Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias

Sierra Wyllie
sierra.wyllie@mail.utoronto.ca
University of Toronto & Vector Institute
Toronto, Canada

Ilia Shumailov
University of Oxford
UK

Nicolas Papernot
University of Toronto & Vector Institute
Torotno, Canada

ABSTRACT

Model-induced distribution shifts (MIDS) occur as previous model outputs pollute new model training sets over generations of models. This is known as *model collapse* in the case of generative models, and *performative prediction* or *unfairness feedback loops* for supervised models. When a model induces a distribution shift, it also encodes its mistakes, biases, and unfairnesses into the ground truth of its data ecosystem. We introduce a framework that allows us to track multiple MIDS over many generations, finding that they can lead to loss in performance, fairness, and minoritized group representation, even in initially unbiased datasets. Despite these negative consequences, we identify how models might be used for positive, intentional, interventions in their data ecosystems, providing redress for historical discrimination through a framework called algorithmic reparation (AR). We simulate AR interventions by curating representative training batches for stochastic gradient descent to demonstrate how AR can improve upon the unfairnesses of models and data ecosystems subject to other MIDS. Our work takes an important step towards identifying, mitigating, and taking accountability for the unfair feedback loops enabled by the idea that ML systems are inherently neutral and objective.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; *Learning settings.*

KEYWORDS

Machine learning, algorithmic bias, fair machine learning, Intersectionality, artificial intelligence, distribution shift

ACM Reference Format:

Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil*. ACM, New York, NY, USA, 35 pages. https://doi.org/10.1145/3630106.3659029

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659029

1 INTRODUCTION

Fairness feedback loops have posed problems for both machine learning practitioners' models and societies' policies for some time. One example are the 1930s Home Owner Loan Corporation Security Maps rediscovered by historian Kenneth T. Jackson in the 1980s. These depict 'redlining,' where minoritized communities (especially Black and Jewish people) were discriminated against in housing in the United States [2, 36, 51]. These maps were likely used by government and banks to determine which neighborhoods should be provided programs and loans, feeding a feedback loop of segregation, limited Black home ownership, environmental racism, and increasing median household income gap between Black and white families [23, 61]. More recently, automated systems used for policies such as loan eligibility and approval prediction [72] risk further entrenchment of inequitable feedback loops [18].

In the machine learning fairness community, this effect is known as performative prediction [55] or fairness feedback loops [46], where the errors and behaviors of a model influence its future inputs, causing runaway unfairness [24]. In economics this is known as the performativity thesis, where economic theories attempting to describe markets instead shape them [11]. Increased attention to these effects and the proliferation of generated content on the internet, has created terminology for a dataset 'ecosystem.' These ecosystems may suffer from 'synthetic data spills,' such as unrealistic AI-generated images of baby peacocks polluting and dominating the image search results for real peachicks [60]. Despite the harms that models might cause their data ecosystems, practitioners lack understanding of the mechanics of these distribution shifts. This can lead to unawareness of the distribution shifts, especially when multiple models participate in the data ecosystem, and a lack of understanding of the fairness and equity harms that may arise.

In this work we introduce *model-induced distribution shift* (MIDS) to describe model-induced changes to the data ecosystem. There are several phenomena in existing literature that we re-specify as MIDS; a subset of distribution shifts which are caused by past generations of model (mis)behaviours (in)advertently impacting successive generations. Each MIDS entails a model causing a gradual change in the data ecosystem; such as when synthetic data is published to the web and re-scraped to form new training sets. Once re-scraped, this polluted data becomes the ground truth for future generations of models, and the MIDS continues. We first unify several existing MIDS into a common framework, allowing a more nuanced understanding of their common causes and enabling analysis even where multiple MIDS occur in the same data ecosystem. We analyze the model behavior and fairness impacts of MIDS continuing over generations of models; including supervised classification models trained with labels sourced from their predecessors' predictions, and generative models trained from synthetic data created by their predecessors' outputs. We evaluate the performance and a variety of properties that may indicate the fairness of models in each generation; finding disproportionately negative impacts on minoritized groups. We find that chains of generative models eventually converge to the majority and amplify model mistakes that eventually come to dominate and degrade the data until little information from the original distribution remains, causing representational disparity between sensitive groups.

In contrast to these harms, we study a conceptual framework introduced in Davis et al. [18] called *algorithmic reparation* (AR). While not strictly limited to algorithmic changes, AR uses ML models to provide redress for past harms to people with marginalized Intersectional identities. For example, if attempting reparative predictive policing for Black communities, AR interventions in models could include re-weighting marginalized peoples' records in a dataset to compensate for over-representation in policing, increasing the model threshold for the detrimental prediction, or perhaps removing the predictive system entirely [34]. Because AR operates in settings where the 'ground truth' is of questionable validity (due to current and historical discrimination), it provides a valuable avenue to provide reparation, and also to counter the injustices of MIDS. Of course, AR interventions also create model-induced distribution shifts; AR co-opts the mechanics of other MIDS to promote equity. To that effect, we simulate the effects of AR interventions through progressive Intersectional categorical sampling, showing how prioritising representation lessens the unfair impacts of coexisting MIDS and their discriminative data ecosystems. In summary, we make the following contributions:

- We define a new term, model-induced distribution shift (MIDS), to unify several distribution shifts under one concept, and explore empirical settings to illustrate their impact. This unification draws attention to the common causes of MIDS and enables analysis even where MIDS co-occur.
- We use our settings to evaluate the impact of the fairness feedback loop and model collapse MIDS in several datasets, including face `CelebA` and `FairFace` datasets. We find that MIDS can lead to poor performance within a few generations of models, causing class imbalance, a lack of minoritized group representation, and unfairness. For example, our experiments on `CelebA` undergoing model collapse and performative prediction leads to a 15% drop in accuracy and complete erasure of the minoritized group.
- We position algorithmic reparation as an intentional MIDS that used model impacts to promote equity and justice in the broader data setting. We create an algorithm, STratified AR (STAR) to simulate AR interventions by making training representative of Intersectional identities. These simulations demonstrate how AR interventions can lessen disparate impact between sensitive groups and combat the unfair effects of other MIDS.

2 BACKGROUND

Several terms in existing literature describe distribution shifts perpetuated by models. We provide an overview of these MIDS, their enablers, and their relationships to fairness, then also connect algorithmic reparation to MIDS. A background on bias, fairness in ML (FML, acronym from Davis et al. [18]), and critiques of FML may be found in Appendix B.

2.1 What are MIDS?

In Table 1, we organize three phenomena from the literature into MIDS by determining how the model changes the data ecosystem: 1) *Performative prediction* occurs when a model's predictions influence outcomes, such as when recommender models influence and change a person's preferences [55, 70]. This is also known as *fairness feedback loops* when the outcomes of model predictions entrench bias or discrimination, as in redlining [26]. 2) *Model collapse* may occur due to a similar phenomenon for generative models. If synthetic outputs are used to train a new generative model, over the course of several generations of models, the data distribution loses its tails and converges to a point estimate [4, 62]. 3) *Disparity amplification* occurs due to poor performance on a group of users. These negatively impacted users disengage from the data ecosystem, causing representational disparity. If trained upon, the altered data ecosystem could lead to even worse performance disparity [31]. While all of these effects cause distribution shift after deploying one model, the changes to the data ecosystem become entrenched as the ground truth if used to train the next generation of models. Throughout the remainder of the paper, we refer to generations, lineages, or sequences of generative and classifier models to indicate the teacher–student (see [54]) model chains underlying these MIDS.

There are other effects, which we refer to as enablers, that provide signal to data ecosystems undergoing MIDS. If the enabler misrepresents the training distribution to a model, this may bias its behavior and outputs. Enablers are *not* innately MIDS, and can include sampling, data annotation, generative feedback, and pseudo-labelling methods (for background material on these concepts, see Appendix C.2). These enablers can also permit MIDS to co-occur: a pseudo-labeling model may annotate synthetic data created from generative models to use for supervised training, allowing model collapse and fairness feedback loops to co-occur. Furthermore, if the classifier resulting from the supervised training then impacts humans (or the non-synthetic portion of the data ecosystem), the next generations of any of these models may also be subject to disparity amplification. We model these MIDS and their interactions in Section 3. For a review of MIDS and enablers with examples, see Appendix C.

2.2 Algorithmic Reparation

Algorithmic reparation (AR), introduced in Davis et al. [18], proposes to substitute traditional frameworks of fairness in ML with a reparative approach to the design, development, and evaluation of machine learning systems for social interventions. AR is primarily inspired by Intersectionality theories, and seeks to promote justice in the broader data ecosystem through interventions from carefully-trained models. These actions are not restricted to algorithmic changes; a truly reparative approach requires transdisciplinary collaboration and a shift of economic, legal, and societal incentives. While AR specifically operates in machine learning, it encourages reflection on whether use of ML or computation in

MIDS	Model action/property	Δ Data Ecosystem
Fairness Feedback Loops Performative Prediction	Model predictions	Predictions become outcomes, future labels
Model Collapse	Generated outputs	Synthetic data in ecosystem become new inputs
Disparity Amplification	Poor utility for marginalized groups	Marginalized groups leave

Table 1: MIDS in the existing literature as organized by the model action that induces the MIDS and the effect on the data ecosystem.

general may be inappropriate; and if so, advocates for eliminating these systems.

AR is set as an alternative framework to FML, which generally attempts to equate model properties such as accuracy or positive prediction rate over sensitive groups. Instead, an AR approach focuses not on an equal distribution of resources and benefits, but on a potentially uneven allocation targeted to benefit marginalized intersectional identities. This arises from AR's basis in Intersectionality theories, which acknowledges that harms compound at intersecting marginalized identities (see [59] for an overview and [15] for a prominent example). AR rejects the notion that equality necessarily begets equity and rejects that technology, including machine learning, can be neutral and objective (see Kapania et al. [38] for a detailed discussion of representational thinking, algorithmic idealism, and algorithmic objectivity). Therefore, AR inherently questions the validity of the 'ground truth' data used when training an ML system; this makes it a critical framework for addressing model-induced changes to the data ecosystem.

In this paper, we empirically simulate how intersectional interventions at each model generation may constitute AR, harm reduction, and better representation. In this perspective, where AR functions as a MIDS, AR provides data ecosystem maintenance with a focus on reparative justice.

2.3 Related Work

We overview several pertinent related works that study MIDS and how they connect to our work. Performative prediction, from Perdomo et al. [55], is detected by comparing the data generating distribution before and after a distribution shift caused by a function of the model's parameters. A performatively optimal model minimizes risk on the data distribution that manifests after its own deployment, and can be approached by methods such as repeated risk minimization and stochastic gradient updates [29, 55]. These discussions are usually constrained to the impacts of a model after one generation, which we extend over several generations and consider alongside other MIDS.

Another work, Taori and Hashimoto [69], observe data feedback loops caused as model predictions contaminate datasets. They provide an upper bound for bias amplification depending on the amount of synthetic predictions and on whether the model has the same label bias as the original dataset. This second criteria is met in classifiers that have high uncertainty over the true labels, which follows from distributional generalization. We build on this work by considering bias amplification due to a changing data ecosystem subject to model collapse and performative prediction, as well as considering fairness impacts beyond remaining faithful to dataset label bias. While in our results some of our metrics converge and stabilize, we do not intentionally aim for distributional generalization.

Discussions of model collapse and the impact of generative models on future training sets are frequent in the natural language processing (NLP) literature, which concludes that removing this data ensures better future performance as NLP models improve [57]. More recent work answers questions on how synthetic data impacts downstream tasks. [32] finds worse downstream classifier performance when training from a synthetic dataset instead of the original. Evaluation settings with multiple, connected generative models have since been investigated in [4, 48, 62]. Each of these works finds negative impact to utility if there is a sufficient lack of non-synthetic data. We depart from all of these works by considering the impacts of model collapse on fairness and equity; we combine the downstream performance task of [32] with the model collapse evaluation scheme of [48] and add fairness considerations, as well as co-occuring MIDS.

The work that introduces disparity amplification, Hashimoto et al. [31], considers fairness cases where sensitive information is unavailable. To minimize the risk that the minoritized group incurs high loss and disengages from the dataset ecosystem, they use distributionally robust optimization (DRO). As mentioned in their discussions, DRO might not protect minoritized groups so much as some worse-off group (as in Rawlsian justice), which for our focus on algorithmic reparation and Intersectionality is inappropriate.

3 METHODOLOGY

In this section we introduce two settings, sequences of classifiers and sequences of generators, to allow for the observation and evaluation of MIDS. In these settings, each new model in the sequence is (at least partially) trained using the outputs of its predecessor(s) as inputs and/or labels. As that model is deployed and used, it propagates MIDS through its own properties and outputs, potentially affecting both the synthetic and non-synthetic portions of the data ecosystem. We also position AR as an intentional MIDS aiming for justice for historical discrimination and oppression. These settings provide an understanding of model impact over many generations, enabling informed maintenance of model and data ecosystem 'health,' and reinforcing accountability for model impacts.

3.1 Modeling Assumptions

Measuring distribution shift requires comparison between the current and the reference distribution, which represents the data ecosystem before the presence of any MIDS. The original reference distribution is given by $\mathcal{H} = \mathcal{X} \times \mathcal{L} \times \mathcal{S}$, where $\mathcal{X}$ represents the inputs, and $\mathcal{L}$ and $\mathcal{S}$ are annotations for the labels and sensitive attribute(s).

Sampling from $\mathcal{H}$ gives dataset $D = X \times L \times S$. (*i.e.,* via disparity amplification). To compare the current data ecosystem against the original, we would need access to the original input distribution $\mathcal{X}$ and oracles for $\mathcal{X} \to \mathcal{L}$ and $\mathcal{X} \to \mathcal{S}$. Instead, we approximate these with a generative model G_0 and classifiers A_L and A_S, all trained from D. We use these to approximate data from the original distribution and to annotate the class and group of generated data.

These oracles provide an infinite data source that may be used to train all of the downstream models in our settings. Therefore, if the oracles misrepresent the distribution, the models trained from their outputs will experience MIDS relative to the original training distribution. These oracle approximations are not strongly limiting as we are primarily interested in the effects of MIDS relative to some distribution, be it the original or its approximation. These oracles resemble an online query model for annotations. Alternatively, for an offline setting, these oracles could be replaced by three pools of data, corresponding to samples of inputs, inputs with sensitive attribute annotations, and inputs with label annotations. The intersection of these pools provides an annotated dataset "generated" offline. Using classifiers to annotate generated samples has been used for the fair training of generative models, though A_L and A_S could also represent human annotators conducting manual data annotation [27, 28, 32, 43]. The initial generator, G_0, is also relevant for scenarios where synthetic data is preferred over human-generated data for a downstream task, which may sometimes arise in FML and privacy [25, 53, 64, 73]. Sampling from G_0, as opposed to the training set, also allows a chance at sampling from groups that might not otherwise be well-represented in the dataset, as in [73].

3.2 Sequential classifiers

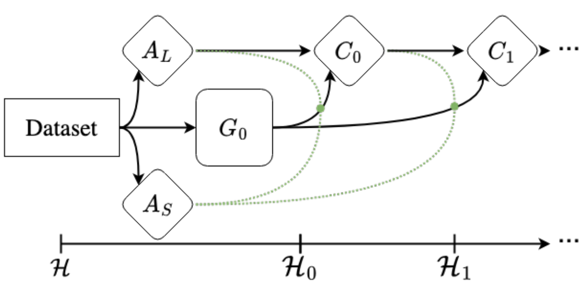


Figure 1: A high-level depiction of sequentially training classifiers (SeqClass setting) for MIDS such as performative prediction and runaway feedback loops. The oracle models, A_L, A_S, and G_0 provide an infinite source of labels, sensitive group annotations, and inputs. We use these to train classifiers C_i, where C_i is trained using labels from C_{i-1}. To alleviate the harms caused from sequentially training, cla-STAR may be used to incorporate sensitive attribute data from A_S, as shown by the narrowly-dashed green lines.

The sequential classifier (SeqClass) setting permits us to pursue the study of MIDS such as fairness feedback loops and performative prediction, where distribution shift is mediated by classifier predictions becoming the ground truth of the next generation, as shown in Figure 1. In the first generation, we train a classifier C_0 by sampling inputs from G_0 and labeling these with oracle A_L. In subsequent generations $i = 1 \dots n$, the classifier C_i is trained on data sampled from G_0 but labeled by the preceding classifier C_{i-1}.

Disparity amplification can be modeled by taking non-synthetic samples $h_i \sim \mathcal{H}_i$ to train C_i, where $\mathcal{H}_i$ is the non-synthetic data distribution after models from generation i were deployed. We assume that disparity amplification has already influenced the label and sensitive group balances of $\mathcal{H}_i$ via C_{i-1}. Therefore, to get h_i in practice, we inference C_{i-1} on a held-out subset of D, recording label prediction frequencies over the categories formed from the Cartesian product of the sensitive attribute values and the possible labels. We use this to define a categorical distribution which we use to perform quota sampling on D. Quota sampling refers to partitioning a population into strata (in our case defined by label and group intersections) and selecting from each partition until we reach its quota, which is given by the categorical distribution multiplied by the total number of samples we wish to select. Henceforth we refer to this categorical distribution as a strata. In a nutshell, if C_{i-1} often assigns negative predictions to a minoritized group, then h_i will contain a proportional number of minoritized group samples with the negative label. Therefore, C_i may be influenced by C_{i-1} in two ways: 1) through data labeled by C_{i-1} and 2) through non-synthetic data undergoing disparity amplification due to prediction disparity in C_{i-1}. When training C_0, we sample $h_0 \sim D$ as disparity amplification has not yet occurred.

The formulations are shown below, where $\mathcal{T}_C(\cdot;\cdot)$ is the classifier training algorithm trained from the data in its first argument(s) as labeled by its second argument(s), $\mathcal{T}_\mathcal{G}$ is the generator training algorithm, and `Sample(·; ·)` samples from its first argument according to a property (such as group representation) of its second argument(s). The terms causing performative prediction and disparity amplification are in **red bold** and **teal bold** face:

$$C_i = \mathcal{T}_C(g_i, \boldsymbol{h_i}; \boldsymbol{C_{i-1}}), \text{ where } C_0 = \mathcal{T}_C(g_0, h_0; A_L),\ G_0 = \mathcal{T}_\mathcal{G}(X), \text{ and } g_i \sim G_0 \tag{1}$$

$$\text{and } h_i = \texttt{Sample}(D; \boldsymbol{C_{i-1}}), \text{ where } h_0 = \texttt{Sample}(D). \tag{2}$$

3.3 Sequential generators and classifiers

The sequential generator (SeqGen) setting primarily investigates the model collapse MIDS, where distribution shift occurs as synthetic data is used for training new models, as shown in Figure 2. For this setting, we train generators G_i sequentially from the samples of the preceding generator G_{i-1}, where the first generator G_0 is trained from the original dataset. This chain of generators is the same setting as used by Shumailov et al. [62]. Departing from them, we also train a downstream classifier C_i by sampling inputs from G_i and labels from either A_L or from the preceding classifier C_{i-1}. The former case is the sequential generator and non-sequential classifier setting (henceforth SeqGenNonSeqClass), and the latter the sequential generator sequential classifier setting (SeqGenSeqClass). In SeqGenSeqClass, the classifiers are chained together and suffer the MIDS described in the SeqClass setting. These downstream classifiers allow us to initiate the study of downstream classifier performance and FML fairness metrics while also tracking the devolution of minoritized group representation and model collapse.

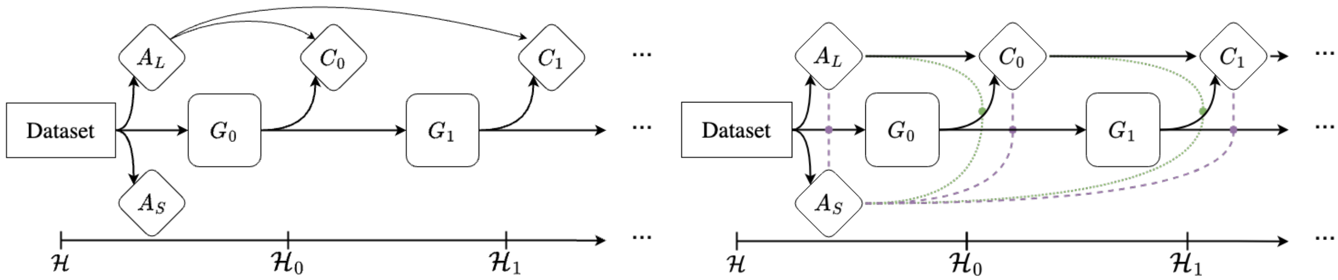


Figure 2: A high-level depiction of sequentially training generators with and without sequential classifiers (*left:* SeqGenNonSeqClass, *right:* SeqGenSeqClass). The oracle models, A_L, A_S, and G_0 provide an infinite source of labels, sensitive group annotations, and inputs. We train a lineage of generative models (G_i) and train classifiers C_i from these, where C_i is trained using labels from A_L (*left*) or C_{i-1} (*right*). To alleviate the harms caused from sequentially training, cla-STAR and gen-STAR may be used to incorporate sensitive attribute data from A_S, as shown by the narrowly-dashed green lines and the broad-dashed purple lines.

Additionally, disparity amplification due to C_{i-1} and G_{i-1} may impact the non-synthetic data distribution, $\mathcal{H}_i$, that may be used when training G_i and/or C_i. Similarly to the SeqClass setting, we find the categorical distribution of C_{i-1} over the sensitive attributes and labels and quota sample proportionally from D to form h_i. While this directly impacts G_i, it also impacts C_i since it trains from G_i.

