A Homogenization Metrics

A.1 Group Homogenization Metrics

In §3.2, we introduced our group level homogenization metrics. Specifically, we note the design decision of how to weight groups and the three weightings we consider: **average**, **uniform**, and **worst**. Here, we provide the full mathematical definition of these metrics. For convenience, let the frequency of group g in a specific dataset D^i by denoted as $p^i(g)$, and the joint probability of the group across all datasets be denoted as $p(g) \triangleq \prod_{i=1}^{k} p^i(g)$.

$$H_{\textbf{avg}}(h^1,...,h^k) \triangleq \frac{\sum_g \left[\frac{p(g)}{\sum_{g'} p(g')} \prod_i \text{FAIL}_g(h^i) \right]}{\prod_i \text{FAIL}(h^i)} \tag{7}$$

$$H_{\textbf{unif}}(h^1,...,h^k) \triangleq \frac{\mathbb{E}_g \left[\prod_i \text{FAIL}_g(h^i) \right]}{\prod_i \text{FAIL}(h^i)} \tag{8}$$

$$H_{\textbf{worst}}(h^1,...,h^k) \triangleq \frac{\max_g \left[\prod_i \text{FAIL}_g(h^i) \right]}{\prod_i \text{FAIL}(h^i)} \tag{9}$$

A.2 Relating Individual and Group Homogenization

In §4, we demonstrate empirically that outcomes can be more homogeneous for individuals than for racial groups. Here we provide two scenarios that demonstrate circumstances where individual-level outcome homogenization is greater than, and is less than, group-level outcome homogenization. (Of course, they can also be equal.) In both settings, we will have two applications and two groups, where each group is comprised of two individuals. As a result, $H_{\textbf{avg}} = H_{\textbf{unif}}$, so we can compare either to $H^{\textbf{individual}}$.

In both settings, we will say Alice and Angelique are members of Group 1 and Bob and Bernardo are members of Group 2.

Scenario 1. Let Alice and Bob be misclassified in Application 1 but not 2, and Angelique and Bernardo be misclassified in Application 2 but not 1. No one is misclassified by both models, hence the number of observed systemic failures is 0 at the individual level, hence $H^{\textbf{individual}} = 0$. However, since there is a failure within Group 1 for both applications (and for Group 2 as well), the number of observed systemic failures is nonzero at the group level, hence $H_{\textbf{avg}} = H_{\textbf{unif}} > 0$. Thus, in this scenario, we have seen that individual-level outcome homogenization can be less than group-level outcome homogenization.

Scenario 2. Let Alice and Bob be misclassified in both applications, and Angelique and Bernardo be misclassified in neither application. At the individual level, there are 2 systemic failures, so the observed rate of systemic failure is $\frac{2}{4} = 0.5$. The overall error rate for each application is 0.5, so the expected rate of systemic failure is $0.5 \times 0.5 = 0.25$. Therefore, $H^{\textbf{individual}} = \frac{0.5}{0.25} = 2$. At the group level, the observed rate of systemic failures is 0.25 for both groups. The overall error rate for each application is still 0.5, so the expected rate of systemic failure is still $0.5 \times 0.5 = 0.25$. Therefore, $H_{\textbf{avg}} = H_{\textbf{worst}} = \frac{\frac{1}{2}(0.25+0.25)}{0.25} = 1$. Thus, in this scenario, we have seen that individual-level outcome homogenization can be greater than group-level outcome homogenization.

A.3 Generalizing Individual-Level Metric

In §3.3, we note that our individual-level framing assumes that every individual j produces inputs x_j^i for every company i. The formalism in the main paper already permits these inputs to be different across companies for the same individual (e.g. Bob may submit different resumes when applying to Microsoft and Google). However, the formalism does not support two further general concepts: (i) multiple inputs per company and (ii) no inputs for some company (e.g. Bob does not apply to Amazon).

To accommodate the former, we note that the notion of failure can be modified depending on how the outcomes for the multiple inputs should be aggregated (e.g. a failure of a search engine may be determined by some fraction of search queries producing poor results for the user).

To address the latter concern, we introduce notation c_j to indicate the subset of companies that individual j interacts with, i.e. $c_j \subseteq \{1,...,k\}$. That is, any companies $i \in \{1,...,k\}$ that are not in c_j are those that individual j does not interact with. Accordingly, we definite $H^{\textbf{individual}}$ as:

$$H^{\textbf{individual}}(h^1,...,h^k) \triangleq \frac{\mathop{\mathbb{E}}_{j}\left[\prod_{i \in c_j} I^i(x_j^i)\right]}{\mathop{\mathbb{E}}_{j}\left[\prod_{i \in c_j} \text{FAIL}(h^i)\right]} \tag{10}$$

Notably, when $\forall j,\, c_j = [k]$, the denominator simplifies to $\prod_{i \in [k]} \text{FAIL}(h^i)$, which matches Equation 3.

A.4 Alternative Metrics

In §3, we introduce the metrics we use to quantify outcome homogenization. Of course, much like the many mathematical expressions that have been used to measure bias and fairness, there are many ways to reasonably measure homogenization. Fundamentally, given the underlying construct of outcome homogenization is largely new, we begin by recognizing our understanding of the concept is incomplete and likely will require study in real systems to truly identify the precise desiderata for a measure.

