DADI: Dynamic Discovery of Fair Information with Adversarial Reinforcement Learning

Michiel A. Bakker
Massachusetts Institute of Technology
MIT-IBM Watson AI Lab
bakker@mit.edu

Duy Patrick Tu*
Massachusetts Institute of Technology
Ludwig-Maximilians-Universität München
patrick2@mit.edu

Humberto Riverón Valdés
Massachusetts Institute of Technology
MIT-IBM Watson AI Lab
hriveron@mit.edu

Krishna P. Gummadi
Max Planck Inst. for Software Systems
gummadi@mpi-sws.org

Kush R. Varshney
IBM Research
MIT-IBM Watson AI Lab
krvarshn@us.ibm.com

Adrian Weller
University of Cambridge
Alan Turing Institute
aw665@cam.ac.uk

Alex 'Sandy' Pentland
Massachusetts Institute of Technology
MIT-IBM Watson AI Lab
pentland@mit.edu

Abstract

We introduce a framework for dynamic adversarial discovery of information (DADI), motivated by a scenario where information (a feature set) is used by third parties with unknown objectives. We train a reinforcement learning agent to sequentially acquire a subset of the information while balancing accuracy and fairness of predictors downstream. Based on the set of already acquired features, the agent decides dynamically to either collect more information from the set of available features or to stop and predict using the information that is currently available. Building on previous work exploring adversarial representation learning, we attain group fairness (demographic parity) by rewarding the agent with the adversary's loss, computed over the final feature set. Importantly, however, the framework provides a more general starting point for fair or private dynamic information discovery. Finally, we demonstrate empirically, using two real-world datasets, that we can trade-off fairness and predictive performance.

1 Introduction

There are two parties involved in information transfer: a *data owner* who has ownership over its own data or data it holds on behalf of others and a *data collector* who is tasked with collecting the most informative set of data, often to maximize the performance of some predictor downstream. Intentionally or otherwise, this process of data collection and prediction can lead to biases that unfairly favor one protected subgroup over another. Numerous recent studies have shown that naively optimizing for predictive performance can lead to unfair prediction outcomes in high-stake domains such as criminal justice, credit assessment, recruiting, and healthcare [12, 2, 8, 19].

Consequently, the data owner faces a critical decision: if it cannot trust the data collector, which information should it share to ensure fair decision making? While the optimal strategy to maximize

*Duy Patrick Tu did this work while visiting MIT from LMU Munich.

Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. (non-archival)

predictive performance is to naively share all the data available, the data owner has to be more careful when it wants to ensure that the predictions downstream are fair. Removing the sensitive attribute is the most obvious strategy, but is ineffective when the attribute is redundantly encoded in other features [3]. Another strategy is to first apply *fair feature selection* in which one formulates an optimization problem to select a subset of features that maximizes accuracy, given a maximum unfairness constraint [6]. This strategy, though effective, is inefficient as it removes each feature simultaneously for all individuals, ignoring any differences in the underlying conditional dependencies. For example, for individuals that live in Chicago, the most racially segregated city in America, zipcode will be highly correlated with race and using this feature can thus lead to racially biased predictions [15]. In contrast, if an individual lives in Irvine, California, America's most racially integrated city, zipcode alone will not reveal an individual's race. Removing zipcode for all individuals is therefore an effective but inefficient strategy to ensure fairness.

Motivated by this problem, we propose the DADI (Dynamic Adversarial Discovery of Information) framework as a general sequential information acquisition framework for any task. Our contributions are as follows: to the best of our knowledge, we introduce the first framework for dynamic adversarial discovery of information which we utilize to acquire feature sets that ensure fair decision making. In this framework, we formulate the feature acquisition task as a minimax optimization problem in which a reinforcement learning (RL) agent simultaneously minimizes the classification loss while maximizing the loss of an adversary. We actualize this with a joint framework that simultaneously trains a classifier, an adversary, and an RL agent using deep Q-learning. Building on work on adversarial representation learning, we investigate the effects of two different adversarial reward functions to achieve *demographic parity* [4, 17]. Finally, we demonstrate the effectiveness of our framework with two real-world public datasets.