The formulation for SeqGen with classifiers is shown below, with a substitute model C that may stand for A_L or C_{i-1} depending on whether the classifiers are sequential. The terms for performative prediction, model collapse, and disparity amplification are bolded in **red**, **blue**, and **teal** face.

$$G_i = \mathcal{T}_{\mathcal{G}}(\boldsymbol{g_{i-1}}, \boldsymbol{h_i}), \text{ where } G_0 = \mathcal{T}_{\mathcal{G}}(X) \text{ and } g_i \sim G_i \tag{3}$$

$$C_i = \mathcal{T}_C(g_i, \boldsymbol{h_i}; \boldsymbol{C}), \text{ where } C_0 = \mathcal{T}_C(g_0, h_0; A_L) \tag{4}$$

$$h_i = \mathtt{Sample}(D; \boldsymbol{C_{i-1}}, \boldsymbol{g_{i-1}}), \text{ where } h_0 = \mathtt{Sample}(D). \tag{5}$$

3.4 Simulating Algorithmic Reparation

In our experiments, we measure and simulate equity-oriented interventions as a change in the discrete distribution (strata) formed from the Cartesian product of the label and sensitive attribute values. For example, the strata of the current data ecosystem may be formed from L and S, and the strata of a classifier C_i may be formed from its predictions (on G_i or X) and sensitive group annotations (from A_S or S). We use strata to characterize the training sets used when training models, where these strata change over the generations.

To simulate the effects of AR interventions, we introduce an algorithm called STratified sampling AR (STAR; see Algorithm 1). STAR creates model training batches by taking a biased sample according to these strata categories (within a resampling budget). As we are not experts in AR and do not provide a case study, we use a uniform distribution over the categories to give quotas for the number of samples from each category that should be present in the batch. Using an 'ideal' (finite) distribution may be inappropriate as a distribution cannot neutrally or objectively determine the 'best' mixture of demographics for a task. Due to this concession, we will refer to these measurements as fairness/unfairness (in the FML sense), as we cannot make claims of equity or justice without considering the myriad sources of bias in the ML life cycle [67] and the specific data ecosystem.

In the SeqClass setting, we use the name Classifier-STAR, or cla-STAR, to refer to creating more Intersectionally representative batches for the classifiers in the lineage. This is not the only avenue for AR, but is inspired by work done by the FML community for performative prediction [24]. To train C_i, cla-STAR labels generated outputs from G_0 using C_{i-1} and A_S, then selects a subset of these samples for each training batch such that each label and group category in the batch meets the quota set by the fairness ideal. We update $\mathcal{T}_C$ to $\mathcal{T}_{\mathcal{A},C}$ and show the additional labeling and sensitive group annotations from C_{i-1} and A_S in **green**:

$$C_i = \mathcal{T}_{\mathcal{A},C}(g_i, h_i; \boldsymbol{C_{i-1}}, \boldsymbol{A_S}), \text{ where } C_0 = \mathcal{T}_{\mathcal{A},C}(g_0, h_0; A_L, \boldsymbol{A_S}). \tag{6}$$

STAR in SeqGenSeqClass may occur at all the same points described above, with the addition of interventions taken while training the generators. We examine both classifier-side STAR (cla-STAR, taken while training classifiers, shown in **green**), and generator-side STAR (gen-STAR, while training generators, shown in **purple**). gen-STAR also uses annotations from C_{i-1} and A_S to fill the label and group category quotas set by the fairness ideal. Both cla-STAR and gen-STAR are described in Equation (7) and Equation (8), respectively:

$$C_i = \mathcal{T}_{\mathcal{A},C}(g_i, h_i; \boldsymbol{C}, \boldsymbol{A_S}), \text{ where } C_0 = \mathcal{T}_{\mathcal{A},C}(g_0, h_0; A_L, \boldsymbol{A_S}) \tag{7}$$

$$G_i = \mathcal{T}_{\mathcal{A},\mathcal{G}}(g_{i-1}, h_i; \boldsymbol{C}, \boldsymbol{A_S}), \text{ where } G_0 = \mathcal{T}_{\mathcal{A},\mathcal{G}}(X; \boldsymbol{L}, \boldsymbol{S}). \tag{8}$$

3.4.1 STAR Implementation. We simulate algorithmic reparation using the two variants of STAR introduced in Section 3.4; cla-STAR and gen-STAR. The algorithm uses sampling and pseudo-labelling to create training batches of size b that meet the fairness ideal by having a prescribed number of samples to fill a quota in each category. However, the closeness between this fairness ideal and the resulting strata of the batch is bounded by the reparation budget r, which simulates costs to conducting reparation. For our experiments, we use a uniform distribution as the fairness ideal.

STAR creates a pool of $b + r$ samples from either the previous generator or the original dataset, which is then annotated by A_S and either A_L or C_{i-1}. The fairness ideal, multiplied by b, gives a quota for the number of samples ideally belonging to each category.

STAR then attempts to fill each category to its quota from the pool of samples. If after this initial filling, some of the categories did not meet their quota, the remainder of the batch is filled with randomly selected samples from the remaining pool. This process (henceforth re-sampling) will most likely add samples representative of the majority group and class. There are therefore two barriers to meaningful reparation: 1) we cap the number of samples that may be drawn to form the batch yet attempt to create equal categories from an unequal dataset; and 2) the effects of MIDS. If STAR increases the representation of a minoritized group, then these samples may be re-selected more often than majoritized group peers, increasing their exposure to mislabeling. Additionally, the higher number of generations this data is subjected to may accelerate the model collapse at this area of the distribution. STAR is shown for binary labels and binary sensitive attribute (4 categories in the STRATA) in Algorithm 1.

4 EVALUATION

Now that we have overviewed our methodology, we outline our experimental setup using two main sets of experiments to illustrate MIDS in the SEQCLASS and SEQGENSEQCLASS settings. These experiments are designed to answer the following questions: (Q1) what are the effects of MIDS on performance, representation, and fairness? (Q2) why is it important to be aware of MIDS? (Q3) how do MIDS interact? (Q4) Can AR interventions alleviate the harms of MIDS? We provide brief answers to these questions in Section 4.2.

4.1 Modeling MIDS

4.1.1 Experimental Setup. We provide computer vision experiments for four datasets: texttttMNIST, SVHN, FairFace, and CelebA. We modify the digit recognition datasets MNIST and SVHN into ColoredMNIST and ColoredSVHN by adding color to create binary sensitive groups and by forming two classes for digits < 5 and ≥ 5. We choose the beneficial class as the class converged to by model collapse, and bias the majoritized group towards it. These arbitrary choices simplify our presentation; we vary the class and group balance in Appendix E.1, finding little impact. These datasets differ in their complexity, SVHN is usually the harder dataset to learn, and in their class balance once binarized (see Table 3). Despite their similarities, they also show very different points of model collapse and even opposite behavior when observing the accuracy difference between groups, as in Figure 5 and Figure 20. We use CelebA and FairFace to represent more complex and real-world tasks. For CelebA, our task is to predict attractiveness with gender as the sensitive attribute; these attributes have well-documented errors and disparities [44]. On the other hand, FairFace is chosen for its balance in both race and gender; alongside other metrics, it is simple to measure sensitive group intersections. For FairFace, we attempt to predict gender (2 values) with sensitive attributes race (7 values) and age (which we binarize at < 30, ≥ 30). Note that our FairFace experiments are intersectional, and for A_S we use a different classifier for each sensitive attribute. For CelebA and FairFace, we provide between 5-10 generations,[1] for ColoredMNIST and ColoredSVHN we train for 40 generations. When training each generation, all synthetic data is sampled from the prior generator and/or classifier/annotator, while non-synthetic data is taken from the training distribution.

Further details on the datasets, model architectures, hyperparameters, and compute specifics we use are in Appendix D.[2] Loss values for generators may be found in Figure 13, and accuracies and fairnesses for A_L and A_S are in Table 4. These performances reflect baseline results of training without fairness optimization. For STAR, we set the reparation budget r at 25% of b, the batch size, for all datasets aside from ColoredSVHN, which is set to 33% of b. This budget was tuned by finding the lowest possible value that results in a decrease in the proportion of each batch resampled over generations, indicating that STAR is changing the data ecosystem towards its ideal. We also train the models from a 50-50 mixture of synthetic and non-synthetic data, allowing us to observe the effects of disparity amplification as it co-occurs with performative prediction and model collapse. We use this data mixture to train the classifiers in SEQCLASS, and the generators in SEQGENSEQCLASS where we observe downstream impacts in the classifiers.

MIDS Metrics. To measure MIDS and their fairness impacts, we inspect the original dataset, generated outputs and annotations, and model performances. We use a held-out evaluation set *i.i.d.* from the original training set D. In the classifiers, we measure fairness using demographic parity difference (DP), equalized odds difference (EOdds), and group accuracy gaps, and measure utility with accuracy. DP difference (Definition 1) compares the positive prediction rates between groups. EOdds difference (Definition 2) is the maximum between two values: the difference between the groups' true positive rates, or between their false positive rates. For performance disparity, we report the maximum accuracy gap when comparing all sensitive groups. For these metrics, a lower value (less difference between groups) indicates more fairness. Note that accuracy and EOdds require a ground truth label which may be taken from a biased original distribution. Therefore, to meet other fairness objectives, such as in STAR, EOdds and group accuracy differences may worsen as the label distribution changes. We measure these metrics on the evaluation set, but also between successive classifiers using images from G_0 or G_i with sensitive annotations from A_S and labels from C_{i-1}. The difference between the former (measuring with respect to D) and the latter (measuring with respect to the preceding models) shows how model performances can be misreported if the evaluator is unaware of MIDS. To track the class and group representation of the generators, we generate 1000 samples and annotate class and group with A_L and A_S.

MIDS STRATA. We also observe the STRATA of the original dataset (using X, L, and S), the model training batch STRATA (using G_0 or G_i, A_L or C_{i-1}, and A_S), and the model output STRATA (classifiers using X, C_i, and A_S, generators using (G_i, A_L, and A_S). We also record the Kullback–Leibler (KL) Divergence between these STRATA and the fairness ideal used in STAR. These KL-Divergence results provide a simple way to measure and visualize change in intersectional representation for our experiments; a 'fairness ideal' is otherwise inappropriate (see Section 3.4 for clarity on representational thinking). We also measure the progression of STAR through

[1] Repeatedly training CelebA and FairFace takes days, due to the cost and CO_2 footprint, we terminated these experiments upon realization of the MIDS.

[2] Our code is available publicly on GitHub: https://github.com/cleverhans-lab/FairFeedbackLoops

the STRATA it achieves during batch curation, the amount of resampling required when categories fail to meet their quotas, and the KL-Divergence between the STRATA and the fairness ideal.

4.2 Results

We return to the questions posed at the start of Section 4 and provide brief answers before discussing the full results: **Q1) What are the effects of MIDS on performance, representation, and fairness?** We find that the performative prediction, model collapse, and disparity amplification MIDS lead to a loss of accuracy, fairness, and representation in classes and/or groups. These effects are more pronounced in SeqGenSeqClass, likely because model collapse sometimes results in the beneficial class and majoritized group dominating the generated samples. These effects are more severe in data ecosystems with higher proportions of synthetic data, which we ablate in Appendix E.2.
Q2) Why is it important to be aware of MIDS? We find that unawareness of MIDS results in overstating the accuracy and fairness, which can be observed by measuring the relative performance of classifiers (comparing C_i against labels provided by C_{i-1}). In SeqClass, these relative results show nearly 100% accuracy and 0 EOdds difference, the same holds for SeqGenSeqClass with the addition of mis-reported class and group balances (see Appendix H).
Q3) How do MIDS interact? We compare SeqGenSeqClass and SeqGenNonSeqClass, revealing that the fairness feedback loop in the former allows the classifiers to adapt to distribution shift in the inputs caused by model collapse. This co-operation lessens the rate and degree of classifier performance decline. When training with a mixture of synthetic and non-synthetic data, we observe that the non-synthetic data greatly slows the degree of model collapse, though also enables disparity amplification amongst groups in the non-synthetic data ecosystem.
Q4) Can AR interventions alleviate the harms of MIDS? Our AR interventions using STAR lessen unfair behaviors and achieve better downstream classifier fairness, especially in SeqClass, and also in settings undergoing disparity amplification. In SeqGenSeqClass, GEN-STAR usually performs better than CLA-STAR, likely due to the strength of the model collapse MIDS in deteriorating the data ecosystem.

4.2.1 Sequential classifier setting. Our first experiment suite uses SeqClass as described in Section 3.2; MIDS occur as a classifier's predictions are used to label the next generation's classifier. In `ColoredMNIST` and `ColoredSVHN` (Figures 3 and 15), we observe an accuracy drop of 10-15% over 40 generations, with an increase in both DP and EOdds unfairnesses (in the case of `ColoredSVHN`, both metrics increased by roughly 0.2, where the maximum unfairness gap is 1). `CelebA` immediately suffers near-random classifier performance as G_0 misrepresents D by incurring significant class imbalance towards the detrimental class (Figure 16). `FairFace` classifier accuracies drop from 57% to random accuracy within 10 generations, and also incur an accuracy difference increase of .1, with a .2 jump in EOdds unfairness (Figure 17). Note that these performances would likely worsen with additional generations.
Reparative interventions reduce performance degradation from MIDS. In `ColoredMNIST`, `ColoredSVHN`, and `CelebA` CLA-STAR lead to a significant reduction in DP and EOdds unfairness, and lessened the gap between CLA-STAR's fairness ideal and the data ecosystem STRATA (Figures 3, 15, and 16). Across most datasets, there is far less variance compared to results without reparation, where variance grows with the number of generations (all figures report the 95% confidence interval). For `FairFace`, classifier STRATA without reparation are constituted primarily of older white males (where younger white males are the plurality of the dataset, see Figure 7). With reparation, the representation of young white non-males increases, but as G_0 fails to adequately generate samples from the other races, there are still large performance disparities and high unfairnesses (Figure 17). See Appendix G.1.1 for detailed figures on the representation of classes and groups in training batches. Due to the high representational disparity between the white race and the other 6 races, CLA-STAR did not lead to better fairness. We also observe tension between FML metrics: for `ColoredMNIST` (Figure 3), the EOdds and accuracy difference increase after generation 15, while both DP and the KL-divergence continue to decrease. As EOdds is satisfied when error rates (relative to a potentially biased dataset) are similar across groups, meeting the STAR fairness ideal leads to an 'unfair' allocation of beneficial labels to the minoritized class, and of detrimental labels to the majoritized class.
Non-synthetic data slows MIDS, but weakens classifier-side reparations. We also trained classifiers on an even mixture of synthetic and non-synthetic data as described in Section 3.2, see Figure 28. Unsurprisingly, adding non-synthetic data greatly improved the performance of classifiers compared to results with 100% synthetic data and no reparation (see a full ablation of the amount of synthetic data in Appendix E). While we were able to further increase this fairness by using CLA-STAR, the impact was far less than on the 100% synthetic results, with FML unfairness metrics converging to higher values at around the 25$^{\text{th}}$ generation. Because the non-synthetic data lessens the impact of the pseudo-labelling enabler in the fairness feedback loop, it likewise lessens the impact of CLA-STAR.

4.2.2 Sequential generator and classifier setting. These experiments refer to the SeqGenSeqClass setting described in Section 3.3 and depicted in Figure 2, which we use to depict model collapse and performative prediction, with additional results including the effects of disparity amplification. In our 100% synthetic training experiments, we observe model collapse deteriorates the data, leading to class imbalance (`ColoredMNIST` Figure 19, `CelebA` Figure 5) and/or to group imbalance (`ColoredSVHN` Figure 4, `CelebA` Figure 20). As model collapse progresses, the downstream classifiers either perform with random accuracy or constantly predict the beneficial class label, leading to poor fairness. Refer to Appendix F to see generated samples undergoing model collapse.

If we judge model collapse at the point when the generated data ceases to have any downstream utility, model collapse occurs at generation 15 for `ColoredMNIST`, 5 for `ColoredSVHN`, and between generations 1-5 for `CelebA` and `FairFace`. These values correspond to the increasing difficulty of the datasets' tasks, which is correlated with heavy-tailedness in their distributions [50]. A small sample size may be able to represent a concentrated distribution, but finite samples of a heavy-tailed distribution will likely be biased, leading to faster distribution shift. The steep decline to random accuracy is likely due in part to the amount of synthetic data used to train each

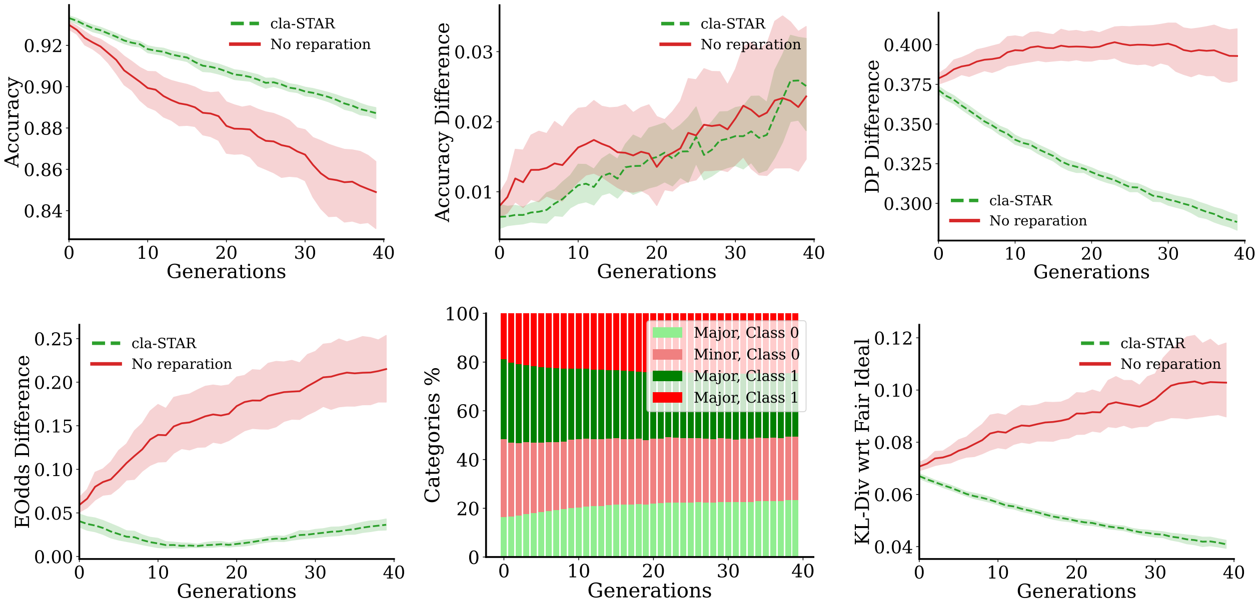


Figure 3: `ColoredMNIST` results for SeqClass on the evaluation set. *Top:* accuracy, accuracy difference, and demographic parity difference. Better fairness (lower fairness difference) and higher accuracy is achieved with cla-STAR. *Bottom:* equalized odds difference, the strata created during cla-STAR, and the KL-divergence between cla-STAR fairness ideal and classifier strata. The KL-Divergence decreases with cla-STAR, indicating more fairness, and the batches become more balanced across group and class. The accuracy difference and EOdds difference between groups is small, but increases during reparation due to metric tension with cla-STAR.

generation. As the amount of synthetic training data decreases, so too does the rate of accuracy decline and beneficial class dominance, as shown in Appendix E.2.

Unpredictable convergence of model collapse. As model collapse progresses, it is difficult to predict the category of the strata in the generators that dominates. For example, in generation 40 of `ColoredMNIST`, both majoritized and minoritized groups of the beneficial class each constitute around 40% of each batch, whereas in `ColoredSVHN`, the majoritized and beneficial category alone constitutes 60% of each batch (Figures 19 and 4). This is in part due to the initial class and label balances (see Table 3), but also due to how the model collapse manifests. For example, `ColoredMNIST` eventually converges to samples that resemble an '8', or all the digits superimposed (see samples from model collapse in Appendix F). As this happens to fall in the advantaged class, it becomes dominant in the generators, without strongly impacting the group balance. Meanwhile, in `CelebA`, model outputs become dominated by the majoritized group, yet these same samples are also classified into the detrimental class by A_S, which is counter to the original label balance in the dataset. Eventually, minoritized group representation in `CelebA` falls to 0% (see Figure 5). `FairFace` has much deterioration in the data but maintains both class and group balance. These observations support a growing consensus that the features of the original data preserved by generative models in synthetic data is difficult to predict, or highly data dependent (for private synthetic data [64], and at the intersection of fairness and privacy in synthetic data [13]).

Performative prediction adapts to model collapse. We also uncover co-operation between MIDS by evaluating the role of sequential classifiers in SeqGenSeqClass and SeqGenNonSeqClass (see Appendix E.3). In `ColoredMNIST` and `ColoredSVHN` (Figures 11 and 12): the non-sequential classifiers converge to accuracies 10-20 percentage points lower than the sequential classifiers. However, the sequential classifiers have considerably more unfairness (in the case of `ColoredMNIST`, by 0.6), likely due to their participation in fairness feedback loops. Performative prediction among sequential classifiers allows C_{i-1} to provide meaningful labels for training C_i from G_i. For the non-sequential classifiers, A_L cannot adequately support the distribution represented by G_i once the i^{th} distribution substantially differs from the original. The inherited knowledge of $\mathbb{P}(Y|X)$ passed through the sequential classifiers allows them to preserve a more accurate map from the changing distribution to the classes.

Generator-side AR improves fairness and minoritized representation. Between gen-STAR and cla-STAR, the former leads to more preservation of the group and label balance in all four datasets. This result fits intuitively as the biased sampling enables these generators to maintain more balanced representations across the categories. For example, gen-STAR leads to better fairness than cla-STAR in `ColoredMNIST` and `CelebA`, though with cost to accuracy. However, because gen-STAR results in oversampling minority (in terms of population) categories relative to the original dataset, it may also expose these areas of the data distribution more to model collapse. In `ColoredSVHN`, for example, gen-STAR results in more balanced strata in both the generators and classifiers (compared to CLA-STAR), but the classifier STRATA are still dominated by the (minoritized, detrimental) and (majoritized, beneficial) categories, leading to worse DP and EOdds fairness (see Appendix G.2.1). In the case of FairFace, a combination of oracle model bias and unrepresentative generators causes large disparities between races as GEN-STAR cannot adequately sample from the smallest Intersectional minorities. As FairFace has relatively balanced races and genders, these disparities indicate that algorithmic reparation should be considered when collecting data, and might require action beyond collecting balanced quotas of data from various groups. Overall, CLA-STAR did not show consistent performance across datasets, achieving worse or equivalent performance to the non-reparative results, likely due to the strength of the model collapse MIDS.