In the interim, it is difficult to assess if the metric has *structural fidelity* [Loevinger, 1957], i.e. does the metric's structure faithfully captures outcome homogenization? Further, it is unclear if the metric has sufficient predictive validity to predict long-term outcomes (e.g. longitudinal harms arising from outcome homogenization) or how useful it is for testing specific scientific and social hypotheses [Jacobs and Wallach, 2021]. Ultimately, we believe the key test for the metric will be its *consequential validity* [Messick, 1987]: will the metric yield positive social impact as it "both reflects structure in the world and imposes structure upon the world" [Hand, 2016].

To facilitate understanding, we transparently discuss other metrics we considered and why they may be preferable in some circumstances. Ultimately, we worked with the metrics we describe in the paper as we found them to be the simplest and preferred their probabilistic interpretations, but we include reasons to prefer alternative as we describe them.

A.4.1 Alternative Metrics in the Binary Setting

Covariance and Pointwise Mutual Information. When $k = 2$, i.e. there are two companies, we note that our metric bears a very close resemblance to the *covariance* between the (indicator) random variables I^1 and I^2. In particular, the covariance is the difference of the quantities that define the ratio for our homogenization metric. Similarly, our metric is the *pointwise mutual information* (PMI) evaluated at $(1,1)$ up to the $\log$.

$$H^{\textbf{individual}}(h^1,h^2) = \frac{\mathop{\mathbb{E}}_{j}\left[I^1 I^2\right]}{\mathop{\mathbb{E}}_{j}[I^1]\mathop{\mathbb{E}}_{j}[I^2]} \tag{11}$$

$$\text{Cov}(I^1,I^2) = \mathop{\mathbb{E}}_{j}\left[I^1 I^2\right] - \mathop{\mathbb{E}}_{j}\left[I^1\right]\mathop{\mathbb{E}}_{j}\left[I^2\right] \tag{12}$$

$$\text{PMI}(I^1=1,I^2=1) = \log\left(\frac{\mathop{\mathbb{E}}_{j}\left[I^1 I^2\right]}{\mathop{\mathbb{E}}_{j}[I^1]\mathop{\mathbb{E}}_{j}[I^2]}\right) = \log\left(H^{\textbf{individual}}(h^1,h^2)\right) \tag{13}$$

With respect to the covariance, we prefer that our metric is more naturally comparable across settings where the failure rates of social systems vary, whereas the covariance is more directly tied to the absolute scale of the failure rates. With respect to the pointwise mutual information, we note that we are simply looking at the behavior of the social system in a special case where all models fail (which is one of the 2^k possible outcomes an individual could receive overall), whereas the overall

PMI considers all of them and is invariant to symmetries that are significant in our setting. Further, both are traditionally studied in the binary setting, whereas we study behavior in settings where $k > 2$.

Pearson Correlation. Building on the relationship with the covariance, we note that our metric therefore also resembles the *Pearson correlation*.

$$H^{\textbf{individual}}(h^1,h^2) = \frac{\mathbb{E}\left[I^1 I^2\right]}{\mathbb{E}[I^1]\mathbb{E}[I^2]} \tag{14}$$

$$\text{Corr}(I^1,I^2) = \frac{\mathbb{E}\left[I^1 I^2\right] - \mathbb{E}\left[I^1\right]\mathbb{E}\left[I^2\right]}{\sqrt{(\mathbb{E}[I^1](1-\mathbb{E}[I^1]))(\mathbb{E}[I^2](1-\mathbb{E}[I^2]))}} \tag{15}$$

In particular, when dealing with accurate models that are homogeneous (i.e. $\mathbb{E}\left[I^1 I^2\right] >> \mathbb{E}\left[I^1\right]\mathbb{E}\left[I^2\right]$), the Pearson correlation coefficient approximates our metric up to the square root in the denominator. Arguments can be made in favor and against this square root (and more generally a k-th root for $k > 2$); for simplicity we favor our metric that does not introduce such normalization but acknowledge this normalization may prove to be more favorable as the metric is further stress-tested.

A.4.2 Alternative Metrics beyond the Binary Setting

As we note in Footnote 2, our formalism and metrics are designed to be general, meaning that they can accommodate settings where the models h^i correspond to different tasks or scenarios. (We make use of this generality in our experiments (§5) for vision and, especially, language.) To permit this generality, we (reductively) binarize outcomes as either failures or not in our use of the indicator functions I^i. In particular, for arbitrarily different tasks, the outcome spaces and their consequences on individuals may not be (easily) related.

