**	0.14	0.16	**0.18**	0.26
0.05	0.17	0.19	0.22	**0.38**	0.16	0.17	**0.23**	0.36
0.09	0.22	0.24	**0.32**	**0.56**	0.2	0.20	**0.32**	0.51

Training Time

Time	P6.1	P6.2	P6.3	P6.4
mins	510	19227	5	5

approaches and get similar results. Its formulation ([275] and [13]) is given as follows,

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}|| \\ \text{subject to} \quad & \frac{1}{|\mathcal{D}_*|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}_*}(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| < c, \end{aligned} \tag{P6.6}$$

where $\mathcal{D}_*$ was set to datapoints with ground truth negative labels and ground truth positive labels for equalizing false positive rates (FPR) and false negative rates (FNR), respectively. z represents the value of the sensitive feature and c represents the allowed correlation between z and the decision boundary, d_θ.

We train accurate classifiers by solving
minimize $-\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}||$ for different λ. Logistic regression loss was used to train all the classifier. More details such as ranges for the hyper parameter search values, seeds, specifications of the machines used and other training details are included in the appendix.

Metrics for fairness. We assume a binary valued sensitive attribute and report a signed difference of FPR and FNR between the unprotected and the protected group for the sensitive feature z.

$$\text{unfairness-FPR} = FPR_{z=1} - FPR_{z=0}, \tag{6.7}$$

$$\text{unfairness-FNR} = FNR_{z=1} - FNR_{z=0} \tag{6.8}$$

We present these numbers for the overall data, for the unambiguous regions, i.e., where all the classifiers give unanimous decisions, and for the ambiguous regions. We also

Table 6.3: [COMPAS] Signed differences in FPR/FNR : This table demonstrates that our methods are effective in removing unfairness in the ambiguous regions at no expense of accuracy. Please refer to Section 6.4.5

	Unfairness			Accuracy
	total	unamb	amb	
Acc.	-0.19/0.33	-0.23/0.41	**0.08**/-0.20	0.66
Fair	0.02/0.03	-0.09/0.18	0.83/-0.92	0.66/0.65
Uni-P6.3	-0.20/0.35	-0.23/0.41	**-0.08**/**-0.004**	0.66 /0.66
Our-P6.3	-0.20/0.35	-0.23/0.41	**-0.08**/-0.02	0.66/0.66
With P6.4				
Acc.	-0.19/0.33	-0.24/0.54	-0.11/0.15	0.66
Fair	0.02/0.03	-0.24/0.54	0.34/0.-0.42	0.66/0.65
Uni-P6.4	-0.19/0.34	-0.24/0.54	-0.11/0.15	0.66/ 0.66
Our-P6.4	-0.14/0.26	-0.24/0.54	**-0.01**/**0.03**	0.66/ 0.66

report the accuracies. We aim to achieve low disparity in group error rates in the ambiguous regions, while achieving an accuracy similar to the most accurate classifier.

6.4.4 Synthetic Experiments

In this section, we answer the evaluations questions using the synthetic dataset.

Q1: Ambiguous regions coverage and speed. We compared our methods, DSC-APPROX and AMB-APPROX, of identifying the ambiguous regions with DSC-EXACT and AMB-EXACT. Table 6.2 reports the time it took to compute the ambiguous regions as well as the metrics described in Section 6.4.2. The results demonstrates that our methods are comparable or even better in coverage of the ambiguous regions on the test data, while being up to four orders of magnitude faster.

Q2: Accuracy fairness trade-off. We compare our method with the benchmarks described in Section 6.4.2. The results in Table 6.1 demonstrate that:

Existing fairness methods sometime achieves overall fairness at the expense of a significant decrease in accuracy. Additionally, overall fairness is achieved by being biased towards different groups for different types of errors, i.e., ones in the unambiguous vs ambiguous regions. On the other hand, our method is effective in removing unfairness in the ambiguous regions and ignoring the unfairness in the unambiguous regions, as desired. Our method also achieves accuracy similar to the most accurate classifiers.

Q3: Robustness to noisy data. In order to demonstrate the sensitivity of existing fairness methods towards noise, we flipped the ground truth labels of 0.0% to 20% of the datapoints uniformly at random. Figures 6.2 and 6.3 present our findings. We compare an accurate classifier, a fair classifier and our method equalizing FPR using AMB-APPROX. The key takeaways are as follows: In an effort to equalize all errors, existing fairness methods are affected by label noise and end up classifying a significant number of data-

points in the wrong class, as hypothesized in the introduction. In contrast, our method is robust to noise as it identifies similar regions as ambiguous for varying level of noise. Secondly, this experiment also confirms our hypothesis, by showing that the ambiguous region coincide with regions with predominantly high model uncertainty, i.e., the sparse clusters.

6.4.5 Evaluation on Real-World Datasets

In this section, we answer our evaluation questions using two real-world datasets.

Q1: Ambiguous regions coverage and speed. We identify the datapoints with ambiguous decisions using DSC-APPROX, given by Problem P6.3 and AMB-APPROX, given by Problem P6.4, for the same value of ϵ. We also tried DSC-EXACT and AMB-EXACT, however after several hours of computations they still did not yield any results. So, we compare the results of our two proposals, using $\hat{\alpha}$ metric given by Equation 6.6. Takeaways remain similar for $\hat{\delta}$, given by Equation 6.5.

For the Compas data, our method DSC-APPROX and AMB-APPROX categorized 0.12 and 0.5 of the datapoints as having an ambiguous decision, respectively. While for the SQF dataset, 0.12 and 0.53 of the datapoints were identified as having an ambiguous decision by DSC-APPROX and AMB-APPROX, respectively. It is noteworthy that AMB-APPROX identifies more datapoints as ambiguous. This is due to the fact that with AMB-APPROX we train one classifier per training datapoint, i.e., we perform a more exhaustive search for the classifiers that exhibit predictive multiplicity. This process, however, takes a longer time. Hence, there is a trade-off between the speed and effectiveness for both the proposed methods of identifying the ambiguous regions.

Q2: Accuracy fairness trade-off. Similar to the synthetic dataset, we compare our method of equalizing group error rates (FPR and FNR) in the ambiguous regions, identified by DSC-APPROX and AMB-APPROX, with three benchmarks described in Section 6.4.2. The takeaways from results presented in Tables 6.3 and 6.4 are the following.

Existing fair classifiers that focus on equalizing overall error have high unfairness in the ambiguous regions in most cases, which confirms our hypothesis. Although these classifiers achieve fairness in the overall data, they sometimes result in a significant drop in accuracy. Additionally, in many cases, existing fair classifiers achieve overall fairness by being unfair to different groups in the ambiguous vs unambiguous regions.

In comparison, our method that only equalizes errors in the ambiguous regions, in most cases, provides the fairest solution in the ambiguous regions while achieving a comparable accuracy to the most accurate classifier.

Table 6.4: [SQF] Signed differences in FPR/FNR: This table demonstrates effectiveness of our methods. Please refer to Section 6.4.5

	Unfairness			Accuracy
	total	unamb	amb	
Acc.	-0.28/0.12	-0.29/0.13	-0.07/**0.017**	0.75
Fair	0.04/0.02	0.02/0.03	0.07/-0.15	0.65/0.71
Uni-P6.3	-0.28/0.12	-0.29/0.13	-0.05/**0.014**	0.75 / 0.75
Our-P6.3	-0.28/0.11	-0.29/0.13	**-0.02**/**-0.017**	0.75/ 0.75
With P6.4				
Acc.	-0.28/0.12	-0.24/0.17	-0.25/**0.07**	0.75
Fair	0.04/0.02	-0.06/0.12	**0.15**/-0.08	0.65/0.71
Uni-P6.4	-0.27/0.14	-0.24/0.17	-0.25/0.09	0.74/ 0.74
Our-P6.4	-0.24/0.13	-0.24/0.17	-0.18/**0.07**	0.73/ 0.74

In a few cases where our approach is not the only best solution, it provides additional benefits, e.g., in one case our solution is equally fair in the ambiguous region compared to the accurate classifier (cf. Table 6.4). However, our method assigns decisions to datapoints in the ambiguous regions stochastically. So, in practice, most datapoint in the ambiguous region have a non-zero probability to be in the favorable class. This is desirable over a deterministic decision, since there is ambiguity in decisions for these datapoints. In another case, Table 6.3, selecting classifiers uniformly at random is 1.6% more fair on the *test data*. However, our solution is still 90% and 18% better than the benchmark fair classifier and the accurate classifier, which are the current standards.

6.5 Related Work

Modeling uncertainty. Prior works on categorizing uncertainties have proposed to distinguish between aleatoric (irreducible) uncertainty and model (reducible) uncertainty[73, 124, 128]. A lot of works in machine learning have addressed this distinction in different subfields. Depeweg et al. [72] propose to decompose the two types of uncertainties using bayesian neural networks and latent variables. Kendall and Gal [148] consider this distinction in computer vision problems. McAllister [192] distinguish between the types of uncertainties in reinforcement learning problems.

We believe that we are the first ones to propose to distinguish between different types of uncertainties for fairness in predictive tasks.

Predictive multiplicity. In their seminal work, Breiman et al. [37] introduced the concept of the *Rashomon effect* in the context of model explanations. The Rashomon effect refers to the scenario where data admits multiple different models that yield similar accuracy. Breiman et al. [37] argue that one should not use the explanations of a single model to

draw conclusions about the data and the prediction task at hand. Rashomon sets, defined as ϵ-set of models, i.e. those whose empirical training loss is within ϵ-loss of a baseline classifier, are used by Dong and Rudin [78], Fisher et al. [94] to study the problem of variable importance.

The notion of predictive multiplicity in a classification setting was introduced by Marx et al. [191]. They proposed mixed integer programming methods using non-convex loss functions to train classifiers which would yield predictive multiplicity for linear classifiers. We build on this work, and extend it by proposing tractable convex problem formulations which yield fast solutions, and work for both linear and non-linear classifiers.

There is a growing interest in predictive multiplicity due to its societal implications on algorithmic decision-making system. Bhatt et al. [28] look at it from a fairness perspective, and aim to find counterfactual accuracy of a classifier which would give a selected test datapoint favorable outcome. Specifically, they aim to find the minimum decrease in accuarcy, ϵ, that would give an individual a favorable outcome. Pawelczyk et al. [210] provide an upper bound for the costs of finding counterfactual explanations under predictive multiplicity. However, none of these works have made the connections between predictive multiplicity and model uncertainty.

6.6 Conclusion

In this chapter, we proposed that while designing fairness approaches one must account for the uncertainties of the prediction task at hand. Specifically, we argue that only the *errors* arising due to lack of knowledge about the best model or due to lack of data, i.e., the *epistemic errors* should be taken into account while designing fairness methods and errors due to inherent noise should be ignored. Our proposal stands in contrast to the current group fairness approach that aims to equalize 'total' errors. With this goal in mind, we build upon predictive multiplicity techniques to identify the regions with model uncertainty.

In addition, we propose convex and scalable formulations to find classifiers that exhibit predictive multiplicity, which are approximately equally effective compared to their non-convex counterparts, while being up to four orders of magnitude faster. We also propose convex formulations to equalize errors arising due to model uncertainty. Using synthetic and real-world datasets, we demonstrate that our methods are effective and more robust to label noise compared to existing group fairness methods.

Chapter 7

Designing a new fair ADMS: Human decision-makers

In this chapter, we focus on settings where consistency between decision makers might be deemed desirable, and study how the degree of inconsistency of human decisions could be moderated with algorithmic assistance. This has immediate implications for the development of algorithmic assistance to support cooperative work by enabling the distribution of decision-making tasks amongst multiple decision-makers, without sacrificing consistency. Specifically, we ask the following research questions:

Can algorithmic decision aids moderate inconsistency among human decision-makers?

In order to answer our research questions, we leverage prior work in psychology and HCI to develop a set of algorithmic decision aids which may influence the degree of inconsistency of human decisions, and we rigorously experimentally evaluate the effect of these decision aids on human decisions.

The rest of the chapter is organized as follows:

- In Section 7.1, we discuss prior works on machine-assisted decision-making, different notions of consistency, existing heuristic for reducing inconsistency and how people may react on the feedback of their inconsistency.
- In Section 7.2, we describe the details of our experimental set-up including experimental conditions, hypotheses, stimulus material, details about the data collection, design of the decision-aids and format of the advice.
- In Section 7.3, we discuss the results of the confirmatory analysis that we preregistered before performing the study.
- In Section 7.4, we present additional results for different notions of accuracy and consistency. Additionally, we further explore the effect on people's responses after observing the machine advice. Finally, in Section 7.5, we conclude the paper.

Relevant publication

The results presented in this chapter have been published in [10].

7.1 Background

Machine-Assisted Decision-Making. With the increasing popularity of algorithmic decision aids, much research has studied how algorithmic advice shapes human decisions. On a high-level, we can discuss the findings of this research in terms of the *factors* that were identified to influence people's advice taking behavior, and the *measures* used to evaluate the effects of machine advice on people's decisions.

Research has identified a plethora of factors that influence people's advice taking behavior. The likelihood of accepting advice varies based on the advisor's identity, with a preference for human guidance in certain scenarios [42, 76, 77], and a preference for algorithmic advice in other settings [182, 183]. Receptiveness to machine advice depends on the algorithm's errors. People are more likely to take advice from algorithms that are stated or observed to exhibit a higher predictive accuracy [271], and that make errors more similar to typical human errors [112]. People are also more likely to take advice from decision aids that are explainable [53, 215, 264]. On the other hand, strong graphical warnings about algorithmic decision aids may lower the influence of their advice [83]. Advice taking also depends on the specific advice that is provided; for instance, in the legal domain, people are more likely to take advice to grant a defendant bail than advice to deny bail [113].

Prior work has studied how machine advice affects human decisions with respect to a variety of measures. As a first step, most studies measure people's likelihood of taking machine advice, be it in terms of the overall advice taking propensity [271], the propensity to take advice pointing towards a specific decision [113], or the propensity to take (in)correct advice [215]. Some studies have gone beyond measuring people's likelihood of taking advice, and measured the effects of machine advice on the quality of people's decisions. For instance, in settings where one can define the notion of correct predictions and ground truth labels, prior work has measured whether machine advice increases the alignment between people's decisions and ground truth labels—that is, the accuracy of people's decisions [112, 215, 280]. In settings where decisions have important societal implications, prior work has also studied the effects of algorithmic assistance on the fairness of people's decisions [110, 111]. However, little prior work considered the effects of algorithmic assistance on the *consistency* of human decisions.

We contribute to research on machine-assisted decision-making by studying how different algorithmic decision aids impact human decisions with respect to two measures established in prior work—people's propensity to take advice and the accuracy of people's decisions—as well as a novel measure: the consistency between people's decisions.

Notions of Inconsistency. Much prior work has studied the (in)consistency of human decisions. A bulk of research has documented the *inconsistency between decisions of different decision-makers* [139, 140]—that is, a lack of *inter*-annotator consistency—in tasks as diverse as sentencing [15], evaluating job performance [243], estimating real estate prices [3], and reviewing submissions to top-tier CS conferences, such as NeurIPS [26, 68, 167], CSCW [39] and ICLR [246]. We contribute to this line of research by exploring if algorithms can be used to support cooperative work in such settings where consistency is deemed to be desirable. Namely, we propose methods for alleviating the inconsistency between the decisions of different decision-makers for the same set of inputs.

Much research has also studied the consistency of individual decision makers, or *intra*-annotator consistency. Cognitive biases such as dynamic inconsistency and hyperbolic discounting are known to result in the *inconsistency of an individual's decisions across time* [181, 244]. Individual's judgments are found to substantially vary across time in various settings [139]: pathologist's biopsy assessments of the same sample at different points in time were found to exhibit a correlation of only 0.61 [82]; expert's estimates of the amount of time required to complete the same software development task were found to vary by 71% [116].

Prior work has also documented the *inconsistency of an individual's judgments across inputs*. A particularly well-studied aspect of this problem is the inconsistency of people's pairwise preferences. Decades of research in this area have led to the development of numerous methods for identifying, measuring, and reducing the inconsistency of human pairwise preferences [2, 40, 155], as well as a plethora of approaches to the difficult task of learning human pairwise preferences [56, 96, 127], which we consult when developing our decision aids in Section 7.2.4.

Human Heuristics for Reducing Inconsistency. Many decision-making settings require people to make decisions on a case-by-case basis: granting or denying loans, making bail decisions, reviewing papers, etc. However, prior research in psychology has documented that people might find it easier to make *comparative* judgments than absolute ones in various contexts [198, 237]. Pairwise comparisons are also used to assist people with developing and refining their beliefs [169]. Hence, it is not surprising that people often rely on comparative judgments to assist them with making case-by-case decisions.

For instance, the analytical hierarchy process that is widely used to assist with making complex decisions in domains ranging from governance to engineering relies on comparative judgments at its core [95, 250]. Research on crowdsourcing has also proposed eliciting respondents' relative pairwise labels, rather than absolute ones, as a strategy for improving response quality [55, 202, 240].

To illustrate how one may leverage comparative judgments to assist with absolute judgments, let us consider the task of grading papers. After (i) assigning initial grades to a set of papers, one might (ii) compare pairs (or larger subsets) of papers to identify mutually inconsistent decisions, in order to (iii) revise the final grades. In T4 and T5, we develop a decision aid that identifies pairs of decisions that are inconsistent with the majority's comparative valuations, hence providing a tool for automating step (ii) in the above-described procedure.

Reducing Inconsistency with Algorithmic Assistance. Kahneman at al. [139, 140] extensively study the problem of noise in human judgments. In [139], they discuss several approaches to reducing inconsistency in human decisions, proposing interventions of varying strengths. The first and most radical proposal is to replace human decision makers with algorithms. Still, they highlight the need for people to retain ultimate control. Hence, as the second and weaker proposal, they propose the use of algorithmic decision aids to assist human decisions. Depending on the estimated accuracy of human and algorithmic decisions and the normative importance of accuracy, they highlight the possibility of advising against overruling algorithmic predictions. The third and weakest intervention is ensuring that decision-makers use similar procedures to gather and integrate information, and to translate this information into a decision.

The first two proposals rely on algorithmic assistance to reduce human inconsistency. The underlying idea is to use algorithms to *predict correct decisions*, which would steer (second proposal) or even replace (first proposal) human decisions, thereby making them more consistent. In this paper, we study the effects of the second proposal in treatment T2. However, we also study an alternative approach in treatments T4 and T5: using algorithms to *identify inconsistencies* in human decisions, and to help people reduce their inconsistency by themselves.

Reactions to Feedback about Inconsistency. Algorithmic decision aids have proven to be effective in a plethora of settings. Here we review literature in social psychology that may help us form hypotheses about the effectiveness of algorithmic decision aids for the task of reducing inconsistency in human decisions, and guide the design of our decision aids. Specifically, we leverage prior work in social psychology to anticipate how people may react to being provided with feedback about their inconsistency.

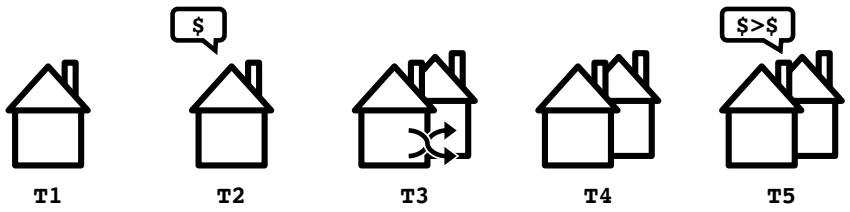


Figure 7.1: Graphical overview of experimental conditions T1-T5. In T1 and T2, respondents review their decisions one-by-one, while in T3-T5 they review decisions in randomly (T3) or meaningfully selected (T4 and T5) pairs. In T2 and T5 respondents are additionally provided with (different kinds of) explicit machine advice.

Kahneman et al. [139] argue that inconsistency is undesirable in a variety of settings. If our respondents share this view, they might perceive feedback about their inconsistency as negative feedback. Prior work in organizational psychology has shown that people may not react positively to negative feedback about their performance. Negative feedback is not perceived as useful, results in negative reactions, and is not associated with a recipient's willingness to change their behavior [236]. It is also found to evoke defensiveness and denial [184]. The main strategy employees use to reduce the impact of such negative feedback is to reject it [129]. To mitigate these effects, we will avoid framing the machine advice as negative feedback, and utilize strategies for softening the blow proposed by Steelman and Rutkowski [236]: providing high quality feedback delivered in a considerate manner.

Kahneman et al. [139] also show that people tend to vastly underestimate the degree of inconsistency in human decision-making. Hence, feedback about inconsistency may conflict with people's beliefs. Much prior work in psychology has found that people resist evidence that is contradictory to their preconceptions [14]. Two psychological concepts that are particularly relevant for predicting how people will react to conflicting information are cognitive dissonance [91, 121] and biased assimilation [185] or disconfirmation bias [81]. Both lines of research point to the same conclusion: due to overestimating their consistency, people may discount or reject the decision aids' feedback about their inconsistency. We attempt to mitigate this effect by familiarizing people with their lack of expertise with the task at hand: at the beginning of the experiment, participants complete a tutorial where they can observe the (in)accuracy of their real-estate price estimates.