Disparity amplification reduced with classifier-side reparation. Recall that we model disparity amplification by sampling non-synthetic data using the STRATA of the classifiers, which we use for half of the generator training data (the other half is sampled from G_{i-1}). We evaluate this setting for ColoredMNIST. Similarly to SeqClass, the non-synthetic data slows MIDS caused by synthetic data spills, including model collapse. We evaluate the fairness performance of GEN-STAR and CLA-STAR, finding substantially better performance and fairness with CLA-STAR (subject to an increase in accuracy disparity due to increased false negatives) than with GEN-STAR (see Appendix I). The GEN-STAR generator STRATA never achieve the fairness ideal as randomly sampling from initially biased generators leads to unfair classifiers which propagates disparity amplification in the non-synthetic data, which incidentally only protects the majority group and class categories from model collapse. Meanwhile, CLA-STAR trains classifiers with more ideal STRATA that reverses disparity amplification, providing balanced non-synthetic data to the generators and protecting all classes and groups equally from model collapse. We may see this by comparing the STAR STRATA for both algorithms, see Figure 30.

4.3 Limitations

We discuss three main limitations of our work. Firstly, we do not provide a specific use case and cannot fully evaluate algorithmic reparation, nor make any claims that our achievements in fairness lead to equity or justice. Secondly, we study a worst case where all the synthetic data used to train a model is sampled only from the immediately preceding model(s); without provenance information the synthetic data may be sourced from any number of other models, including multiple predecessor models. Thirdly, in CelebA and FairFace, we rely on race annotations which might fail to represent the various skin tones within groups, an important consideration in computer vision tasks, and for better representation (as discussed in [9]). Additionally, racial categorizations are not universally consistent, and so these datasets provide a simplification that may be inappropriate.

5 CONCLUDING REMARKS

In this paper, we introduced model-induced distribution shifts (MIDS) and created empirical settings enabling the evaluation of their harms. With these settings we found that MIDS, both on their own and co-occurring with enablers such as data annotation, lead to major degradation in utility, fairness, and minoritized group representation. While MIDS can be intentional or unintentional, unawareness of their existence can lead to grossly overstating model utility and fairness. Based on these harms, we discussed how algorithmic reparation (from the literature of critical theory in ML) may act as an intentional MIDS with goals of equity and justice. By simulating the impacts of algorithmic reparation at various points in our settings, we saw a lessening in harms.

We would also like to acknowledge that throughout this work, we have implied that models cause model-induced distribution shift. This is not the case; agency over MIDS rests primarily on model owners, data publishers and collectors, and model users. Several related works, including Shumailov et al. [62] and Hardt and Mendler-Dünner [29], have also considered the power or advantage given to an entity that has more control over the amount of synthetic data spillage or has more access to non-synthetic data. As we have found that MIDS are of imminent concern in data ecosystems undergoing synthetic data spills; we now turn to solutions to MIDS and methods for taking accountability of them. One possible solution, also mentioned in Davis et al. [18], is an archival perspective on data curation as introduced in Jo and Gebru [37]. Specifically, adopting the tenets of archival description codes could enable gathering of high-quality provenance information, and adopting the moral obligations underlying many an archives' raison d'être could help to identify and repair structural and historical bias [17, 37, 56]. Another solution motivated by our results is the importance of non-synthetic data and human data annotation to prevent or slow the rate of MIDS. We therefore advocate for more attention to the often-unseen and underappreciated labor of human data workers. We end with a call for safer conditions for data workers given their current and increasing importance in our data ecosystems.

STATEMENTS

Ethical Considerations

We recognize that technical solutions are never disjoint from their societal impacts, and have striven towards a more sociotechnical framing for this work. We navigate several definitions and frameworks for algorithmic fairness and equity by considering multiple definitions of fairness and their contrasts with algorithmic reparation. However, we primarily focus on group-based fairness metrics, including when those groups are formed intersectionally, which we acknowledge can reinforce the ideologies behind them. The representational and allocative harms on minoritized groups arising from MIDS have been a major ethical motivation and consideration underlying our work and contributions to algorithmic reparation. Examples of harms of MIDS include minoritized group erasure and disparately poor performance.

Positionality

We are researchers usually operating within the more technical areas of machine learning. Throughout our time working in this area, we have turned to *Data Feminism* by D'Ignazio and Klein [22] to inform our discussions around fairness and equity, and to inform some of our language choices. We also rely heavily on Davis et al. [18] and Kapania et al. [38] for their critiques of 'representationalist thinking,' which has been admittedly ubiquitous in our education,

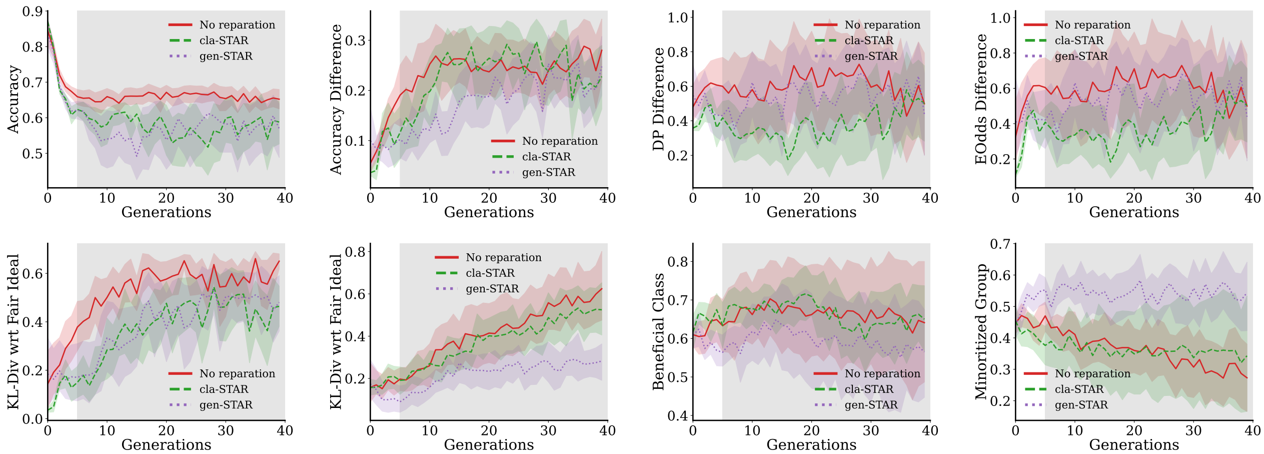


Figure 4: `ColoredSVHN` results for SeqGenSeqClass. *Top:* shows accuracy, accuracy difference, demographic parity difference, and equalized odds difference. For the latter three, lower values are better. *Bottom:* KL-Divergence between fairness ideal and classifiers, and between fairness ideal and generator strata, the class balance, and group balance. Shading shows collapsed generations. We observe that gen-STAR provides more minoritized group representation. While model collapse causes outputs to eventually resemble a '3,' which moves class balance towards the beneficial class, gen-STAR also maintains the original dataset imbalance of 60%.

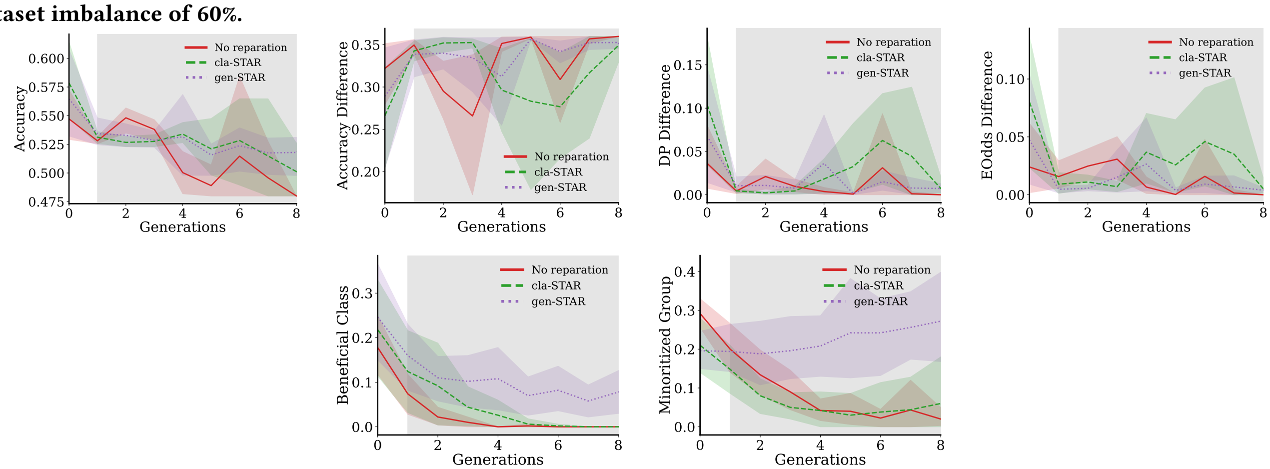


Figure 5: `CelebA` results for SeqGenSeqClass. *Top:* shows accuracy, accuracy difference, demographic parity difference, and equalized odds difference. For the latter three, lower values are better. *Bottom:* shows the class balance the and group balance. Shading shows collapsed generations. The performance of the classifiers was initially low, though STAR is moderately better in later generations. gen-STAR provides better class and group balance compared to cla-STAR and results without reparation, but is still unable to achieve uniform representation due to the strength of model collapse.

and which seemingly appears commonly in ML research (including ML fairness research).

Adverse Impact

We note that this work might be used to fuel despair over the 'long-term' existential harms of models, especially generative models. We advise readers to think critically about the systems of power behind machine learning and consider the current harms these permit, continue, and worsen. We also acknowledge that due to our lack of a specific use case to fully evaluate algorithmic reparation, we risk representing it as a mathematical or technical definition to be satisfied or optimized for. This runs counter to the tenets of AR (see Davis et al. [18]) and we have been careful with our language around this area (for example in Sections 3.4 and 4.3, and in Appendix B).

ACKNOWLEDGMENTS

We would like to acknowledge our sponsors, who support our research with financial and in-kind contributions: CIFAR through the Canada CIFAR AI Chair program and the Catalyst grant program, Microsoft, and NSERC through the Discovery Grant and COHESA Strategic Alliance. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of

Canada through CIFAR, and companies sponsoring the Vector Institute. We would like to thank members of the CleverHans Lab for their feedback. We additionally thank David Glukhov, Syed Ishtiaque Ahmed, and Ramaravind K. Mothilal for feedback on earlier versions of this work.

REFERENCES

[1] 90th United States Congress. 1968. 82 Stat. 73 - An Act to prescribe penalties for certain acts of violence or intimidation, and for other purposes. https://www.hud.gov/sites/dfiles/FHEO/documents/fairhousingact.pdf

[2] Federal Housing Administration. 1938. *Underwriting Manual: Underwriting and Valuation Procedure Under Title 2 of the National Housing Act.* Department of Housing and Urban Development. https://www.huduser.gov/portal/sites/default/files/pdf/Federal-Housing-Administration-Underwriting-Manual.pdf

[3] Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 161–170. https://proceedings.mlr.press/v97/aivodji19a.html

[4] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2023. Self-Consuming Generative Models Go MAD. arXiv:2307.01850 [cs.LG] https://arxiv.org/abs/2307.01850

[5] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization. arXiv:1907.02893 [stat.ML] https://arxiv.org/abs/1907.02893

[6] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

[7] Samuel James Bell and Levent Sagun. 2023. Simplicity Bias Leads to Amplified Performance Disparities. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 355–369. https://doi.org/10.1145/3593013.3594003

[8] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *FAT*. https://api.semanticscholar.org/CorpusID:3298854

[10] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops (ICDM)*. Miami, Florida, USA. https://ieeexplore.ieee.org/document/5360534

[11] Michel Callon. 1998. Introduction: The Embeddedness of Economic Markets in Economics. *The Sociological Review* 46, 1_suppl (1998), 1–57. https://doi.org/10.1111/j.1467-954X.1998.tb03468.x arXiv:https://doi.org/10.1111/j.1467-954X.1998.tb03468.x

[12] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. 2020. Curriculum Labeling: Self-paced Pseudo-Labeling for Semi-Supervised Learning. *CoRR* abs/2001.06001 (2020). arXiv:2001.06001 https://arxiv.org/abs/2001.06001

[13] Victoria Cheng, Vinith M. Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. 2021. Can You Fake It Until You Make It? Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 149–160. https://doi.org/10.1145/3442188.3445879

[14] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:1610.07524 [stat.AP]

[15] The Combahee River Collective. 1977. The Combahee River Collective Statement.

[16] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[17] DACS 2023. *Describing Archives: A Content Standard (DACS), an Implementation of General International Standard Archival Description (ISAD(G))*. Standard. Society of American Archivists' Technical Subcommittee on Describing Archives: A Content Standard (TS-DACS). https://github.com/saa-ts-dacs/dacs

[18] Jenny L. Davis, Apryl Williams, and Michael W. Yang. 2021. Algorithmic reparation. *Big Data & Society* 8, 2 (2021), 20539517211044808. https://doi.org/10.1177/20539517211044808 arXiv:https://doi.org/10.1177/20539517211044808

[19] Li Deng. 2012. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* 29, 6 (2012), 141–142. https://ieeexplore.ieee.org/document/6296535

[20] Northpointe Inc. Research Department. 2016. COMPAS Risk Scales : Demonstrating Accuracy Equity and Predictive Parity Performance of the COMPAS Risk Scales in Broward County. https://api.semanticscholar.org/CorpusID:51920414

[21] Detroit Deomgraphics 1955. The Non-White Popultion of Metropolitan Detroit. https://hdl.handle.net/2027/mdp.39015060547265?urlappend=%3Bseq=21%3Bownerid=13510798897484245-29

[22] Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.

[23] Erin Einhorn and Olivia Lewis. 2021. Built to keep Black from white: Detroit segregation wall still stands, a stark reminder of racial divisions. *NBC News* (2021). https://www.nbcnews.com/specials/detroit-segregation-wall/

[24] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Eduardo Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway Feedback Loops in Predictive Policing. *CoRR* abs/1706.09847 (2017). arXiv:1706.09847 http://arxiv.org/abs/1706.09847

[25] Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. 2022. Robin Hood and Matthew Effects: Differential Privacy Has Disparate Impact on Synthetic Data. In *ICML*. 6944–6959. https://proceedings.mlr.press/v162/ganev22a.html

[26] Ben Green. 2020. The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 594–606. https://doi.org/10.1145/3351095.3372869

[27] Aditya Grover, Kristy Choi, Trisha Singh, Rui Shu, and Stefano Ermon. 2019. Fair Generative Modeling via Weak Supervision. *arXiv preprint arXiv:1910.12008* (2019). https://arxiv.org/abs/1910.12008

[28] Aditya Grover, Jiaming Song, Alekh Agarwal, Kenneth Tran, Ashish Kapoor, Eric Horvitz, and Stefano Ermon. 2019. Bias Correction of Learned Generative Models using Likelihood-Free Importance Weighting. https://proceedings.neurips.cc/paper/2019/file/d76d8deea9c19cc9aaf2237d2bf2f785-Paper.pdf

[29] Moritz Hardt and Celestine Mendler-Dünner. 2023. Performative Prediction: Past and Future. arXiv:2310.16608 [cs.LG]

[30] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems (NIPS)* 29 (2016), 3315–3323. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[31] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 1934–1943. http://proceedings.mlr.press/v80/hashimoto18a.html

[32] Ryuichiro Hataya, Han Bao, and Hiromi Arai. 2022. Will Large-scale Generative Models Corrupt Future Datasets? *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2022), 20498–20508. https://api.semanticscholar.org/CorpusID:253523513

[33] Lily Hu and Yiling Chen. 2020. Fair Classification and Social Welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAccT* '20)*. Association for Computing Machinery, New York, NY, USA, 535–545. https://doi.org/10.1145/3351095.3372857

[34] J.D. Humerick. 2019. Reprogramming fairness: Affirmative action in algorithmic criminal sentencing. *Columbia Human Rights Law Review* (2019). https://hrlr.law.columbia.edu/files/2020/04/8-Humerick_FINAL.pdf

[35] Yerlan Idelbayev. 2018. Proper ResNet Implementation for CIFAR10/CIFAR100 in PyTorch. https://github.com/akamaster/pytorch_resnet_cifar10. Accessed: 2023-07-26.

[36] Kenneth T. Jackson. 1985. *Crabgrass frontier: the suburbanization of the United States*. Oxford University Press.

[37] Eun Seo Jo and Timnit Gebru. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 306–316. https://doi.org/10.1145/3351095.3372829

[38] Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A Hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 133, 15 pages. https://doi.org/10.

1145/3544548.3580645
[39] Maximilian Kasy and Rediet Abebe. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 576–586. https://doi.org/10.1145/3442188.3445919
[40] Patrik Joslin Kenfack, Daniil Dmitrievich Arapovy, Rasheed Hussain, S. M. Ahsan Kazmi, and Adil Mehmood Khan. 2021. On the Fairness of Generative Adversarial Networks (GANs). *Arxiv* abs/2103.00950 (2021). arXiv:2103.00950 https://arxiv.org/abs/2103.00950
[41] Jon Kleinberg. 2018. Inherent Trade-Offs in Algorithmic Fairness. *SIGMETRICS Perform. Eval. Rev.* 46, 1 (jun 2018), 40. https://doi.org/10.1145/3292040.3219634
[42] Kimmo Kärkkäinen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1547–1557. https://doi.org/10.1109/WACV48630.2021.00159
[43] Jie Li, Yongli Ren, and Ke Deng. 2022. FairGAN: GANs-Based Fairness-Aware Learning for Recommendations with Implicit Feedback. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 297–307. https://doi.org/10.1145/3485447.3511958
[44] Bryson Lingenfelter, Sara R. Davis, and Emily M. Hand. 2022. A Quantitative Analysis of Labeling Issues in the CelebA Dataset. In *Advances in Visual Computing: 17th International Symposium, ISVC 2022, San Diego, CA, USA, October 3–5, 2022, Proceedings, Part I* (San Diego, CA, USA). Springer-Verlag, Berlin, Heidelberg, 129–141. https://doi.org/10.1007/978-3-031-20713-6_10
[45] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*. https://ieeexplore.ieee.org/document/7410782
[46] Kristian Lum and William Isaac. 2016. To Predict and Serve? *Significance* 13, 5 (10 2016), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x arXiv:https://academic.oup.com/jrssig/article-pdf/13/5/14/49106470/sign960-sup-0001-s1.pdf
[47] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the Applicability of Machine Learning Fairness Notions. *SIGKDD Explor. Newsl.* 23, 1 (may 2021), 14–23. https://doi.org/10.1145/3468507.3468511
[48] Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. 2023. Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet. arXiv:2306.06130 [cs.AI]
[49] G. J. McLachlan. 1975. Iterative Reclassification Procedure for Constructing an Asymptotically Optimal Rule of Allocation in Discriminant Analysis. *J. Amer. Statist. Assoc.* 70, 350 (1975), 365–369. https://doi.org/10.1080/01621459.1975.10479874 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1975.10479874
[50] Xuran Meng and Jeff Yao. 2023. Impact of classification difficulty on the weight matrices spectra in Deep Learning and application to early-stopping. *Journal of Machine Learning Research* 24, 28 (2023), 1–40. http://jmlr.org/papers/v24/21-1441.html
[51] Robert K. Nelson, LaDale Winling, Richard Marciano, Nathan Connolly, and et. al. 2020. Mapping Inequality. https://dsl.richmond.edu/panorama/redlining/#loc=5/39.1/-94.58
[52] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf
[53] NIST. 2018. 2018 Differential Privacy Synthetic Data Challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic,2018a.
[54] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. 2017. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HkwoSDPgg
[55] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. *CoRR* abs/2002.06673 (2020). arXiv:2002.06673 https://arxiv.org/abs/2002.06673
[56] RAD 2008. *Rules for Archival Description (RAD)*. Standard. Bureau of Canadian Archivists Planning Committe on Descriptive Standards. https://archivescanada.ca/wp-content/uploads/2022/08/RADComplete_July2008.pdf
[57] Spencer Rarrick, Chris Quirk, and William D. Lewis. 2011. MT Detection in Web-Scraped Parallel Corpora. In *Machine Translation Summit*. https://api.semanticscholar.org/CorpusID:2289219
[58] Rashida Richardson, Jason Schultz, and Kate Crawford. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. https://ssrn.com/abstract=3333423
[59] Mary Romero. 2017. *Introducing intersectionality*. John Wiley & Sons.
[60] Chirag Shah and Emily M. Bender. 2023. Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? *Under Review at non-double blind venue. September 1 version.* (2023). https://faculty.washington.edu/ebender/papers/Envisioning_IAS_preprint.pdf
[61] Abas Shkembi, Lauren M. Smith, and Richard L. Neitzel. 2022. Linking environmental injustices in Detroit, MI to institutional racial segregation through historical federal redlining. *Journal of Exposure Science and Environmental Epidemiology* (2022). https://doi.org/10.1038/s41370-022-00512-y
[62] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv:2305.17493 [cs.LG] https://arxiv.org/abs/2305.17493
[63] W. So, P. Lothia, R. Pimplikar, A.E. Hosoi, and C. D'Ignazio. 2022. Beyond Fairness: Reparative Algorithms to Address Historical Injustices of Housing Discrimination in the US. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533160
[64] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic Data – Anonymisation Groundhog Day. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 1451–1468. https://www.usenix.org/conference/usenixsecurity22/presentation/stadler
[65] A.K Subramanian. 2020. PyTorch-VAE. https://github.com/AntixK/PyTorch-VAE.
[66] Shiva Kanth Sujit. 2019. VAE-Pytorch. https://github.com/shivakanthsujit/VAE-PyTorch/tree/master.
[67] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (–, NY, USA) *(EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, Article 17, 9 pages. https://doi.org/10.1145/3465416.3483305
[68] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM. https://doi.org/10.1145/3465416.3483305
[69] Rohan Taori and Tatsunori Hashimoto. 2023. Data Feedback Loops: Model-driven Amplification of Dataset Biases. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 33883–33920. https://proceedings.mlr.press/v202/taori23a.html
[70] Ding Tong, Qifeng Qiao, Ting-Po Lee, James McInerney, and Justin Basilico. 2023. Navigating the Feedback Loop in Recommender Systems: Insights and Strategies from Industry Practice. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1058–1061. https://doi.org/10.1145/3604915.3610246
[71] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv:2306.07899 [cs.CL]
[72] Wanjun Wu. 2022. Machine Learning Approaches to Predict Loan Default. *Intelligent Information Management* 14, 5 (2022), 157–164. https://www.scirp.org/journal/paperinformation.aspx?paperid=120102
[73] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 325–333. https://proceedings.mlr.press/v28/zemel13.html
[74] Zhaowei Zhu, Tianyi Luo, and Yang Liu. 2022. The Rich Get Richer: Disparate Impact of Semi-Supervised Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=DXPftn5kjQK

A TABLE OF NOTATIONS

To aid in clarity, we present common notation used throughout the paper in Table 2.