In some settings, specifically those where the tasks that constitute the social system are sufficiently similar, we may instead prefer a more graded *loss* in the place of the binary notion of failures. For each deployment by company i, denote the associated loss function as $\mathcal{L}^i$ such that $\mathcal{L}^i(h^i(x^i_j),y^i_j) = \ell^i_j$ is the loss experienced by individual j when interacting with model h^i. In these settings, where the loss achieved across different applications is comparable, we can consider additional measures for homogenization in terms of this loss.

$$H^{\textbf{individual}}(h^1,...,h^k) = \frac{\mathop{\mathbb{E}}_j\left[\prod_i \ell^i_j\right]}{\prod_i\left[\mathop{\mathbb{E}}_j \ell^i_j\right]} \tag{16}$$

$$\text{MinExp}(h^1,...,h^k) = \frac{\mathop{\mathbb{E}}_j\left[\min_i \ell^i_j\right]}{\min_i\left[\mathop{\mathbb{E}}_j \ell^i_j\right]} \tag{17}$$

$$\text{ExpExp}(h^1,...,h^k) = \frac{\mathop{\mathbb{E}}_j\left[\min_i \ell^i_j\right]}{\mathop{\mathbb{E}}_i\left[\mathop{\mathbb{E}}_j \ell^i_j\right]} \tag{18}$$

$$\tag{19}$$

In words, the MinExp definition is the ratio of the average best-case loss for individuals with the loss of the best model h^{best} and the ExpExp definition is the same but the denominator is the average loss of the models rather than the best loss. These definitions bear close resemblance to the MaxMin and lexicographic (leximax and leximin) fairness definitions studied in the fairness literature [Dubins and Spanier, 1961, Henzinger et al., 2022]; the group-weighted analogue that use the **worst** group weighting recovers the MaxMin definition in the denominator. (Note that the naming conventions are reversed since we define metrics in terms of loss whereas work in fairness and equitable allocations generally defines metrics in terms of utility.)

When the loss is the 0-1 classification loss, the MinExp definition and $H^{\textbf{individual}}$ are very similar: the numerators are the same (as the product of indicator variables is the same as their minimum) and

the denominators are precisely $z = \frac{\prod_{i=1}^{k} \text{FAIL}(h^i)}{\text{FAIL}(h^{\text{best}})}$ factors of each other (i.e. the failure rate of all of the models except the best one). Consequently, the MinExp yields values in the range $[0,1]$ whereas $H^{\textbf{individual}}$ is in $[0,\infty)$. Independent behavior across models in $H^{\textbf{individual}}$ is guaranteed to be 1 and maximal systemic failure in MinExp is guaranteed to be 1; symmetrically, independent behavior in MinExp is z (which is non-constant) and maximal systemic failure in $H^{\textbf{individual}}$ is $\frac{1}{z}$.

More broadly, the indicator variables we use in the main paper can be seen as a special case when using the 0-1 classification loss whereas arbitrary loss functions can be mapped to indicators by thresholding the loss (which does not require the loss to comparable, as different thresholds can be applied for different deployments). As a result, both $H^{\textbf{individual}}$ and MinExp have clear merits; we believe MinExp may especially be a more natural definition where individuals have *choices* on which model h^i to interact with of the k possible models. In these settings (e.g. picking which voice assistant to use, or more generally consumer products), the losses will naturally be comparable and it may suffice for the individual to have one good option, which is more smoothly encoded by the minimum of the loss rather than the product of the indicators.

We encourage future work to explore whether the MinExp or $H^{\textbf{individual}}$ is preferable: in many settings we expect they will be strongly correlated given they are scalings of each other that depend not on the correlated nature of errors but the overall error rates, but they may be some settings where they diverge and one is clearly preferable to the other. Currently, we recommend work to also consider MinExp when the losses are comparable, but to default to $H^{\textbf{individual}}$ since this is not required and $H^{\textbf{individual}}$ has a simple probabilistic interpretation.

B Reproducibility

All of the code required to train the models, group inputs for group-based metrics, measure homogenization, and generate visualizations will be released upon acceptance. Additionally, to facilitate accessibility we will release a simple tool to compute our homogenization metrics. We provide further experimental details below.

B.1 Census Experiments

B.1.1 Data

We work with the **ACS PUMS** data introduced by Ding et al. [2021], which contains US Census survey data for 3.6 million individuals. Ding et al. [2021] introduce the dataset to facilitate research into algorithmic fairness and the measurement of harms associated with algorithmic systems. For each individual in the Census, 286 features are recorded (e.g. self-reported race and sex, occupation, average hours worked per week, marital status, healthcare status). Ding et al. [2021] construct several classification tasks using this data, where each task uses some of an individual's features as inputs and one of their features as the label for the prediction task (e.g. predicting if an individual's income exceeds $50000). Of these tasks, we work with three in particular: **ACSEmployment** (predict if an individual is employed), **ACSIncomePovertyRatio** (predict an individual's income normalized by the poverty threshold), and **ACSHealthInsurance** (predict if an individual has health insurance). Following Ding et al. [2021], we split the data 80/20 for train/test. We access this data through their `folktables` package.[10] We use the data for all of the US (they provided state-level data as well) from 2018, which is the primary data they analyze in their paper. We select these tasks because Ding et al. [2021] do not impose any filtering constraints in selecting data for these tasks, meaning all three tasks are posed for the same underlying individuals. See Ding et al. [2021] for more details. License information is provided at `https://github.com/zykls/folktables#license-and-terms-of-use` and our use of the dataset adheres to these terms of service. The data clearly can be personally identifying given it has census records for particular individuals, but there is no offensive content.