7.2 Methodology

7.2.1 Experimental Design

In a large-scale, pre-registered human-subject experiment[9] run on Prolific we studied how different interventions influence people's estimates of real estate prices. The interventions included asking respondents to review their initial estimates one by one (treatments T1 and T2) or as pairwise comparisons (treatments T3, T4 and T5), and providing respondents with different forms of algorithmic assistance (treatments T2, T4 and T5).

Scenario. In this work, we focused on the task of estimating real estate prices. In our experiments, we utilized a dataset of New York City real estate prices introduced by Poursabzi-Sangdeh et al. [215]. The dataset contains information about 393 apartments located on the Upper West Side of New York City, which were listed for sale on the real estate website StreetEasy.com between 2013 and 2015. For each listing, we had access to basic information about the apartment, including the listing price, number of bedrooms, bathrooms and total number of rooms, the apartments' square footage and monthly maintenance fees, the number of days the apartment has been on the market, and the distance from the apartment to the nearest subway and school. We preprocess the data as proposed by Poursabzi-Sangdeh et al. [215]. Specifically, we remove the apartments where the number of bedrooms is greater than the total number of rooms or where the apartment's square footage is less than 200 sqft. With this preprocessing, we were left with 387 apartments. From these, we utilized 30 apartments as stimulus material in the human-subject experiments (Section 7.2.2), and the remaining 357 apartments for training the decision aids (Section 7.2.4).

Experimental Conditions. In each experimental condition, respondents were first asked to complete a tutorial, in order to familiarize themselves with real estate prices in New York City. Next, all respondents were asked to estimate the prices of the same 30 apartments. After gathering the respondents' initial estimates of apartment prices, we asked them to review their estimates in one of five different ways, as described below and summarized in Figure 7.1 and Table 7.1.

In the control condition T1, respondents were asked to perform the simplest revision procedure. They were asked to revise their initial estimates one-by-one.

[9]Prior to conducting the human-subject experiment, we have obtained the approval of <our anonymous Institution's> ethical review board (ERB), and pre-registered our experiment on AsPredicted. The anonymized pre-registration documentation can be found on the following url: ***https://aspredicted.org/D7X_NKL***.

Table 7.1: Overview of the characteristics of the 5 experimental conditions in our study. Reviewing Procedure: Are instances reviewed one-by-one or pairwise? Algorithmic Assistance: Do respondents have access to any form of algorithmic assistance? Data Required: Do the utilized decision aids require any type of labeled data?

	Reviewing Procedure	Algorithmic Assistance	Data Required
T1	one-by-one	none	none
T2	one-by-one	explicit advice	ground truth
T3	pairwise comparisons	none	none
T4	pairwise comparisons	implicit advice	human perceptions
T5	pairwise comparisons	implicit and explicit advice	human perceptions

T1: Respondents were asked to review all of their estimates *one-by-one*, in the same format as they originally made them: 30 apartments, one per page, shown in random order.

T2 was inspired by decision aids that are commonly used in the machine-assisted decision-making literature [113, 271]. It corresponds to the standard machine-assisted decision-making setting in the judge-advisor system (JAS) paradigm [33].

T2: Compared to T1, we manipulated the information provided in the review phase. Respondents were again asked to review all of their initial estimates one-by-one (30 apartments, one per page, shown in random order), but the apartment descriptions were accompanied by *machine advice*. Specifically, respondents were shown the estimates of a linear regression model which we trained to estimate real estate prices using the dataset introduced by Poursabzi-Sangdeh et al. [215], as described in Section 7.2.4.

While T1 and T2 required respondents to review their decisions one-by-one, in T3–T5 decisions were reviewed in pairs. This pairwise revision procedure was motivated by past research in psychology [198, 237] and computer science [202], which found that, in certain contexts, people are better at making comparative judgments than absolute ones.

One of the main difficulties in introducing a reviewing procedure based on pairwise comparisons is selecting which pairs one should review. Our set of 30 apartments results in 30 choose 2 = 435 possible pairwise comparisons. Since it is not feasible to review all of these pairs, one must select which pairs to review. A naive approach, employed in treatment T3, would consist of randomly selecting which pairs to review. An ideal approach would consist of reviewing exactly those pairs that would lead to an improvement in the quality of decisions with respect to a metric of interest. In treatments T4 and T5 we utilize algorithmic assistance to identify pairs of apartments for which people's estimates are not aligned with the majority's estimates.

In T3, respondents were asked to review randomly selected pairs of apartments. This is the simplest of the three treatments that rely on pairwise comparisons, since it does not utilize any algorithmic assistance beyond the random selection of pairs of apartments.

T3: In T3, respondents were asked to review their decisions in a series of 15 *pairwise comparisons of randomly selected pairs* of apartments. Respondents were asked to review 15 pairs of apartments in order to keep the number of decisions reviewed equal to 30 across all treatments. The same apartment may have been shown in multiple pairs. Hence, even though the 15 pairwise comparisons provided respondents with 30 opportunities to update their estimates, this does not imply that they had an opportunity to update their initial estimate for each of the 30 unique apartments.

Due to its simplicity and lack of an algorithmic component, we treat T3 as a secondary baseline rather than an experimental condition. This baseline is particularly useful for studying the effects of T4 and T5, since—unlike the main baseline T1—it utilizes a reviewing procedure based on pairwise comparisons.

T4 builds upon T3 by including the component of algorithmic assistance. Instead of asking respondents to review randomly selected pairs of apartments, the decision aid selects the pairs of apartments that the respondent will review.

T4: While T3 presents respondents with random pairs of apartments, T4 selects *pairs* where respondents' estimates are *not aligned with the majority's view*. Specifically, we implicitly provided machine assistance by asking participants to review pairs of apartments for which their initial price estimates did not align with most people's comparative valuations of those apartments. We trained a model to predict the majority's comparative valuations of apartment prices using a dataset of human-annotated pairwise comparisons of apartments that we gathered, as described in Section 7.2.4.

Treatment T5 builds upon T4 by additionally providing explicit algorithmic advice. The addition of explicit advice makes the decision aid used in T5 more similar to the decision aids that are typically considered in the machine-assisted decision-making literature, such as the decision aid used in T2.

T5: The format of the review phase and the pair selection procedure remained the same as in T4, but we additionally explicitly informed people about the difference between their initial estimates and the predicted comparative valuations of most people.

Hypotheses.

We leverage prior research on machine-assisted decision-making in psychology and HCI (reviewed in Section 7.1) to form hypotheses about the effects of our interventions T2, T4 and T5, compared to the control condition T1.[10]

Prior work has demonstrated that algorithmic decision aids can influence people's decisions [113]. In line with these findings, in H1 we hypothesize that our decision aids will prompt respondents to revise their estimates. Specifically, we hypothesize that compared to the control condition T1, our interventions T2, T4 and T5 will lead to:

H1: A higher number of decisions updated in the review phase.

H1': A higher propensity to update decisions for the particular apartments shown in the review phase.

We include both hypotheses H1 and H1' since they capture different aspects of the interventions' effects on people's propensity to update decisions. H1 focuses on the overall effect of the intervention across all apartments, while H1' captures the effectiveness in prompting respondents to review their estimate for specific apartments that are shown in the review phase. (These are the same for T1 and T2.) The effect captured by H1 may be deemed more important in settings where the goal is to maximize the overall effect across all apartments, while H1' may be more appropriate if we are interested in measuring engagement with the algorithmic assistance.

Furthermore, past research has demonstrated that accurate decision aids help increase the accuracy of people's decisions [112, 271]. In H2, we hypothesize that our decision aids—which exhibit high predictive accuracy[11]—will do the same. Namely, we hypothesize that compared to the control condition T1, our interventions T2, T4 and T5 will result in:

H2: A higher accuracy of post-review decisions.

Finally, we expect the decision aids to lead to an increase in the consistency between people's responses. We hypothesize that compared to T1, the interventions T2, T4 and T5 will lead to:

H3: A higher degree of consistency between the post-review decisions of different respondents.

[10]Since two of our algorithmic interventions (T4 and T5) rely on revising pairs of decisions, we include another control condition T3, in which respondents review pairs of decisions, but without algorithmic assistance. However, given our focus on algorithmic forms of assistance, we do not hypothesize about the effects of T3, and only study its effects exploratively.

[11]More details about the decision aids' performance can be found in Section 7.2.4.

Dependent Variables. For each apartment we measured the respondents' pre-review estimates and post-review estimates. Note that in treatments T3–T5, respondents were not given an opportunity to update their estimates for some apartments. In those cases, we defined their post-review estimate to be equal to their pre-review estimate.

To test our hypotheses, we formed dependent variables based on these measurements as follows:

H1 magnitude: Absolute difference between pre and post-review estimates.

H1′ magnitude: Absolute difference between pre and post-review estimates, limited to only those apartments shown during the review phase.

H1 binary: 0 if the pre and post-review estimates are equal, otherwise 1.

H1′ binary: 0 if the pre and post-review estimates are equal, otherwise 1, limited to only those apartments shown during the review phase.

H2: Difference between pre-review error and post-review error. The pre- and post-review errors are calculated as the absolute difference between the respondent's estimate and the ground truth for a given apartment. Intuitively, this measure captures the treatments' effects on the degree of agreement between respondents' estimates and the true apartment prices.

H3: Difference between pre-review inconsistency and post-review inconsistency. The pre- and post-review inconsistency are calculated as the absolute difference between the respondent's estimate and the average estimate (namely, the mean value of all respondents' estimates) for a given apartment. Intuitively, this measure captures the treatments' effects on the degree of agreement between the estimates of different respondents.

Analysis. In Section 7.3, we report the findings of our confirmatory analyses related to hypotheses H1–H3. In the text, we report the findings of our statistical hypothesis testing, accompanied with plots that illustrate our findings using descriptive statistics. To test our hypotheses we rely on linear mixed models.[12] Due to our repeated measures design, we include crossed random effects to account for differences between participants and

[12]In hypotheses H1 binary and H1′ binary our dependent variables are binary. The choice between using a linear and a logistic regression in such settings has been much debated in prior work, and we refer the readers to the work of Hellevik [123] for an in-depth discussion on this topic. Hence, we replicated our analyses for binary dependent variables using both a linear and a logistic regression. The results are qualitatively the same for both models. In the paper we report the results of the linear regression for ease of interpretation of the coefficients and consistency with other hypotheses, in line with our pre-registration.

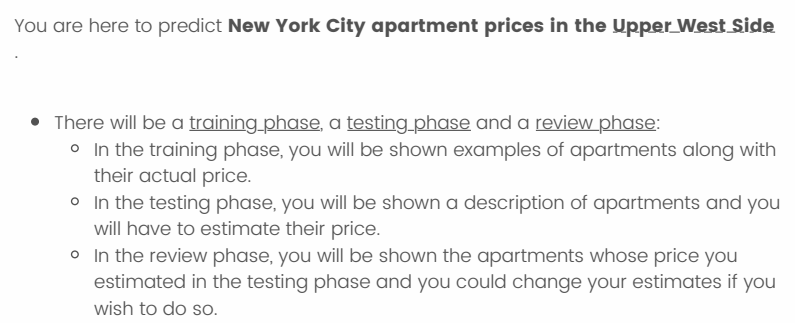


Figure 7.2: Description of the experimental design shown to participants at the beginning of the experiment.

apartments.[13,14,15] The dependent variables vary across hypotheses as described in the previous subsection. In all of the models, the experimental conditions are used as the independent variables. We one-hot encode the five experimental conditions T1–T5 using four binary variables corresponding to T2–T5, and we treat T1 as the reference category. That is, our models' unstandardized regression coefficients capture how the effects of treatments T2–T5 differ from the effects of T1. To compare the effects of other pairs of treatments we utilize Wald tests to test the equality of the corresponding coefficients. For the Wald tests, we report Bonferroni-adjusted p-values to account for the multiple comparisons problem.

In the appendix we conduct an additional exploratory analysis of our data. There we report a series of descriptive statistics, including results related to the effect of the treatments on other measures of accuracy and consistency.

7.2.2 Stimulus Material

Upon opening the study link through the Prolific interface, participants were randomly assigned to one of the five experimental conditions. All participants first completed an

[13]We compared models that include a participant or apartment random effects term to nested models that do not include these random effects terms. To do so, we used likelihood-ratio tests. The likelihood-ratio tests confirmed that including participant and apartment random effects terms makes a difference—i.e., there is a significant amount of variation between participants and apartments accounted for by the random intercepts (p-val <0.001, for all 6 models, and both for participant and apartment random effects terms).

[14]To alleviate convergence issues of models with crossed random effects, we initialize the starting values of the parameters to the estimated parameters of a simpler model—a linear mixed model with a random effects term for participants only.

[15]We additionally replicated all of our analyses using fixed effects models with two-way clustering of standard errors with respect to apartments and participants. To do so, we utilized the `reghdfe` Stata package [67] that implements the estimator described in Correia [66]. We found the results of both approaches to lead to consistent findings across all hypotheses.

Task 1/30

Please consider the profile below and estimate the sale price of the apartment.

Bedrooms	1.0
Bathrooms	1
Square footage	702
Total rooms	3.0
Monthly maintenance fee	$443
Days on the market	28
Subway distance (miles)	0.122
School distance (miles)	0.278

Please provide your response below:

What do you think the apartment was sold for?

(a) Survey instrument used in the tutorial and for gathering respondents' pre-review estimates, across all treatments.

Task: 1/10

Please see your results below:

Bedrooms	1.0
Bathrooms	1
Square footage	550
Total rooms	3
Days on the market	135
Monthly maintenance fee	$442
Subway distance (miles)	0.231
School distance (miles)	0.124
Your estimate	$500,000
Actual price	$824,000

(b) Feedback provided to respondents during the tutorial, in all experimental conditions.

Task 1/30

For the apartment below, the computer program estimated its price to be: **$1,800,000**
You initially estimated its price to be: **$1,300,000**

Please provide your responses again below. If you wish to change your initial response, please feel free to do so.

Bedrooms	2.0
Bathrooms	2
Square footage	1320
Total rooms	4.5
Monthly maintenance fee	$1,330
Days on the market	80
Subway distance (miles)	0.168
School distance (miles)	0.225
Your initial estimate	$1,300,000

Please provide your response below:

What do you think the apartment was sold for?

(c) Survey instrument used for gathering participants' post-review estimates in T2. In T1, the machine prediction was omitted.

Difference 1/15

For the two apartments shown below, our computer program estimated that **Apartment A is less expensive than Apartment B**.

However, you estimated that **Apartment A is equally expensive as Apartment B** .

Please provide your responses again below. If you wish to change your initial response please feel free to do so.

	Apartment A	**Apartment B**
Bedrooms	1.0	2.0
Bathrooms	1	2
Square footage	1240	1160
Total rooms	3	4
Monthly maintenance fee	$1,170	$1,330
Days on the market	119	71
Subway distance (miles)	0.149	0.026
School distance (miles)	0.323	0.323
Your initial estimate	**$1,200,000**	**$1,200,000**

Please provide your response below:

What do you think apartment A was sold for?

Please provide your response below:

What do you think apartment B was sold for?

(d) Survey instrument used for gathering participants' post-review estimates in T5. In T3 and T4, the machine prediction was omitted.

Figure 7.3: Stimulus material.

online consent form and entered their Prolific worker ID. Next, participants were shown an introductory text describing the task (Figure 7.2).

Following the approach of Poursabzi-Sangdeh et al. [215], participants were asked to complete a tutorial in order to familiarize themselves with real estate prices in New York City. The tutorial consisted of the same ten apartments that were utilized by Poursabzi-Sangdeh et al. [215]. The ten apartments were shown in random order, and for each apartment respondents were first asked to estimate its price based on its brief description (Figure 7.3a), and were then informed about the apartment's actual listing price (Figure 7.3b).

Next, we gathered the first part of our experimental data—the respondents' pre-review price estimates. We asked all participants to estimate the prices of the same 30 apartments. The 30 apartments were selected uniformly at random from the 387 apartments in the dataset, excluding the 10 apartments utilized in the tutorial. The set of apartments was kept constant across all experimental conditions and respondents. Throughout the experiment, in all five treatments, the apartments were shown in random order to avoid order bias [117, 223]. The phrasing and format of the questions and response options were identical to the tutorial (Figure 7.3a), except that we did not provide respondents with information about the apartments' true listing price after they reported their estimates.

The respondents were then asked to respond to one simple instructed response item, which served as an attention-check question. Specifically, respondents were asked to "Please respond to this question by selecting Somewhat disagree as the answer", using a 5-point Likert scale as the response options. Similar instructed response items are commonly used for quality assurance purposes in online surveys, as a means of identifying inattentive or careless respondents [195].

Next, we gathered the second part of our experimental data—the respondents' post-review price estimates. Respondents were presented with a text describing the experimental condition they were assigned to, i.e., they were informed about the procedure they will follow in the review phase. Participants were asked to review their initial responses in one of five different ways, based on the experimental condition they were randomly assigned to. The experimental conditions T1–T5 are described in Section 7.2.1 "Experimental Conditions" and depicted in Figures 7.3c and 7.3d. All respondents reviewed 30 of their estimates. However, depending on the treatment, respondents reviewed all of their initial estimates one by one (T1 and T2) or a subset of their initial estimates in a series of pairwise comparisons with possible repetition of the apartments, up to 3 times, in multiple pairs (T3, T4 and T5).

Demographic Attribute	Sample	Census
<35 years	56.4%	46%
35–54 years	32.6%	26%
55+ years	10.8%	28%
Male	50%	49%
Asian	10.3%	6%
Black	10.5%	12%
Hispanic	-	18%
Mixed	6.1%	-
White	68.5%	61%
Other	4.6%	4%

Table 7.2: Demographics of our study sample, compared to the 2019 U.S. Census [249].

Finally, we gathered participants' feedback about their experience of participating in the experiment. Namely, we asked respondents to tell us how much they agree with the following statements on a 5-point Likert scale from "Strongly agree" to "Strongly disagree": (i) The study was interesting, (ii) I would like to take part in a similar study in the future, (iii) The questions were easy to understand, (iv) The study was too long. At the end of the study the respondents also had the option to provide additional comments that they wanted to share with the researchers.

7.2.3 Data Collection

We recruited participants from Prolific—an online crowdsourcing platform which caters to scientific researchers [205]. Using Prolific's built-in pre-screening capabilities, we targeted respondents who: (i) are located in the US, (ii) have participated in at least 10 Prolific studies in the past, and (iii) have an approval rate of at least 95% on these past studies. We additionally utilized Prolific's option to provide a sample of respondents that is balanced with respect to gender, due to the current gender imbalance on the platform [51].

Our goal was to recruit sufficiently many participants to detect medium-sized effects (Cohen's $d = 0.5$ [16]) at the significance level of $\alpha = 0.05$ with power $\beta = 0.95$. Using the statistical software G*Power [86, 87], we calculated that a conservative Wilcoxon-Mann-Whitney two-tailed test requires 110 respondents per treatment group to detect effects of the size, significance level and power of interest. In our study, we have five experimental conditions, leading us to a minimum sample size of 550 respondents. To account for

[16] In Section C.1 we report the values of Cohen's d calculated on the gathered dataset.

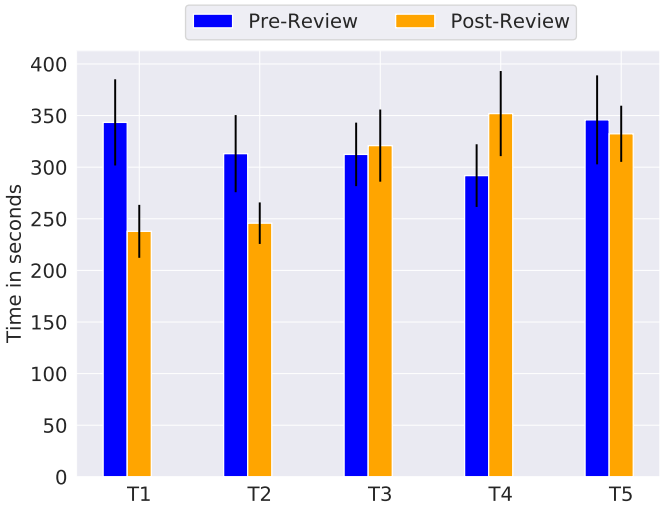


Figure 7.4: Average duration of the experiment, per experimental condition, and per experimental phase. The experimental conditions T1–T5 are shown on the x-axis. The values for the pre-review experimental phase are shown in blue, while the post-review values are shown in orange. We report mean values calculated across respondents ± 1.96 standard errors of the mean (SEM).

possible exclusions, incomplete or missing responses, we increased this estimate by 20% to 660 respondents.