B FAIRNESS IN MACHINE LEARNING (FML)

Suresh and Guttag [67] describe several sources of bias and oppression that may be encoded into a model as a result of the processes for gathering and encoding data, training, evaluating, and deploying the model. These sytems may then commit several types of harms, including allocative (where resources are withheld from certain groups, such as in redlining) and representational (where groups are stigmatized and stereotyped). Likewise, there are many different biases that may co-occur such as historical, representational,

Symbol		Description
$\mathcal{H}$	$\triangleq$	The original reference distribution before MIDS occur. $\mathcal{H} = \mathcal{X} \times \mathcal{L} \times \mathcal{S}$, where $\mathcal{X}$ represents the inputs, $\mathcal{L}$ the labels, and $\mathcal{S}$ the sensitive attribute(s).
D	$\triangleq$	Dataset sampled from $\mathcal{H}$, where $D = X \times L \times S$.
G_0	$\triangleq$	The generative model that represents a generative oracle for $\mathcal{X}$. G_0 is trained from D.
A_L	$\triangleq$	The labelling oracle model that represents the mapping $\mathcal{X} \to \mathcal{L}$. Model is trained from D.
A_S	$\triangleq$	The sensitive attribute oracle model that represents the mapping $\mathcal{X} \to \mathcal{S}$. Model is trained from D.
C_i	$\triangleq$	A classification model corresponding to generation i, trained using algorithm $\mathcal{T}_C(\cdot)$.
G_i	$\triangleq$	A generative model corresponding to generation i, trained using algorithm $\mathcal{T}_G(\cdot)$ It may be used to generate samples g_i.
$\mathcal{H}_{\rangle}$	$\triangleq$	The non-synthetic data distribution arising from MIDS in generation i. Samples h_i may be drawn to train models C_i and/or G_i.
		STAR parameters
STRATA	$\triangleq$	A categorical distribution over the Cartesian product of sensitive attribute values and label values. If multiplied by the number of samples to draw, it provides a quota for the number of samples to draw from each sensitive attribute and label category.
b	$=$	The size of the batch created by STAR.
r	$=$	The reparation budget. STAR draws a pool of $b + r$ samples, then attempts to meet the quotas for each category as given by $b\times$STRATA for a total of b samples in the returned minibatch.

Table 2: Table of notations.

measurement, aggregation, learning, evaluation, and deployment biases [68]. These often arise from misrepresenting a complex feature (e.g., treating gender or sex as a binary), mis-measuring features, stripping data of its context (e.g., regional or dialectal language heteroglossia), and from historical oppression influencing the data modelling processes. There are several frameworks for defining and addressing issues of fairness; calibration (used when the sensitive identities have impact on the decision task), anti-classification (used when sensitive data is unavailable or illegal to use), individual fairness ("similar individuals should be treated similarly"), and classification parity.

In group fairness, protected attributes are often chosen from legally-protected attributes such as race or gender, and encoded into categorical features to determine *sensitive groups*. In this paper we use the terms *majoritized* and *minoritized* as in D'Ignazio and Klein [22] to emphasize the impact of a model's behavior on a group. Note that the majority population might not correspond with the majoritized (benefit-receiving) group; for example the Black population is a minoritized majority in the COMPAS dataset [20]. This grouping often splits the dataset into two groups, delineated by one attribute with two possible values (e.g., 'male' vs 'female' or 'white' vs 'people of color') which may misrepresent or inappropriately group populations and ignores the compounding impact of possessing multiple marginalized identities.

In classification parity, there are a variety of metrics that aim for some equality of rates between these groups, such as accuracy, positive selection rate, or error rates. In this work, we use accuracy difference, demographic parity difference, and equalized odds difference to cover a multitude of differing priorities model owners may value. It is often impossible to satisfy multiple fairness metrics simultaneously, so they are ideally chosen based upon the task [14, 41]. The binary classification and binary grouping versions of these metrics are presented below.

Definition 1 (Demographic Parity (DP) [10]). *A classifier $\hat{Y}$ satisfies Demographic Parity with respect to the sensitive attribute s if:*

$$\mathbb{P}(\hat{Y} = 1 | s = 0) = \mathbb{P}(\hat{Y} = 1 | s = 1) \qquad \forall 0, 1 \in s.$$

In this work we consider demographic parity difference, which is the absolute value of the difference between the two terms equated above. Each term is also the selection rate, or rate of positive prediction, for the group.

Definition 2 (Equalized Odds (EOdds) [30]). *A classifier $\hat{Y}$ satisfies Equalized Odds with respect to the sensitive attribute s if for ground truth L:*

$$\mathbb{P}(\hat{Y} = 1 | L = l, s = 0) = \mathbb{P}(\hat{Y} = 1 | L = l, s = 1) \qquad \forall l \in \{0, 1\}, \forall 0, 1 \in s.$$

We also use equalized odds difference. This is formulated as $\max[|\mathbb{P}(\hat{Y} = 1|L = 0, s = 0) - \mathbb{P}(\hat{Y} = 1|L = 0, s = 1)|, |\mathbb{P}(\hat{Y} = 1|L = 1, s = 0) - \mathbb{P}(\hat{Y} = 1|L = 1, s = 1)|]$, or the larger of the absolute value differences between the false and true positive rates for the groups. These metrics may be used for multiple groups by taking each difference between every pair of groups and reporting the maximal disparity. Similarly, these groups may be formed by intersecting multiple sensitive attributes.

Due to the biases that may exist in data collection, training, evaluation, and deployment, adherence to or achievement of any of these fairness metrics does not guarantee fairness or equity. For example, these fairness metrics assume that the cost of an error borne by a person of any group is the same, when in practice the costs and benefits may differ greatly depending on identity [47]. There are several works that focus on the trade-offs between meeting a decision maker's FML criteria and the potentially inequitable social outcomes which cast doubt on the suitability of FML metrics for societal welfare [16, 33, 39]. This is in part a byproduct from FML's reliance upon algorithmic idealism, where computation assumes a meritocratic society whereby equalizing demographic disparities must therefore lead to fairness at a societal level [18, 26, 39]. Additionally, FML may also engage in or reinforce two main biases: 1) automation bias, that machines are objective and are less biased than humans, and 2) that automation invites justice without regard for the objective and purpose of the models [18, 26].

C MIDS IN LITERATURE

This section provides a more detailed review of the MIDS and enablers described in Section 2.

C.1 MIDS

Performative prediction. *Performative prediction* is a distribution shift that occurs when a model's predictions impact the outcome. For example, when economists publish forecasts, they may influence the behavior of others in the market, causing the market to fit the forecast in a self-fulfilling prophesy [29, 55]. In this case, the model's predictions leak into the data ecosystem as they become outcomes. If this data ecosystem is trained upon, these outcomes are treated as the ground truth, and the MIDS continues into another generation of models.

Fairness feedback loops. The fairness community has studied performative prediction and fairness feedback loops in the context of risk assessment systems, including mortgaging and predictive policing [26]. We provide two notable examples:

1) Figure 6 shows the 1939 HOLC Residential Security Maps for Detroit alongside 1955 demographic information of non-white communities. This map is a historical multiclass classification model that reflects the values and priorities of the individuals and institutions responsible for its creation; specifically the white male gaze of Depression-era professional realtors [22, 36, 63]. These values encoded into this model and the MIDS from the model itself has contributed to housing discrimination in Detroit, as seen in the right of Figure 6.

2) Work in predictive policing found that policing locations may converge towards over-policing low-income non-white communities [46, 58]. Theoretical follow-ups find that the degree of runaway feedback may be moderated with careful training set weighting, but cannot be negated entirely [24].

Model Collapse. Where performative prediction and runaway feedback loops generally refer to classification models, model collapse describes the same effect for generative models. *Model collapse* occurs when new generative models are trained on samples created by their predecessor over many generations, as introduced in Shumailov et al. [62] and concurrently in Alemohammad et al. [4]. This leads to new models forgetting the original data distribution as they recreate and amplify the failures of their ancestors. There are two error sources that contribute: 1) functional approximation error due to an inadequately expressive generator, and 2) statistical approximation error from finite sampling. Model collapse begins with a loss of information from the tails of the data distribution. In late-stage model collapse, the model mixes the modes of the original distribution, converging to a point estimate of some mean betwixt them. There are also concerns over the effects of model collapse on fairness, as model failure on "low-probability events" may have negative effects on minoritized groups when datasets have poor representation [68].

Disparity amplification. Unlike the aforementioned MIDS, disparity amplification arises from human-model interaction. If a model suffers from problems derived from representational bias, it may have overall high performance but low performance on minoritized groups (performance disparity). This can lead to *disparity amplification*, where minoritized users who suffer high error rates may choose to disengage from the model, shifting the future dataset towards the majoritized group, and increasing the representational disparity of the data ecosystem [31].

C.2 MIDS Enablers

We describe several enablers in more detail here. Note that even sampling is an enabler, and indeed it informs our approach to algorithmic reparation in our experiments.

Pseudo-labelling. *Pseudo-labelling* generally refers to using a model to assign labels to unlabeled samples in a dataset, so that this data may also be used for supervised or semi-supervised training [12]. This may occur just once [12] or iteratively [49]. Incentives for pseudo-labeling may arise in cases where manual labeling and/or annotation is too expensive for vast swathes of data. The fairness impact of using pseudo-labelling for self-supervised learning was discussed in Zhu et al. [74], finding that groups with high initial accuracy benefit whereas groups with low initial accuracy may see a degradation in performance.

Feedback and Data Annotation. Similarly to pseudo-labelling, feedback (whether human or model-based) is often used for labeling and annotating data for supervised training or for providing feedback on generative outputs. Reliance on human annotation can lead to unfairness arising from individual annotator bias and instructions for annotating [68]. Recent work has found indicators that some human data annotators use LLMs or other models, which may lead to MIDS if these models are updated and retrained on the data they labeled [71]. AI feedback is also used in methods such as Constitutional AI, which uses a succession of fine-tuned supervised models to provide RLAIF (reinforcement learning from AI feedback) for training 'harmless' AI assistants [6].

D EXPERIMENTAL DETAILS

D.1 Datasets

We evaluate the fairness effects of model collapse and algorithmic reparation on several datasets; adapted versions of `MNIST` and `SVHN`, as well as `CelebA` and `FairFace`.

`ColoredMNIST` `MNIST` is a single-channel handwritten digit recognition dataset [19]. We use 50000 images for training, 10000 for validation, and another 10000 for testing. We adapt `MNIST` to a binary classification scheme (determining if a digit is in [0..4] for class 0 or [5..9] for class 1). The class label is switched with a uniform probability of 5% to add label noise, as in Arjovsky

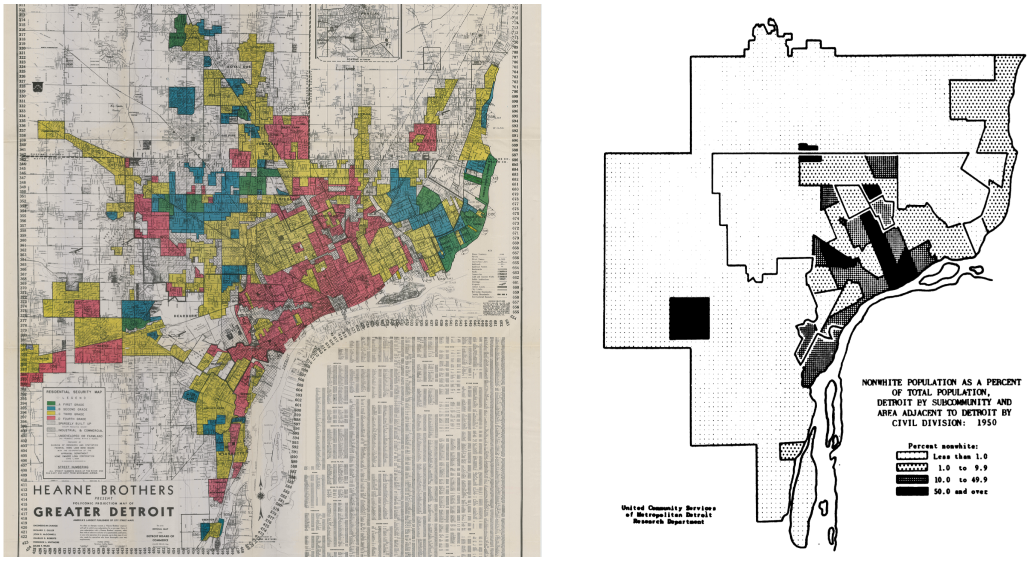


Figure 6: *Left:* 1939 Home Owner Loan Corporation Security Map for Detroit. Red areas are Grade D, or 'hazardous' locations due to the presence of racial and religious minorities [51]. Banks of the time were likely privy to these otherwise secret maps and in surveys stated that only high-graded neighborhoods would be offered loans [36, 51]. Redlining continued *de jure* until the Fair Housing Act in 1968 [1]. *Right:* Census data on the non-white population in metropolitan Detroit from 1955 [21].

ColoredMNIST	Class	
Group	Beneficial	Detrimental
Majoritized	0.350	0.150
Minoritized	0.150	0.350

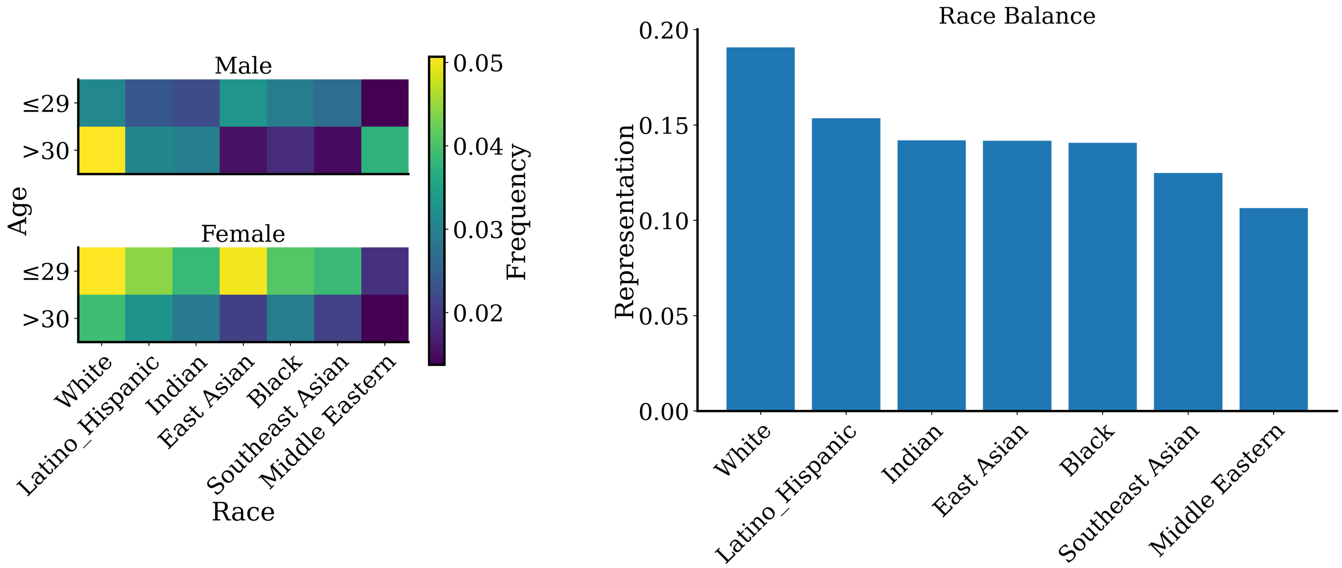
ColoredSVHN	Class	
Group	Beneficial	Detrimental
Majoritized	0.424	0.118
Minoritized	0.183	0.275

CelebA	Class	
Group	Beneficial (Attractive)	Detrimental
Majoritized (Not Male)	0.396	0.184
Minoritized	0.117	0.302

Table 3: Class and group demographics of training datasets. Values show the proportion of that group–class category in the training dataset (and therefore sum to 1). *Top:* ColoredMNIST and ColoredSVHN the majoritized is group skewed towards the beneficial class with probability 0.7, and to the detrimental class with probability 0.3. *Bottom:* CelebA.

et al. [5]. We also adapt MNIST to have binary groups by coloring the sample either red or green, where green is treated as the 'majoritized' group. We skew the majoritized group to the beneficial class, such that $\mathbb{P}(S = \text{majoritized}|L = \text{beneficial}) = 0.7$, $\mathbb{P}(S = \text{majoritized}|L = \text{detrimental}) = 0.3$. In this case, both classes and groups are balanced, as seen in the dataset composition matrix in Table 3. Ablations for this skew and class and label balances are in Appendix E.

ColoredSVHN SVHN (Street View House Numbers) is a digit recognition dataset composed of house numbers sourced from Google Street View [52]. We use 52327 images for training, 20930 for validation, and another 26032 for testing. For SVHN, we adapt to a binary task similarly as in MNIST. We binarize the classification task to determining if a digit is in [0..4] for class 1 or [5..9] for class 0, where class 1 is the beneficial class as converged to by model collapse. The class label is swapped with a uniform probability of 5% to add label noise, as in Arjovsky et al. [5]. Unlike in ColoredMNIST, class 1 is the lower numbers as SVHN converges to small numbers over the course of model collapse, as seen in Figure 14. This causes class imbalance; class 1 composes 60.7% of the data. We also add sensitive groups by converting the images to grayscale and then coloring the samples either red or green as in ColoredMNIST. The green group again serves as the majoritized group, and is skewed towards the beneficial class at rates $\mathbb{P}(S = \text{majoritized}|L = \text{beneficial}) = 0.7$, and $\mathbb{P}(S = \text{majoritized}|L = \text{detrimental}) = 0.3$, leading to group imbalance with the majoritized group as 54.3%. A matrix showing the composition of the ColoredSVHN training distribution is shown in Table 3.



Age	≤ 29	> 30
	0.541	0.459
Gender	Male	Female
	0.531	0.469

Figure 7: Class (gender) and group (race and binarized age) balance for the `FairFace` training set.

`CelebA` `CelebA` is a dataset of celebrity faces [45]. We use an 80/10/10 train/validation/test split of the 202599 cropped and aligned images. The binary classification task is to predict attractiveness, where sensitive groups are given from gender ('Male,' 'not Male'). The group and class balance is 58.% 'not male' and 51.3% 'attractive.' The composition of the dataset is shown in Table 3. This dataset is criticized for the inclusion of subjective features such as 'attractive,' and there are many instances of incorrect labeling and annotation [44]. In the other datasets, we chose the beneficial class as the class most converged to by model collapse. Interestingly, against the class imbalance, model collapse converges to 'unattractive,' which would benefit the 'Male' group more. However, model collapse also converges to 'Not Male.' We therefore use 'attractive' as the beneficial class and 'Not Male' as the majoritized group.

`FairFace` `FairFace` is a face dataset that is balanced by both gender (two values) and race (seven values) [42]. The composition, including the intersections of age, race, and gender, are shown in Figure 7. We use an 80/10/10 train/validation/test split of the 108501 images. For binary classification, we predict gender ('Male,' 'not Male'). For sensitive features, we use S_1 as race, which has seven potential values ('White', 'Latino Hispanic', 'Indian', 'East Asian', 'Black', 'Southeast Asian', 'Middle Eastern'), and S_2 as age, which we binarize as above and below 30 years (see Figure 7 for the class and group balances). For the group annotation oracles, we have two separate classifiers corresponding to these features. The beneficial class is 'not Male,' as model collapse increases the representation of this class, even through they are a minority in the original dataset class balance. As there are 14 categories created from the intersection of age and race, we instead track the representations of both of these attributes alone and the largest and smallest categories among these over time. Note that these choices for the beneficial and majoritized annotations are arbitrary for these experiments as we do not make claims of justice (recall Section 3.4).

Justification of dataset choice. We choose `ColoredMNIST` and `ColoredSVHN` due to their similarity. Both detect and classify digits, and may be easily adapted into a binary classification and binary fairness grouping task. They differ in their complexity, `SVHN` is usually the harder dataset to learn, and also in their class balance once binarized (see Table 3). These two datasets, despite their similarities, show very different points of model collapse and even opposite behavior when observing the accuracy difference between groups, as in Figure 5 and Figure 20. `CelebA` is chosen to represent a more complex and real-world dataset with well-documented disparities between the class and grouping we use for its task [44]. On the other hand, `FairFace` is chosen for its balance in both race and gender; alongside other metrics, it is simple to measure sensitive group intersections.