B.1.2 Models

Following Ding et al. [2021], we train logistic regression models using `sk-learn` [Pedregosa et al., 2011] with default hyperparameters. As a sanity check, we compared average model performance and saw it matched/exceeded what is reported in Ding et al. [2021]. For each setting (**fixed**, **disjoint**) and

[10] `https://github.com/zykls/folktables`

amount of training data, we trained 25 models across 5 random subsamples of training and 5 random seeds for model run per subsample, for each of the three tasks. In aggregate, all of the models we trained took approximately 10 hours across 5 NVIDIA Titan Xp GPUs (or 50 hours on 1 NVIDIA Titan Xp GPU), with additional experiments/debugging that is unreported in the paper taking approximately an additional 400 NVIDIA Titan Xp GPU hours.

B.1.3 Groupings

We consider racial groups, which are already provided in the dataset based on the self-identified category individuals chose in providing their information to the US Census. The specific racial categories used are: White alone, Black/African American alone, American Indian alone, Alaska Native alone, American Indian and Alaska Native, Asian alone, Native Hawaiian and Other Pacific Islander alone, other unspecified race, two or more races.

B.2 Vision Experiments

B.2.1 Data

We work with the **CelebA** dataset [Liu et al., 2015], which is a widely used dataset of celebrity faces paired with annotations for facial attributes. For each face image, given the associated attributes, we define two tasks (**Earrings**, **Necklace**) that involve predicting whether the individual is wearing the specific apparel item. Attribute prediction in CelebA has been studied previously in work on fairness and robustness [Sagawa* et al., 2020, Khani and Liang, 2021, Wang et al., 2021]. Given recent documentation of significant issues with computer vision datasets [e.g. Birhane and Prabhu, 2021, Birhane et al., 2021], we emphasize that we use the dataset solely for analytic reasons to study homogenization. Further, given works like GenderShades [Buolamwini and Gebru, 2018] that highlight the harms of face recognition, we emphasize that we do not use the dataset for face recognition, but instead consider apparel prediction tasks where each apparel item/accessory is clearly observable in the face image. In addition to **Earrings** and **Necklace**, the dataset also contains attributes for **Eyeglasses** and **Neckties**. We initially included these tasks, but observed no individual was misclassified for all four tasks since these two tasks were very easy and models rarely produced any errors (e.g. the error rate for **Eyeglasses** was generally less than 1%). To be able to present non-trivial results for individual-level outcome homogenization, we therefore removed these tasks from consideration so that a nonzero number of systemic failures could be observed. We downloaded the CelebA data from `https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`. We resized images to 224-by-224, and then apply the same augmentations as in CLIP [Radford et al., 2021], before feeding the image into the ViT-B/16 (which is what Radford et al. [2021] do as well). License information is provided here: `https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`. Our use of the dataset is consistent with the requirements for non-commercial research use. The data clearly can be personally identifying given it has face images for particular individuals, but there is no offensive content.

B.2.2 Models

Following Radford et al. [2021], we either use their released 150M parameter ViT-B/16 CLIP model (the largest publicly available CLIP model at the time of writing) or a randomly initialized model with the same architecture. We used code from the official CLIP repository at `https://github.com/openai/CLIP`. Since Radford et al. [2021] modify the standard Vision Transformer [Dosovitskiy et al., 2021], we use their modified version for our *scratch* models. As a sanity check of our implementation, we confirmed that our average finetuning accuracy were comparable to prior work (Sagawa* et al. [2020] paper considers the task of predicting Blonde hair, with an ImageNet pretrained ResNet-50, they get 94.8% and we get 95.9% in this particular task, i.e. when we also do the task of predicting Blonde hair with the same ImageNet pretrained ResNet-50). Further, on the **Wearing Earrings** task, we also confirmed that the CLIP pretrained ViT-B/16 did better than a ResNet-50 (both ImageNet pretrained and CLIP pretrained). For each setting (*scratch*, *probing*, *finetuning*), we trained 5 model runs with different seeds, for each of the tasks. The final learning rates we used were 0.003 (training from scratch), 0.01 (linear probing), 0.000003 (fine-tuning), and they were selected by grid searching the learning rates on the **Earrings** task and we sanity checked that choosing a higher or lower learning rate led to lower accuracy. We also train each approach for 10 epochs to ensure similar computational resources are provided to each approach. In aggregate, all of the models we trained took approximately 1000 hours on one NVIDIA Titan Xp GPU.

B.2.3 Groupings

We consider groups based on hair color. The dataset provides five hair-related annotations of Black, Brown, Blonde, Grey, and Bald. Since the Bald category was quite small, we collapsed the category with all examples that lacked a hair annotation (e.g. the hair color is obscured due to a hat) into an "Other" category to yield five total categories. In addition, we measure outcome homogenization for individuals based on whether they have a *beard*. While the **CelebA** dataset also contains annotations for race and gender, we chose to not look at these groups given we were concerned that the gender/race was being inferred by an annotator (crowdworker) from the face rather than being self-identified [Liu et al., 2015].