2 Related Work

Fairness Recent years have seen an explosion in academic work that seeks to define and obtain fairness in automated decision making systems. At a high level, this literature has focused on two families of definitions: *statistical* notions of fairness and *individual* notions of fairness [3, 22]. Most of the literature, including this work, focuses on statistical or group definitions of fairness, in which we require parity of some statistical measure to hold across a small number of protected subgroups. In contrast, individual fairness definitions have no notion of protected subgroups, but instead formulate constraints that bind on pairs of individuals [3, 11]. Both families of definitions have strengths and weaknesses; statistical notions are easy to verify but do not provide any guarantees to individuals, while individual notions do give individual guarantees but are difficult to implement in practice and are ambiguous with respect to the agreed-upon distance function.

In this work, we focus on *demographic parity*, requiring parity of the positive classification rate across groups, i.e. $P(\hat{y} = 1 \mid b = 0) = P(\hat{y} = 1 \mid b = 1)$, where $\hat{y} \in \{0, 1\}$ is the binary prediction of a model that classifies feature set $\mathbf{x}$ and $b \in \{0, 1\}$ is the sensitive attribute. The usefulness of demographic parity is limited when the base rate differs across groups, i.e. $P(y = 1 \mid b = 0) \neq (y = 1 \mid b = 1)$ where $y \in \{0, 1\}$ is the ground truth label. In that case, the metric can be generalized by conditioning on the ground truth label, yielding equal false negative rates (*equal opportunity*) or equal false negative and false positive rates (*equal odds*) as measures of fairness [7]. We demonstrate the effectiveness of our framework using demographic parity, but note that alternative adversarial objectives have been introduced that can be combined with our framework to achieve equal opportunity or equal odds [17].

Adversarial learning Adversarial learning for deep generative models was introduced by [5], framing the learning as a two-player game between a generator and a discriminator. The generator aims to fool the discriminator by generating fake data that resembles data from a dataset X while the discriminator is trained to distinguish between 'real' data from and 'fake' data generated by the generator. Learning proceeds using a minimax optimization where the generator and discriminator are optimized jointly. At each iteration, the discriminator improves its ability to discriminate between real and fake which, in turn, forces the generator to generate fake data that better resembles the real data.

Adversarial learning was first applied in the context of fairness by [4], proposing adversarial learning to ensure that multiple distinct data distributions from different demographic subgroups are modeled as a single representation. The discriminator aims to distinguish between subgroups while an encoder aims to map each data distribution to a single representation to fool the discriminator. Subsequently, these representations can be safely shared with a data collector while ensuring demographic parity for predictions downstream. [1] further explores this approach in the context of demographically imbalanced data. Finally, [25] and [17] extend this body of work by connecting multiple statistical notions of fairness to different adversarial objectives. Whereas the method presented in [25] predicts the sensitive attribute from the prediction of the classifier, our work is closer to the method in [17], working directly with the learned representation. This allows for transferable representations that ensure fair outcomes for other third-party classifiers downstream.

Although this work is similar in spirit to adversarial representation learning, we aim to dynamically collect a fair subset of features instead of learning to map the full feature set to a fair representation. The ability to collect raw features instead of mapping to a representation is crucial for integration with current information systems where the collected information is used or audited by both human and machine decision makers downstream. If we consider our the example of credit assessment, a bank not only wants to collect a low-level abstract representation for the purpose of the initial creditworthiness prediction but also wants to store the applicant's information in a database to provide other services downstream or allow for audits.