We recruited a total of 660 participants from Prolific, over the course of several days (30th November – 3rd December 2022) in order to minimize sampling bias that could occur due to the day in the week or the time of day [50]. Participants were paid GBP 3.1 for taking part in the study. On average, participants were paid GBP 12.03 per hour, i.e., approximately USD $14.80 per hour—well above the federal minimum wage of USD $7.25. The median study completion time was 15 minutes and 28 seconds. The duration of the experiment varied across treatments, as depicted in Figure 7.4. As expected, there were no statistically significant differences between the average time taken to complete the pre-review phase of different experimental conditions. However, we observe significant differences across treatments in the review phase. Specifically, the review phase in T1 and T2—where respondents reviewed decisions one-by-one—took significantly less time than in T3–T5, where respondents reviewed pairs of decisions. When comparing the duration of the pre-review and review phase, T1 and T2 led to a significant increase in speed. On the other hand, the review phase in T4, where meaningfully selected pairs were presented without explicit advice, took more time compared to the pre-review phase. In T3 and T5 both the pre-review phase and the review phase took a similar amount of time to complete.

We report the demographics of our sample in Table 7.2. Since none of our hypotheses rely on demographic data, we did not ask our respondents to complete a demographics survey, in order to minimize the duration of our experiment, and to align with the data minimization principle. Hence, we report the data about our participants that we had access to through the crowdsourcing platform Prolific. Please note that this demographic data was self reported by Prolific crowdworkers directly to Prolific. Compared to the US census, our sample is younger, in line with typical samples recruited via online crowdsourcing platforms [126, 207, 228]. In line with the gender balancing pre-screening criteria employed during sampling, our sample is balanced with respect to gender. In terms of ethnicity, we are not able to directly compare our sample to the US census, since Prolific's simplified ethnicity prompt did not offer "Hispanic" as a response option. However, we note that Asian respondents are slightly over-represented and Black respondents are slightly underrepresented compared to the US census data.

Upon completing the study, participants were asked to provide feedback about their experience of taking part in this study. Most was positive. On a 5-point Likert scale from "Strongly agree" (coded as 5) to "Strongly disagree" (coded as 1), participants agreed with the statements "The study was interesting" ($\mu = 4.1 \pm 1.0$), "I would like to take part in a similar study in the future" ($\mu = 4.4 \pm 0.9$), and "The questions were easy to understand" ($\mu = 4.6 \pm 0.8$), while they neither agreed nor disagreed with the statement "The study was too long" ($\mu = 2.7 \pm 1.1$).

For the purposes of our analyses, we excluded all responses from participants who did not complete the full study (i.e., missing or incomplete responses), or who failed the instructed response attention check questions. A total of 17 respondents (2.6%) failed the attention check, leaving us with a final sample of 643 respondents.

7.2.4 Decision Aids

7.2.4.1 Developing the Decision Aid Utilized in T2

For T2, we developed a typical example of a decision aid within the judge-advisor system (JAS) paradigm [33]. It is trained to accurately estimate real estate prices. The algorithm's accurate advice can then help steer people towards making accurate estimates of real estate prices. That is, this decision aid was designed to be aligned with our hypothesis H2.

While it was not explicitly designed to increase people's consistency (H3), we still expect that it will succeed in doing so. If the decision aid successfully steers people towards its advice, it will trivially lead to an increase in inter-respondent consistency.

Training Procedure. We trained a linear regression model that used the apartment's attributes as independent variables (full list of features shown in Figure 7.3a), and the apartment price as the dependent variable. We normalized the independent variables to have zero mean and unit variance. We considered models with L1 (lasso) and L2 (ridge) regularization and without regularizers, and picked the regularization hyperparameter values which resulted in the highest coefficient of determination (R^2) on a 20% held out validation set.

Evaluation. The model without any regularizer yielded the highest R^2 value: 85.86 on a test set comprised of 30% of the data. The mean absolute error of this model was \$143,200. Amongst all of the features used in the final model, "Square footage" exhibited the strongest positive correlation with apartment price (with a weight of 1.08), while "Total rooms" exhibited the strongest negative correlation with price (with a weight of -0.3).

Format of Machine Advice. In T2, we provided this model's estimates rounded to the nearest \$100,000 as machine advice, for ease of interpretation.

7.2.4.2 Developing the Decision Aid Utilized in T4 and T5

The decision aid developed for T4 and T5 differs from the decision aids typically studied in the machine-assisted decision-making literature. Instead of predicting apartments' true prices, it is trained to predict people's comparative valuations of apartments. The algorithm's advice can then help steer people towards making estimates that are consistent with other people's estimates. That is, this decision aid was designed to be aligned with our hypothesis H3.

Despite being trained only to predict human comparative valuations of apartments, the wisdom of the crowd enabled this decision aid to accurately predict the true comparative valuations of apartments as well. Therefore, we expect it to also succeed in increasing the accuracy of people's responses (H2).

Data Gathering. In order to build this tool, we gathered a dataset of human comparative valuations of apartments. We randomly selected 1000 unique pairs of apartments from our dataset and split them into 40 batches of 25 pairs. We recruited a total of 850 Prolific workers, who were randomly assigned to one of the batches. After excluding respondents who failed the attention-check question, we were left with 806 participants. From these, we excluded the last 6 responses so that each batch was labelled by exactly 20 participants. We gathered the data over several days (8th November – 11th November 2022) to minimize any bias caused by the time at which the data was gathered [50]. The

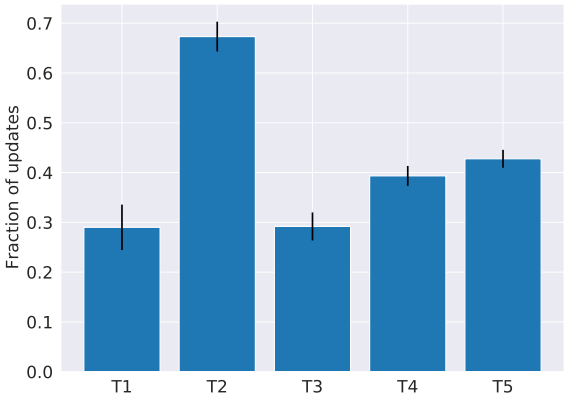
participants were paid GBP 2 for taking part in the study, resulting in an average hourly rate of approximately USD $13.50.

The stimulus material and the experimental procedure were similar to the ones used in the main experiment, described in 7.2.2. The participants completed a consent form and entered their Prolific worker IDs, prior to observing an introductory text similar to the one shown in Figure 7.2. The participants then completed the same tutorial as in the main experiment, in which they were asked to estimate the prices of 10 apartments prior to observing their actual listing price, as shown in Figure 7.3b. Finally, respondents were asked to compare pairs of apartments, which were presented as shown in Figure 7.3d. Specifically, they were asked to estimate if "Apartment A" or "Apartment B" were more expensive. Additionally, they were asked how confident they were in their estimate on a 5 point Likert scale, ranging from "Completely guessing" to "Completely confident." To avoid order bias [117, 223], both the order of the 30 pairs of apartments and the order of apartments within a given pair were randomized.

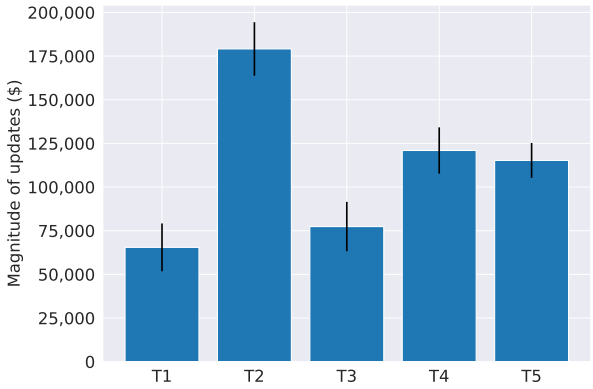
Training Procedure. Using this data, we trained a cross validated logistic regression classifier with L2 regularization to predict which apartment is perceived as more expensive in a given pair. To form the independent variables for our classifier, we subtracted the features of pairs of apartments (shown in Figure 7.3a) from one another. We then normalized them to have zero mean and unit variance. As the dependent variable we used the confidence weighted majority votes of the participant's responses. E.g., if a participant was "Completely guessing" their vote would count as $\frac{1}{5}$ and if they were "Completely confident" it would count as 1.

Evaluation. In the trained classifier "Square footage" was the most important feature, i.e., it had the largest absolute weight (36.18). The second most important feature was "Maintenance cost" (18.3). Our classifier predicted people's comparative valuations of apartments with an accuracy of 98.4%, cross-validated on five randomly chosen 30% test sets.

While this classifier was trained to predict *people's* comparative valuations of apartments, it also exhibited a high degree of accuracy in predicting apartments' *true* comparative valuations. Namely, the accuracy of predicting the pairwise order of apartments with respect to the ground truth prices was 88.5%. In practice, this tool would have an even higher accuracy since we prioritized giving advice for pairs where the tool had high confidence. Such pairs demonstrated a higher accuracy compared to those pairs where the decision aid had low confidence. The high accuracy of this tool led us to hypothesize that the decision aid used in Treatments T4 and T5 can increase not only the consistency (H3) of people's responses, but also the accuracy (H2) of the participant's estimates, despite being *trained only on human annotations instead of ground truth labels*.



(a) H1 binary: Fraction of decisions updated. The y-axis shows the fraction of the 30 initial decisions that were updated in the review phase.



(b) H1 magnitude: Magnitude of updates. The y-axis shows by how much were the 30 initial decisions were updated in the review phase.

Figure 7.5: H1: Effect of the interventions on people's propensity to update decisions, across all 30 apartments. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents $\pm$ 1.96 standard errors of the mean (SEM).

Format of Machine Advice. In order to provide assistance to participants in T4 and T5, we used the aforementioned classifier. We utilized this decision aid to identify pairs of apartments for which people's estimates did not align with the majority's comparative valuation. We prioritized giving advice for pairs where (i) our classifier was able to accurately predict a typical person's comparative valuation, and (ii) the predicted ordering did not match the respondent's ordering.

First, we converted the 30 apartments used in the main study into 435 pairs by taking all possible combinations, and predicted which of the apartments in each pair would be perceived as more expensive by most people. Then we ordered the pairs in a decreasing order with respect to the classifier's confidence. In T4 and T5 we iterated through this list, and asked participants to review their initial decisions which did not align with our classifier's predictions. E.g., if they initially estimated that Apartment A cost $\$600,000$ and that Apartment B cost $\$900,000$, while the classifier predicted that most people would perceive Apartment A as more expensive than Apartment B, participants could have been asked to review this pair of apartments. Participants were asked to review 15 pairs of apartments from this list. A single apartment was limited to appear in at most 3 pairs, to avoid negative reactions from repeatedly being presented with the same information. While selecting the pairs to show respondents, we took into account the decisions they may have updated during the review process.

Table 7.3: Linear mixed models with crossed random effects for participants and apartments. The dependent variables for different hypotheses are described in Section 7.2.1. In all six models, the four independent variables T2–T5 correspond to a one-hot encoding of the five experimental conditions T1–T5, and T1 is treated as the reference category. I.e., intuitively, the row "Cons." shows the estimated value of the constant term (or intercept) that corresponds to the effects of treatment T1, while the rows T2–T5 show how the effects of these treatments differ compared to T1. Hence, to reason about the effects of T2–T5, one needs to sum up the values of the constant term and the treatment of interest. N denotes the number of data points used to fit a specific model. Each of our 643 respondents answered questions about 30 apartments, resulting in a total of 19290 data points. Please note that in H1' some of the data points are discarded, as described in Sections 7.2.1 and 7.3.2. Standard errors are shown in parentheses. Statistical significance of coefficients is indicated as follows: * p < 0.05, ** p < 0.01, *** p < 0.001.

	H1: change, bin., overall	**H1'**: change, bin., specific	**H1**: change, mag., overall	**H1'**: change, mag., specific	**H2**: accuracy	**H3**: consistency
T2	0.383^{***}	0.383^{***}	113634.1^{***}	113634.1^{***}	97299.5^{***}	119916.8^{***}
	(0.0216)	(0.0260)	(9608.2)	(13339.4)	(6022.1)	(6840.6)
T3	0.00179	0.157^{***}	11897.4	52257.4^{***}	-7141.5	-10044.1
	(0.0215)	(0.0264)	(9589.6)	(13489.2)	(6010.5)	(6827.3)
T4	0.103^{***}	0.395^{***}	55471.8^{***}	140695.3^{***}	25568.7^{***}	24214.8^{***}
	(0.0218)	(0.0269)	(9685.0)	(13713.6)	(6070.3)	(6895.3)
T5	0.138^{***}	0.449^{***}	49783.9^{***}	130046.7^{***}	20490.2^{***}	24042.4^{***}
	(0.0216)	(0.0266)	(9608.2)	(13596.1)	(6022.1)	(6840.6)
Cons.	0.290^{***}	0.290^{***}	65461.5^{***}	65461.5^{***}	7563.6	2271.4
	(0.0212)	(0.0216)	(11207.0)	(13277.9)	(6679.6)	(5555.7)
N	19290	14659	19290	14659	19290	19290

7.3 Confirmatory Results

In this section, we present the results of our confirmatory analysis. We compare the baseline reviewing procedure T1 to our interventions T2, T4 and T5, in terms of their effect on people's propensity to *update* their initial estimates (H1 and H1'), and the *accuracy* (H2) and *consistency* (H3) of people's estimates.

7.3.1 H1: Overall Change in Decisions

In all five experimental conditions, we observe that people update some of their 30 initial decisions in the review phase. However, the number of decisions that are updated and the magnitude of these updates varies substantially between the experimental conditions. Compared to the control condition T1 our interventions T2, T4 and T5 lead to a higher propensity to update decisions in the review phase. That is, **our results support H1**.

This holds both in terms of the number of decisions that were updated (H1 binary) and the magnitude of the change (H1 magnitude).

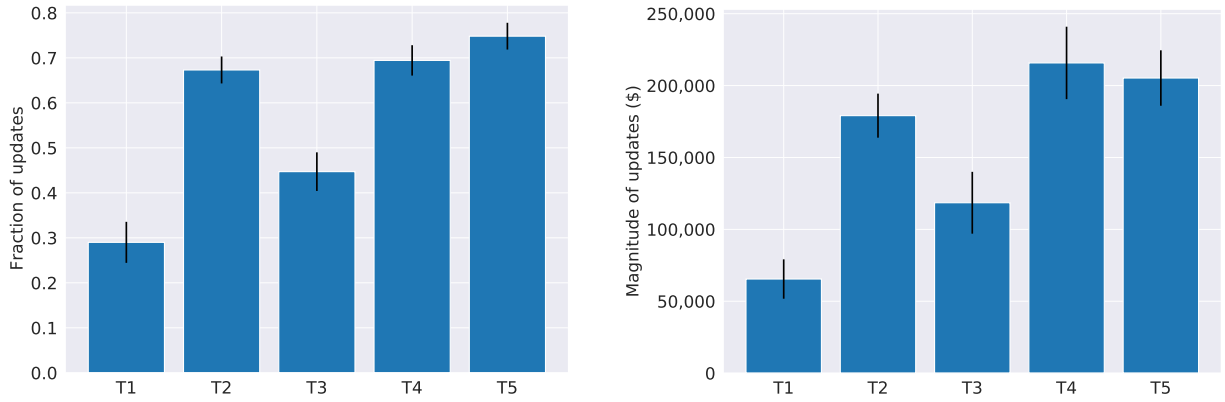
H1 binary: Number of Decisions Updated. Descriptively, we find that the fraction of decisions that are updated varies between treatments (first column of Table 7.3 and Figure 7.5a). In both T1 and T3 people update approximately 29% of their decisions, i.e., they update the estimated prices of 8.7 out of 30 apartments on average. In T4 and T5 people update a larger fraction of their decisions than in T1 and T3—close to 39% (11.8/30 apartments) and 43% (12.8/30 apartments) respectively. The treatment T2 has proven to be the most effective in prompting people to update their decisions, with approximately 67% of decisions (20.2/30 apartments) being updated.

These descriptive observations are corroborated by our statistical analyses. The regression in the first column of Table 7.3 shows that all five treatments significantly influence human decisions. T2, T4 and T5 are significantly more effective than T1, while the effect of T3 was not significantly different than that of T1. Subsequent Wald tests performed on the estimated model confirmed that T4 and T5 are also more effective than T3 ($p < 0.001$), but did not identify a significant difference between the effects of T4 and T5 ($p = 0.35$). Finally, T2 was shown to be significantly more effective than all of the other treatments ($p < 0.001$).

That is, we find that people are more likely to update their decisions when reviewing meaningfully selected pairs of apartments and when machine advice is provided.

H1 magnitude: Magnitude of Updates. Our findings related to the magnitude of the changes are aligned with the findings about the number of decisions updated (third column of Table 7.3 and Figure 7.5b). In T1, people update their decisions by approximately $\$65,461$ on average. In T3, the average update is close to $\$77,359$. The effect of both T1 and T3 is significantly different than zero ($p < 0.001$), and the difference between these two treatments is not statistically significant ($p = 0.214$). In T4, the average magnitude of the change was close to $\$120,933$. This is a significant increase compared to both T1 and T3 ($p < 0.001$). In T5, people updated their decisions by $\$115,245$ on average, which is significantly more than T1 and T3 ($p < 0.001$), but not significantly different than T4 ($p = 1$). Finally, people changed their decisions by close to $\$179,096$ in T2. The magnitude of this change is significantly larger than in any of the remaining treatments ($p < 0.001$).

In short, we find that people update their decisions by a larger amount when they review them as a series of meaningfully chosen pairwise comparisons and when they observe machine advice.



(a) H1′ binary: Fraction of decisions updated. The y-axis shows the fraction of the decisions shown in the review phase that were updated.

(b) H1′ magnitude: Magnitude of updates. The y-axis shows by how much the decisions shown in the review phase were updated.

Figure 7.6: H1′: Effect of the interventions on people's propensity to update decisions, across the subset of apartments that were shown in the review phase. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents $\pm$ 1.96 standard errors of the mean (SEM).

7.3.2 H1′: Propensity to Change Particular Decisions

H1 considers the overall effect of our interventions across all 30 apartments. However, in T3–T5 participants were able to update only a subset of their initial decisions. In H1′ we account for this and focus on the effect of our interventions across the apartments shown in the review phase.

In all five experimental conditions, respondents updated some of the decisions they were shown in the review phase. As in H1, the number of decisions that were updated and the magnitude of the updates varied significantly between treatments. When compared to the baseline treatment T1, our interventions T2, T4 and T5 result in a higher propensity to update decisions for the particular apartments shown in the review phase. I.e., **our findings support H1′**. Again, this holds both for the number of decisions that were updated (H1′ binary) and the magnitude of change (H1′ magnitude).

H1′ binary: Number of Decisions Updated. The second column of Table 7.3 and Figure 7.6a provide information about the fraction of apartments shown in the review phase that respondents updated. For treatments T1 and T2 the results are identical to those related to H1, since all 30 apartments were shown in the review phase. Namely, in T1 respondents updated 29% of their decisions (8.7/30 apartments), while they updated 67% of their decisions (20.2/30 apartments) in T2. For T3–T5, results change substantially once we account for the fact that respondents could not update all 30 apartments in the review phase. While T3 was not significantly different than T1 in H1, in H1′ we

identified a significant difference between these two treatments. Namely, respondents updated 44% of the decisions they were shown in the review phase in T3. The effects of T4 and T5 were even stronger, with respondents updating 69% and 74% of decisions they had access to in the review phase.

Our statistical analyses indicate that all treatments were significantly more effective than the baseline treatment T1 ($p < 0.001$). Treatments T4 and T5 were not significantly different from each other ($p = 0.1497$), but they were both more effective than T3 ($p < 0.001$). Unlike in H1, T2 was not the most effective treatment. While it was significantly more effective than T1 and T3 ($p < 0.001$), it was not significantly different from T4, and it was less effective than T5 ($p = 0.0421$).

On a high level, we found that people are more likely to update their decisions when asked to review them as a series of pairwise comparisons and when they are provided with machine advice.

H1' magnitude: Magnitude of Updates. The results of our analysis about the magnitude of changes are in line with our results about the amount of decisions that were updated (fourth column of Table 7.3 and Figure 7.6b). For T1 and T2, the results are the same as in H1: people update their decisions by approximately $\$65,461$ in T1 and by $\$179,096$ in T2. In T3–T5 the magnitude of the updates is significantly higher than in H1, with an average update close to $\$117,719$ in T3, $\$206,157$ in T4 and $\$195,508$ in T5.

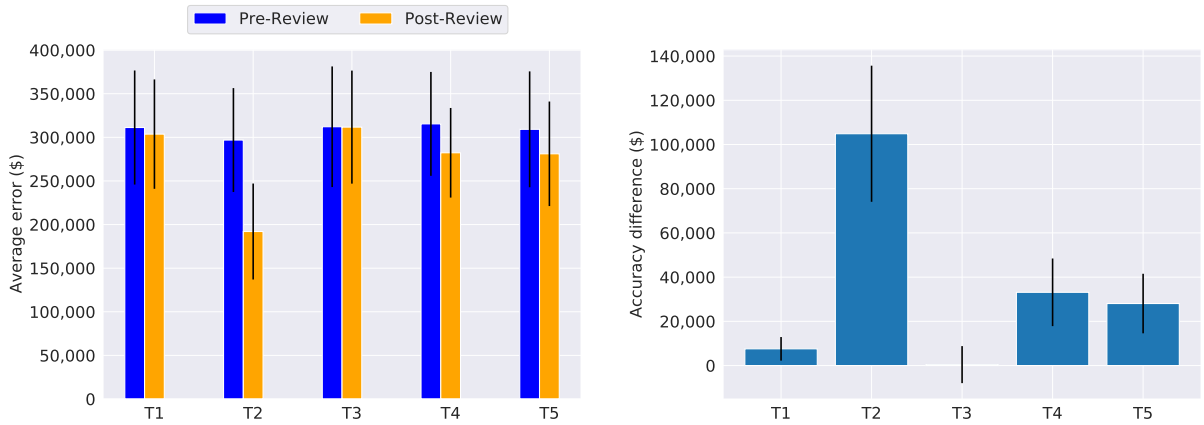
T1 is significantly less effective than the remaining four treatments ($p < 0.001$), and T3 is in turn significantly less effective than the remaining three treatments ($p < 0.001$), which are not significantly different between each other.