B.3 Language Experiments

B.3.1 Data

We use the **IMDB**, **AGNews**, **Yahoo**, and **HateSpeech18** datasets. For **IMDB**, we were unable to find formal license information. The data may contain some PII, but it is unlikely there is significant offensive content. For **AGNews**, we were unable to find formal license information but found information indicating it should be used non-commercially, which we adhere to see.[11] The data may contain some PII, but it is unlikely there is significant offensive content. For **Yahoo**, we were unable to find formal license information. The data may contain some PII, especially given its nature, but it is unlikely there is significant offensive content. For **HateSpeech18**, we adhere to the license provided here: `https://github.com/Vicomtech/hate-speech-dataset#license`. The data likely contains some PII given it is from forums, and certainly contains offensive content. We access this data through Hugging Face Datasets [Lhoest et al., 2021].[12] The associated papers describe how the data was collected or scraped. We tokenize the data using the RoBERTa [Liu et al., 2019] tokenizer provided in Hugging Face Transformers [Wolf et al., 2020].

B.3.2 Models

For all models we produced, we adapt RoBERTa-base [Liu et al., 2019] using the weights provided through Hugging Face Transformers [Wolf et al., 2020]. For all models, we use the default hyperparameters in the Trainer provide in the Transformer library, with the only change being a fairly standard setting of the learning rate to 2e-5. As a sanity check of our implementation, we confirmed that our accuracy matches those provided in standard scripts/tutorials provided in Transformers and are quite similar to other works that work with these standard datasets [e.g. Gururangan et al., 2019]. For each setting (*probing*, *finetuning*, *BitFit*), we trained 5 model runs with different seeds, for each of the four tasks. In aggregate, all of the models we discuss in the paper took approximately 36 hours across 5 NVIDIA Titan Xp GPUs (or 180 hours on 1 NVIDIA Titan Xp GPU), with additional experiments/debugging that is unreported in the paper taking approximately an additional 2000 NVIDIA Titan Xp GPU hours.

B.3.3 Groupings

Since we consider four deployments that are largely unrelated to each other, there are no annotations of individuals or groups available that apply across all four datasets. Consequently, we group inputs by (binary) gender, as this grouping applies across the four datasets.[13] Specifically, for each input we identify whether the input contains more references to the female gender (e.g. uses of words like "she"), the male gender, or no reference to an explicitly gendered term is made. We acknowledge that this treats gender as a binary as part of an unfortunate trend in NLP of works involving gender using binaries [Cao and

[11] See `http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html`.

[12] `https://huggingface.co/datasets`

[13] We also considered grouping by *names*, given recent works showing systemic behavior in NLP models for names [Shwartz et al., 2020, Romanov et al., 2019], and by *race*, using names that are strongly statistically associated [Tzioumis, 2018, Garg et al., 2018]. However, we found few systemic failures that we traced to the underlying groups: very few names appear in every dataset (often because the fictional movie characters and actors in **IDMB** are not discussed in the rest).

Daumé III, 2020]. We use the peer-reviewed list of gender terms from Garg et al. [2018] and in accordance with the recommendations of Antoniak and Mimno [2021]. In the event that the same number of male and female gender terms are mentioned (possibly zero for both) in an input, we grouped the input in a third "Other" category. While we did not extensively test, we did observe that the findings were not sensitive to small perturbations (i.e. random deletions of words from each list) of the lists we used.
Following Antoniak and Mimno [2021], we provide the exact lists below.

Male words = {"he", "son", "his", "him", "father", "man", "boy", "himself", "male", "brother", "sons", "fathers", "men", "boys", "males", "brothers", "uncle", "uncles", "nephew", "nephews"}
Female words = {"she", "daughter", "hers", "her", "mother", "woman", "girl", "herself", "female", "sister", "daughters", "mothers", "women", "girls", "femen", "sisters", "aunt", "aunts", "niece", "nieces"}

C Additional Experiments

Summary. In §4, we report results on the **ACS PUMS** dataset. Here, we supplement those findings to clarify whether the findings generalize across model families and to other datasets. Qualitatively, across these additional evaluations, we do find the findings transfer: (i) the **fixed** partition of the data, where there is strictly greater data-sharing, reliably yields greater homogenization and (ii) when group-level data is available, individual-level homogenization exceeds group-level homogenization in magnitude.

Datasets. In §4, we report results on the **ACS PUMS** dataset. Here, we replicate the experiments performed in that section, but vary the dataset to clarify if the qualitative trends generalize to other datasets. Recall that the structure of the data we deal with is fairly unusual: we are interested in datasets where each input is associated with multiple outcomes (i.e. the traditional multi-task learning setting) as we will share the training data across the models for each task. However, because of our interests in social outcomes, we would further like each input to be meaningfully associated with a person (e.g. arbitrary multi-task learning datasets could be used but are unideal if they don't additionally have this human-centric structure).

To identify additional relevant datasets, we survey datasets for multi-task learning, datasets for fairness in ML (which are generally human-centric as desired), and work at the intersection of fairness and multi-task ML. Of these, we looked at Zhang and Yang [2017] for multi-task learning as well as `https://paperswithcode.com/task/multi-task-learning`, which provided a list of 51 datasets. We also looked at the 15 fairness datasets surveyed by Le Quy et al. [2022]. Finally, Wang et al. [2021] initiated the study of multi-task fairness, considering four datasets.[14]

From all of these, we arrived at four datasets with the structure we wanted: **CelebA**, **UCI Adult**, **LSAC** [Wightman et al., 1998], and **GC** [German Contracts; Dua and Graff, 2017]. All of these datasets have the desired structure: each input x_j^i is directly associated with an individual j and corresponds to multiple outcomes y_j^1,...,y_j^k. Of them, we report results for **CelebA** in §5.1, and we use the **ACS PUMS** dataset of Ding et al. [2021] that was explicitly designed to supersede the similar US Census-based **UCI Adult**. Hence, we turn our attention to **GC** and **LSAC**.