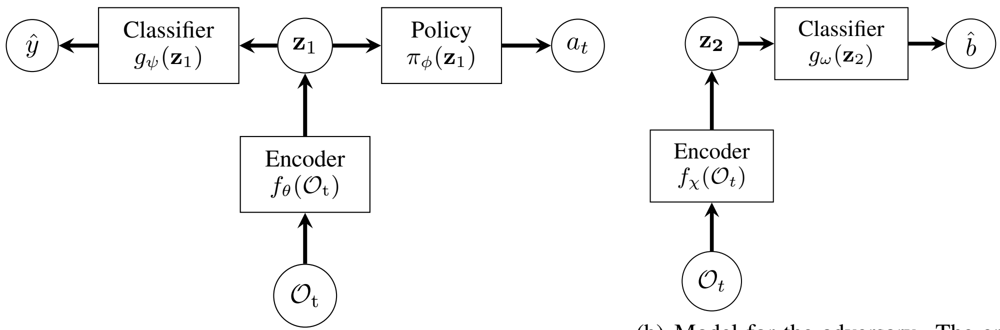
Active feature-value acquisition Different from *active learning*, active feature-value acquisition (AFA) is concerned with feature-wise active learning for each instance. AFA is of great need in cost-sensitive applications where the data collector needs to balance an available information budget with predictive accuracy. A traditional AFA system consists of three components: 1) a classifier that can handle partially observed feature sets, 2) a strategy for determining which feature to select next based on the features that are already collected, and 3) a stopping criterion for determining when to stop acquiring more features and make a final prediction.

First, there are different ways a classifier can handle a partial features set. Generative models handle missing features naturally by first integrating out the variables while in discriminate models feature imputation or expectation-maximization can be used to first replace the missing values with estimates. In this work, we use a set encoder based on [23] to encode arbitrary subsets of features. Second, to determine which feature to select next, we need a method that estimates the value of each of the unselected features based on the features that we have already collected. A recent approach, Efficient Dynamic Discovery of High-Value Information (EDDI), uses a partial variational autoencoder to represent the set of already acquired features. It then computes the mutual information between the current representation and each of the available features to select the feature that minimizes this information [16]. Finally, a stopping criterion is not specified in EDDI and most other AFA methods. However most prior work assumes a fixed feature budget per individual after which the process terminates [13]. The *active fairness* framework extends this to group-specific budgets that are found to attain equal opportunity (equal false-positive or false-negative rates) [18].

To effectively trade-off fairness and accuracy, we need a unified optimization framework that includes both the acquisition strategy and the stopping criterion. We adopt the framework from [21] and model the feature acquisition process as a Markov decision process (MDP) where the action space consists of the set of unselected set of features and an additional STOP action[21] which, upon selection, terminates the acquisition process. To ensure fairness, we formulate a reward function that balances low classification loss with a high adversarial loss.

3 Adversarial Discovery of Fair Information

Problem setup The setup of our framework most follows the joint active feature acquisition and classification framework in [21]; however, we extend their framework for use with an adversary. Let $(\mathbf{x}^{(i)}, y^{(i)}, b^{(i)}) \sim P$ be individual i in P represented by a d-dimensional feature vector $\mathbf{x}^{(i)} \subseteq \mathbb{R}^d$, a binary label $y^{(i)} \in \{0, 1\}$, and a binary sensitive attribute $b^{(i)} \in \{0, 1\}$. We acquire the features in sequential order starting with an empty set $\mathcal{O}_0 := \emptyset$ at time $t = 0$. At every later timestep t, we choose a subset of features from the unselected set of features, $\mathbf{S}_t^{(i)} \subseteq \{1, \ldots, d\} \setminus \mathcal{O}_{t-1}^{(i)}$. After each new acquisition step, the classifier will have access to feature values in $\mathcal{O}_t^{(i)} := \mathcal{S}_t^{(i)} \cup \mathcal{O}_{t-1}^{(i)}$. We keep acquiring features up to time $T^{(i)}$ when we meet a stopping criterion. At that point, we will classify $\mathbf{x}^{(i)}$ using only the set of features in $\mathcal{O}_{T^{(i)}}^{(i)}$. Note that the specific set of selected features $\mathcal{O}_{T^{(i)}}^{(i)}$ will



(a) Joint model for the label classifier and policy. The encoder f_θ maps the set of observed features $\mathcal{O}_t$ to the latent representation $\mathbf{z}_1$. From this the classifier g_ψ predicts $\hat{y}$ while the policy π_ϕ predicts the action a_t (the next feature to select or STOP).