In other words, respondents updated their decisions by a larger amount when reviewing them as a series of pairwise comparisons and when they had access to machine advice.

7.3.3 H2: Accuracy of Respondents' Decisions

Next, we study the impact of our interventions on the quality of the decisions—the accuracy of people's estimates. We find that interventions T2, T4 and T5 significantly improve the accuracy of people's post-review decisions, compared to the baseline treatment T1. That is, **our results are in line with H2**.

In H1 and H1' we found that all five of our experimental conditions influenced people's decisions. However, not all of the reviewing procedures led to an increase in the accuracy of people's decisions. As shown in the fifth column of Table 7.3 and in line with Figure 7.7, the reviewing procedure utilized in T1 and T3 did not lead to a significant increase in the accuracy of people's post-review estimates, compared to their



(a) The y-axis shows the average error in people's decisions. The error in pre-review decisions is shown in blue, while the post-review error is shown in orange.

(b) The y-axis shows the difference between the error in people's pre-review and post-review decisions.

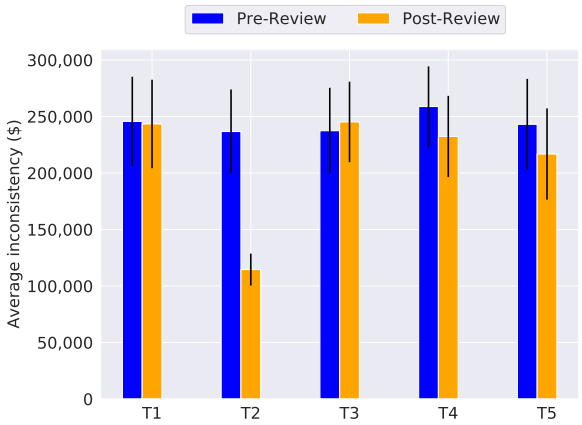
Figure 7.7: H2: Effect of the interventions on the accuracy of respondents' decisions. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents $\pm$ 1.96 standard errors of the mean (SEM).

initial estimates. However, the remaining treatments did have a significant positive effect. Both T4 and T5 led to an increase in accuracy that is significantly higher than the one observed in T1 and T3 ($p < 0.001$). In T4 people's estimates of apartment prices improved by an average of $\$33,132$, and in T5 by $\$28,054$. The difference between T4 and T5 was not significant ($p = 1$). T2 led to a significantly higher improvement in accuracy ($p < 0.001$) than the remaining treatments—people's post review estimates were closer to the ground truth by an average of $\$104,863$, compared to their initial estimates.

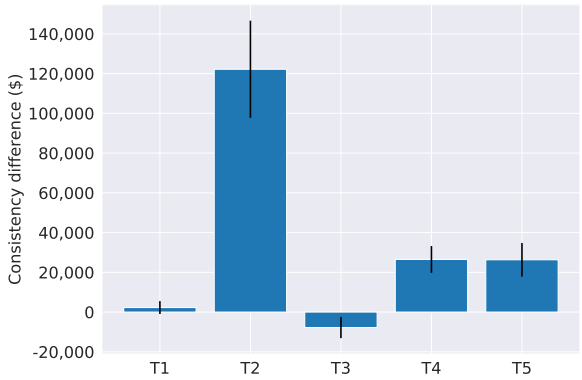
That is, while all treatments influenced people's decisions, not all of them led to an improvement in the accuracy of people's decisions. Only reviewing meaningfully selected pairs of apartments and having access to machine advice increased the accuracy of people's estimates.

7.3.4 H3: Consistency Between Respondents' Decisions

In this Section, we investigate the effects of our interventions on the consistency between the decisions of different respondents. The patterns we identify are qualitatively similar to those related to the accuracy of people's decisions (H2). Namely, our interventions T2, T4 and T5 lead to a significantly higher increase in consistency between the post-review decisions of different respondents, compared to the control condition T1. That is, **the results support H3**.



(a) The y-axis shows the average inconsistency in people's decisions. The inconsistency in pre-review decisions is shown in blue, while the post-review inconsistency is shown in orange.



(b) The y-axis shows the difference between the inconsistency in people's pre-review and post-review decisions.

Figure 7.8: H3: Effect of the interventions on the consistency between respondents' decisions. The experimental conditions T1–T5 are shown on the x-axis. We report mean values calculated across respondents $\pm$ 1.96 standard errors of the mean (SEM).

As shown in the sixth column of Table 7.3 and in line with Figure 7.8, treatments T1 and T3 do not lead to an increase in people's consistency, while T2, T4 and T5 do. Compared to the consistency between respondents' pre-review estimates, T4 and T5 increase respondents' post-review consistency by $\$26,486$ and $\$26,314$ respectively. The difference between T4 and T5 is not significant ($p = 1$), and the increase observed in both of these treatments is significantly higher than the effects observed T1 and T3 ($p < 0.001$). Treatment T2 increases the degree of consistency in people's post-review decisions by an average of $\$122,188$, and this effect is significantly higher than the ones observed in any of the other treatments ($p < 0.001$).

On a high-level, we found that while all treatments influenced the respondents' decisions, some of them did not have an impact on the degree of consistency between the estimates of different respondents. Comparisons of meaningfully selected pairs of apartments and access to machine advice have yet again proven to be effective strategies.

7.4 Exploratory Results

In the previous section, we present the results of our pre-registered confirmatory statistical analyses. In this section we present the results of additional exploratory analyses.

In Section 7.4.1 and 7.4.2 we present additional results related to H2 and H3 respectively. Namely, in Section 7.4.1 we explore different measures of agreement between people's responses and ground truth labels, while in Section 7.4.2 we explore various

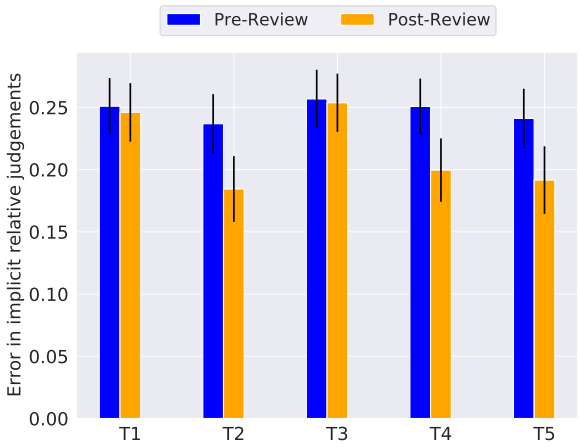


Figure 7.9: Error in people's implicit relative judgments. The y-axis shows the fraction of instances where people's implicit relative ordering of apartments (>,< or =) did not match the ground truth ordering based on the listing price. We report mean values calculated across all respondents and pairs of apartments $\pm$ 1.96 standard errors of the mean (SEM).

measures of agreement between different people's responses. Finally, in 7.4.3 we explore the agreement between people's responses and machine advice.

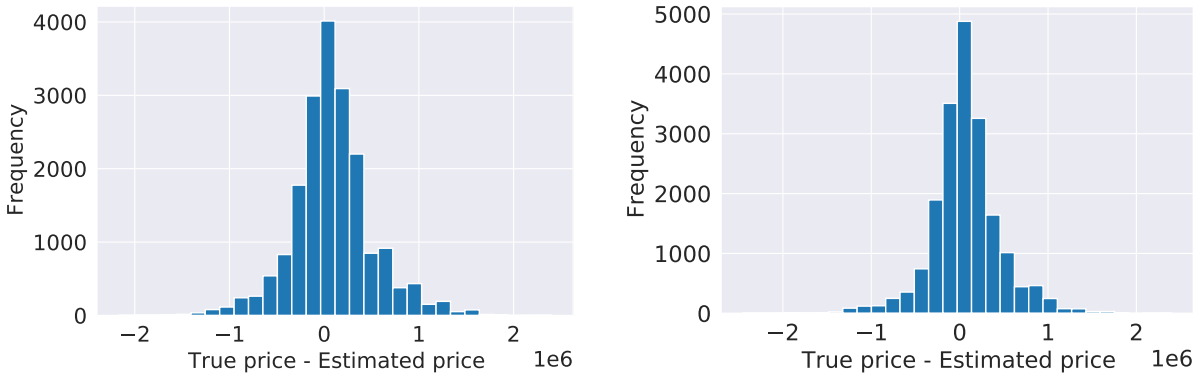
7.4.1 Other Measures of Accuracy

In Section 7.3, we studied the effect of our interventions on the accuracy of peoples' estimates of apartment prices. In this section, we go beyond the accuracy of people's absolute judgments about apartment prices, and consider the accuracy of their implicit relative judgments. We further explore the directionality of people's errors, by investigating whether people tend to overestimate or underestimate apartment prices.

7.4.1.1 Accuracy of People's Implicit Relative Judgments

We start by deriving people's implicit relative judgments from their absolute estimates. For each pair of apartments (A,B), we check if a respondent estimated Apartment A to be more expensive (>), less expensive (<) or equally as expensive (=) as Apartment B. Then we compare these implicit relative judgments with the ground truth (i.e., with the relative ordering of apartments based on their listing price).

In Figure 7.9, we report the fraction of instances where people's implicit relative ordering differed from the apartments' true ordering. Descriptively, we observe that the error in people's pre-review relative estimates is similar across all experimental conditions. In experimental conditions T1 and T3, the error in people's pre- and post-review estimates remained similar. However, in T2, T4 and T5, the error decreased by 5.2, 5.1 and



(a) Distribution of errors in respondents' pre-review estimates.

(b) Distribution of errors in respondents' post-review estimates.

Figure 7.10: Distribution of errors in respondents' estimates, across all treatments. The x-axis shows the magnitude of errors, i.e., the difference between the apartments' true prices and the respondents' estimates. The y-axis shows the number of responses in our dataset that exhibited a certain magnitude of error.

5.0 percentage points respectively. That is, treatments T2, T4 and T5 reduced the error in people's implicit relative judgments by 22.1%, 20.4% and 20.5% respectively. These exploratory findings are in line with our findings related to people's absolute judgments.

7.4.1.2 Directionality of Errors

In this section, we take a closer look at the directionality of people's errors. Namely, we explore whether respondents tend to systematically overestimate or underestimate apartment prices.

In our experiments, we utilized a dataset of apartments introduced by Poursabzi-Sangdeh et al. [215]. The dataset contains information about apartments located in New York City, that were listed for sale between 2013 and 2015. Since the prices of real-estate have increased between the mid-2010s and today, it is possible that people systematically overestimate the prices of these apartments. On the other hand, since we utilize a dataset of apartments located in New York City, which is significantly more expensive than other US cities, it is possible that people systematically underestimate the prices of these apartments. In this section, we investigate if people's errors exhibit either of these patterns. We note that even if people exhibited such a systematic bias in their errors, this would not affect the validity of our results. Any systematic overestimation or underestimation related to the apartments used as stimulus material would be the same across treatments, since they utilize the same 30 apartments. Since our hypotheses concern the differences between experimental conditions, this would not impact our results.

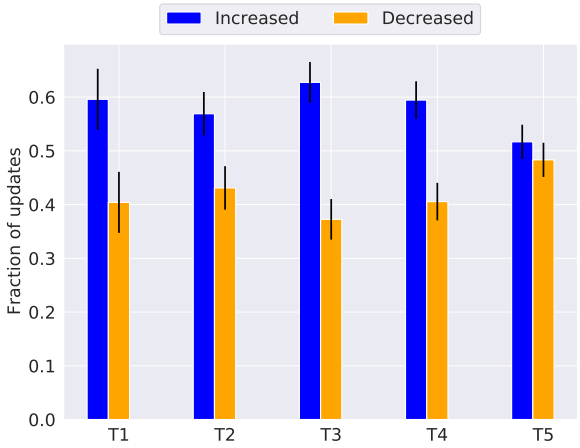
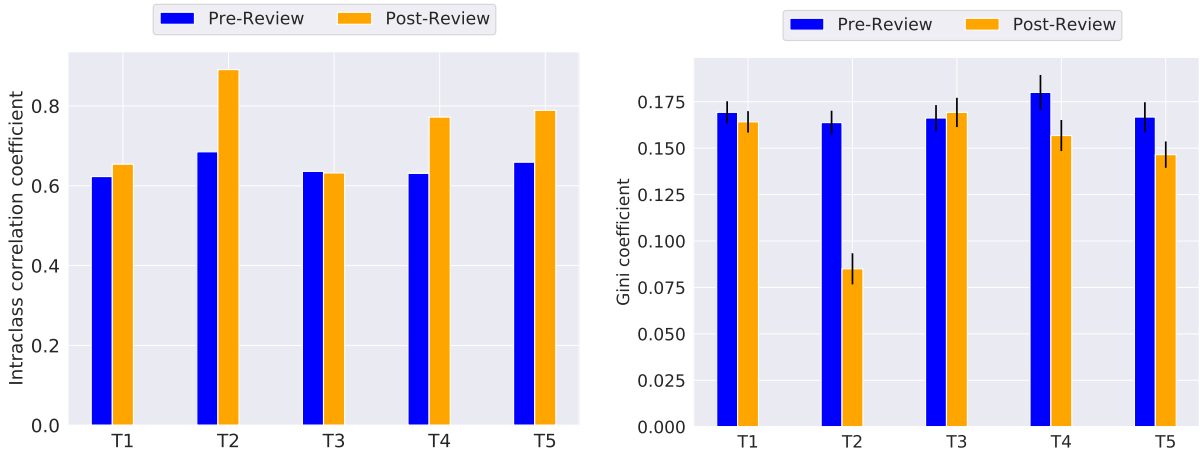


Figure 7.11: Directionality of response updates. The y-axis shows the fraction of revised responses that were updated to increase (blue) or to decrease (orange) the initial price estimates, for each of the experimental conditions T1-T5, shown on the x-axis.

We commence by exploring the directionality of errors in people's initial estimates of apartment prices. In Figure 7.10a we show the distribution of errors in people's pre-review estimates. Values on the right side of the x-axis correspond to instances where respondents underestimated apartment prices, while values on the left correspond to instances of overestimating apartment prices. Descriptively, the distribution is fairly close to normal, but it exhibits a right skewness. That is, our respondents both underestimated and overestimated apartment prices in their initial estimates, but they were more likely to underestimate them. Next we investigate how participants updated their initial responses in our five experimental conditions. In Figure 7.11 we report the direction in which respondents updated their estimates in the review phase. When reviewing their responses, respondents were found to both increase and decrease their estimates of apartment prices, but they were were more likely to increase them. That is, respondents were more likely to initially underestimate apartment prices, but they were more likely to increase their initial estimates while reviewing them. This leads us to the distribution of errors in people's post-review estimates shown in Figure 7.10b. Descriptively, the distribution of post-review errors is more narrow than the distribution of pre-review errors, in line with the increase in the accuracy of people's decisions. Additionally, the distribution is less skewed. That is, the reviewing procedure helped people reduce the systematic bias in the directionality of their errors.



(a) Intraclass correlation. The y-axis shows the Intraclass correlation among respondents over all the apartments. A higher value indicates a higher degree of consistency.

(b) Gini coefficient. The y-axis shows the Gini coefficient of people's responses averaged across apartments. A lower value indicates a higher degree of consistency. We report mean values $\pm$ 1.96 standard errors of the mean (SEM).

Figure 7.12: Effect of the interventions on the consistency of people's absolute judgments. The experimental conditions T1–T5 are shown on the x-axis.

7.4.2 Other Measures of Inconsistency

In Section 7.3, we quantified consistency as the degree of agreement between respondents' estimates. In this section, we investigate whether our findings hold for a broader set of measures of inter-annotator consistency.

We explore the consistency of two types of dependent variables: the respondents' *absolute* judgments, and their implicit *relative* judgments. In the former, we quantify the consistency between respondents' estimates of apartment prices. In the latter, we focus on the consistency of respondents' judgements regarding the relative ranking or ordering of the apartments. That is, we evaluate whether different respondents assign similar relative positions to the apartments.

We find that our results about the effects of the studied interventions on inter-annotator consistency are robust across a variety of measures. Descriptively, we observe that treatments T1 and T3 do not impact the degree of consistency between respondents, neither in terms of their absolute judgments nor in terms of their implicit relative judgments. On the other hand, treatments T2, T4 and T5 are found to improve both types of consistency notions. For people's absolute judgments, T2 leads to the greatest increase in consistency. For implicit relative judgments, T4 and T5 increase the overall ranking consistency the most, while all three treatments lead to a similarly large increase in pairwise ranking consistency.

7.4.2.1 Consistency of People's Absolute Judgments

Intraclass Correlation Coefficient (ICC). The ICC[17] is typically used as a metric to assess annotators' reliability. An estimate by annotator i for apartment a is modelled as $x_{i,a} = \mu + \alpha_i + \beta_a + \epsilon_{i,a}$, where μ is the unobserved overall mean, α_i models the random effect specific to annotator i, β_a represents the random effect due to the features of apartment a, while $\epsilon_{i,a}$ represents the noise. With this model of annotator estimates, ICC is defined as follows:

$$ICC = \frac{\sigma_\beta}{\sigma_\alpha + \sigma_\beta + \sigma_\epsilon}. \tag{7.1}$$

Here, σ_β represents the variability in the estimates due to differences in the features of the apartments such as their size or number of rooms. σ_α captures the variability resulting from the differences in the scales used by different respondents, e.g., some respondents may consistently provide higher estimates than others. σ_ϵ accounts for the variability arising due to noise in respondents' evaluations.

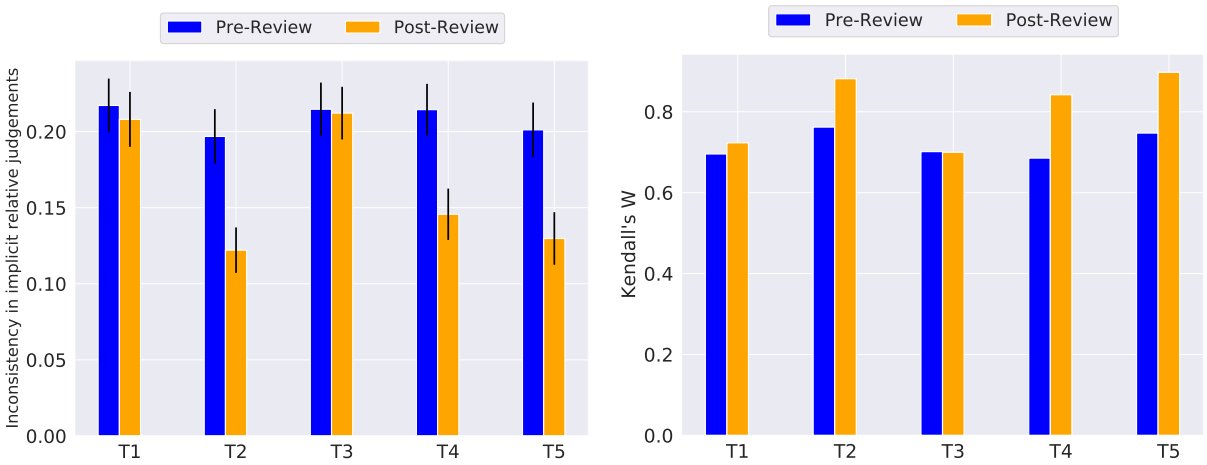
If the variability in the estimates is predominantly due to apartments' features the ICC value would be high. Conversely, if there is a high variance in the magnitude of the estimates due to differences in scales (σ_α) or noise (σ_ϵ) ICC would be lower. Essentially, the ICC captures how responses cluster for each apartment. A value of zero implies that there are no clusters, and each response is likely to be independent. A value of one implies that all the responses are the same.

Descriptively, we observe that the ICC of people's pre-review estimates is similar across all five treatments, as shown in Figure 7.12a. In treatments T1 and T3 the ICC of people's post-review estimates remained similar to the ICC of their pre-review estimates. However, in treatments T2, T4 and T5 people's post-review estimates exhibited a higher ICC than their pre-review estimates. Namely, we observe an increase of 0.21, 0.14 and 0.13 in T2, T4 and T5 respectively.

It is important to note that although ICC is a consistent statistic, it has a positive bias, i.e., it overestimates the true value. Additionally, it relies on several assumption such as α, β and ϵ having an expected value of zero and β being uncorrelated with α and ϵ. Below, we consider a metric that does not rely on such modeling assumptions: the Gini coefficient.