D.2 Compute and codebase

Experiments were performed on Ubuntu 18.04.6 LTS using 4 Intel Xeon CPU cores per GPU. We use the following GPUs per random seed: for `ColoredMNIST` we use 1 NVIDIA T4 GPU with 16 gigabytes of memory; for `ColoredSVHN` we use 2 NVIDIA RTX6000s with 40 gigabytes each; for `CelebA` we use 2 NVIDIA A100s with

40 gigabytes each (split GPUs); for `FairFace` we use 2 NVIDIA A40s with 48 gigabytes each. Our codebase is in Python 3.9 with PyTorch and fairness metrics from Fairlearn [8]. Existing code was adapted for our experiments: VAE models from Subramanian [65], `ColoredMNIST` from Arjovsky et al. [5], SVHN with fairness from Kenfack et al. [40], and ResNets from Idelbayev [35].

D.3 Models

The architectures and hyperparameters differ based on dataset. The performance of the annotator models A_L and A_S are shown in Table 4. All generative models were trained with ADAM (weight decal 1×10^{-5}), and all classifiers with SGD and cross-entropy loss.

`ColoredMNIST`

- Classifiers and annotators are 6 layer CNNs with 2 convolution layers and ReLU activations. Learning rate 0.1, batch size 256, for 30 epochs.
- Generators (VAEs) are mirrored encoder and decoder CNNs. Each is 2 convolution layers with ReLU activations. Uses BCE Loss with KL-divergence term, a latent space dimension of 20, and a variational beta of 1. Learning rate 0.001, batch size 256, for 30 epochs.

`ColoredSVHN`

- Classifiers and annotators are 32-layer ResNets adapted from Idelbayev [35]. Learning rate 0.001, batch size 32, for 30 epochs.
- Generators are the deep convolutional VAE adapted from Sujit [66], using MSE loss with KL-divergence term, a latent space dimension of 32, and a variational beta of 1. Learning rate 0.0005, batch size 128, for 30 epochs.

`CelebA`

- Classifiers and annotators are 110-layer ResNets adapted from [35]. Learning rate 0.001, batch size 128, for 15 epochs.
- Generator VAEs are composed of a 5-layer CNN encoder and 6-layer upsampling CNN decoder with LeakyReLU activations. Loss is BCE with KL-divergence term, a latent space dimension of 500, with a variational beta of 5×10^{-6}. Learning rate 0.005, batch size 64, for 30 epochs.

`FairFace`

- Classifiers and annotators are pretrained 50-layer ResNets adapted from [35]. Pretraining is on Imagenet, using the version 1 weights from PyTorch. Learning rate 0.001, batch size 256, for 30 epochs.
- Generator VAEs are composed of a 5-layer CNN encoder and 6-layer upsampling CNN decoder with LeakyReLU activations. Loss is MSE with KL-divergence term, a latent space dimension of 500, with a variational beta of 1×10^{-6}. Learning rate 0.0001, batch size 256, for 50 epochs.

D.4 STAR Algorithm

Algorithm 1 shows an example of STAR for binary group and binary sensitive attributes.

Algorithm 1: Training with algorithmic reparation batches.

Input: Sample-providing generator G, batch size b, reparation budget r, label annotator C (either C_{i-1} or A_L), sensitive attribute annotator A_S.
Output: Reparation batch

for batch in number batches **do**
  Ideal = $[b/4, b/4, b/4, b/4]$ ▷ Ideal category sizes
  Batch = $[b_{L=0,S=0}, b_{L=0,S=1}, b_{L=1,S=0}, b_{L=1,S=1}] = [0, 0, 0, 0]$ ▷ Initialize batch categories
  Temporary batch = Sample $b + r$ times from G ▷ Initial batch from uniform sampling
  Annotate temporary batch using C and A_S
  Categorize batch depending on L and S values from annotations
  Populate Batch until Ideal$_i$ =Batch$_i$
  To_resample =sum(Ideal − Batch) ▷ Get amount to sample to fill deficient categories
  Batch.append(Sample To_resample times from G, annotate with C and A_S) ▷ Refill batch
Update model on Batch.
return Batch

E APPENDIX: ABLATION STUDIES

In this appendix we provide experiments to demonstrate the effects of MIDS over several ablated variables: the sensitive group imbalance, class imbalance, and amount of synthetic training data. We provide results for both `ColoredMNIST` and `ColoredSVHN`.

E.1 Class and group imbalance

In these studies we varied the class balance or group balance. The study was carried out on `ColoredMNIST` with 5 seeds in the sequential generator and classifier setting. For group imbalance, the groups were equally likely to belong to the beneficial class, though their populations were varied. For class imbalance, the majoritized group was skewed towards the beneficial class in the same manner as discussed in Appendix D.1, and the class population varied. For this task, the variations in balance did not strongly effect the generated population or downstream classifier performance. The generator class and group balances are shown for varied group balance in Figure 8 and for varied class balance in Figure 9. The results in Section 4.2.2 use datasets with a mixture of class and group imbalance which better elucidate the effects of MIDS.

E.2 Amount of synthetic data

In this study we varied the amount of original training data (drawn randomly from the training set) in each batch for training generators in the sequential generator and classifier setting. These experiments were carried out on `ColoredMNIST` for 5 seeds. There is a substantially higher accuracy cost and accuracy disparity between groups, as shown in Figure 10. Note that even with 0% synthetic data (i.e., training each generator from the original training set) there is still an accuracy loss over time due to the effects of the sequential classifiers. While the group balance is not hugely effected (as in

Table 4: Performances (with standard deviations) of the label and sensitive attribute annotator models A_L and A_S. Performances are shown for each dataset. The performance of A_S is high for `ColoredMNIST` and `ColoredSVHN` as determining sample color is an easy task. The fairness metrics for A_S should be close to 1, as these models should assign class based on the sensitive attribute alone. Reported accuracies are all macro-averaged. Note the high accuracy disparity for `FairFace` A_{S_1}, although racial groups are roughly balanced, we observed far higher accuracy for 'white' than any other group, perhaps due to simplicity bias [7].

	ColoredMNIST		ColoredSVHN		CelebA	
	A_L	A_S	A_L	A_S	A_L	A_S
Accuracy	0.928 ± 0.003	1 ± 0	0.849 ± 0.080	1 ± 0	0.816 ± 0.005	0.976 ± 0.002
Δ Accuracy	0.009 ± 0.005	0 ± 0	0.052 ± 0.111	0 ± 0	0.029 ± 0.008	0.015 ± 0.008
Δ DP	0.367 ± 0.008	1 ± 0	0.151 ± 0.193	1 ± 0	0.440 ± 0.022	0.951 ± 0.004
Δ EOdds	0.032 ± 0.022	1 ± 0	0.163 ± 0.240	1 ± 0	0.271 ± 0.037	0.971 ± 0.008

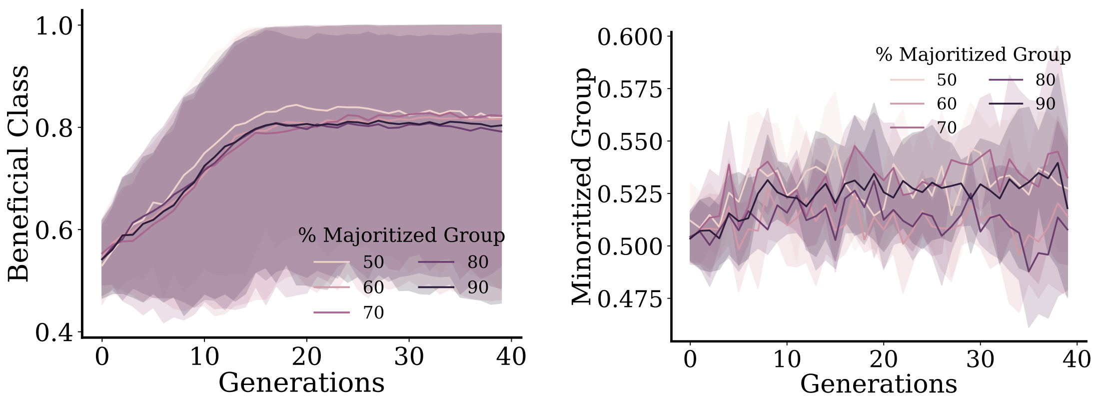


Figure 8: `ColoredMNIST` class and group balance while varying the group balance in SeqGenSeqClass.

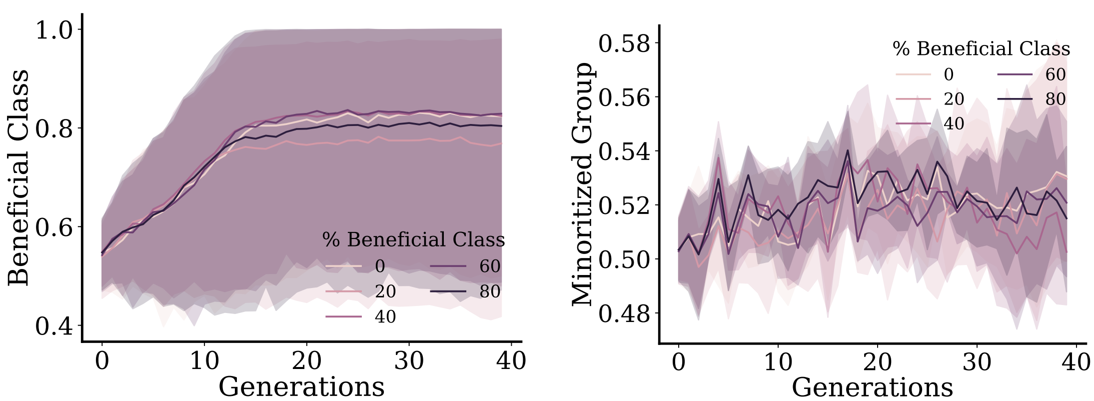


Figure 9: `ColoredMNIST` class and group balance while varying the class balance in SeqGenSeqClass.

the other ColoredMNIST results in Section 4.2), the class balance skews towards the beneficial class over the generations, fueling an increase of equalized odds difference with more synthetic data, see Figure 10.

In practice, there may be several generations of synthetic data present when drawing from a corpus of polluted data. For example, when training G_2, samples from G_0 and G_1 might also be present. In this case, the compounded artefacts of model collapse will be lesser in these early generations. In this study, the synthetic data is only pooled from the most recent generator, and so these results may overstate the effect of model collapse in the aforementioned case.

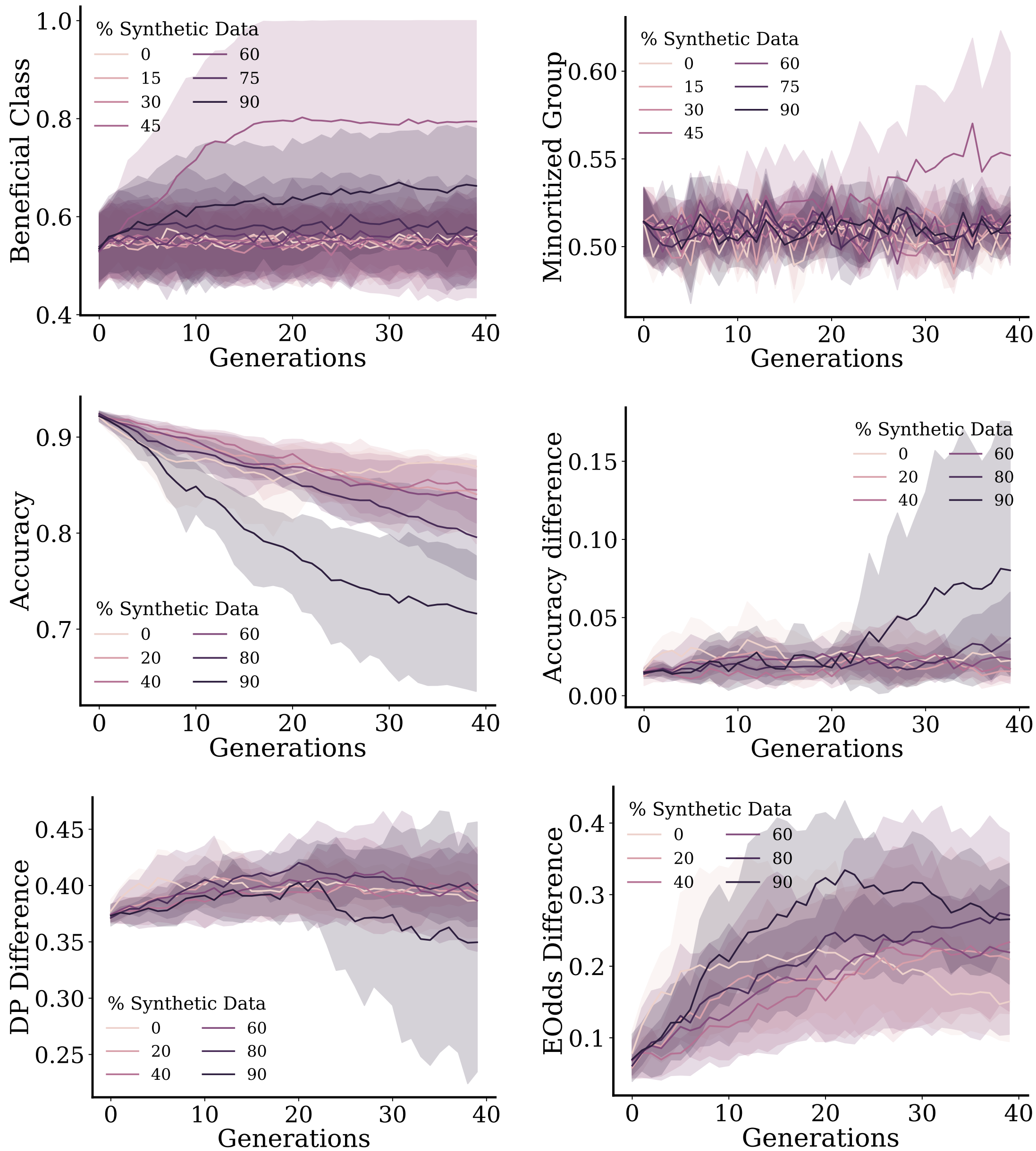


Figure 10: `ColoredMNIST` metrics while varying the amount of synthetic data in SeqGenSeqClass. *Top:* beneficial class balance and group balance. *Center:* accuracy, and accuracy difference between groups. *Bottom:* DP and EOdds difference. These generally show worse performance and fairness with more synthetic data, with larger variances.

	`FairFace`		
	A_L	A_{S_1}	A_{S_2}
Accuracy	0.884 ± 0.013	0.617 ± 0.010	0.789 ± 0.017
Δ Accuracy	0.104 ± 0.011	0.428 ± 0.088	0.251 ± 0.094
Δ DP	0.360 ± 0.023	0.443 ± 0.057	0.683 ± 0.031
Δ EOdds	0.308 ± 0.014	0.438 ± 0.139	0.254 ± 0.075

E.3 Sequential versus non-sequential classifiers in SeqGenSeqClass and SeqGenNonSeqClass

In this study we demonstrate the impact of sequential classifiers in SeqGenSeqClass. These experiments were conducted for `ColoredMNIST` and `ColoredSVHN` for 25 and 10 seeds, respectively.

The non-sequential classifiers are more sensitive to changes in the distribution, as seen in the accuracy over generations and selection rate graphs in Figure 11 and Figure 12 for ColoredMNIST and ColoredSVHN, respectively. This is consistent with the intuition that the sequential classifiers are 'adapting' their mapping of $\mathcal{X} \rightarrow \mathcal{L}$ to changes in $\mathcal{X}$ caused by model collapse. This allows the sequential classifiers to experience more generations of utility compared to the non-sequential classifiers, as seen in their higher accuracies. As model collapse causes strong imbalance towards the beneficial class (as determined by A_L), the non-sequential classifiers eventually only predict the positive label, decreasing DP and EOdds unfairness as both groups receive the same predictions and error rates (in ColoredMNIST, EOdds difference drops to 0 as error rates from only giving the positive prediction are identical due to group and class balance). Meanwhile, the sequential classifiers for ColoredMNIST andColoredSVHN instead evolve to only give a beneficial prediction to a majoritized sample, increasing unfairnesses (see Section 4.2.2).

Note that the achievement of higher fairness in the non-sequential classifier case indicates higher fairness with respect to the original distribution. This may be undesirable in some cases, particularly those applicable to algorithmic reparation, which specifically notes that equality of of model outputs to base rates in a dataset does not guarantee equity. This is especially true if the dataset is collected with any biases, including compounding Intersectional biases which these experiments do not inform upon [18].

F MODEL COLLAPSE IN GENERATORS

We show the losses with respect to the parent generator loss ($\mathcal{L}(G_i, G_{i-1})$) over generations as they undergo model collapse and while subject to GEN-STAR (which causes a minor adjustment). Intuition would suggest that the distribution collapses to be increasingly easy-to-learn, such that successive generators inherit simplified versions (due to finite sampling of their parents) of the problem and so perform better. We observe this effect with the smoothly decreasing loss curves of ColoredSVHN and FairFace in Figure 13.

However, for both ColoredMNIST and CelebA, we see the exact opposite curve. The child generators are faced with an increasingly hard-to-learn distribution. We hypothesize that this may be due to one of two causes. 1) We do not perform hyper-parameter tuning for the generators at each generation, and perhaps ColoredMNIST and CelebA experience hyperparameter instability. 2) Perhaps this is simply a quirk of model collapse, finite sampling of heavy-tailed distributions may lead to enough bias and noise to significantly complicate the learning task. We propose to investigate the stability of model collapse in future work.

We also provide some examples generated by generators undergoing model collapse in Figure 14.

G ADDITIONAL RESULTS

G.1 SeqClass Results

We provide full suites of figures for ColoredMNIST, ColoredSVHN, CelebA, and FairFace on SeqClass. See Figures 3, 15, 16 and 17, respectively.

G.1.1 *CLA-STAR Batch Balances.*

We report the composition of the batches used when training the classifiers with and without CLA-STAR in the SeqClass setting in Figure 18. These figures show the STRATA CLA-STAR uses to train classifiers, the resulting classifier STRATA, and the STRATA of classifiers trained without any reparation. Usually, the CLA-STAR STRATA are the most balanced, followed by the STRATA of classifiers that received reparation. In FairFace, the batches are mostly older white males, and sometimes younger white non-males; the least populated categories are usually younger people from the Indian, South East Asian, and Hispanic/Latino races.

G.2 SeqGenSeqClass Results

We provide full suites of figures for ColoredMNIST, ColoredSVHN, CelebA, and FairFace on SeqGenSeqClass. See Figures 19, 4, 5 and 20, respectively.

G.2.1 *STAR Batch Balances.*

We compare the composition of the batches used when training the classifiers with and without STAR in the SeqGenSeqClass setting. For ColoredMNIST, see Figure 21, ColoredSVHN, see Figure 22, and FairFace, see Figure 23. These figures show the STRATA STAR uses to train models, the resulting model STRATA, and the STRATA of models trained without reparation. In FairFace, the batches are mostly older white males, and sometimes younger white non-males; the least populated categories are usually people from the Indian, Middle Eastern, South East Asian, Black, and Hispanic/Latino races.

H RELATIVE PERFORMANCES OF MIDS

If the model trainer is unaware of MIDS occurring over time, they may see only the relative performances (*i.e.,* performance of generation i measured *w.r.t.* generation $i-1$) of each generation compared to its prior generation. In this case, when each generation of models is trained to have relatively high performance, it may look as though the models are performing well, though not when compared to the original data distribution. This may lead to overstating the model's performance, which for the FML metrics results in fairwashing the model due to inadequate validation and testing [3]. For ColoredMNIST and ColoredSVHN, we report results on the testing set for the 'actual' results (classifiers measured against the testing set) and for the relative results (classifiers measured against the previous generation's classifier predictions on the testing set inputs). We choose not to present these two graphs on the same plots to prevent confusion as they measure two different properties.

For reference, ColoredMNIST plots are in Figure 24 for SeqClass and Figure 26 for SeqGenSeqClass. ColoredSVHN plots are in Figure 25 for SeqClass and Figure 27 for SeqGenSeqClass.

These results also demonstrate how even when training each new classifier with a small tolerance for unfairness can accrue to high unfairness. For example, consider the relative equalized odds results in Figure 24 which on average stay below 0.06 for each generation accrue to over 0.2.