The **GC** dataset contains information on 1000 German contracts, including credit history, credit amount, and the corresponding credit risk for that individual [Dua and Graff, 2017]. Following Wang et al. [2021], the two prediction tasks we consider are (i) predicting if the individual receives a good or bad loan and (ii) predicting whether their credit amount exceeds 2000.[15] We additionally featurize in the same way as Wang et al. [2021], using 16 attributes as features. The data is accessed through the `UCI Machine Learning Repository`[16] and we use an 80/20 train-test split like Wang et al. [2021].

[14] Following our work, Fabris et al. [2022] introduced a search engine for fairness datasets that confirms our selection process for datasets was comprehensive.

[15] We filter any individuals where only one of the outcomes is reported.

[16] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

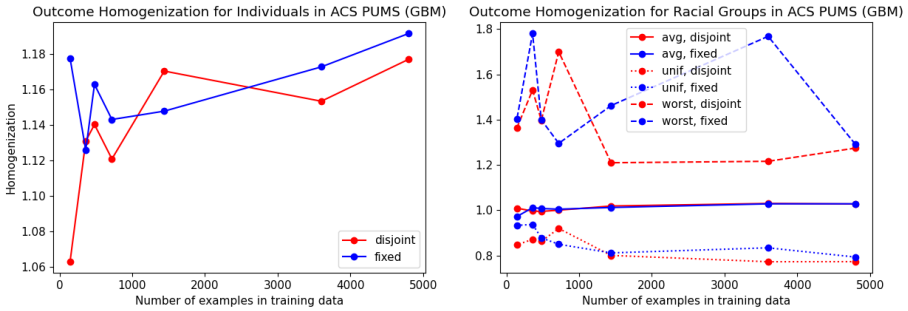

Figure 4: Results for data-sharing experiments on **ACS PUMS** with gradient boosted classifiers showing homogenization (y) as a function of training dataset size (x). Training across tasks on the same data (**fixed**) yields more homogeneous outcomes than on non-identical but identically distributed data (**disjoint**), especially for small datasets.

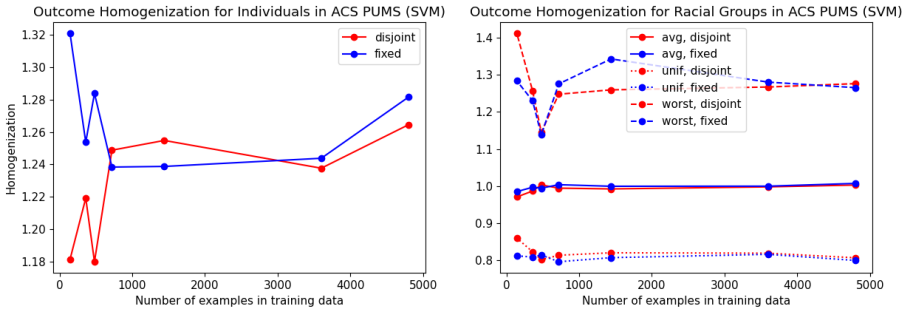

Figure 5: Results for data-sharing experiments on **ACS PUMS** with support vector machines showing homogenization (y) as a function of training dataset size (x). Training across tasks on the same data (**fixed**) yields more homogeneous outcomes than on non-identical but identically distributed data (**disjoint**), especially for small datasets.

All other experimental conditions match what we describe for the data-sharing experiments in §4 with further details given in §B.1.

The **LSAC** dataset was generated by the Law School Admission Council in the United States [Wightman et al., 1998]. This dataset contains information on 21,790 law students such as their entrance exam scores (LSAT) and their undergrad grade-point average (GPA) collected prior to law school. From this, the two prediction tasks we consider are predicting (i) whether they pass the bar exam and (ii) whether their law school GPA exceeds the mean, directly following Wang et al. [2021].[17] The data is accessed through the `tempeh`[18] package with the default train-test split. All other experimental conditions match what we describe for the data-sharing experiments in §4 with further details given in §B.1.

Model Families. In §4, we report results using logistic regression as the model family. Here, we replicate the experiments performed in that section, but vary the model family to clarify the influence of the model family in the homogenization of outcomes. Specifically, we consider three additional model families: gradient boosted decision tree classifiers (GBM),[19] support vector machines, and neural networks. All of these models are implemented using `sk-learn` [Pedregosa et al., 2011] with default parameters.

Results and Analysis. We report additional results for **ACS PUMS** in Figures 4–6, for **LSAC** in Figures 7–9, and for **GC** in Figures 10–1. Across these figures, we first establish that our core findings are upheld: (i) homogenization is reliably greater when there is more homogenized (**fixed**) than less (**disjoint**) across datasets and model families and (ii) homogenization is reliably greater when

[17] We filter any students where only one of the outcomes is reported.