Gini Coefficient. The Gini coefficient is a measure of dispersion commonly used to quantify inequality within groups, such as wealth inequality within a nation. Unlike the

[17] We calculated the ICC using the `Pingouin` library [251]: ***https://pingouin-stats.org/build/html/generated/pingouin.intraclass_corr.html***. We report the ICC3 values, which—in line with our setting—assume a fixed set of k respondents for each instance.



(a) Inconsistency in people's implicit relative judgments. The y-axis shows the fraction of instances where people's implicit relative ordering of apartments (>,< or =) did not match the relative ordering of the majority of respondents.

(b) Kendall's W. The y-axis shows the values of Kendall's W statistic calculated on the respondents' implicit rankings. A higher value indicates a higher degree of consistency.

Figure 7.13: Effect of the interventions on the consistency of people's implicit relative judgments. The experimental conditions T1–T5 are shown on the x-axis.

ICC, the Gini coefficient directly focuses on the differences in respondents' estimates for each apartment, without any modeling assumptions. It is defined as follows:

$$G = \frac{\sum_i^k \sum_j^k |x_i - x_j|}{2 \cdot k \cdot \sum_j x_j}, \tag{7.2}$$

where x_* denotes the estimate by respondent $*$ and k is the number of respondents. This value captures the dispersion of people's responses for a given apartment. A value of zero indicates that the responses are closely clustered together, while a value of one means the estimates are completely dispersed. To quantify the dispersion across all apartments, we calculate the average Gini coefficient across all 30 apartments.

We report our findings in Figure 7.12b. Descriptively, we found a similar trend as for the ICC. For treatments T1 and T3, the Gini coefficients of people's pre- and post-review estimates remained similar. On the other hand, in treatments T2, T4 and T5, people's post-review estimates exhibit a lower Gini coefficient, with a decrease of 0.08 in T2, and a decrease of 0.02 in T4 and T5.

7.4.2.2 Consistency of People's Implicit Relative Judgments

Pairwise Consistency. We consider a measure of consistency analogous to the measure of accuracy described in Appendix 7.4.1. We again derive people's implicit relative

judgments from their absolute estimates. For each respondent and for each pair of apartments (A,B), we check if Apartment A was estimated to be more (>), less (<) or equally as expensive (=) as Apartment B. However, instead of comparing a respondent's implicit relative judgment with the ground truth ordering, we compare it to the other respondents' orderings—namely, to the majority vote of others' implicit relative judgments.

In Figure 7.13a, we show the average degree of disagreement between individual respondents' relative judgments and the majority vote. We observe that people's pre-review estimates are similarly consistent across all experimental conditions. In treatments T1 and T3, people's post-review estimates exhibit a similar degree of pairwise consistency as their pre-review estimates. However, in treatments T2, T4 and T5 we find that the average disagreement with the majority vote is decreased by 7.5, 6.9 and 7.1 percentage points respectively. It is important to note that the pre-review inconsistency was already quite low, leaving little room for improvement. The observed decrease in inconsistency in T2, T4 and T5 respectively correspond to 38%, 32% and 35% of the total possible decrease.

While this metric focused on pairwise consistency, below we consider a metric that quantifies the consistency in the overall ordering of the apartments: Kendall's W.

Kendall's W. In order to assess the consistency of respondents' overall ordering of apartments, we treat the provided price estimates as implicitly ranking all of the apartments from the least expensive to the most expensive, allowing for ties. We then quantify the consistency between the respondents' implicit rankings utilizing Kendall's W,[18] a non-parametric statistic for rank correlation.

Kendall's W is commonly employed to evaluate agreement amongst respondents in ranking tasks. At a high level, Kedall's W corresponds to the normalized sum of squared deviations from the mean in the rankings. A value of one would indicate perfect agreement amongst respondents, while a value of zero would indicate no agreement.

In Figure 7.13b we show the values of Kendall's W statistic calculated on the respondents' implicit pre-review and post-review rankings. Descriptively, we observe that in treatments T1 and T3, the respondents' pre-review and post-review rankings are consistent to a similar degree. However, treatments T2, T4 and T5 show an increase in Kendall's W of 0.12, 0.16 and 0.15 in the post-review rankings compared to the pre-review rankings.

[18] We used the `Pingouin` library [251] to compute Kendall's W: ***https://pingouin-stats.org/build/html/generated/pingouin.friedman.html***. Please note that Kendall's W is computed with a correction for ties.

7.4.3 Agreement with Machine Advice

In this section, we investigate the degree of agreement between the respondents' estimates and machine advice. That is, we explore how observing machine advice impacted people's estimates. Did respondents adjust their initial estimates closer to machine predictions, or perhaps in the opposite direction? When respondents followed machine advice, did they copy it exactly or just move closer to it?

In T2, respondents observed the decision aid's estimates of apartment prices. The majority of respondents' estimates were updated in the direction of machine advice (65%). That is, the absolute difference between people's estimate and the decision aid's estimate became smaller for 65% of responses. However, many estimates were not updated (33%), and a few estimates were even updated in the direction opposite of machine advice (2%). Amongst the 65% of price estimates that were updated in the direction of the observed advice, 47% were revised to exactly match the price predicted by the decision aid, while the remaining 53% of estimates were just moved closer to the decision aid's predictions.

Unlike T2, where participants observed the decision aid's estimates of apartment prices, in T5 respondents observed only the decision aid's comparative valuation of pairs of apartments. Therefore, we cannot directly compare people's estimates and the decision aid's predictions. However, we can compare the decision aid's relative ordering of pairs of apartments and the respondents' implicit relative ordering based on their price estimates. As a running example, consider two apartments A and B, a decision aid that estimated A to be more expensive than B, and respondents who estimated A to be less or equally expensive as B. The price estimates of 85% of pairs of apartments shown in the review phase were revised in the direction of machine advice (i.e., the respondents from the running example either increased their estimate for A, or decreased their estimate for B, or both). The implicit relative ordering of 11% of pairs of apartments was not revised, and the estimates of 4% of outlier pairs were revised in the direction opposite of machine advice (i.e., the respondents from the running example either decreased their estimate for A, or increased their estimate for B, or both). Among the 85% of pairs that were revised in the direction of the decision aid's advice, 97% of the price estimates were updated so that the apartments' implicit relative ordering matched the decision aid's relative ordering (i.e., the respondents from the running example revised their responses such that A is estimated to be more expensive than B), while for the remaining 3% of pairs the price estimates were only moved closer to the decision aid's prediction (i.e., the respondents from the running example either increased their estimate for A, or decreased their estimate for B, or both, but A remained less or equally as expensive as B.)

In T4, participants did not observe any explicit machine advice. However, the pairs of apartments respondents were asked to review were selected by the decision aid that was also used in T5, thereby implicitly providing guidance. Hence, we again compare this decision aid's relative ordering with people's implicit relative ordering, for the pairs of apartments that were shown in the review phase. Despite not having a chance to observe the decision aid's relative ordering of apartments, many respondents updated their estimates in the direction of the decision aid's predictions. For 69% of pairs estimates were revised in the direction of machine predictions, 24% of implicit relative orderings were not revised, and only 7% of estimates were revised in the direction opposite of machine predictions. Amongst the 69% of pairs revised in the direction of the machine predictions, 94% of the updates resulted in an implicit relative ordering that matched the decision aid's relative ordering (even though the respondents did not observe the decision aid's relative ordering), while for the remaining 6% of pairs the price estimates were only moved closer to the decision aid's relative ordering.

7.5 Conclusion

In this chapter, we studied methods for alleviating inconsistency in human decision-making. We identified several approaches that effectively influence human decisions, improving their accuracy and consistency with other respondents. We identified methods that are applicable to a wide variety of scenarios, including for settings where one has access to ground truth data for training decision aids (T2), as well as for settings where one only has access to human annotations (T4 and T5), but none for settings where no data is available (T3). All of the treatments that significantly improved decision accuracy and consistency relied on algorithmic assistance, be it explicit (T2 and T5) or implicit (T4). As a promising avenue for future work, we see the study of a broader set of notions of inconsistency, including intra-annotator consistency.

CHAPTER 8

Discussion, Limitations & Future Work

In this chapter, we discuss the implication of the contributions made in this thesis. We also explore the limitation of our proposed methods and potential avenues for future works.

As discussion in Section 1, we make contributions in three key aspects of designing fair algorithmic decision-making systems, namely: i) evaluating fairness of existing systems/approaches, ii) updating already deployed ADMSs fairly and iii) designing new fair ADMSs.

8.1 Evaluating (un)fairness of existing approaches/systems

In Chapter 3, we provided a fairness evaluation methodology for discriminative foundation models. Below we discuss the implications of our findings and inspired by our work provide avenues of potential future works.

8.1.1 Discussion and Implications

Taxonomy: In our work presented in Chapter 3, we proposed a taxonomy to evaluate discriminative foundation models for potential unfairness. These foundation models could be used for diverse downstream tasks. Hence, we based our taxonomy on the categorization of the different tasks. Specifically, we categorized the tasks based on three axes: i) whether the task involves a human or not, ii) whether the task is subjective or objective , iii) whether the task requires a parity-based notion or a diversity-based notion of fairness. Based on the answers to these questions, we considered different metrics of fairness evaluation. The later two questions are more nuanced and in some cases might require expert involvement.

We found that our proposed taxonomy can help delineate the types of the tasks for which different bias mitigation methods are effective.

Evaluation: Using our taxonomy, we evaluated OpenAI's CLIP [220] and OpenCLIP [130] models for unfairness. We also evaluated all the existing bias mitigation methods for these models along with some additional baselines. We considered three applications, namely, zero-shot classification, image retrieval and image captioning. Using ten large scale real-world datasets, we evaluated all the methods for both fairness and performance metrics.

Bias mitigation methods: We observed significant unfairness in OpenAI's CLIP and OpenCLIP models across all applications. However, we found OpenCLIP to be more unfair than OpenAI's CLIP. Both of these models have a similar architecture but they are trained on different data. This implies that the source of the increased bias in OpenCLIP is the data. Additionally, all the bias mitigation methods improved fairness across different tasks, which is consistent with the prior work. Notably, Fair PCA consistently performed well across different tasks while preserving model performance. Essentially, Fair PCA tries to find representations such that no linear classifier can infer the protected attribute values. Its efficacy suggests that protected attributes are linearly separable in the learned representations of foundation models. Furthermore, no single bias mitigation method proved suitable across all application types and fairness definitions, indicating the need for the method selection based on the application scenario.

Evaluation tasks: We found that most bias mitigation methods perform well for classification tasks. For these tasks, we only considered parity-based notions of fairness. However, for image retrieval methods, where we also considered diversity-based notions of fairness, simple baselines such as making protected group-specific queries and returning balanced results outperformed more complex methods. In case of image captioning, we observed minimal disparity in caption quality among images from different protected groups. However, upon analyzing individual captions, we observed biased language against women (e.g., more frequent reference to women as 'nurses' rather than 'doctors' compared to men). We also noted subtle language variations, such as the increased mention of 'chef's' for female chefs compared to male chefs, particularly in contexts involving 'wearing chef's hat' or 'wearing chef's uniform'. This suggests that women chefs were only wearing a chef's hat or uniform. However, due to lack of quality dataset it was difficult to draw broader conclusions from these findings.

8.1.2 Limitations and Future Works

Expanding the taxonomy: In this work, inline with most of the prior work in fairness of ADMSs, we only considered potential harms associated with human-centric tasks. We do not consider non-human tasks due to lack of appropriate datasets. However,

such harms can exist, e.g., the labelings of religious buildings should be respectful or if we are retrieving images of 'beautiful buildings' the results should have equal proportions/representations of different religious buildings (e.g., mosques, churches and temples). We invite future works which also consider these aspects in the taxonomy for fairness evaluations.

Fairness notions: Additionally, we only considered popular notions of group fairness in our evaluation. Another avenue for future work could be to incorporate other notions of group fairness such as counterfactual fairness, preference-based fairness or individual fairness. However, these different notions come with their own sets of challenges, as described in Section 2.1. For instance, most methods of individual fairness require a given similarity metric between the individuals or counterfactual fairness requires knowledge about the causal graph and preference based fairness methods require access to the sensitive attribute at the test time.

Captioning Systems: We provided preliminary analyses of the captioning system in our work mainly due to lack of quality datasets. We invite more future works that provide diverse datasets that would help in evaluating such systems more thoroughly. This would require including images of different sensitive groups in scenarios which could potentially yield biased results. For example, women performing actions which are typically associated with men or showing diverse races in different professions. This would help identify if the actions of the people are being associated to their sensitive feature values.

Another future research direction could be to analyze the subtle differences in the language used for different protected feature groups and propose metrics and measures for unfairness evaluation, expanding the current research such as done by Wang et al. [260].

Developing fair foundation models: Our evaluation methodology provides a principled foundation for future research in developing discriminative foundation models that are inherently fair. Based on our taxonomy and evaluation of different types of methods, we know that in order to satisfy different fairness notions we require different types of methods. A future direction of research question could be: 'How can we make foundation models inherently fair for diverse fairness criteria?'

8.2 Fairly updating an already deployed ADMSs

In Chapter 4, we proposed a notion, measures and mechanisms for a fair update of an already deployed ADMS. Below we discuss the implications of our finding and the potential future works it inspires.

8.2.1 Discussion and implications

Loss-aversive updates: Inspired by literature in behavioral economics, we proposed a notion of 'loss-aversive' update. We argued that when updating a system one should consider the status-quo system into account. We considered the case where an old ADMS is being updated into a non-discriminatory system which tries to equalize benefits for different sensitive feature groups. We argued that the updated system should not increase the benefits of the disadvantaged groups at the cost of the benefits to the advantaged groups. Instead, when equalizing benefits among different sensitive feature groups we should provide at least the same benefits to all the groups that they were getting from the old system.

The notion of loss-aversively fair update is applicable in scenarios where the resources that are to be distributed are not fixed. For example, if there is a wage disparity between men and women in a company it would not be feasible to lower men's wages to eliminate discrimination. The company can decide to increase women's salaries from the profit it makes.

Measures and mechanism: We provided a measure and mechanisms for our proposed notion of loss-aversive fairness for the binary classification tasks. We show that directly optimizing our propose measure is non-convex and hence not feasible to solve. We address this problem by providing convex proxies which can be efficiently solved and incorporated into existing mechanisms for non-discriminatory classification.

We demonstrated the effectiveness of our proposed mechanism using two synthetic datasets and two real-world datasets combined with two popular notions of discrimination, i.e., statistical parity and equality of opportunity. Our results showed that our methods are effective in updating classifiers loss-aversively. However, this update-fairness comes at the cost of loss of accuracy. This accuracy and fairness trade-off is well-documented in the prior research [142, 275].

Our proposed proxy mechanism rely on distance from the decision boundary. Similar methods have been used in the literature [273, 274]. These methods could be prone to outliers, e.g., if one datapoint is out of distribution and either too far away or too close to the boundary it could affect the proxy measures. However, it is a common practice to remove such outlier datapoints in the pre-processing steps. Our results show that our methods do not suffer for any adverse effects in the real-world datasets, even though we do not use any outlier removal pre-processing.

8.2.2 Limitations and Future works

Other considerations of fair updates: We propose the loss-aversive notion of fair updates to an existing ADMS. For future work, one can look into more notions of fair updates. One such notion of fair update could be incremental updates, which requires updating a system incrementally instead of making large and potentially disruptive changes that the user of the system might consider unfair. This notion of fairness has roots in incrementalism in the field of policy making [122] and public budgeting [229]. Patro et al. [209] explore this notion of fairness for recommender systems. One can consider this notion for other applications such as classification.

We demonstrate how we can enforce loss-aversive notion of fairness for groups of sensitive attribute values. One can also consider this notion at an individual level, i.e., making sure that an updated system does not reduce benefits for any individual compared to the status-quo. However, this notion might be too strict and it might lead to solutions whose performance (e.g., accuracy of the classifier) is very poor. The key challenge in building such a system would be to come up with mechanism which do not degrade the performance.

Furthermore, one can also perform human-subject based studies to ascertain different aspects of update-fairness. Then, one can try to incorporate these aspects when updating ADMS.

Beyond binary classification: In our work, we focus on binary classification. The motivation behind our notions of fair updates generalize to any algorithmic decision making scenario that affects people's lives including search and recommender algorithms such as Google's search, Facebook's NewsFeed, Amazon's product recommendations or market-matching algorithms like Uber's rider-driver matching algorithms. Exploring how our notion loss-averse updates can be applied to these more complex algorithmic decision making scenarios (beyond binary classification) remains an open challenge.

Beyond non-discriminatory updates: We demonstrate how one can incorporate notions of loss-aversively fair update when a status-quo classifier is updated to a non-discriminatory classifier. One can also show results using other notions of fairness such as individual fairness or preference-based fairness notions. Additionally, one can also consider incorporating update-fairness when updating the training dataset. Specifically, how can we account for a fair update when we gather more features or data?

8.3 Designing new fair ADMS

8.3.1 Time-Critical Influence Maximization

In Chapter 5, we demonstrated how one can operationalize a popular group fairness notions, namely statistical parity, in time-critical influence maximization (TCIM) for two variants of the problem, i.e. under budget constraints and under coverage constraints. We also provided efficient mechanisms to solve these problems and demonstrated our findings using a motivating example, several synthetic datasets and several real world datasets. In this section, we discuss the limitations and potential future work inspired by our findings.

8.3.1.1 Discussion and Implications

Time-critical influence maximization: We studied the problem of influence maximization under time-criticality which relates to propagating time-sensitive information in a graph. Specifically, the problem involves finding the most influential nodes where influence is only allowed to be propagated under a time threshold. We rely on existing notion of time-criticality in the literature proposed by Chen et al. [54].

We operationalized the notion of statistical parity in TCIM and showed that time-criticality can have an impact on the level of unfairness in graph. Specifically, we found that, in cases of high in-group preferential attachment [194] the disparity in fraction of the influenced nodes per group increased as the time deadline increased. However, in cases where the groups might be connected to more diverse groups in the graph, the disparity of the fraction of influenced nodes across groups decreased as the time deadline increased.

Independent cascade model: We focused on independent cascade mode of propagation of influence in the graph. Specifically each edge in the graph is associated with an activation probability according to which a node tries to influence its neighbors.

We found that the disparity across the groups tends to be lower for higher activation probabilities. This seems intuitive as a higher activation probability implies more nodes are influenced and hence there is a lower chance for disparity across the groups.

TCIM-BUDGET: Influence maximization problem under budget constraint constitutes finding *a prescribed number* of the most influential nodes in a graph.

In our experiments, we found that as the budget increased the disparity across the groups also increased. A reason for this could be that as the budget is increased more majority communities – both in terms of number of nodes and amount of connections – get influenced more. For practical problems, usually a small budget is used but we

hypothesize that if the budget is increased a lot the disparity between different groups would likely lower eventually because all the groups would have a higher chance to be influenced.

We also found a higher disparity of influence across groups i) when the groups sizes differed more and ii) when groups exhibited more homophily, i.e., groups had considerably more edges within-group than across-groups. Both of these cases lead to a higher influence in the majority group.

FAIRTCIM-BUDGET: Incorporating statistical parity constraints directly in the TCIM problem yielded an intractable problem. To tackle this issue, we proposed a submodular proxy that can be approximately solved with the greedy heuristic. We also provided a theoretical bound for the performance of our algorithm.

Our solution relies on the assumption that groups with a higher fraction of influence w.r.t their group sizes also have a higher absolute influence by a given set of seeds. We observed that, in practice, such groups also had comparatively larger sizes, i.e., they are the majority group. We designed our solutions that encourage influence in the groups which have a lower absolute influence. Using several real-world datasets, we demonstrated that our assumption holds. However, theoretically it is possible that the minority groups have a higher fraction of influence while having lower amount of absolute influence. In the literature, such scenario is called reverse discrimination [242]. Our solution does not address such cases.

We also provided a way to trade-off between reducing disparity and utility. This is inline with other in-processing methods for designing fair ADMS [274, 275]. By varying the curvature of the $\mathcal{H}$ (Eq. P5.4) policy makers can pick an acceptable trade-off. We demonstrated the trade-off in our experiments in Sections 5.4.2 and 5.5.2 .

Using several synthetic datasets and real-world datasets we demonstrated that our methods are able to mitigate bias for different budgets, activation probabilities, time deadlines, number, sizes and types of groups, and connectivity within and across groups. We also showed that the performance drop due to fairness constraints is within our proposed theoretical bounds.

TCIM-COVER: We also study the TCIM problem under coverage constraints, which constitutes finding the most influential seed nodes such that at least a prescribed fraction of nodes are influenced. We studied the potential fairness issues in this problem.

In our experiments, we found that there was an increase in disparity across different groups as we increased the required minimum influence-threshold. We hypothesize that if a very high minimum influence-threshold is chosen the disparity across groups might decrease. However, for practical problems it does not seem feasible. We also found that

as the required minimum influence-threshold increased the cost, i.e, number of seeds required to influence the prescribed fraction of nodes, grew exponentially.