(b) Model for the adversary. The encoder f_χ maps the set of observed features $\mathcal{O}_t$ to the latent representation $\mathbf{z}_2$ from which the classifier g_ω predicts the sensitive attribute $\hat{b}$.

Figure 1: Joint framework for dynamic adversarial discovery of information (DADI)

generally be different for each individual i. To learn the model that minimizes classification loss while maximizing the loss of the adversary we formulate the following optimization problem.

$$\underset{\psi,\theta,\omega,\chi}{\text{minimize}} \frac{1}{|P|} \sum_{i \in P} (1-\gamma)\mathcal{L}_C \left(g_\psi(f_\theta(\mathcal{O}_T^{(i)}), y^{(i)} \right) - \gamma \mathcal{L}_A \left(g_\omega(f_\chi(\mathcal{O}_T^{(i)}), b^{(i)} \right) \tag{1}$$

Where $\mathcal{L}_C$ and $\mathcal{L}_A$ are the suitable losses for the label classifier and the adversary. The encoder f_θ feeds into a classifier g_ψ for the label prediction $\hat{y}$ while f_χ and g_ω are the encoder and classifier for the sensitive attribute prediction $\hat{b}$. Hyperparameter γ specifies the desired balance between classification performance and fairness. When clear from context, we drop the superscript (i).

Markov decision process We define a Markov decision process (MDP) to find the set of features $\mathcal{O}_T^{(i)}$ that minimizes the objective in Eq. (1). For each episode, the state at time t is represented by the set of selected features $\{x_j\}_{j \in \mathcal{O}_t}$. The size of the state space is 2^d, the powerset of the feature set. At each timestep t, the action space consists of the set of unselected features $\{1, \ldots, d\} \setminus \mathcal{O}_{t-1}$ and an additional STOP action which, upon selection, stops the acquisition process after which the rewards are computed. The agent's reward function, computed at end of the episode for individual i, corresponds to

$$r(\mathcal{O}_T^{(i)}) = -(1-\gamma)\mathcal{L}_C\big(g_\psi(f_\theta(\mathcal{O}_T^{(i)}), y^{(i)})\big) + \gamma \mathcal{L}_A\big(g_\omega(f_\chi(\mathcal{O}_T^{(i)}), b^{(i)})\big) \tag{2}$$

(3)

where the first reward encourages accurate classification and the second reward encourages low mutual information between the feature set and the sensitive attribute. Now, if we now consider a policy π_ϕ^*, parametrized by ϕ, that is optimal for this MDP, then π_ϕ^* is also the optimal solution to the objective in Eq. (1). We can proof this by maximizing the aggregated reward in Eq. (2) over the population P

$$\arg\max_{\phi} \frac{1}{|P|} \sum_{i \in P} -(1-\gamma)\mathcal{L}_C\big(g_\psi(f_\theta(\mathcal{O}_{T_\phi}^{(i)}), y^{(i)})\big) + \gamma \mathcal{L}_A\big(g_\omega(f_\chi(\mathcal{O}_{T_\phi}^{(i)}), b^{(i)})\big) \tag{4}$$

$$= \arg\min_{\phi} \frac{1}{|P|} \sum_{i \in P} (1-\gamma)\mathcal{L}_C\big(g_\psi(f_\theta(\mathcal{O}_{T_\phi}^{(i)}), y^{(i)})\big) - \gamma \mathcal{L}_A\big(g_\omega(f_\chi(\mathcal{O}_{T_\phi}^{(i)}), b^{(i)})\big) \tag{5}$$

which is equivalent to the minimization objective in Eq. (1).