[18] https://github.com/microsoft/tempeh

[19] Also considered by Ding et al. [2021] in their work with **ACS PUMS**.

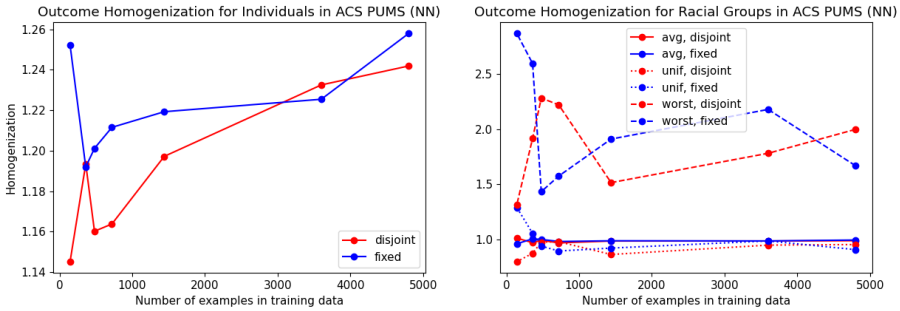

Figure 6: Results for data-sharing experiments on **ACS PUMS** with neural network classifiers showing homogenization (y) as a function of training dataset size (x). Training across tasks on the same data (**fixed**) yields more homogeneous outcomes than on non-identical but identically distributed data (**disjoint**), especially for small datasets.

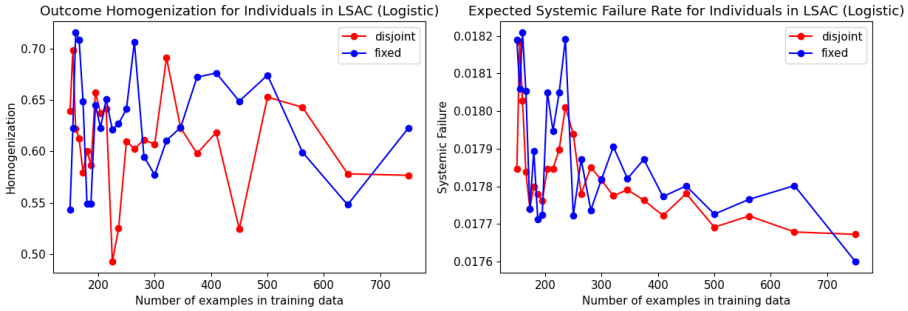

Figure 7: Results for data-sharing experiments on **LSAC** with logistic regression classifiers showing homogenization (**left**) and expected systemic failure rate ($\prod_{i\in[k]} \text{FAIL}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

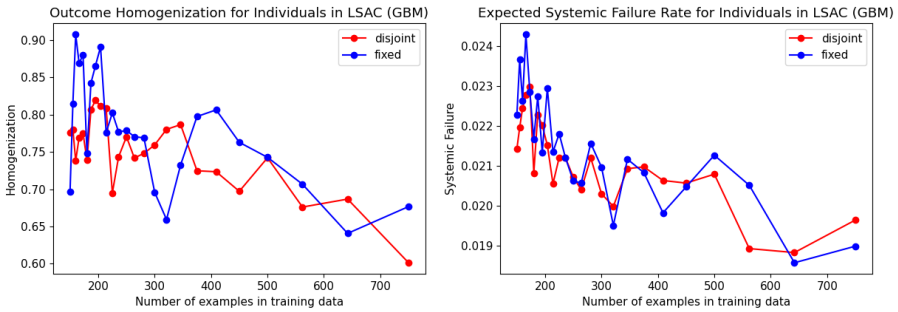

Figure 8: Results for data-sharing experiments on **LSAC** with gradient boosted classifiers showing homogenization (**left**) and expected systemic failure rate ($\prod_{i\in[k]} \text{FAIL}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

contrasting individual-level measures with the appropriate racial group-level measures.[20] However, we do note the absolute scale of homogenization is quite different across the datasets: we note that this is not surprising given the datasets are quite different, as are the relationship between the prediction tasks within a dataset. Therefore, we highlight that our hypotheses make predictions about relative change (i.e. sharing increases homogenization), but the underlying homogenization and absolute quantities will also depend significantly on the structure of the data and relationship between prediction tasks.

[20] **ACS PUMS** is the only dataset where we have race metadata.

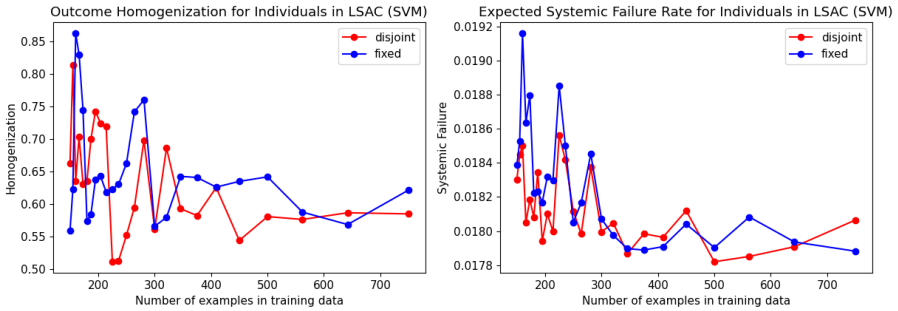

Figure 9: Results for data-sharing experiments on **LSAC** with support vector machines showing homogenization (**left**) and expected systemic failure rate ($\prod_{i \in [k]} \text{FAIL}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