Figure 26 shows the point of model collapse in the SeqGenSeqClass setting in ColoredMNIST (collapse by generation 15) can be seen in the relative accuracy plot and in the increase in variance in the other relative plots. Similarly as found in the SeqClass plots discussed above, low relative equalized odds difference and

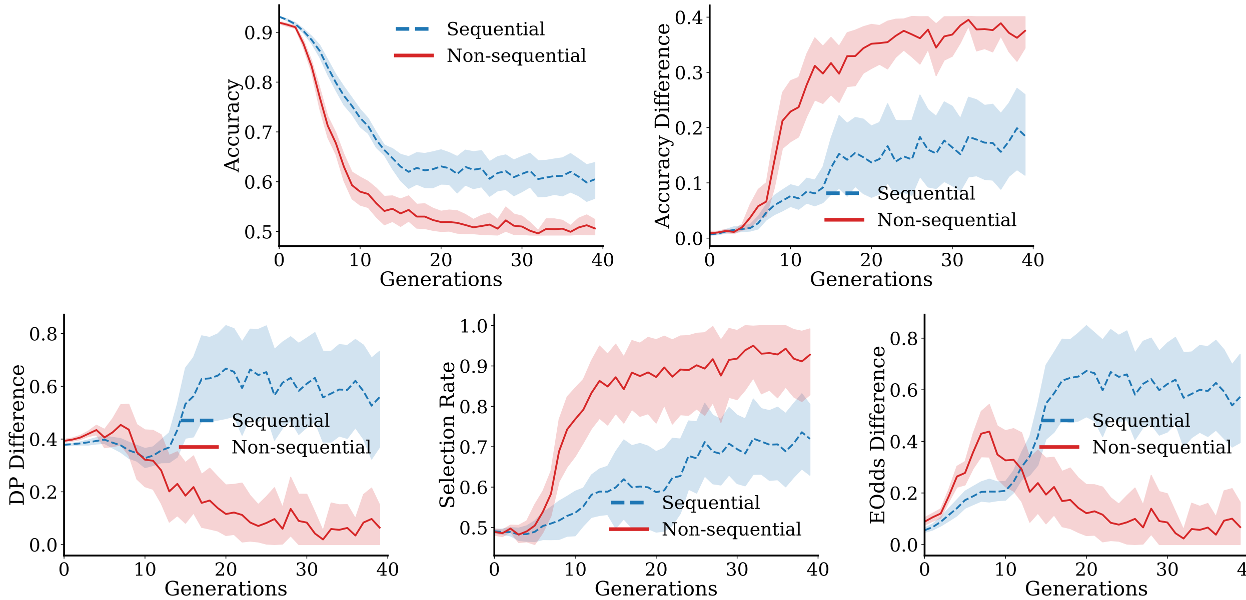


Figure 11: `ColoredMNIST` results for sequential versus non-sequential classifiers in SeqGenSeqClass and SeqGenNonSeqClass. *Top:* accuracy and accuracy difference between groups. *Bottom:* demographic parity difference, selection rate, and equalized odds difference.

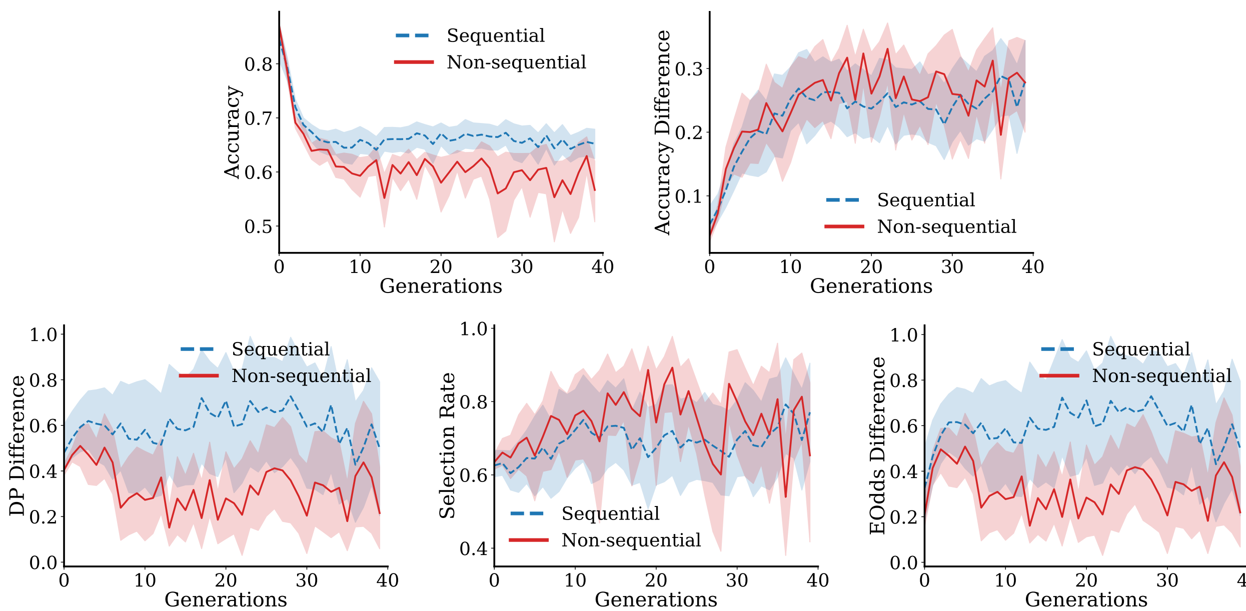


Figure 12: `ColoredSVHN` results for sequential versus non-sequential classifiers in SeqGenSeqClass and SeqGenNonSeqClass. *Top:* accuracy and accuracy difference between groups. *Bottom:* demographic parity difference, selection rate, and equalized odds difference.

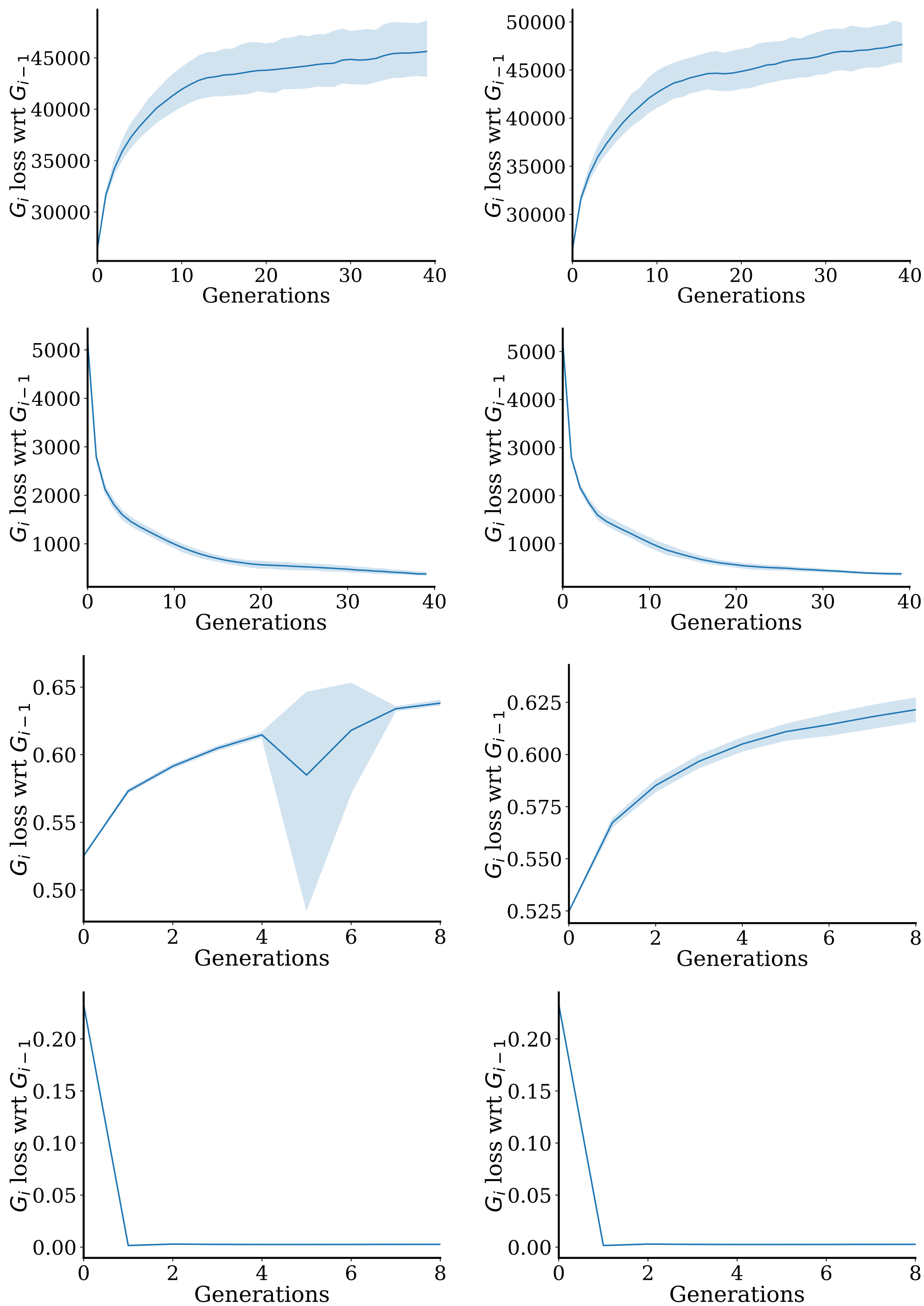


Figure 13: ***Top row*: `ColoredMNIST`, *Second row:* `ColoredSVHN`, *Third row:* `CelebA`, *Bottom row:* `FairFace`. *Left*: model collapse losses, *Right:* model collapse with GEN-STAR losses.**

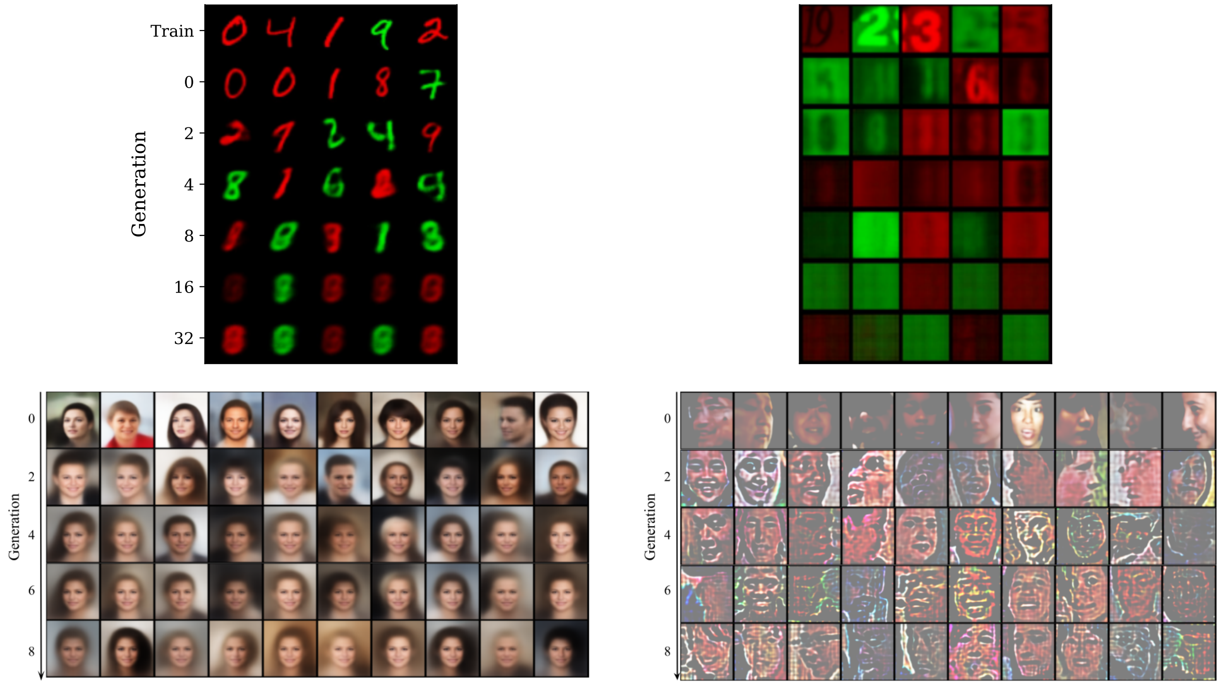


Figure 14: Samples from generators undergoing model collapse in SeqGenSeqClass for `ColoredMNIST` (*top left*), `ColoredSVHN` (*top right*), `CelebA` (*bottom left*), and `FairFace` (*bottom right*).

a relatively balanced minoritized group under-report the actual unfairness and imbalance.

I CO-OCCURING MIDS

In this section we present two sets of experiments to showcase how disparity amplification can co-occur with performative prediction, and also with model collapse. We evaluate these experiments for `ColoredMNIST`, where Figure 28 shows the SeqClass case, Figure 29 shows the SeqGenSeqClass case, and Figure 30 shows the strata for models trained in both settings.

For the SeqClass setting, we train each classifier in the lineage from a 50/50 mixture of data from G_0 and from the original training set. In the SeqGenSeqClass setting, the generators are trained from this data mixture, though the downstream classifiers are trained entirely from their corresponding generator's synthetic outputs. The inclusion of human-generated data moderates the degree of model collapse to showcase other effects.

For disparity amplification to co-occur, we use stratified sampling on the original training set portion of the data mixture, where the strata are determined by the classifier's label distribution over the groups. Note that this is not disparity amplification as discussed in Hashimoto et al. [31], which is due to performance failures, but instead due to label and group representation. This approximates the effects of the classifiers on the human-generated data distribution, and shows how this effects feeds into the other MIDS. We conduct additional experiments to showcase the effects of AR at the classifiers or generators, in isolation from and in combination with the disparity amplification sampling strategy.

Note that at the limit where there is no synthetic data, we recover the promising technical question of how to create a biased sampling mechanism that begets fairness in a downstream model trained from a generator.

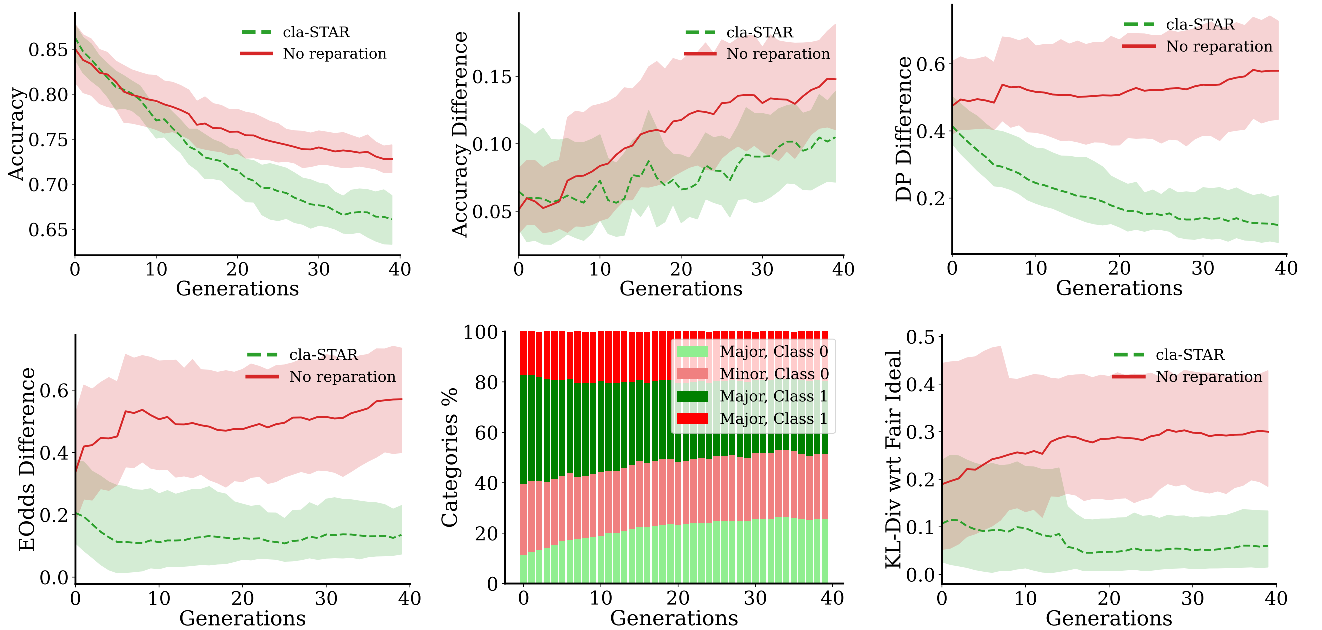


Figure 15: `ColoredSVHN` results for SeqClass on the evaluation set. *Top:* accuracy, accuracy difference, and demographic parity difference. We observe lower fairness differences with cla-STAR, with a cost of more inaccuracy. *Bottom:* equalized odds difference, the strata created during cla-STAR, and the KL-divergence between cla-STAR fairness ideal and classifier strata. The KL-Divergence decreases with cla-STAR as the batches become more evenly balanced across group and class.

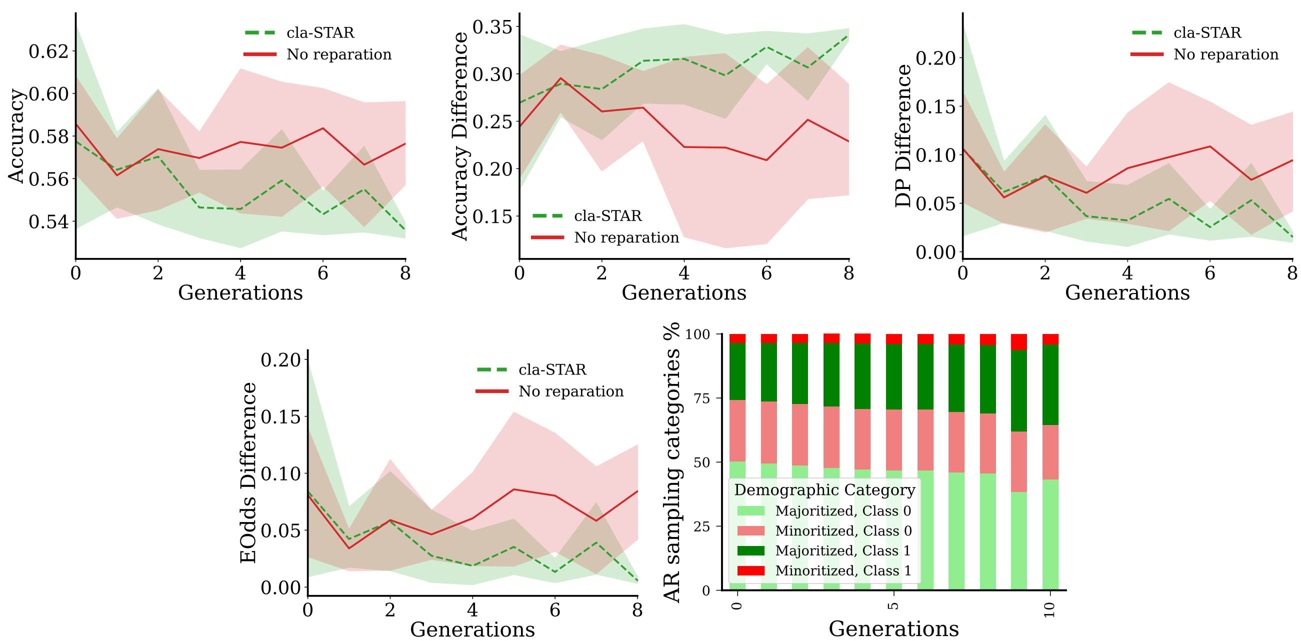


Figure 16: `CelebA` results for SeqClass on the evaluation set. *Top:* accuracy, accuracy difference, and demographic parity difference. Better fairness (lower fairness difference) and higher accuracy is achieved with cla-STAR. *Bottom:* equalized odds difference and the strata created during cla-STAR. The batches become more evenly balanced across group and class.

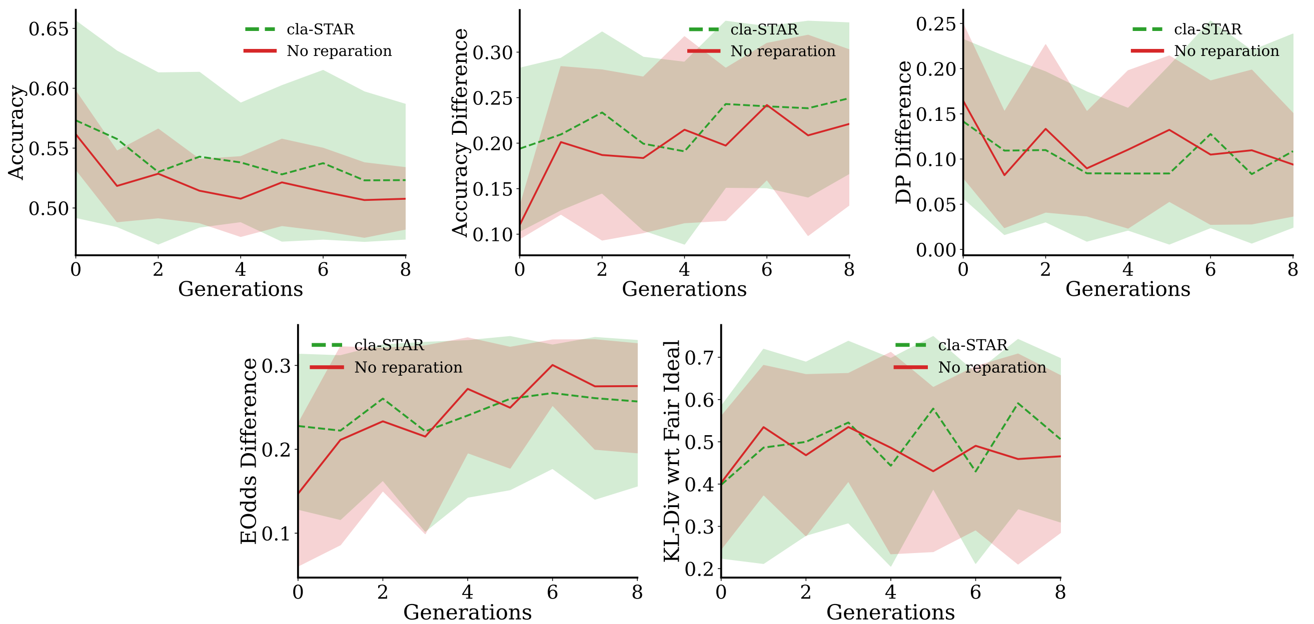


Figure 17: `FairFace` results for SeqClass on the evaluation set. *Top:* accuracy, accuracy difference, and demographic parity difference. We observe lower fairness differences with cla-STAR, with a cost of more inaccuracy. *Bottom:* equalized odds difference, and the KL-divergence between cla-STAR fairness ideal and classifier strata. We do not report the strata formed by cla-STAR as there are 28 categories; instead, refer to the strata of the classifiers in Appendix G.1.1.