Generalized framework The generalized framework in Fig. 1 consists of two parts: the first part in Fig. 1a seeks to learn a representation of the set of observed features $\mathbf{z_1} = f_\theta(\mathcal{O}_t)$ capable of classifying the label $\hat{y} = g_\psi(\mathbf{z_1})$ and estimating the optimal next action $a_t = \pi_\phi(\mathbf{z_1})$. The model has two heads that share the same encoder which leads to improved performance over a model with two separate encoders [21]. In parallel, the second network in Fig. 1b seeks to learn a related but separate representation $\mathbf{z_2} = f_\chi(\mathcal{O}_t)$, which is fed to a classifier g_ω that predicts the sensitive attribute $\hat{b}$. Crucially and different from prior work on adversarial representation learning, the second adversarial classification task cannot have a shared encoder with the first two tasks as this could encourage the encoder to mask the unfairness of features directly which, in turn, would not lead to selecting a set of fair features that generalize to any downstream task. While in adversarial representation learning the adversarial loss is backpropagated directly through a gradient reversal layer to update the encoder [5, 4], our agent learns to fool the adversary by selecting the set of features that maximize the adversarial classification loss.

We realize f_θ, g_ψ, f_χ, g_ω and π_ϕ as neural networks parametrized by θ, ψ, χ, ω, and ϕ, which are optimized using alternating gradient descent steps. To facilitate encoding of partially observed feature sets, we adopt a feature-level set encoder [21]. Each observed feature x_i is first mapped to a memory vector $\mathbf{m}_i$ after which an LSTM processes the set of memory vectors repeatedly while an attention layer improves the set embedding. The attention step ensures the input is order-invariant. The final set embedding $\mathbf{z}_1$ is fed to both the classifier and the policy network. A second independent set embedding $\mathbf{z}_2$ is fed to the adversary. We refer to the SM for details on the set encoding process and the implementation.

Adversarial fairness We compare two different loss functions to compute the rewards for the adversary. First, earlier work on adversarial fair representation learning for demographic parity has shown that using binary cross-entropy (CE) loss for both the classifier and the adversary encourages fair and high-value representations [4, 1]

$$\mathcal{L}_A^{CE} = -\left(b \log(g_\omega(f_\chi(\mathcal{O}_T))) + (1-b)\log(1 - g_\omega(f_\chi(\mathcal{O}_T)))\right) \tag{6}$$

where the adversary only has access to the final feature set $\mathcal{O}_T$ obtained after stopping. Though effective, $\mathcal{L}_A^{CE}$ fails to account for demographically unbalanced training data. To address this problem, [17] introduces group-normalized L_1 (GNL_1) loss as a more natural relaxation of demographic parity, which we adopt to compute the rewards from the adversary

$$\mathcal{L}_A^{GNL_1} = \frac{|P|}{2|P_b|}|g_\omega(f_\chi(\mathcal{O}_T)) - b| \tag{7}$$

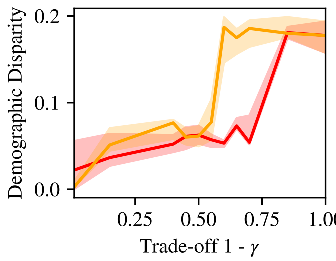
where P_0 and P_1 are the protected subgroups with respectively attributes $b = 0$ and $b = 1$. As neural networks have difficulty learning with L_1 loss [10], we continue to use cross-entropy loss to train the adversary but use $\mathcal{L}_A^{GNL_1}$ to compute the final rewards for the agent. We refer to [17] for the theoretical properties of both loss functions.

4 Experiments

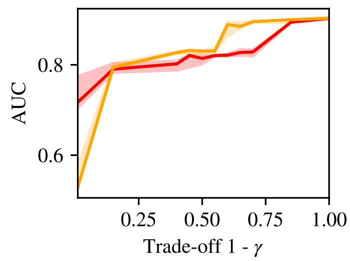
DADI seeks to select the subset of features that can be used by third parties with the assurance that their trained classifiers are both fair and accurate. As exact demographic parity is hard to enforce in practice, we use demographic disparity $|P(\hat{y} = 1 \mid b = 0) - P(\hat{y} = 1 \mid b = 1)|$ as measure for the degree of unfairness. The performance of the classifier is measured using the Area Under the Receiver Operating Characteristics curve (AUC) to account for the imbalanced label distributions.