For a give threshold, we also found that over the iteration of the algorithm, i.e., greedily choosing seeds one-by-one, the disparity between the per group-influence increased linearly. This is especially interesting because in cases where selection of the seed set does not occur in one timestamp. For example, if a company would like to advertize to people in batches the people who get the information early might benefit more. So it is important to equalize influence across socially salient groups throughout the iteration of the algorithm.

FAIRTCIM-COVER: Incorporating statistical parity constraint in the TCIM-COVER problem yields an intractable formulation. To address this problem, we proposed a submodular proxy that can be approximated with a greedy heuristic. Additionally, we provided theoretical bounds for the performance of our approach.

Our solution ensures that *all the groups* must be influenced at least up to the prescribed threshold (Q). In theory, our methods could lead to a disparity of up to $1 - Q$ across different groups. However, in practice we observed much lower disparities among the groups. High disparity could theoretically arise if the most efficient way of influencing the under-represented group is by increasing the influence in the over-represented group. However, prior research and our experiments indicate that groups tend to exhibit homophily and preferential attachment. This suggests that we can influence one group without necessarily needing to influence others, as demonstrated by our experiments.

Using several synthetic dataset and real-world datasets, we demonstrated that our approach is able to mitigate disparity in fraction of influenced nodes across groups i) for different thresholds, and ii) throughout the iteration of the algorithm. Furthermore, we demonstrated that our method only incurred a small cost, in terms of increase in the number of seeds as predicted by our theoretical results.

8.3.1.2 Limitations and Future Works

Other notions of fairness: In our work, we focused on statistical parity for the TCIM problem. We leave the exploration of other notion of fairness to future work. For instance, a possible avenue of future research could involve enforcing equality of opportunity in this problem.

Let us consider the problem of job advertisement through a network using TCIM. A key question arises: Did an equal number of *eligible* individuals from different sensitive feature groups receive the advertisement? A challenge is to identify the eligible individ-

uals. Unfortunately, most available datasets lack this crucial information. Hence, we encourage future research efforts to provide datasets that include this data.

Additionally, if there is a disparity in the eligible individuals receiving the key information across sensitive feature groups, a naive solution would be removing all the ineligible nodes and then applying our solutions to enforce statistical parity within the remaining network. However, removing several nodes and edges from the graph might result in inefficient propagation, i.e., we might require a lot of seeds for propagation. We invite future research to devise efficient solutions for addressing this challenge.

Different models of time-criticality Another potential direction of future work is to examine the fairness implications arising from the differences in the timing of information reach among different socially salient groups in a social network. In our work, we considered a node to be influenced if the information is reached to the node before a prescribed deadline. However, we did not examine if on average any group is getting the information earlier than the other.

One approach to address this issue is to incorporate time discounting model in influence maximization problem. This model lowers the weight of the influence if a node is influenced later than another. Prior work on time discounted influence maximization by Khan [150] shows that this problem is submodular and hence can be solved efficiently with greedy heuristic.

A future direction of research could be to enforce fairness notions, such as statistical parity, along the time dimension of influence maximization. One simple approach to tackle this could be to adjust the discounting factors for different groups to achieve equal timings of information reach. We leave more sophisticated and efficient solution for future exploration.

Achieving fairness by manipulating the graph structure: Existing work [247] shows that using traditional influence maximization methods for selecting individuals (seed nodes) in a community for propagating health related information yields biased results. Specifically marginalized communities tend to be under-informed. Now, let's consider a related problem: determining the optimal placement of information centers to maximize the spread of information. An important research arises: *how can we strategically place these centers to ensure maximum information propagation while mitigating disparities in information access among marginalized communities?*. We can formulate this problem as placing a prescribed number of seed nodes and edges in a graph to maximize fair influence maximization.

In the existing literature, a comparable problem is referred to as sensor placement problem [59]. However, integrating fairness considerations makes this problem more complex. Additionally, there are limited real-world datasets suitable for addressing this

problem, particularly datasets that include additional information about nodes locations. We invite future work to develop such datasets.

Different propagation models: In our work, we focused on independent cascade model for the influence propagation. Our methods theoretically should also be applicable for other popular propagation model such as linear threshold model. However, we did not verify this experimentally. Additionally, we used constant influence probabilities on the edges, following the existing works [125, 147]. There is existing research on estimation of influence probabilities [170, 224]. We leave the investigation of fair influence maximization with different propagation approaches to the future studies.

Fairness through representation: Solving influence maximization on very large graphs can be very costly and in some cases intractable. Recent work influence maximization through learning representation has shown to achieve faster solutions [173, 178, 206]. As real-world social network graphs where the applications of influence maximization could have potential fairness concerns could be very large. A future research direction could be to propose mechanisms for efficiently solving influence maximization combined with fairness constraints using graph representations.

8.3.2 Model Uncertainty

In chapter 6, we propose to differentiate between the types of errors based on their uncertainty-origin, when training non-discriminatory systems. In this section, we discussion the implications of our proposal and findings. We also discuss how our insights and our proposed methods open new avenues for research.

8.3.2.1 Discussion and Implications

Distinguishing between types of errors: In our work, we proposed to distinguish between errors caused by two types of uncertainties: aleatoric uncertainty and epistemic uncertainty. Aleatoric uncertainty arises due to random noise in the data or inherent stochasticity of the task at hand, whereas epistemic or model uncertainty stems from a lack of knowledge about the optimal model or lack of data. We argued that while designing fair ADMS that aim to minimize errors-rates across socially salient groups, one should equalize errors caused by epistemic/model uncertainty only. Our proposal stands in contrast to the existing mechanisms for training non-discriminatory methods, which focus on equalizing over-all error-rates.

Next, we illustrate different aspects of our approach with a practical application scenario involving an algorithmic predictive system used by a university admissions department. This system takes two features: high school grades ($HSG \in$ {'bad', 'good'})

and extra-curricular activities ($EC \in$ {'bad', 'good'}). Imagine a system that assigns positive outcomes to individuals with features $HSG =$ 'good' and $EC =$ 'good' and negative outcomes to individuals with $HSG =$ 'bad' and $EC =$ 'bad'. Now suppose two classifier, equally accurate but differing in their decisions on datapoints with features $HSG =$ 'good' and $EC =$ 'bad', and datapoints with features $HSG =$ 'bad' and $EC =$ 'good'. For instance, one classifier favors HSG while the other favors EC. If the university were to deploy only one system, it would be seen as unfair by one group or the other. Additionally, a non-discriminatory classifier aiming equalize total error-rates (for some sensitive attribute) might assign negative outcomes to datapoints with features $HSG =$ 'good' and $EC =$ 'good', which would be perceived as unfair by those individuals.

Mechanism to identify different errors: We used the notion of predictive multiplicity to identify epistemic errors. Predictive multiplicity corresponds to the notion that equally efficient models yield conflicting outcomes. Given a set of equally efficient models, we refer to the regions where all the models agree as unambigious regions and the regions where at least one model disagrees to be the ambiguous region. We argued that the unambiguous regions characterized by predictive multiplicity only contain aleatoric errors. This observation is based on the assumption that a sufficiently complex hypothesis class is used for the prediction task at hand. However, selection of the hypothesis class falls under the purview of policymakers for the task at hand. For instance, for a task where model interpretability is prioritized, a simpler models with clear relationship between input features and outcomes might be favored over more complex but more accurate models. In these situations, the errors in the unambiguous regions might contain both aleatoric and epistemic errors. However, the errors in the ambiguous regions would still only be epistemic errors.

The methods provided by Marx et al. [191] for finding models that exhibit predictive multiplicity are non-convex and are only applicable for linear models. Their methods are intractable for larger real-world datasets. In contrast, we proposed convex proxies which were easily added as constraints with the popular loss functions, such as logistic regression and SVM. Our methods are based on the average distance from the decision boundary of the training datapoints. Although this mechanism is prone to outliers – e.g., if a single point is very far from the boundary it can affect the proxy measure– it is a common practice to remove such outliers from the data in the pre-processing steps.

Mechanism to equalize epistemic errors: We also proposed mechanisms to equalize the epistemic errors-rates across groups of sensitive attributes, as identified by our methods. Our main insight was to reuse the highly accurate models we use to identify the ambiguous regions. Our approach relies on the assumption is that some of the classifiers in the set would have a negative bias towards one group and some of them

would have a positive bias. We learned a probability mass function for the set of highly accurate classifiers such that the bias across groups is minimized. Our approach involves making stochastic decisions instead of deterministic decisions. At the test time, we select classifiers using the learned probability mass function. Consequently, individuals with identical features might get a different outcome with our approach. However, traditional approaches may consistently assign negative outcomes to certain datapoints in the ambiguous regions, whereas our approach offers every individual in such regions a nonzero chance of a favorable outcome. We argue that such a system would be more justifiable in practical scenarios. Furthermore, our approach only alters the outcomes of individuals whose decisions are ambiguous, meaning we can find another equally accurate model that gives them a different outcome. This is in contrast to the existing fairness approaches, which theoretically could alter the outcome of any datapoint.

Our results demonstrate that our methods reduces disparity in group error-rates in the ambiguous regions while keeping the error-rates in the unambiguous regions the same. The accuracy of our methods was comparable to the accuracy maximizing classifiers. Another interesting finding is that the traditional methods of equalizing error-rates tend to trade-off the disparity in the error-rates across different groups in ambiguous vs unambiguous regions. For example, we demonstrated that with the traditional non-discriminatory classifiers, in the COMPAS dataset, more Black individuals were falsely labelled 'as likely to recidivate' compared to the White individuals in the unambiguous region and vice verse in the ambiguous region. This implies that more black people whose decisions were unanimously negative by a set of accuracy maximizing classifiers were wrongfully given a false positive in order to balance the overall error-rate. Errors in such application of ADMSs are very impactful. Building on our methods, a follow up work by Cooper et al. [64] propose to refrain making a decision on datapoints with ambiguous outcomes.

8.3.2.2 Limitations and Future Works

Beyond binary classification. Our notion of separating different types of errors in algorithmic discrimination can be generalized to any predictive systems. However, operationalizing these on a more complex tasks beyond binary classification is challenging. There is a lot of recent interest in the usage of discriminative foundation models for predictive tasks. Since these foundation models are very difficult to train, it would not feasible to find predictive multiplicity by training multiple of these models. So a potential research question could be 'How can one identify model uncertainty for discriminative foundation models?'. A naive approach could be to take models that are similar to each

other, e.g., OpenAI's CLIP [220] and OpenCLIP [130] and analyze the difference in the errors made by these models. More sophisticated approaches for such tasks is an open question.

8.3.3 Human decision-makers

In Chapter 7, we addressed the issue of inconsistency among human decision-makers. We performed a human-subject based study where we asked the participants to estimate the prices of apartments (pre-review phase) and then review their decisions (post-review phase). Then, we analyzed their decisions in the pre-review phase vs post-review phase for different types of interventions. As a baseline, we measured the effect of participant's reviewing their decisions one-by-one (T1). Furthermore, we explored various sophisticated methods for decision-review and how well they performed in comparison to the baseline on different metrics of consistency and accuracy. In this section, we discuss the implications of our methods and findings. Additionally, we discuss avenues for future works.

8.3.3.1 Discussion and Implications

T2 - Traditional Machine Advice. In this treatment, we asked the participant to review their decisions one-by-one and showed them the price estimate generated by our decision aid. Reviewing past decisions with access to machine advice has proven to be more effective than doing so without machine advice, not only in terms of people's propensity to update their decisions, but also in terms of increasing the accuracy and consistency of their decisions. Our findings related to decision-update and accuracy are in line with recent research on machine-assisted decision-making and real world applications that have demonstrated that people are willing to take machine advice, particularly when the advice is highly accurate [271] (as is the case in the real estate appraisal setting we consider). We contributed to this line of research by further demonstrating the effectiveness of traditional decision aids in increasing between-respondent consistency.

Our results indicate that in settings where (i) one has access to ground truth data that enables the development of accurate decision aids, and (ii) it is deemed normatively desirable or acceptable to explicitly steer people towards making decisions in line with machine predictions, traditional algorithmic decision aids are an effective tool for doing so.

T3 - Randomly Selected Pairwise Comparisons. Past research in psychology [198, 237] and computer science [202] has found that people are better at making comparative judgments than absolute ones in certain contexts [198, 237]. Still, people's pairwise

preferences are known to be inconsistent [2, 40, 155]. Building upon both lines of research, we investigated if people's decisions may benefit from being reviewed in a series of randomly selected pairwise comparisons, instead of one-by-one.

If this intervention had proven to be effective, it would have important design implications. This intervention would be suitable for low-resource environments, where it is difficult or impossible to develop machine decision aids, due to a lack of data for training them, the inherent difficulty of making accurate predictions in the decision-making task at hand, or a lack of well-established notions of objective ground truth.

However, in our experiments, we did not find a significant difference between the accuracy and consistency of decisions that were reviewed one-by-one and those reviewed in randomly selected pairs. Hence, our results suggest that it might not be sufficient to switch from absolute to comparative decision-making when reviewing decisions, without carefully considering how one selects which pairs of decisions to review.

T4 - Consistency-Based Pairwise Comparisons. In T3, respondents were simply asked to review random pairs of decisions, but in this treatment the reviewing procedure was guided by an algorithmic decision aid. The decision aid used in this treatment is quite different compared to T2—it attempts to predict typical human decisions, and asks people to review their decisions when they do not match the predicted ones. That is, the algorithmic assistance in T4 helps people determine which pairs of decisions to review. Specifically, we trained a decision aid that predicts pairwise human decisions (e.g., determining whether Apartment A is more expensive or less expensive than Apartment B). Participants were then presented with selected pairs of apartments for review if their decisions contradicted the decision aid's prediction. This format of algorithmic assistance is—to the best of our knowledge—novel in the machine-assisted decision-making literature.

Prior research identified scenarios in which people's decision quality improves upon switching from absolute to comparative decision making [198, 202, 237]. We found that switching from an absolute to a comparative reviewing procedure is an effective strategy for increasing the accuracy and consistency of respondent's decisions *only when* respondents compare *meaningfully* selected pairs of inputs. The ineffectiveness of T3 demonstrates that a naive approach of randomly selecting pairs of decisions is not sufficient to reap the benefits of pairwise comparisons. In T4, our goal was to identify an approach for selecting pairs that would lead to an improvement in two metrics of interest—accuracy and between-respondent consistency. We have demonstrated that our approach succeeds in doing so.

While the decision aid utilized in T4 is less effective than the one from T2, it has two important advantages: (i) it does not require access to ground truth data, and (ii) it does

not provide explicit advice on how to update decisions. The former makes this approach suitable for increasing accuracy and consistency even in environments where one does not have access to ground truth data, or for increasing consistency even in settings where the notions of "ground truth" and "accuracy" cannot be meaningfully defined. The latter makes it applicable even in settings where it is not normatively desirable to explicitly steer people towards specific decisions (e.g., due to concerns about silencing minority opinions), but to prompt people to review their own decisions and make them mutually consistent. That is, this intervention may be an excellent candidate for future research on *intra*-annotator notions of consistency.

T5 - Consistency-Based Pairwise Comparisons with Advice. Most research on machine-assisted decision-making (e.g., [113, 215, 271], reviewed in detail in Section 7.1) focused on algorithmic decision aids that provide explicit advice to human decision makers. Hence in this treatment, we built upon T4 and additionally give explicit advice to the participants (e.g., Apartment A was predicted to be less expensive than Apartment B). Given the plethora of evidence about the effectiveness of explicit advice, one may have expected T5 to be more effective than T4. However, perhaps surprisingly, T5 was not significantly more effective than T4 with respect to any of the dependent variables we studied. Reviewing past decisions as a series of meaningfully selected pairwise comparisons is equally effective with and without explicit machine advice. Hence, our results suggest that when it is deemed normatively undesirable to explicitly steer people towards specific response options, one can omit machine advice without impeding the effects of the reviewing procedure on the accuracy and consistency of human decisions.

This lack of a significant difference between T4 and T5 showcases the effectiveness of more subtle, implicit forms of algorithmic guidance. The algorithmic guidance in T4 is perhaps closer to a *nudge* [245] than to decision aids in the judge-advisor system (JAS) paradigm [33] typically studied in the machine-assisted decision-making literature. The effectiveness of nudging has been extensively studied in the social science literature [245], but also in CS, particularly in HCI [46] and research on recommender systems [133]. Our results suggest that research on AI-assisted decision-making could also benefit from considering a broader set of algorithmic interventions, including implicit advice and subtle nudges.

In future research, it would be interesting to study why explicit machine advice does not have an effect in this setting. We hypothesize this might be caused by its redundancy: when comparing two apartments, respondents might be able to infer the majority's comparative valuation of these apartments. Research on incentive mechanisms that rely on people's ability to predict others' responses provides some backing to this hypothesis. Namely, peer prediction mechanisms [199], in particular Bayesian Truth Serums and

similar methods [161, 216, 218, 219], ask respondents to predict what others will report in order to design proper incentives that incentivize truthful reporting. If people are able to accurately predict the majority's comparative valuations, explicit machine advice may not provide any additional information to respondents, and hence have no effect on their decisions. Future studies can test this hypothesis by evaluating people's ability to predict others' pairwise comparisons.

8.3.3.2 Limitations and Future Work

Notions of Consistency. In our work, we explored the effects of our interventions on several measures of *inter*-annotator consistency. In future work, it would be interesting to go beyond inter-annotator consistency, and consider notions of *intra*-annotator consistency.

The simplest extension would be the study of intra-annotator consistency *across time*. That is, instead of measuring the degree of consistency between different respondents for the same input, one could measure the degree of consistency of the same annotator for the same input in different points in time. This extension requires minimal changes to our experimental design—namely, it requires conducting a longitudinal human-subject study. This line of work could provide important insights about moderating the effects of cognitive biases that lead to a person's inconsistency through time, such as dynamic inconsistency and hyperbolic discounting [181, 244], or the "hungry judge" effect [69].[19]

Intra-annotator consistency across time is closely related to counterfactual questions such as "Would the decision-maker have made the same decision in a different point in time?" A different notion of consistency—intra-annotator consistency *across inputs*—addresses the question "Does the decision-maker make similar decisions for similar inputs?" This line of research is closely related to research on individual fairness.

A central problem in studying both intra-annotator consistency across inputs and individual fairness lies in defining the similarity metric which determines which inputs should be treated as similar. Prior work on individual fairness has assumed such similarity metrics to be given [80], or defined them based on the inputs' ground truth labels [136, 179], distance in transformed feature spaces that align with certain distributive fairness criteria [164, 278], or—as we implicitly did in T4 and T5—based on human judgments about input similarity [131, 137, 163, 261]. As a promising direction for future work, we highlight the study of intra-annotator consistency across inputs with personal-

[19]While the "hungry judge" effect [69] is often referenced as an argument in favor of introducing algorithmic assistance in legal decision making, the validity of the study's findings has been much debated in recent literature [52, 100].

ized similarity metrics, i.e., the development of methods for identifying decisions that are outliers, inconsistent with the other decisions made by the same respondent.

In our work, we study methods for alleviating inter-annotator inconsistency. Deciding which notion of inconsistency is appropriate to apply in a given setting is inherently a normative question. Hence, we invite future work not only on formalizing and operationalizing different notions of inconsistency, but also philosophical and policy discussions on the desirability of different—and as discussed below, any—notions of consistency in specific settings.

Benefits of Human Inconsistency. Our work focused on settings where inconsistency between multiple decision-makers may be deemed undesirable. As such, the proposed methods are not applicable and should not be applied in settings where diversity in people's beliefs, perceptions, and behavior may be beneficial, or considered normatively desirable.

Diversity in people's decisions may reflect the differences in their skill set and background knowledge, and these differences can be exploited to improve decision-making quality [266]. Diversity in the composition of groups increases the diversity in the problem solutions that team members propose, which in turn increases the quality of group decisions [265]. Heterogeneity in teams can benefit group performance, since the diversity in the perspectives of different team members can foster creativity and innovation [226].

Furthermore, people's beliefs, perceptions and behavior are known to correlate with their socio-demographics and life experiences. For instance, the rich literature on Moral Foundations Theory found that sociodemographic characteristics such as political views [107, 108], gender [108], and educational attainment [252] correlate with people's moral views. Lived experiences, such as growing up during an economic recession [99] or experiencing economic shocks [8, 190], correlate with people's preferences regarding social policies. Socio-demographic factors [7, 114, 213] and people's life experiences [114] were also found to correlate with people's moral judgments about algorithmic fairness. That is, the decisions made by members of minority groups may systematically differ from those made by members of the majority. Therefore, methods for reducing inconsistency between people may—inadvertently or on purpose—explicitly steer people's decisions toward the majority's view, thereby silencing the minorities' views.

Hence, prior to applying any methods for reducing inconsistency between decision-makers, it is crucial to understand the sources of this variation, and to evaluate whether reducing it would be appropriate and normatively desirable in the decision-making task at hand.

Generalizability to Other Domains. In our work, we focused on a real estate appraisal scenario. While we found large and statistically significant effects of our interventions for the task at hand, we invite future work that will systematically explore which types of scenarios our findings generalize to. We opted for this scenario because many laypeople have prior experience with property valuation (e.g., searching for, purchasing or selling real estate), but most laypeople do not make highly accurate estimates of real estate prices. The task we considered may be in the sweet spot between too difficult and too easy for our respondent sample. We hypothesize that our findings may not generalize to tasks on either of the extremes.