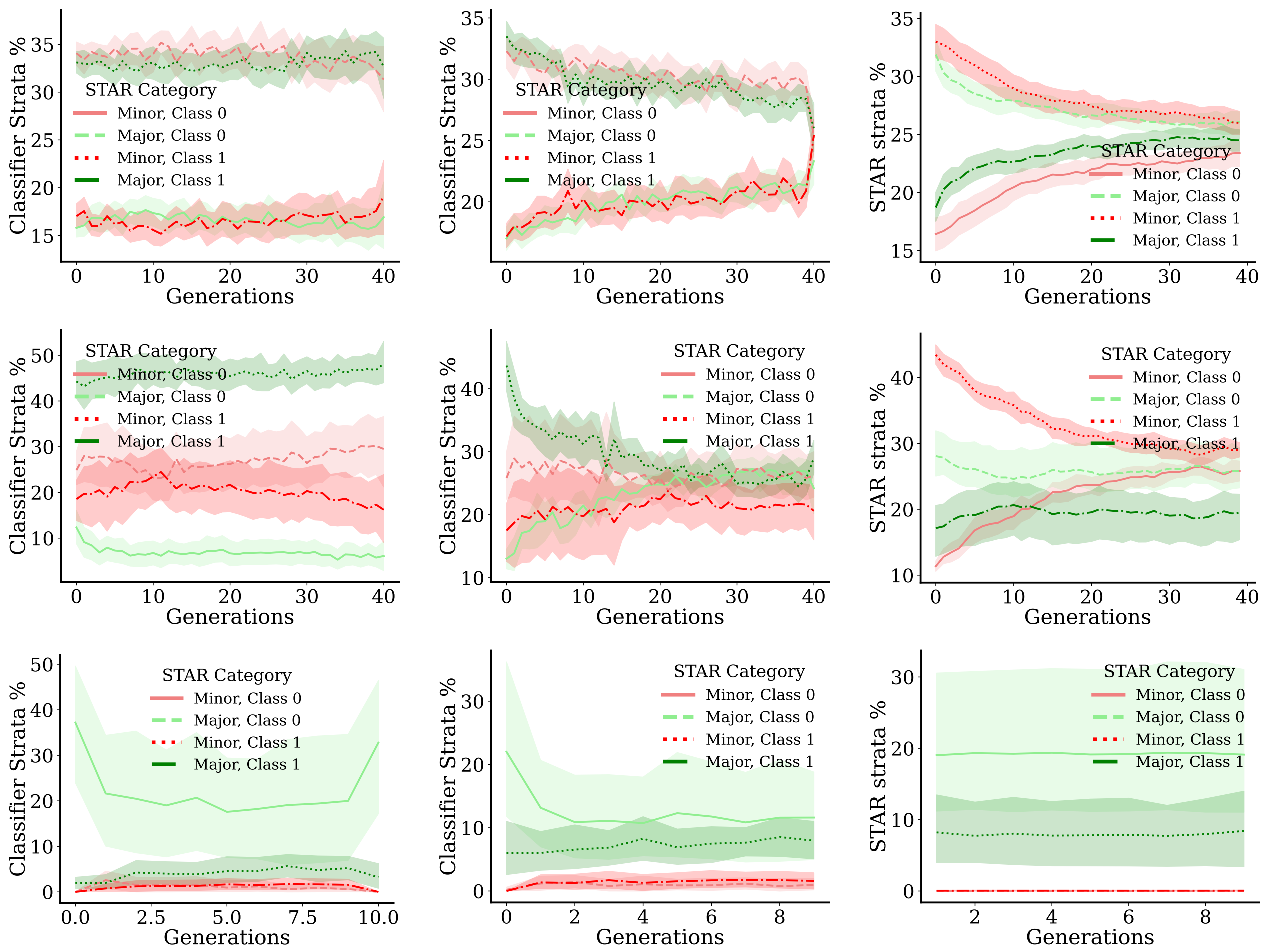


Figure 18: STRATA balances for datasets in SEQCLASS. *Left:* The STRATA of classifiers without reparation. *Center:* The STRATA resulting from classifiers with CLA-STAR. *Right:* The STRATA used to train classifiers with CLA-STAR. *Top:* `ColoredMNIST`. *Second row:* `ColoredSVHN`. *Bottom:* `FairFace`, instead of showing all 28 categories, we choose the two categories per label that are most frequently the largest and smallest portion of the batch across all generations.

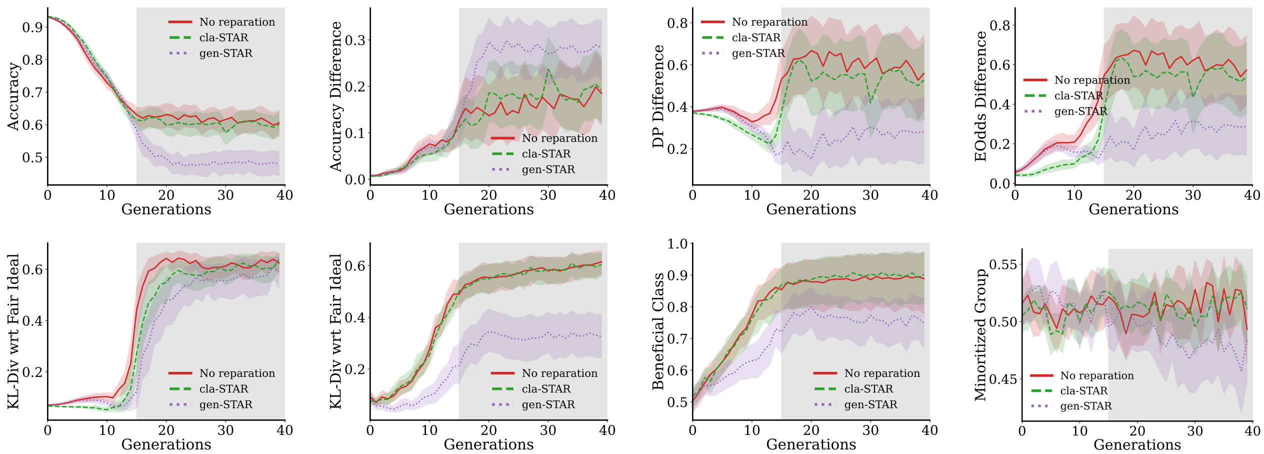


Figure 19: `ColoredMNIST` results for SeqGenSeqClass. *Top:* shows accuracy, accuracy difference, demographic parity difference, and equalized odds difference. For the latter three, lower values are better. *Bottom:* KL-Divergence between fairness ideal and classifiers, and between fairness ideal and generator strata, the class balance, and group balance. Shading shows collapsed generations. We observe that gen-STAR leads to better representation and fairness, though with a cost to the accuracy metrics. Recall that for `ColoredMNIST`, we find metric tension between EOdds, accuracy difference, and the STAR fairness ideal.

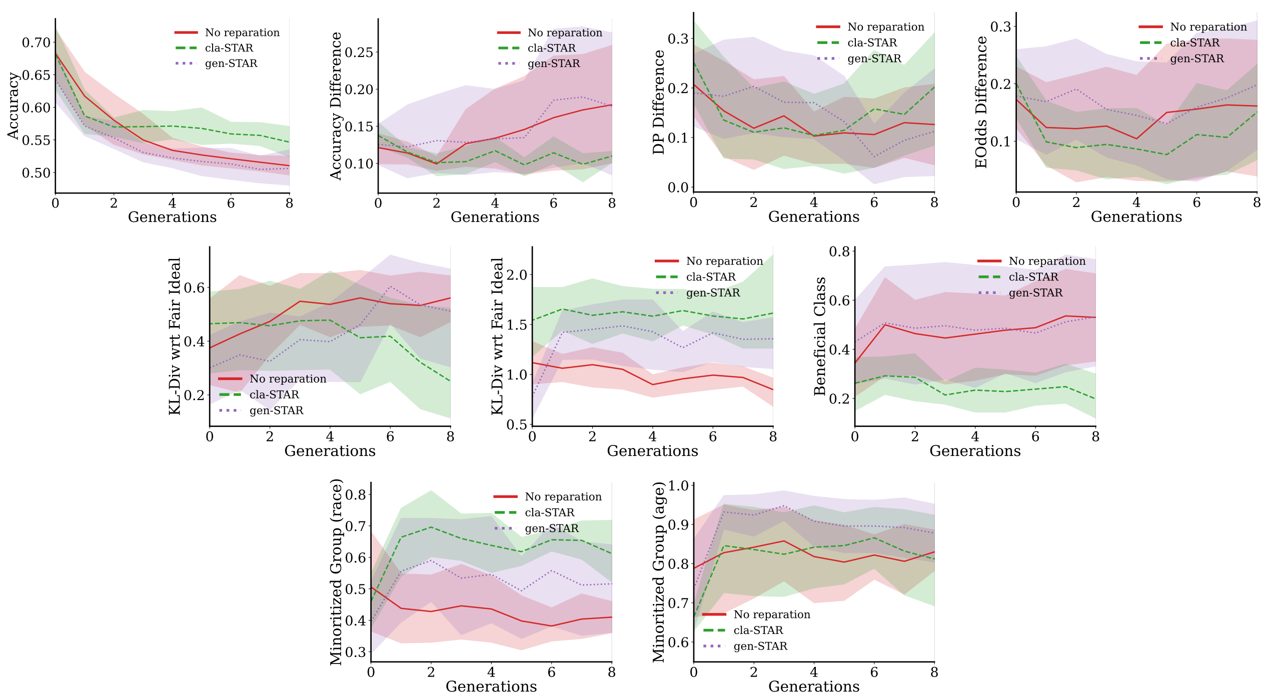


Figure 20: `FairFace` results for SeqGenSeqClass. *Top:* shows accuracy, accuracy difference, demographic parity difference, and equalized odds difference. For the latter three, lower values are better. *Center:* KL-Divergence between fairness ideal and classifiers, and between fairness ideal and generator strata, and the class balance. *Bottom:* shows the group balance for sensitive attributes race and age by reporting the plurality race and age at each generation. Shading shows collapsed generations. For this dataset, the annotator for race (A_{S_1}) as roughly 45% utility on all groups aside from 'white,' where it is 80% accurate. This is despite the near-perfect balance between racial groups. This effects our annotations for race, which, alongside with a similarly biased lineage of generators leads to a lack of samples from non-white races. This thwarts STAR, resulting in similar unfairnesses as results without reparation.

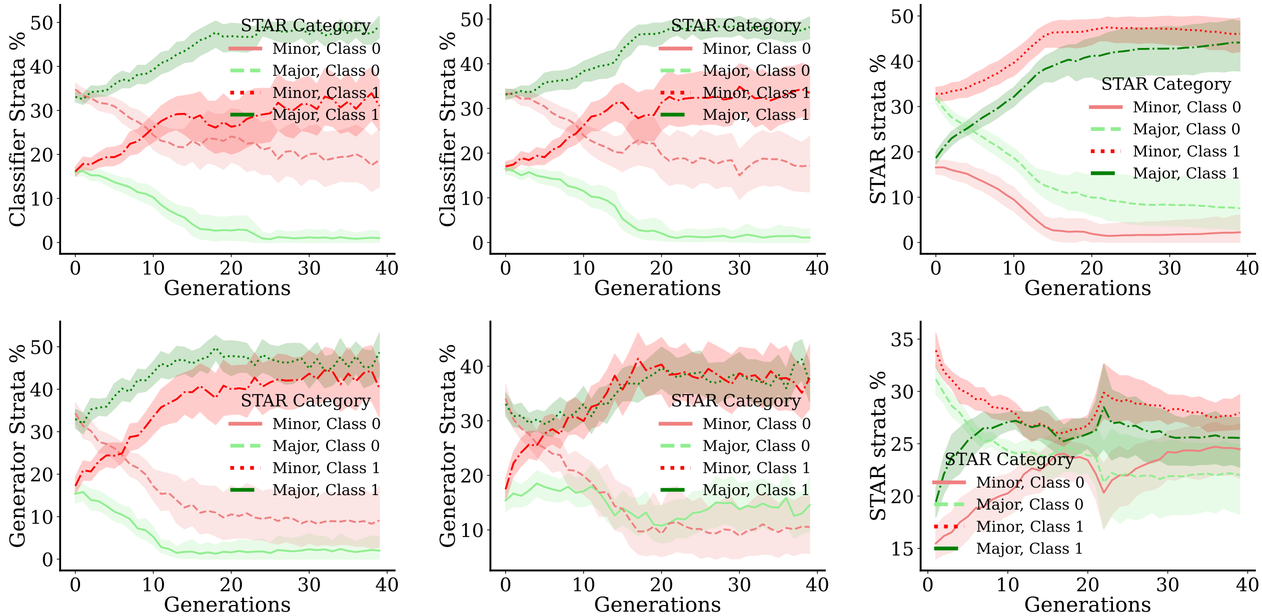


Figure 21: `ColoredMNIST` STRATA balances for datasets in the SeqGenSeqClass setting. *Top:* STRATA resulting from classifiers without reparation, STRATA resulting from classifiers with CLA-STAR, and the STRATA used to train classifiers with CLA-STAR. *Bottom:* STRATA resulting from generators without reparation, STRATA resulting from generators with GEN-STAR, and the STRATA used to train generators with GEN-STAR. We see that GEN-STAR accomplishes a more balanced STRATA than CLA-STAR, but ultimately both are unable to get perfect balance due to model collapse casuing class imbalance.

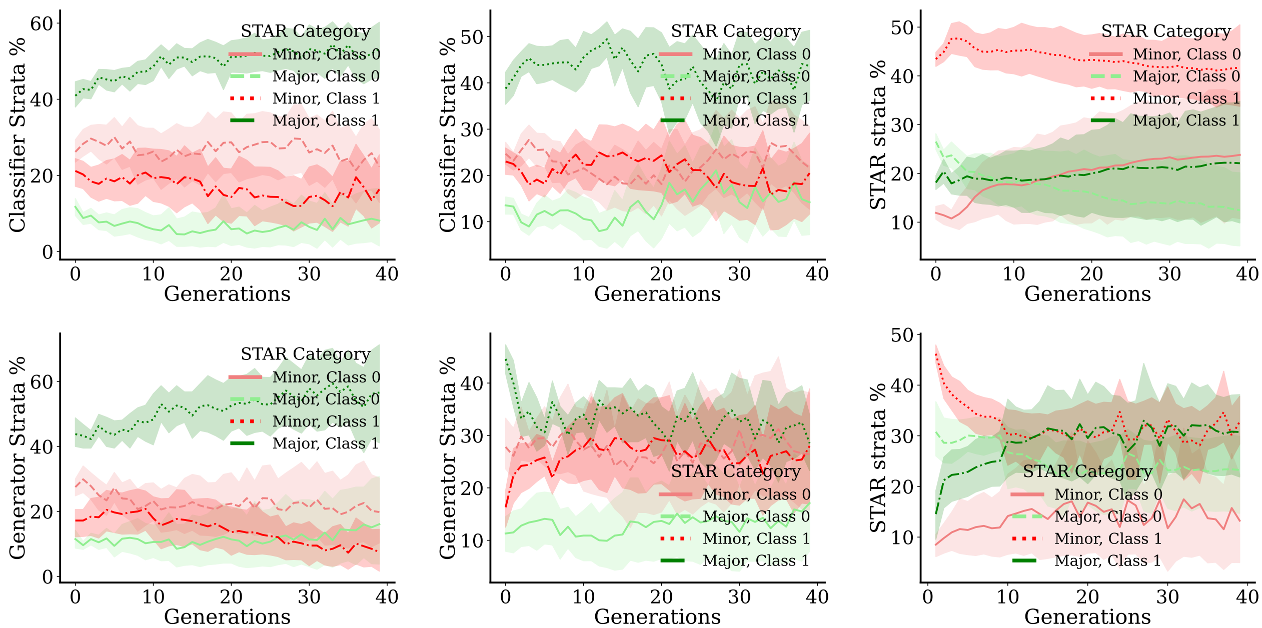


Figure 22: `ColoredSVHN` STRATA balances for datasets in the SeqGenSeqClass setting. *Top:* STRATA resulting from classifiers without reparation, STRATA resulting from classifiers with CLA-STAR, and the STRATA used to train classifiers with CLA-STAR. *Bottom:* STRATA resulting from generators without reparation, STRATA resulting from generators with GEN-STAR, and the STRATA used to train generators with GEN-STAR. We can see that GEN-STAR results in slightly more balanced representation than does CLA-STAR.

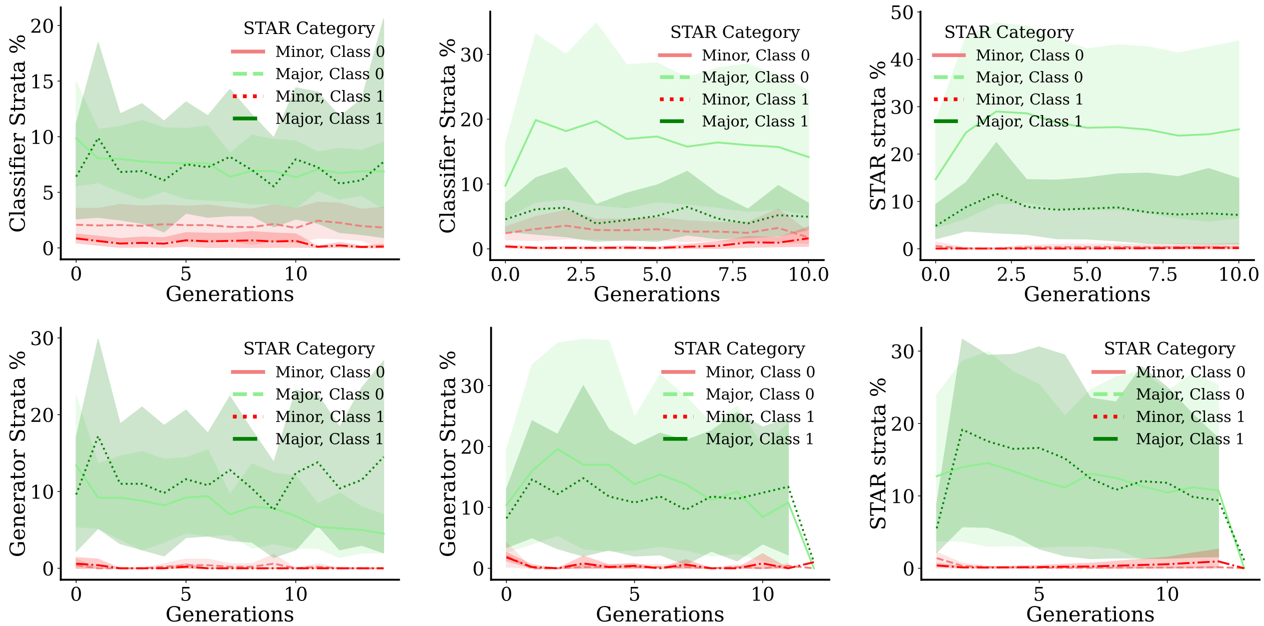


Figure 23: `FairFace` STRATA balances for datasets in the SEQGENSEQCLASS setting. *Top:* STRATA resulting from classifiers without reparation, STRATA resulting from classifiers with CLA-STAR, and the STRATA used to train classifiers with CLA-STAR. *Bottom:* STRATA resulting from generators without reparation, STRATA resulting from generators with GEN-STAR, and the STRATA used to train generators with GEN-STAR. Neither AR simulation is able to achieve balance due to large racial disparities.

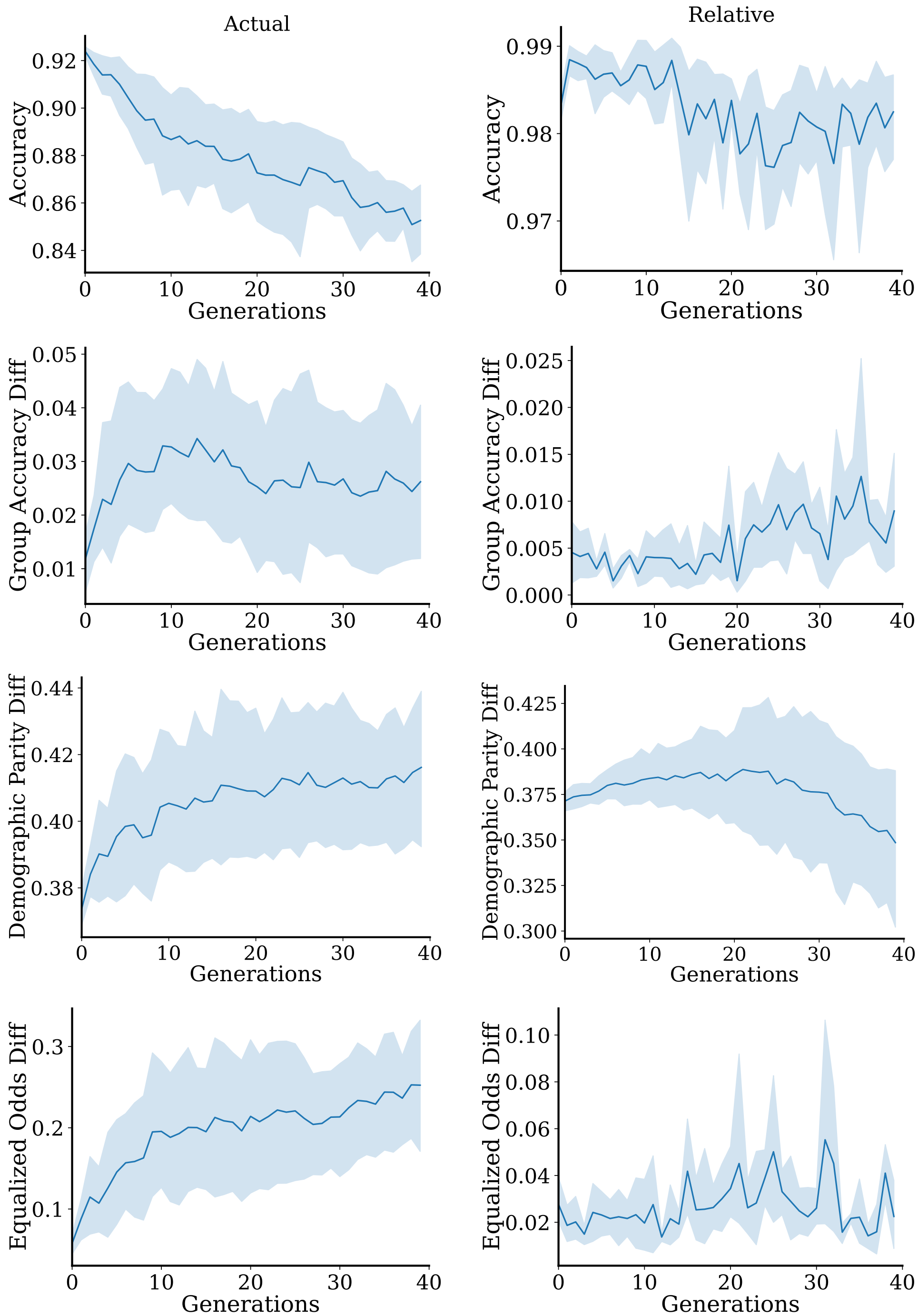


Figure 24: Accuracy, demographic parity difference, and equalized odds difference in SeqClass on `ColoredMNIST`. Higher accuracy is better, but for the FML metrics higher difference is worse. *Left:* Performances on the test set. *Right:* Relative performances between models. The model quality of accuracy and equalized odds in the relative performances is far higher than the actual results. In equalized odds, this shows that even if small unfairnesses were tolerated over while training each classifier, the result over time accrues high unfairness compared to the original testing set.