Datasets We evaluate DADI empirically on the UCI Adult [14] and Mexican Poverty [18] datasets. We use one-hot encoding for categorical features and standardize numerical features. For the mapping to actions, we combine multiple one-hot encoded binary features that stem from the same categorical feature into a single action (e.g. the binary features *marital=divorced, marital=married* and *marital=single* correspond to a single action that acquires these features simultaneously). We use 8-fold cross validation with a random $87.5\%/12.5\%$ train-val/test split. We further split the train-val set into training and validation data using a second random $80\%/20\%$ split.

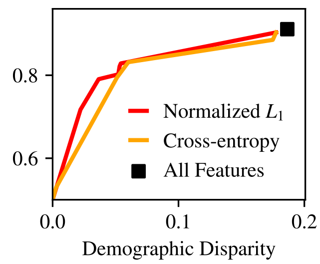
The Adult Dataset from UCI Machine Learning Repository [14] comprises 14 demographic and occupational attributes, which translates after preprocessing into 98 continuous and binary features and 14 actions for 48,842 individuals, with the goal of classifying whether a person's income is above \$50,000 (25% are above). Rows with missing values are omitted resulting in a dataset with 45,222 samples. In line with previous work, we use gender as the sensitive attribute, listed as male or female.



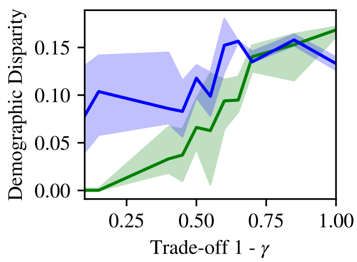
(a) Adult Income: Disparity-$(1-\gamma)$



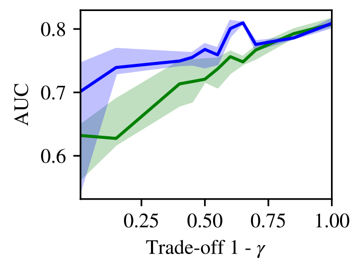
(b) Adult Income: AUC-$(1-\gamma)$



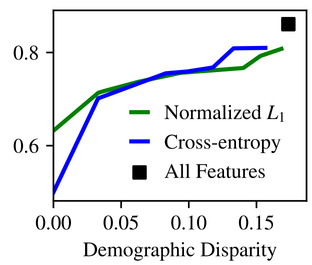
(c) Adult Income: AUC-disparity



(d) Mex. Poverty: Disparity-$(1-\gamma)$



(e) Mex. Poverty: AUC-$(1-\gamma)$



(f) Mex. Poverty: AUC-disparity

Figure 2: DADI for mitigating demographic disparity across subgroups in the Adult and Mexico datasets. Subfigures (a),(b), (d) and (e) show respectively the AUC and disparity for a range of trade-off parameters $1-\gamma$. The lines are plotted using the median with first/third quantile as confidence area computed using 8-fold CV. Subfigures (c) and (f) show the Pareto front along the AUC-disparity trade-off. The black square represents the baseline unfair classifier for which we use the pretrained classifier together with the full feature set. The median AUC and disparity are again computed across the 8 folds.

The Mexican Poverty dataset is extracted from the Mexican household survey 2016, which contains ground-truth household poverty levels and 99 attributes, related to household information such as the number of rooms or the type of heating system [9]. The processed dataset is obtained from [18] and comprises a sample of 70,305 households in Mexico, with 183 continuous and one-hot encoded binary features and 99 actions. Classification is binary according to the country's official poverty line, with 36% of the households having the label poor. The considered sensitive attribute describes whether the head of the household is a senior citizen or not.

Results Fig. 2 shows the results for both datasets. First, Figs. 2(a), 2(b), 2(d) and 2(e) show that increasing $1-\gamma$, i.e., decreasing the relative weight of the adversarial reward γ, leads to an increase in both performance and disparity for both choices of adversarial reward functions. Naturally, as the adversarial reward becomes less important, the agent will have a stronger incentive to maximize the accuracy which, in turn, leads to the collection of more features and thus higher AUC at the cost of a higher disparity.