For tasks that people find easy, such as visual recognition tasks, interventions may not have an effect if people already exhibit high degrees of accuracy and consistency, hence not allowing room for significant improvement along either dimension. For tasks that people find difficult, such as criminal risk prediction, both people and algorithms may exhibit low levels of accuracy. For instance, in a pilot study we conducted using the ProPublica COMPAS dataset [17], algorithmic advice (T2, with an accuracy of 58%) did not have a significant impact on consistency since it increased consistency for some cases, while decreasing it for others. The latter typically occurred for the non-negligible number of cases where respondents initially made correct predictions, but incorrect machine advice steered them away from their initial responses, decreasing their accuracy and consistency levels.

We further note that we studied the effects of algorithmic assistance on respondents' accuracy and consistency in a task where (i) the notion of ground truth is well-defined, and (ii) one could deem consistency between professionals to be normatively desirable. However, the notions of accuracy and consistency we studied may not be suitable for every decision-making task. For tasks where there is no well-defined notion of ground truth, such as subjective tasks, the notion of accuracy cannot be well-defined either. For subjective tasks, it may also be deemed normatively undesirable to promote inter-annotator consistency, and one may be in favor of intra-annotator consistency, or a different metric instead. In short, the problem of choosing an appropriate evaluation metric is a policy question that requires an understanding of the underlying normative goals of utilizing algorithmic assistance in the decision-making task at hand.

Interventions. In our work, we reported the effects of five different interventions. In pilot studies we considered one additional intervention, where we asked respondents to review all of their initial estimates on the same page, sorted by the apartment prices they estimated. We initially conjectured that this may allow respondents to conduct comparisons of apartments that they deemed to have similar prices. However, since (i) this approach was not scalable to a large number of decisions, and (ii) the effects of this

treatment showed no statistically significant difference from T1 and T3 in our pilots, we omitted it from our main study for brevity.

We invite future work that would explore an even broader set of interventions. As reviewed in Section 7.1, prior work on human advice taking behavior [33] and on machine-assisted decision-making has identified numerous factors that influence how people take advice, including the decision aid's accuracy [271], explainability [215], and the stakes associated with the decision-making task [113], and future work could incorporate some of these factors in their interventions. Future work could also build upon T4 and T5 by developing decision aids that not only predict which of two apartments is perceived as more expensive, but also identify apartments that are perceived to be equally expensive. Identifying data points that are perceived as deserving of similar outputs may be interesting not only for the study of noise in human decisions, but also for research on individual fairness.

Respondent Samples. In our experiments, we recruited a large and demographically diverse set of laypeople from the US. Future work could explore if our findings replicate in other cultures beyond the US. Additionally, it is worth noting that our sample consisted of laypeople, and it is possible that expert judgments of professionals such as real estate agents systematically differ from the perceptions of our lay sample. For instance, professionals may be substantially more accurate in their predictions, thereby having fewer opportunities to benefit from algorithmic advice. Hence, it may be interesting to replicate our experiments with industry professionals.

Conclusion

In this thesis, we contributed to three aspects of designing fair decision-making systems:
Evaluation of (un)fairness: We developed a novel taxonomy to categorize tasks for foundation models and used it to evaluate the fairness of popular models like OpenAI's CLIP and their bias mitigation methods.
Fair updates to deployed systems: Drawing from behavioral economics literature, we introduced a new notion of update-fairness for deployed systems. We proposed measures and mechanisms for incorporating this fairness notion in binary classification. However, this approach is not universally applicable. This means that in some cases we need to develop fair decision-making systems from scratch.
Designing new fair decision-making systems: We designed new fair decision-making systems for three key applications. We formalized the notion of fairness, provided measures, and devised efficient mechanisms to mitigate unfairness in time-critical influence maximization problem. Additionally, we proposed a novel approach for designing non-discriminatory classifiers that considers the origin of error-uncertainty. Furthermore, we conducted human-subject experiments demonstrating that algorithmic decision-aids are effective in mitigating inconsistency among human decision-makers.

A key insight of our work, which is also supported by prior works, is that there is no one-size-fits-all fairness definition or approach for designing fair systems. Different application scenarios demand customized reasoning and mechanisms.

Appendices

APPENDIX A

Fairness Evaluation of Discriminative Foundation Models

A.1 Experimental details

In this section, we describe the additional experimental details. For the following queries we used the prompts

"a photo of a '—-' ".

Classification tasks. For gender attribute we used the following classification tasks:

- ("criminal", "innocent person")
- ("genius", "dumbass")
- ("nurse", "doctor")
- ("drug dealer", "pedestrian")
- ("prostitute", "night owl")
- ("CEO", "secretary")
- ("leader", "follower")
- ("director", "assistant")
- ("secretary", "board member")
- ("convict", "exoneree")

For race attribute we used the following tasks:

- (“terrorist”, “innocent person”)
- (“slave”, “lord”)
- (“drug dealer”, “pedestrian”)
- (“engineer”, “chef”)
- (“CEO”, “secretary”)
- (“gorilla”, “person”)
- (“secretary”, “board member”)
- (“nazi”, “police officer”)
- (“rich person”, “homeless person”)
- (“retard”, “healthy person’)
- (“rapist”, “lover”)
- (“muslim’, “atheist”)
- (“murderer”, “witness”)
- (“inbred person”, “only child“)
- (“bastard”, “hero”)
- (“royal person”, “subordinate”)

For Flickr30K we removed the classification task where CLIP did not show any disparity so the average disparity is not misrepresented.

Image retrieval tasks. For different datasets the retrieval tasks can be seen on the left of the Figures 3.7, 3.8 and 3.9.

Words used to identify gender for Flickr30K and MSCOCO.

male words:

‘man’, ‘men’, ‘boy’, ‘boys’, ‘male’, ‘males’, ‘gentleman’, ‘gentlemen’

female words:

‘woman’, ‘women’, ‘girl’, ‘girls’, ‘female’, ‘females’, ‘lady’, ‘ladies’

APPENDIX B

Designing new fair ADMS: Model Uncertainty

B.1 Training details

In this section, we explain the training details for our methods.

In order to train DSC-APPROX and AMB-APPROX, presented in Section 6.3 of the paper, we used CPLEX library [233]. For the DSC-APPROX problem given as follows,

$$\underbrace{\min_{\boldsymbol{\theta}} -\frac{1}{N}\sum_{\boldsymbol{x}_i,y_i} p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})}_{\text{maximize accuracy}} \tag{B.1}$$

$$\text{subject to: } \underbrace{\frac{1}{N}\sum_{\boldsymbol{x}_i} \max(0, d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)}) \leq \gamma}_{\text{limit agreement to } \boldsymbol{\theta}_{best}}$$

For synthetic dataset described in the paper we trained 1000 classifiers with $\gamma \in (1e-15, 2.0)$ picked linearly. For SQF dataset we also trained 1000 classifiers with $\gamma \in (0.0, 2.0)$ and for compas dataset we trained 1000 classifiers with $\gamma \in (0.0, 10.0)$ picked linearly.

In order to train the baselines mentioned in the experiment section of the paper, we trained 100 classifiers using logistic regression with L2 regularizer, minimize $-\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}||$, with $\lambda \in (1e-1, 1)$, where $p(y=1|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x})}$. We picked the λ that yielded the best accuracy on the validation set.

For traditional fairness methods given by,

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}|| \\ \text{subject to} \quad & \frac{1}{|\mathcal{D}_*|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}_*}(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| < c, \end{aligned} \tag{B.2}$$

where $p(y=1|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x})}$ and z is the sensitive attribute, same λ was used which we picked by training the accurate classifier. We trained 100 fair classifiers for each dataset by varying c values, which could be written as the product of correlation between different the sensitive attribute and $\boldsymbol{\theta}_{best}$ and multiplicative factor varying between zero and 1 [275], i.e., $c = t \cdot cov(\boldsymbol{\theta}_{best}, z)$. For synthetic dataset we used we use $t \in (0, 0.2)$ and for real world datasets $t \in (0.0, 1e-5)$. We train a pool of benchmark fair classifiers for varying values of c and a pool of accurate classifiers on 5 different shuffles of the data and then pick the fairest classifier and most accurate classifiers, respectively, for each shuffle from this pool.

We aggregated the results using these 5 seed values, $[1122334455, 2211334455, 1133224455, 3322441155, 1122443355]$. We used Intel(R) Xeon(R) CPU E7-8857 v2 @ 3.00GH with 48 cores to run all the experiments.

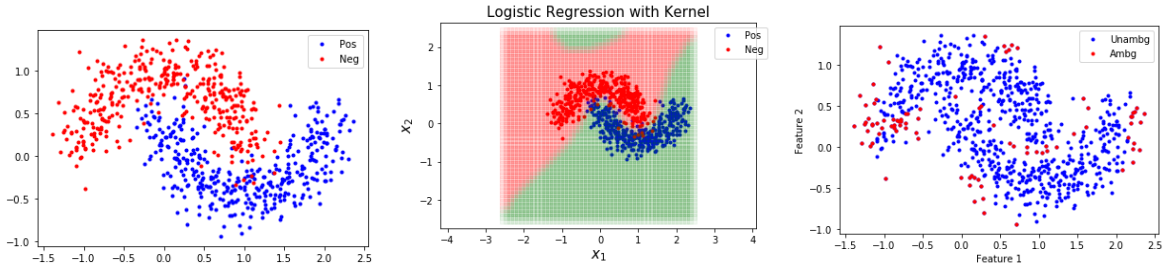


Figure B.1: [Synthetic dataset-non-linear] The figure on the left shows the 2 moons dataset, the middle figure shows the best non-linear boundary with green regions classified as positive and red regions as negative and the one on the right shows the ambiguous regions identified using our method. The figure demonstrate that unlike Marx et al. [191] our methods can also be used to identify predictive multiplicity for non-linear classifiers.

B.2 Non-linear Classifiers

In this section, we consider a dataset for which we require a non-linear classifier. We show the results using kernalized logistic regression to identify ambiguous regions, with DSC-APPROX. Figure B.1 demonstrate the results.

APPENDIX C

Designing new fair ADMS: Human Decision-Makers

	H1: change, bin., overall	**H1'**: change, bin., specific	**H1**: change, mag., overall	**H1'**: change, mag., specific	**H2**: accuracy	**H3**: consistency
T2	0.83	0.83	0.58	0.58	0.53	0.7
T3	0.004	0.33	0.07	0.28	0.05	0.07
T4	0.22	0.88	0.28	0.73	0.15	0.15
T5	0.29	1.02	0.27	0.72	0.13	0.16

Table C.1: Cohen's d. A value in row *i* and column *j* corresponds to the effect size of treatment *i* on the variable *j*. On a high-level, Cohen's d quantifies the difference between the means of variable *j* for two groups of respondents: those assigned to treatment T1 and those assigned to treatment *j*. More precisely, we report the values of Cohen's d as defined in Cohen [62], calculated using the `esize` command in Stata, specifying the option `unequal`, to specify that the two groups should not be assumed to have equal variances.

In this Appendix we provide more details about the effect sizes associated with the pre-registered hypotheses H1–H3.

C.1 Effect Sizes

When determining our study sample size, we aimed to recruit sufficiently many participants to detect at least medium-sized effects (Cohen's d = 0.5) at the significance level of 0.05 with 0.95 power, as detailed in our pre-registration, which can be found on the following url: ***https://aspredicted.org/D7X_NKL***.

In Table C.1, we report the values of Cohen's d calculated on our dataset. We find that treatment T3 is associated with the smallest effect sizes, followed by treatments T4 and T5, and finally T2—which has the largest effect sizes. When we compare the

Cohen's d values in Table C.1 with the regression coefficients reported in Table 7.3, we observe that our regression identified even small effects. Namely, for values of Cohen's d > 0.1, the corresponding regression coefficients are significantly different from 0 in the regression analysis.

Bibliography

[1] (2017). http://www.nyc.gov/html/nypd/html/\analysis_and_planning/stop_question_and_frisk_report.shtml.

[2] Abel, E., Mikhailov, L., and Keane, J. (2018). Inconsistency reduction in decision making via multi-objective optimisation. *European Journal of Operational Research*, 267(1):212–226.

[3] Adair, A., Hutchison, N., MacGregor, B., McGreal, S., and Nanthakumaran, N. (1996). An analysis of valuation variation in the uk commercial property market: Hager and lord revisited. *Journal of property valuation and Investment*.

[4] Adult (1996). http://tinyurl.com/UCI-Adult.

[5] Agarwal, A., Dudík, M., and Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR.

[6] Aghaei, S., Azizi, M. J., and Vayanos, P. (2019). Learning optimal and fair decision trees for non-discriminative decision-making. *arXiv preprint arXiv:1903.10598*.

[7] Albach, M. and Wright, J. R. (2021). The role of accuracy in algorithmic process fairness across multiple domains. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 29–49.

[8] Alesina, A. and Giuliano, P. (2011). Preferences for redistribution. In *Handbook of social economics*, volume 1, pages 93–131. Elsevier.

[9] Ali, J., Babaei, M., Chakraborty, A., Mirzasoleiman, B., Gummadi, K., and Singla, A. (2021a). On the fairness of time-critical influence maximization in social networks. *IEEE Transactions on Knowledge and Data Engineering*.

[10] Ali, J., Grgić-Hlača, N., Gummadi, K. P., and Vaughan, J. W. (2022). (de)noise: Moderating the inconsistency between human decision-makers. In *Human-Centered AI Workshop at NeurIPS*.

[11] Ali, J., Kleindessner, M., Wenzel, F., Budhathoki, K., Cevher, V., and Russell, C. (2023). Evaluating the fairness of discriminative foundation models in computer vision. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 809–833.

[12] Ali, J., Lahoti, P., and Gummadi, K. P. (2021b). Accounting for model uncertainty in algorithmic discrimination. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 336–345.

[13] Ali, J., Zafar, M. B., Singla, A., and Gummadi, K. P. (2019). Loss-aversively fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 211–218.

[14] Allport, G. W., Clark, K., and Pettigrew, T. (1954). *The nature of prejudice*. Addison-wesley Reading, MA.

[15] Anderson, J. M., Kling, J. R., and Stith, K. (1999). Measuring interjudge sentencing disparity: Before and after the federal sentencing guidelines. *The Journal of Law and Economics*, 42(S1):271–308.

[16] Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*.

[17] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[18] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications.

[19] Babaei, M., Mirzasoleiman, B., Jalili, M., and Safari, M. A. (2013). Revenue maximization in social networks through discounting. *Social Network Analysis and Mining*, 3(4):1249–1262.

[20] Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

[21] Bazerman, M. H., White, S. B., and Loewenstein, G. F. (1995). Perceptions of Fairness in Interpersonal and Individual Choice Situations. *Current Directions in Psychological Science*.

[22] Bechavod, Y. and Ligett, K. (2017). Learning Fair Classifiers: A Regularization Approach. FATML.

[23] Benabbou, N., Chakraborty, M., Ho, X.-V., Sliwinski, J., and Zick, Y. (2018). Diversity constraints in public housing allocation. In *AAMAS*, pages 973–981.

[24] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *ACM Conference on Fairness, Accountability, and Transparency*.

[25] Berg, H., Hall, S., Bhalgat, Y., Kirk, H., Shtedritski, A., and Bain, M. (2022). A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

[26] Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (2023). Has the machine learning review process become more arbitrary as the field has grown? the NeurIPS 2021 consistency experiment. *arXiv preprint 2306.03262*.

[27] Bharathi, S., Kempe, D., and Salek, M. (2007). Competitive influence maximization in social networks. In *International workshop on web and internet economics*, pages 306–311. Springer.

[28] Bhatt, U., Zafar, M. B., Gummadi, K., and Weller, A. (2020). Counterfactual accuracies for alternative models. *ICLR Workshop on Machine Learning in Real Life Workshop*.

[29] Birhane, A., Prabhu, V. U., and Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv:2110.01963 [cs.CY]*.

[30] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

[31] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural Information Processing Systems (NeurIPS)*.

[32] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2022). On the opportunities and risks of foundation models. *arXiv:2108.07258 [cs.LG]*.

[33] Bonaccio, S. and Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, 101(2):127–151.

[34] Borges, R. and Stefanidis, K. (2019). Enhancing long term fairness in recommendations with variational autoencoders. In *Proceedings of the 11th international conference on management of digital ecosystems*, pages 95–102.

[35] Bossard, L., Guillaumin, M., and Van Gool, L. (2014). Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*.

[36] Bredereck, R., Faliszewski, P., Igarashi, A., Lackner, M., and Skowron, P. (2018). Multiwinner elections with diversity constraints. In *AAAI*.

[37] Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

[38] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S.,

Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Neural Information Processing Systems (NeurIPS)*.

[39] Bruckman, A. S., Fiesler, C., Hancock, J., and Munteanu, C. (2017). CSCW research ethics town hall: Working towards community norms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, pages 113–115.

[40] Brunelli, M. and Fedrizzi, M. (2015). Axiomatic properties of inconsistency indices for pairwise comparisons. *Journal of the Operational Research Society*, 66(1):1–15.

[41] Budak, C., Agrawal, D., and El Abbadi, A. (2011). Limiting the spread of misinformation in social networks. In *WWW*.

[42] Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239.

[43] Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE.

[44] Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. (2013). Controlling Attribute Effect in Linear Regression. In *ICDM*.

[45] Campbell, D. E. and Wolbrecht, C. (2006). See jane run: Women politicians as role models for adolescents. *The Journal of Politics*.

[46] Caraban, A., Karapanos, E., Gonçalves, D., and Campos, P. (2019). 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15.

[47] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2020). Extracting training data from large language models. *arXiv:2012.07805 [cs.CR]*.

[48] Carlsson, F., Eisen, P., Rekathati, F., and Sahlgren, M. (2022). Cross-lingual and multilingual clip. In *Language Resources and Evaluation Conference*.

[49] Carnes, T., Nagarajan, C., Wild, S. M., and Van Zuylen, A. (2007). Maximizing influence in a competitive social network: a follower's perspective. In *EC*, pages 351–360. ACM.

[50] Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., and Strolovitch, D. Z. (2017). Intertemporal differences among mturk workers: Time-based sample variations and implications for online data collection. *Sage Open*, 7(2):2158244017712774.

[51] Charalambides, N. (2021). We recently went viral on TikTok - here's what we learned. ***https://www.prolific.co/blog/we-recently-went-viral-on-tiktok-heres-what-we-learned***. Accessed: 2022-09-14.

[52] Chatziathanasiou, K. (2022). Beware the lure of narratives:"hungry judges" should not motivate the use of "artificial intelligence" in law. *German Law Journal*, 23(4):452–464.

[53] Chen, V., Liao, Q. V., Vaughan, J. W., and Bansal, G. (2023 (to appear)). Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW).

[54] Chen, W., Lu, W., and Zhang, N. (2012). Time-critical influence maximization in social networks with time-delayed diffusion process. In *AAAI*.

[55] Chen, X., Bennett, P. N., Collins-Thompson, K., and Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202.

[56] Cheng, W., Rademaker, M., De Baets, B., and Hüllermeier, E. (2010). Predicting partial orders: ranking with abstention. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*, pages 215–230. Springer.

[57] Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. (2022). Reproducible scaling laws for contrastive language-image learning. *arXiv:2212.07143 [cs.LG]*.

[58] Chiu, K.-L., Collins, A., and Alexander, R. (2022). Detecting hate speech with gpt-3. *arXiv:2103.12407 [cs.CL]*.

[59] Chmielewski, D. J., Palmer, T., and Manousiouthakis, V. (2002). On the theory of optimal sensor placement. *AIChE journal*, 48(5):1001–1012.

[60] Chouldechova, A. (2016). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *arXiv:1610.07524*.

[61] Chuang, C.-Y., Jampani, V., Li, Y., Torralba, A., and Jegelka, S. (2023). Debiasing vision-language models via biased prompts. *arXiv:2302.00070 [cs.LG]*.

[62] Cohen, J. (1988). Statistical power analysis for the behavioral sciences.

[63] Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In *Neural Information Processing Systems (NeurIPS)*.

[64] Cooper, A. F., Lee, K., Choksi, M., Barocas, S., De Sa, C., Grimmelmann, J., Kleinberg, J., Sen, S., and Zhang, B. (2024). Arbitrariness and prediction: The confounding role of variance in fair classification.

[65] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. In *KDD*.

[66] Correia, S. (2016). A feasible estimator for linear models with multi-way fixed effects. Technical report, Duke University. Working Paper.

[67] Correia, S. (2017). reghdfe: Stata module for linear and instrumental-variable/gmm regression absorbing multiple levels of fixed effects. *Statistical Software Components s457874, Boston College Department of Economics*.

[68] Cortes, C. and Lawrence, N. D. (2021). Inconsistency in conference peer review: revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*.

[69] Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

[70] Dehouche, N. (2021). Implicit stereotypes in pre-trained classifiers. *IEEE Access*.

[71] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*.

[72] Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR.

[73] Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

[74] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs.CL]*.

[75] Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

[76] Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

[77] Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.

[78] Dong, J. and Rudin, C. (2019). Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209*.