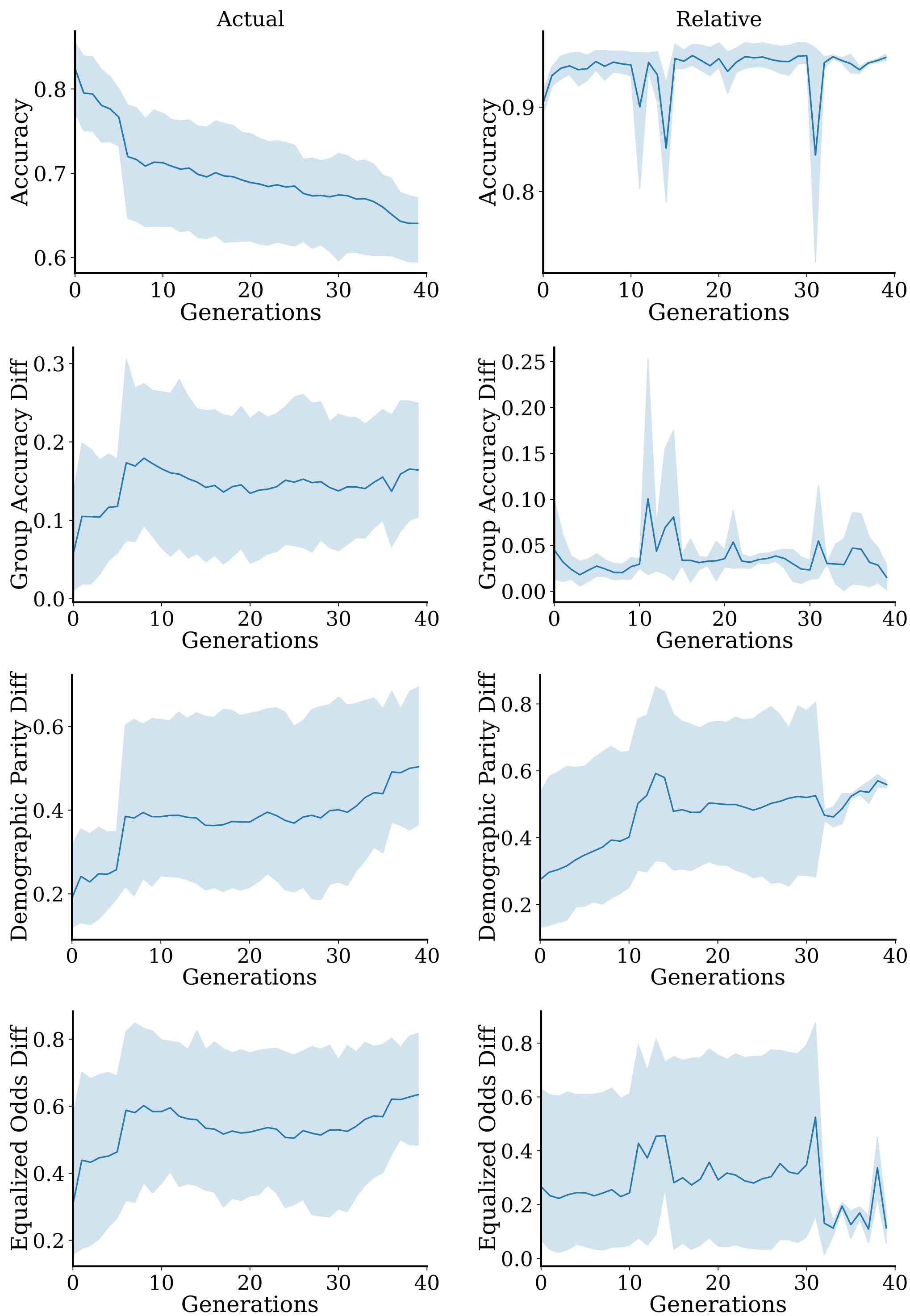


Figure 25: Accuracy, demographic parity difference, and equalized odds difference in SeqClass on `ColoredSVHN`. Higher accuracy is better, but for the FML metrics higher difference is worse. *Left:* Results on the test set. *Right:* Relative performances between models.

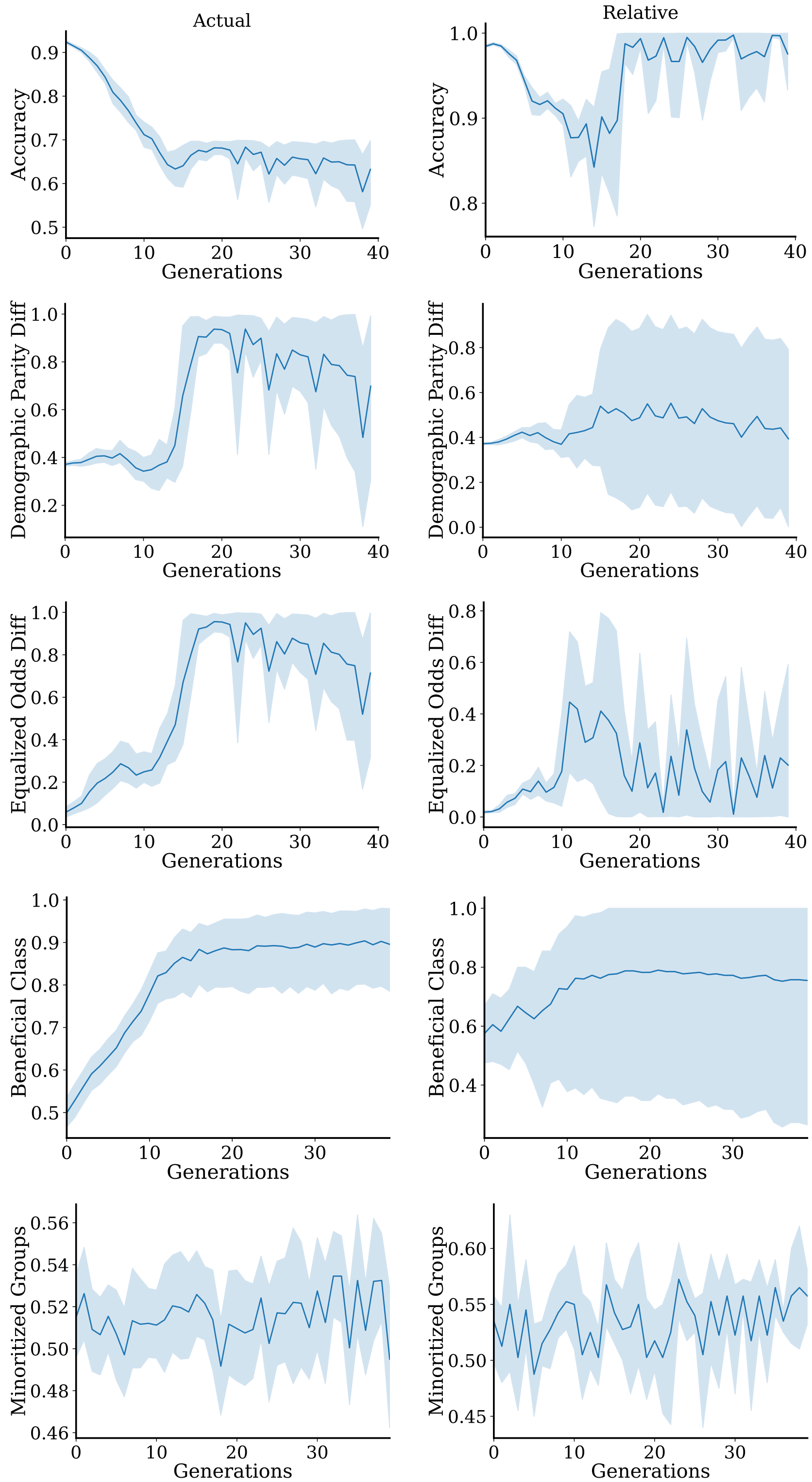


Figure 26: Accuracy, demographic parity difference, equalized odds difference, and rates of the beneficial class and minoritized group in SeqGenSeqClass on ColoredMNIST. Higher accuracy is better, but for the FML metrics higher difference is worse. *Left:* Results on the test set. *Right:* Relative performances between models.

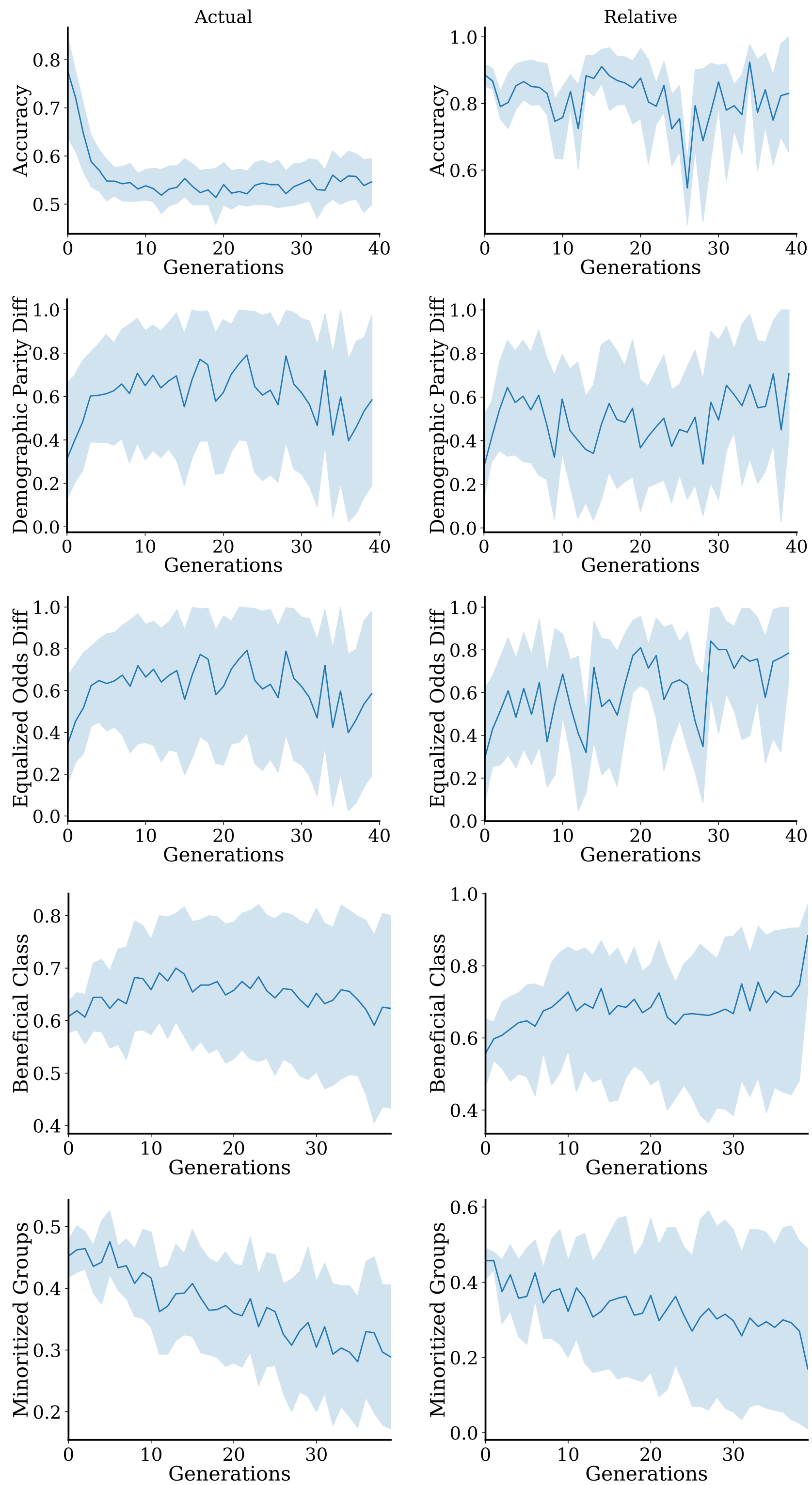


Figure 27: Accuracy, demographic parity difference, equalized odds difference, and rates of the beneficial class and minoritized group in SeqGenSeqClass on `ColoredSVHN`. Higher accuracy is better, but for the FML metrics higher difference is worse. *Left:* Results on the test set. *Right:* Relative performances between models.

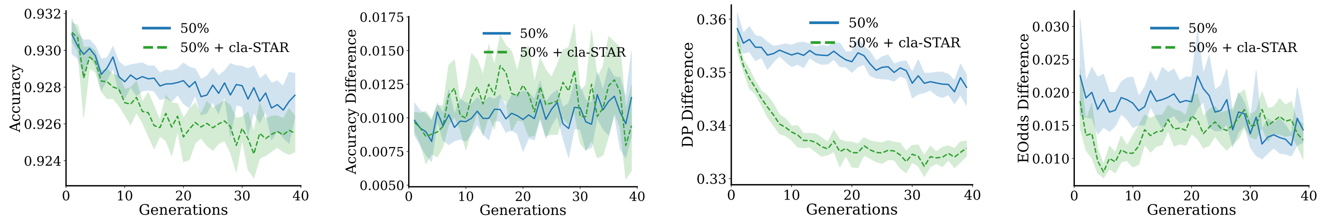


Figure 28: `ColoredMNIST` results for SeqClass when training with half synthetic and half non-synthetic data. Plots show accuracy, accuracy difference, demographic parity difference, and equalized odds difference. Non-synthetic data is sampled according to the strata of the prior classifier to model disparity amplification. cla-STAR leads to more DP and EOdds fairness even while disparity amplification and performative prediction occur.

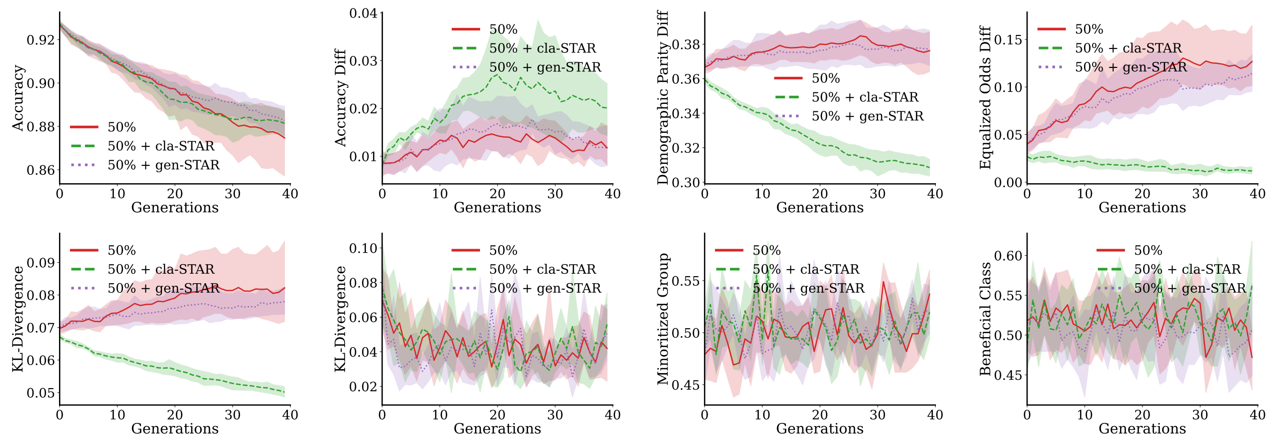


Figure 29: Training on a mixture of synthetic and non-synthetic data, where performative prediction, model collapse, and disparity amplification can co-occur, for `ColoredMNIST` on SeqGenSeqClass. *Top:* accuracy, accuracy difference, demographic parity difference, and equalized odds difference. *Bottom:* KL-Divergence between the STAR fairness ideal and the strata of classifiers and generators, the group balance, and the label balance.

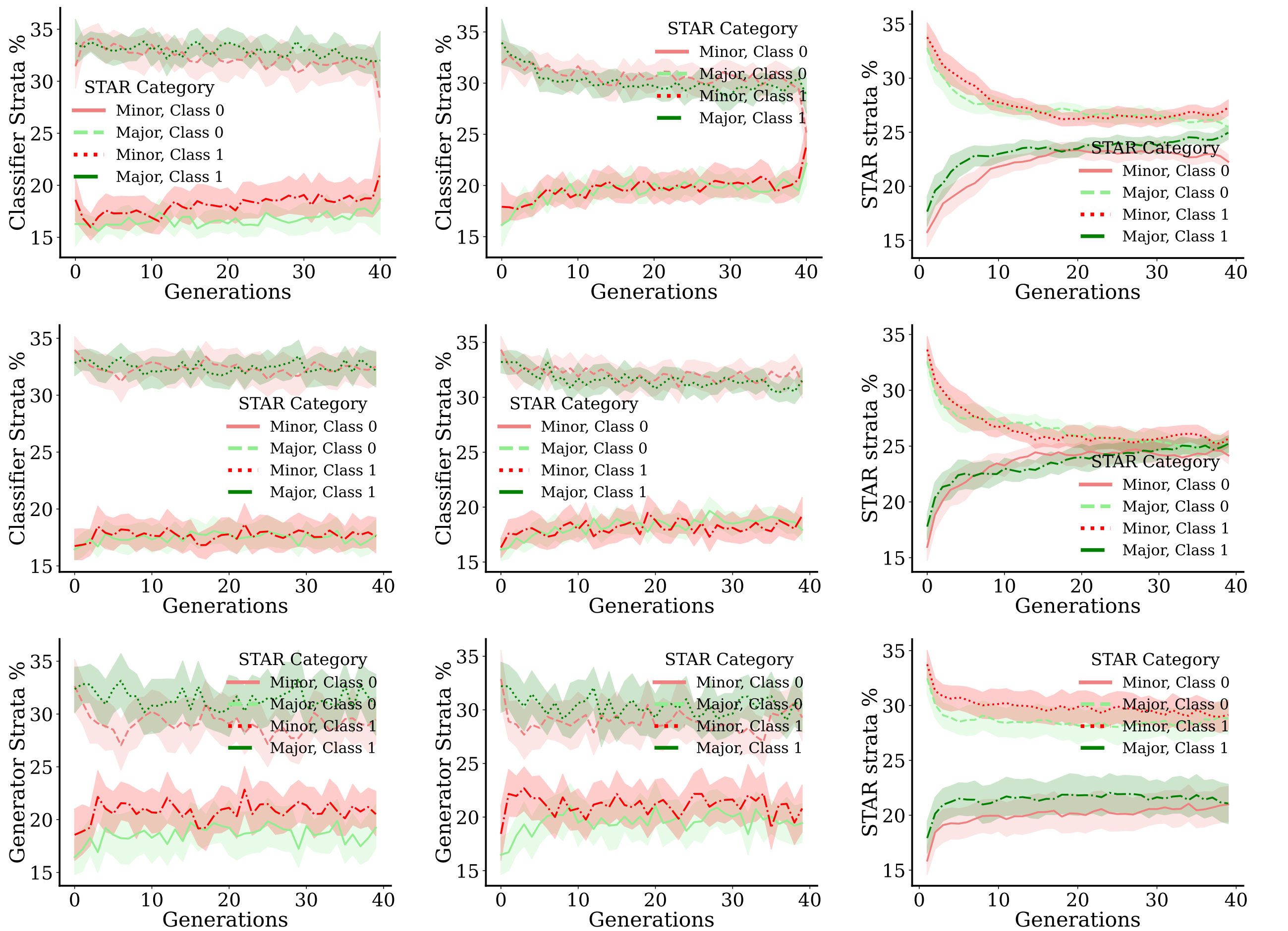


Figure 30: ***Left:*** **The STRATA of classifiers without reparation.** ***Center:*** **The STRATA resulting from models with STAR.** ***Right:*** **The STRATA used to train models with STAR.** ***Top:*** **`ColoredMNIST` in SEQCLASS with a mixture of synthetic and non-synthetic data.** ***Second row:*** **`ColoredMNIST` in SEQGENSEQCLASS with a mixture of synthetic and non-synthetic data, reporting STRATA of the classifiers and using CLA-STAR.** ***Bottom:*** **`ColoredMNIST` in SEQGENSEQCLASS with a mixture of synthetic and non-synthetic data, reporting STRATA of the generators and using GEN-STAR.**