Importantly, however, we observe that while the AUC increases drastically from the start, demographic parity only increases drastically for larger values of $1-\gamma$, allowing for agents that achieve good predictive performance with minimal disparity loss. This conclusion is supported by Figs. 2(c) and 2(f) where we visualize the Pareto front along the AUC-disparity trade-off. These results are encouraging as we show that a data collector can still maintain good performance while only having access to a unique fair subset of features for each data owner. Finally, we observe that the group-normalized L_1 reward generally results in a better trade-off, especially in the most important fairness range for small values of the disparity.

5 Conclusion and Future Work

A number of recent works have focused on adversarially learning fair representations. However, the methods underlying these works, are ineffective when the data owner is required to share raw features, a key aspect in many use cases where features are collected for both human and machine decision making. To tackle this problem, we propose DADI, to our knowledge the first framework for dynamic adversarial discovery of fair information. We frame the data owner's choice as a reinforcement learning problem where an agent selects a subset of features while an adversary critiques potentially unfair feature sets. Experimentally, we demonstrate how our framework guides information discovery for ensuring demographic parity and how it allows the data owner to efficiently trade-off fairness and accuracy.

Importantly, however, our framework is more generally applicable in settings where a data owner may wish to guard itself against a naive or malicious data collector by sharing only a subset of features. First, by changing the adversarial objective function, the framework in [17] demonstrates that one can achieve other notions of fairness such as equal opportunity and equal odds. Second, several recent works have formulated adversarial objectives to attain (differentially) private data representations. These objectives could be adopted using DADI to automate dynamic discovery of private information [24, 20] which could be further extended by encoding features in different levels of precision (such as age by year or age by decade), allowing the agent to select the level of precision that maximizes accuracy while minimizing privacy risk. Finally, adding monetary acquisition costs of features as a penalty at each collection step would allow our agent to holistically trade-off accuracy, information costs, and fairness or privacy [21].

Acknowledgements

The authors want to thank IBM for access to their compute resources and Prasanna Sattigeri for helpful discussions. AW acknowledges support from the David MacKay Newton research fellowship at Darwin College, The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074, and the Leverhulme Trust via the CFI.

References

[1] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[2] Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5), 2016.

[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[4] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *The International Conference on Learning Representations (ICLR)*, 2016.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[7] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, page 3315, 2016.

[8] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856, 2007.

[9] Pablo Ibarrarán, Nadin Medellín, Ferdinando Regalia, Marco Stampini, Sandro Parodi, Luis Tejerina, Pedro Cueva, and Madiery Vásquez. How conditional cash transfers work. 2017.

[10] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *Conference on Theoretical Foundations of Machine Learning (TFML)*, 2017.

[11] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[12] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science (ITCS)*, 2017.

[13] Balaji Krishnapuram, Shipeng Yu, and R Bharat Rao. *Cost-sensitive Machine Learning*. CRC Press, 2011.

[14] Moshe Lichman et al. Uci machine learning repository, 2013.

[15] John Logan. *Diversity and disparities: America enters a new century*. Russell Sage Foundation, 2014.

[16] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. In *International Conference on Machine Learning*, pages 4234–4243, 2019.

[17] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3381–3390, 2018.

[18] Alejandro Noriega-Campero, Michiel Bakker, Bernardo Garcia-Bulle, and Alex Pentland. Active fairness in algorithmic decision making. *Proceedings of AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society*, 2019.

[19] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[20] NhatHai Phan, Ruoming Jin, My T Thai, Han Hu, and Dejing Dou. Preserving differential privacy in adversarial learning with provable robustness. *arXiv preprint arXiv:1903.09822*, 2019.

[21] Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in Neural Information Processing Systems*, pages 1368–1378, 2018.

[22] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[23] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *The International Conference on Learning Representations (ICLR)*, 2015.

[24] Tsung-Yen Yang, Christopher Brinton, Prateek Mittal, Mung Chiang, and Andrew Lan. Learning informative and private representations via generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1534–1543. IEEE, 2018.

[25] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.