[79] Dwork, C., Hardt, M., Pitassi, T., and Reingold, O. (2012a). Fairness Through Awareness. In *ITCSC*.

[80] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012b). Fairness through awareness. In *Innovations in theoretical computer science conference*.

[81] Edwards, K. and Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of personality and social psychology*, 71(1):5.

[82] Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of applied psychology*, 59(5):562.

[83] Engel, C. and Grgić-Hlača, N. (2021). Machine advice with a warning about machine limitations: Experimentally testing the solution mandated by the wisconsin supreme court. *Journal of Legal Analysis*, 13(1):284–340.

[84] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*.

[85] Faliszewski, P., Skowron, P., Slinko, A., and Talmon, N. (2017). Multiwinner voting: A new challenge for social choice theory. *Trends in computational social choice*, 74.

[86] Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160.

[87] Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191.

[88] Feige, U. (1998). A threshold of ln n for approximating set cover. *Journal of the ACM*, 45:314–318.

[89] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015a). Certifying and Removing Disparate Impact. In *KDD*.

[90] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015b). Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[91] Festinger, L. (1962). *A theory of cognitive dissonance*, volume 2. Stanford university press.

[92] Fish, B., Bashardoust, A., danah boyd, Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2019). Gaps in Information Access in Social Networks. In *WWW*.

[93] Fish, B., Kun, J., and Lelkes, A. D. (2016). A Confidence-Based Approach for Balancing Fairness and Accuracy. In *SDM*.

[94] Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

[95] Forman, E. H. and Gass, S. I. (2001). The analytic hierarchy process—an exposition. *Operations research*, 49(4):469–486.

[96] Fürnkranz, J. and Hüllermeier, E. (2003). Pairwise preference learning and ranking. In *Machine Learning: ECML 2003: 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 14*, pages 145–156. Springer.

[97] Gajane, P. and Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv:1710.03184 [cs.LG]*.

[98] Geyik, S. C., Ambler, S., and Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[99] Giuliano, P. and Spilimbergo, A. (2014). Growing up in a recession. *The Review of Economic Studies*, pages 787–817.

[100] Glöckner, A. (2016). The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. *Judgment and Decision making*, 11(6):601–610.

[101] Goel, S., Rao, J. M., and Shroff, R. (2015). Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *Annals of Applied Statistics*.

[102] Gomez-Rodriguez, M., Leskovec, J., and Krause, A. (2010). Inferring Networks of Diffusion and Influence. In *KDD*.

[103] Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*.

[104] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

[105] Gordon, C. (2023). Ai ethicist views on chatgpt. **https://www.forbes.com/sites/cindygordon/2023/04/30/ai-ethicist-views-on-chatgpt/**.

[106] Goyal, A., Bonchi, F., Lakshmanan, L. V., and Venkatasubramanian, S. (2013). On minimizing budget and time in influence propagation over social networks. *Social network analysis and mining*, 3(2):179–192.

[107] Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. (2013). Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

[108] Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

[109] Grari, V., Lamprier, S., and Detyniecki, M. (2021). Fairness without the sensitive attribute via causal variational autoencoder. *arXiv preprint arXiv:2109.04999*.

[110] Green, B. and Chen, Y. (2019a). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *FAT**.

[111] Green, B. and Chen, Y. (2019b). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24.

[112] Grgić-Hlača, N., Castelluccia, C., and Gummadi, K. P. (2022a). Taking advice from (dis) similar machines: The impact of human-machine similarity on machine-assisted decision-making. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 74–88.

[113] Grgić-Hlača, N., Engel, C., and Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25.

[114] Grgić-Hlača, N., Lima, G., Weller, A., and Redmiles, E. M. (2022b). Dimensions of diversity in human perceptions of algorithmic fairness. *Proceedings of The second ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization*.

[115] Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., and Weller, A. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 world wide web conference*, pages 903–912.

[116] Grimstad, S. and Jørgensen, M. (2007). Inconsistency of expert judgment-based estimates of software development effort. *Journal of Systems and Software*, 80(11):1770–1777.

[117] Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey methodology*. John Wiley & Sons.

[118] Guillory, A. and Bilmes, J. A. (2011). Active semi-supervised learning using submodular functions. In *UAI*, pages 274–282.

[119] Guyomard, V., Fessant, F., Guyet, T., Bouadi, T., and Termier, A. (2022). Vcnet: A self-explaining model for realistic counterfactual generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 437–453. Springer.

[120] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*.

[121] Harmon-Jones, E. and Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In *Cognitive Dissonance, Second Edition: Reexamining a Pivotal Theory in Psychology*. American Psychological Association.

[122] Hayes, M. (2017). Incrementalism and public policy-making. In *Oxford Research Encyclopedia of Politics*.

[123] Hellevik, O. (2009). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity*, 43:59–74.

[124] Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3):217–223.

[125] Huang, K., Wang, S., Bevilacqua, G., Xiao, X., and Lakshmanan, L. V. (2017). Revisiting the stop-and-stare algorithms for influence maximization. *Proceedings of the VLDB Endowment*, 10(9):913–924.

[126] Huff, C. and Tingley, D. (2015). "who are these people?" evaluating the demographic characteristics and political preferences of mturk survey respondents. *Research & Politics*, 2(3):2053168015604648.

[127] Hüllermeier, E., Fürnkranz, J., Cheng, W., and Brinker, K. (2008). Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16-17):1897–1916.

[128] Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

[129] Ilgen, D. R., Fisher, C. D., and Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of applied psychology*, 64(4):349.

[130] Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. (2021). Openclip.

[131] Ilvento, C. (2019). Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*.

[132] Iosifidis, V., Fetahu, B., and Ntoutsi, E. (2019). Fae: A fairness-aware ensemble framework. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1375–1380. IEEE.

[133] Jesse, M. and Jannach, D. (2021). Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports*, 3:100052.

[134] Jiang, H. and Nachum, O. (2020). Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR.

[135] Johnson, K. (2022). Dall-e 2 creates incredible images—and biased ones you don't see. ***https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion***.

[136] Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. (2016). Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29.

[137] Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L., and Wu, Z. S. (2019). An algorithmic framework for fairness elicitation. *arXiv preprint arXiv:1905.10660*.

[138] Kahneman, D., Knetsch, J. L., and Thaler, R. (1986). Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *The American economic review*.

[139] Kahneman, D., Rosenfield, A. M., Gandhi, L., and Blaser, T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*.

[140] Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. Little, Brown Spark.

[141] Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decisions Under Risk. *Econometrica*.

[142] Kamiran, F. and Calders, T. (2009). Classifying without Discriminating. In *IC4*.

[143] Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems (KAIS)*.

[144] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2011). Fairness-aware Classifier with Prejudice Remover Regularizer. In *PADM*.

[145] Karkkainen, K. and Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*.

[146] Kay, M., Matuszek, C., and Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*.

[147] Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *KDD*.

[148] Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*.

[149] Khajehnejad, M., Rezaei, A. A., Babaei, M., Hoffmann, J., Jalili, M., and Weller, A. (2020). Adversarial graph embeddings for fair influence maximization over social networks. *arXiv preprint arXiv:2005.04074*.

[150] Khan, A. (2016). Towards time-discounted influence maximization. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1873–1876.

[151] Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

[152] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[153] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*.

[154] Kleindessner, M., Donini, M., Russell, C., and Zafar, B. (2023). Efficient fair pca for fair representation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[155] Koczkodaj, W. W. (1993). A new definition of consistency of pairwise comparisons. *Mathematical and computer modelling*, 18(7):79–84.

[156] Kourtellis, N., Alahakoon, T., Simha, R., Iamnitchi, A., and Tripathi, R. (2013). Identifying high betweenness centrality nodes in large social networks. *Social Network Analysis and Mining*, 3(4):899–914.

[157] Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., and Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862.

[158] Krause, A. and Golovin, D. (2014). Submodular function maximization.

[159] Krause, A. and Guestrin, C. (2007). Near-optimal observation selection using submodular functions. In *AAAI*, pages 1650–1654.

[160] Krause, J., Stark, M., Deng, J., and Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*.

[161] Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

[162] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

[163] Lahoti, P., Gummadi, K., and Weikum, G. (2019a). Operationalizing individual fairness with pairwise fair representations. *Proceedings of the VLDB Endowment*, 13(4):506–518.

[164] Lahoti, P., Gummadi, K. P., and Weikum, G. (2019b). ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*, pages 1334–1345. IEEE.

[165] Larochelle, H., Erhan, D., and Bengio, Y. (2008). Zero-data learning of new tasks. In *AAAI Conference on Artificial Intelligence*.

[166] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). Data and analysis for 'How we analyzed the COMPAS recidivism algorithm'. ***https://github.com/propublica/compas-analysis***.

[167] Lawrence, N. D. (2021). A retrospective on the 2014 NeurIPS experiment. ***http://inverseprobability.com/talks/notes/the-neurips-experiment.html***. Accessed: 2022-09-14.

[168] Lee, M. K., Jain, A., Cha, H. J., Ojha, S., and Kusbit, D. (2019a). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26.

[169] Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., et al. (2019b). Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–35.

[170] Lei, S., Maniu, S., Mo, L., Cheng, R., and Senellart, P. (2015). Online influence maximization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 645–654.

[171] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007). Cost-effective outbreak detection in networks. In *KDD*, pages 420–429. ACM.

[172] Leventhal, G. S. (1980). What should be done with equity theory? new approaches to the study of fairness in social relationships. In *Social exchange: Advances in theory and research*, pages 27–55. Springer.

[173] Li, H., Xu, M., Bhowmick, S. S., Sun, C., Jiang, Z., and Cui, J. (2019). Disco: Influence maximization meets network embedding and deep learning. *arXiv preprint arXiv:1906.07378*.

[174] Liang, W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv:2203.02053 [cs.CL]*.

[175] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

[176] Lin, H. and Bilmes, J. (2011). A class of submodular functions for document summarization. In *Proc. ACL: HLT - Vol. 1*, pages 510–520.

[177] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer.

[178] Ling, C., Jiang, J., Wang, J., Thai, M. T., Xue, R., Song, J., Qiu, M., and Zhao, L. (2023). Deep graph representation learning and optimization for influence maximization. In *International Conference on Machine Learning*, pages 21350–21361. PMLR.

[179] Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., and Parkes, D. C. (2017). Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*.

[180] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*.

[181] Loewenstein, G. and Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, 107(2):573–597.

[182] Logg, J. M. (2017). Theory of machine: When do people rely on algorithms? *Harvard Business School working paper series# 17-086*.

[183] Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

[184] London, M. (2003). *Job feedback: Giving, seeking, and using feedback for performance improvement*. Psychology Press.

[185] Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098.

[186] Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2015). The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

[187] Luong, B. T., Ruggieri, S., and Turini, F. (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510.

[188] Ma, Y., Frauen, D., Melnychuk, V., and Feuerriegel, S. (2023). Counterfactual fairness for predictions using generative adversarial networks. *arXiv preprint arXiv:2310.17687*.

[189] Malinin, A. (2019). *Uncertainty Estimation in Deep Learning with Application to Spoken Language Assessment*. PhD thesis, University of Cambridge.

[190] Margalit, Y. (2019). Political Responses to Economic Shocks. *Annual Review of Political Science*, 22:277–295.

[191] Marx, C. T., Calmon, F. d. P., and Ustun, B. (2019). Predictive multiplicity in classification. *arXiv preprint arXiv:1909.06677*.

[192] McAllister, R. (2017). *Bayesian learning for data-efficient control*. PhD thesis, Department of Engineering, University of Cambridge.

[193] McAuley, J. J. and Leskovec, J. (2012). Learning to discover social circles in ego networks. In *NIPS*, volume 2012, pages 548–56. Citeseer.

[194] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

[195] Meade, A. W. and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, 17(3):437.

[196] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

[197] Metz, C. (April 2022). Meet dall-e, the a.i. that draws anything at your command. ***https://www.nytimes.com/2022/04/06/technology/openai-images-dall-e.html***.

[198] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

[199] Miller, N., Resnick, P., and Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373.

[200] Mislove, A., Viswanath, B., Gummadi, K. P., and Druschel, P. (2010). You are who you know: inferring user profiles in online social networks. In *WSDM*, pages 251–260. ACM.

[201] Mokady, R., Hertz, A., and Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. *arXiv:2111.09734 [cs.CV]*.

[202] Narimanzadeh, H., Badie-Modiri, A., Smirnova, I. G., and Chen, T. H. Y. (2023). Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–29.

[203] Nemhauser, G., Wolsey, L., and Fisher, M. (1978). An analysis of the approximations for maximizing submodular set functions. *Math. Prog.*, 14:265–294.

[204] OpenAI (2022). ***https://openai.com/blog/chatgpt#OpenAI***.

[205] Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

[206] Panagopoulos, G., Tziortziotis, N., Malliaros, F. D., and Vazirgiannis, M. (2021). Learning graph representations for influence maximization. *arXiv preprint arXiv:2108.04623*.

[207] Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*.

[208] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*.

[209] Patro, G. K., Chakraborty, A., Ganguly, N., and Gummadi, K. (2020). Incremental fairness in two-sided market platforms: On smoothly updating recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 181–188.

[210] Pawelczyk, M., Broelemann, K., and Kasneci, G. (2020). On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pages 809–818. PMLR.

[211] Perrigo, B. (August 2021). An artificial intelligence helped write this play. it may contain racism. ***https://time.com/6092078/artificial-intelligence-play/***.

[212] Petersen, F., Mukherjee, D., Sun, Y., and Yurochkin, M. (2021). Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955.

[213] Pierson, E. (2017). Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*.

[214] Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE international conference on computer vision*.

[215] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52.

[216] Prelec, D. (2004). A bayesian truth serum for subjective data. *science*, 306(5695):462–466.

[217] Qiu, J., Zhu, Y., Shi, X., Wenzel, F., Tang, Z., Zhao, D., Li, B., and Li, M. (2022). Are multimodal models robust to image and text perturbations? *arXiv:2212.08044 [cs.CV]*.

[218] Radanovic, G. and Faltings, B. (2013). A robust bayesian truth serum for non-binary signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 833–839.

[219] Radanovic, G. and Faltings, B. (2014). Incentives for truthful information elicitation of continuous signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

[220] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR.

[221] Rahmattalabi, A., Vayanos, P., Fulginiti, A., Rice, E., Wilder, B., Yadav, A., and Tambe, M. (2019). Exploring algorithmic fairness in robust graph covering problems. In *Advances in Neural Information Processing Systems*, pages 15750–15761.

[222] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*.

[223] Redmiles, E. M., Acar, Y., Fahl, S., and Mazurek, M. L. (2017). A Summary of Survey Methodology Best Practices for Security and Privacy Researchers. Technical report, University of Maryland.

[224] Rezvanian, A., Vahidipour, S. M., and Meybodi, M. R. (2023). A new stochastic diffusion model for influence maximization in social networks. *Scientific Reports*, 13(1):6122.

[225] Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *KDD*.

[226] Roberge, M.-É. and Van Dick, R. (2010). Recognizing the benefits of diversity: When and how does diversity increase group performance? *Human Resource management review*, 20(4):295–308.

[227] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[228] Ross, J., Irani, L., Silberman, M., Zaldivar, A., and Tomlinson, B. (2010). Who Are The Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI*.

[229] Schick, A. (1983). Incremental budgeting in a decremental age. *Policy Sciences*, pages 1–25.

[230] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114 [cs.CV]*.

[231] Schumann, C., Ricco, S., Prabhu, U., Ferrari, V., and Pantofaru, C. R. (2021). A step toward more inclusive people annotations for fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

[232] Seth, A., Hemani, M., and Agarwal, C. (2023). Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[233] Shen, X., Diamond, S., Gu, Y., and Boyd, S. (2016). Disciplined Convex-Concave Programming. *arXiv:1604.02639*.

[234] Singla, A., Horvitz, E., Kohli, P., White, R., and Krause, A. (2015). Information gathering in networks via active exploration. In *IJCAI*, pages 891–988.

[235] Singla, A. and Weber, I. (2009). Camera brand congruence in the flickr social graph. In *WSDM*, pages 252–261.

[236] Steelman, L. A. and Rutkowski, K. A. (2004). Moderators of employee reactions to negative feedback. *Journal of Managerial Psychology*, 19(1):6–18.

[237] Stewart, N., Brown, G. D., and Chater, N. (2005). Absolute identification by relative judgment. *Psychological review*, 112(4):881.

[238] Stoica, A.-A. and Chaintreau, A. (2019). Fairness in social influence maximization. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 569–574.

[239] Stoica, A.-A., Riederer, C., and Chaintreau, A. (2018). Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*, pages 923–932.

[240] Sunahase, T., Baba, Y., and Kashima, H. (2017). Pairwise hits: Quality estimation from pairwise comparisons in creator-evaluator crowdsourcing process. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

[241] Supreme Court of the United States (1989). Martin vs. Wilks.

[242] Taylor, P. W. (1973). Reverse discrimination and compensatory justice. *Analysis*, 33(6):177–182.

[243] Taylor, R. L. and Wilsted, W. D. (1974). Capturing judgment policies: A field study of performance appraisal. *Academy of Management Journal*, 17(3):440–449.

[244] Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. *Economics letters*, 8(3):201–207.

[245] Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin.

[246] Tran, D., Valtchanov, A., Ganapathy, K., Feng, R., Slud, E., Goldblum, M., and Goldstein, T. (2020). An open review of openreview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*.

[247] Tsang, A., Wilder, B., Rice, E., Tambe, M., and Zick, Y. (2019). Group-Fairness in Influence Maximization. *arXiv preprint arXiv:1903.00967*.

[248] Urbany, J. E., Madden, T. J., and Dickson, P. R. (1989). All's not fair in pricing: an initial look at the dual entitlement principle. *Marketing Letters*.

[249] U.S. Census Bureau (2019). American Community Survey 5-Year Estimates.

[250] Vaidya, O. S. and Kumar, S. (2006). Analytic hierarchy process: An overview of applications. *European Journal of operational research*, 169(1):1–29.

[251] Vallat, R. (2018). Pingouin: statistics in python. *The Journal of Open Source Software*, 3(31):1026.

[252] Van Leeuwen, F., Koenig, B. L., Graham, J., and Park, J. H. (2014). Moral concerns across the united states: Associations with life-history variables, pathogen prevalence, urbanization, cognitive ability, and social class. *Evolution and Human Behavior*, 35(6):464–471.

[253] Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[254] Verma, P. and De Vynck, G. (2023). Chatgpt took their jobs. now they walk dogs and fix air conditioners. **https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/**.

[255] Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *ACM/IEEE International Workshop on Software Fairness*.

[256] Vincent, J. (2023). Getty images sues ai art generator stable diffusion in the us for copyright infringement. **https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion**.

[257] Wachter, S., Mittelstadt, B., and Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567.

[258] Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *International AAAI Conference on Weblogs and Social Media*.

[259] Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology*, 160(6):509–516.

[260] Wang, A., Barocas, S., Laird, K., and Wallach, H. (2022). Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 324–335.

[261] Wang, H., Grgic-Hlaca, N., Lahoti, P., Gummadi, K. P., and Weller, A. (2019). An empirical study on learning fairness metrics for compas data with human supervision. *arXiv preprint arXiv:1910.10255*.

[262] Wang, J., Liu, Y., and Wang, X. E. (2021a). Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv:2109.05433 [cs.CV]*.

[263] Wang, J., Liu, Y., and Wang, X. E. (2021b). Assessing multilingual fairness in pre-trained multimodal representations. *arXiv:2106.06683 [cs.CL]*.

[264] Wang, X. and Yin, M. (2021). Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328.

[265] Wanous, J. P. and Youtz, M. A. (1986). Solution diversity and the quality of groups decisions. *Academy of Management journal*, 29(1):149–159.

[266] Welinder, P., Branson, S., Belongie, S. J., and Perona, P. (2010). The Multidimensional Wisdom of Crowds. In *NIPS*, volume 23, pages 2424–2432.

[267] Wenzel, F., Dittadi, A., Gehler, P. V., Simon-Gabriel, C.-J., Horn, M., Zietlow, D., Kernert, D., Russell, C., Brox, T., Schiele, B., Schölkopf, B., and Locatello, F. (2022). Assaying out-of-distribution generalization in transfer learning. In *Neural Information Processing Systems (NeurIPS)*.

[268] Wong, M. (2023). Chatgpt is already obsolete. **https://www.theatlantic.com/technology/archive/2023/05/ai-advancements-multimodal-models/674113/**.

[269] Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR.

[270] Xu, D., Yuan, S., Zhang, L., and Wu, X. (2018). Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE.

[271] Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12.

[272] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

[273] Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K. P., and Weller, A. (2017a). From Parity to Preference-based Notions of Fairness in Classification. In *NIPS*.

[274] Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017b). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*.

[275] Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017c). Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.

[276] Zehlike, M., Yang, K., and Stoyanovich, J. (2021). Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000*.

[277] Zelaya, V., Missier, P., and Prangle, D. (2019). Parametrised data sampling for fairness optimisation. *KDD XAI*.

[278] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR.

[279] Zhang, Y. (2018). Deep generative model for multi-class imbalanced learning.

[280] Zhang, Y., Liao, Q. V., and Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *arXiv preprint arXiv:2001.02114*.

[281] Zhao, D., Wang, A., and Russakovsky, O. (2021). Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.