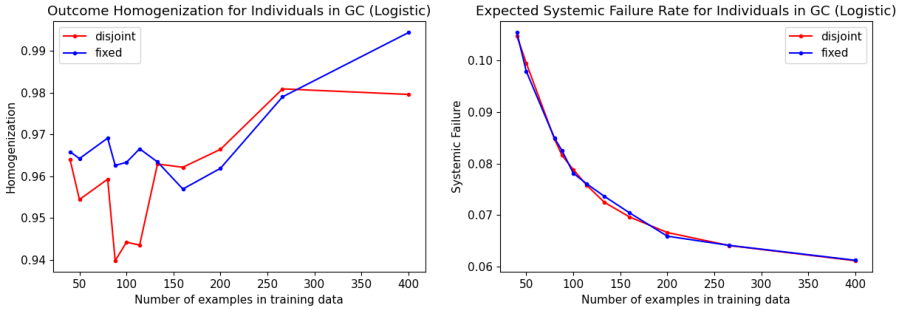

Figure 10: Results for data-sharing experiments on **GC** with logistic regression showing homogenization (**left**) and expected systemic failure rate ($\prod_{i \in [k]} \text{FAIL}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

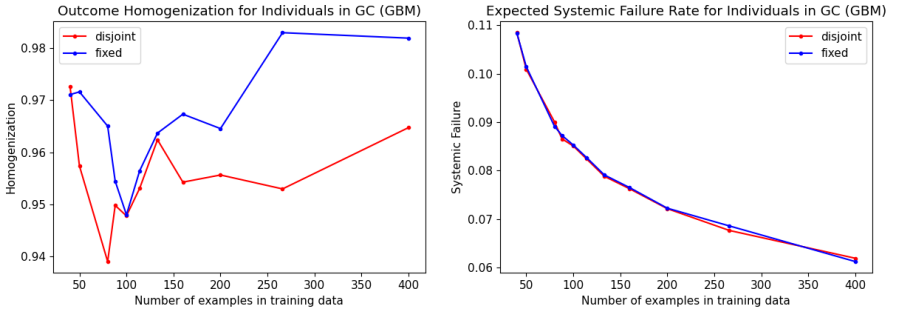

Figure 11: Results for data-sharing experiments on **GC** with gradient boosted classifiers showing homogenization (**left**) and expected systemic failure rate ($\prod_{i \in [k]} \text{FAIL}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

Further, for the **LSAC** and **GC** datasets, we visualize how the expected systemic failure rate (i.e. the product of the error rates of the models, which is the denominator in our homogenization metric) changes as a function of data scale. What we find is for both datasets, even one we have sufficient samples that the systemic failure rates in both the **disjoint** and **fixed** settings are identical (i.e. variance due to finite sample effects in error rates becomes minimal), we see higher homogenization. This demonstrates an important point: as the data grows, the expected systemic failure rate converges (as expected) to the same value for the **disjoint** and **fixed** partitions. That is, if one pays attention only to the accuracies for each task, these systems are the same. But even at this state, we see sizable discrepancies in outcome homogenization (i.e. the *observed* systemic failure rates remain

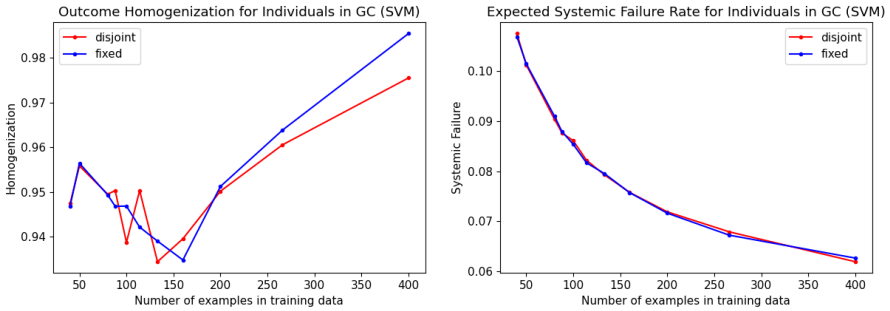

Figure 12: Results for data-sharing experiments on **GC** with support vector machines showing homogenization (**left**) and expected systemic failure rate ($\prod_{i \in [k]} \text{FAIL}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

different) with the **fixed** partition displaying greater homogenization. This drives home the point that data-sharing here has no effect on the accuracies of the resultant models, but that it does yield greater homogenization. This demonstration is reminscent of work that demonstrates[21] the existence of models of (near) equal accuracy but that are simpler [Semenova et al., 2022] or more fair [Marx et al., 2020]. Here we observe a Rashomon effect [Breiman, 2001] at the level of social systems: both systems achieve the same accuracies, but one (**disjoint**) is more homogeneous than the other (**fixed**).

[21] Generally these works show the existence of such models is theoretically guaranteed. We encourage future work to provide similar guarantees for the systems we describe, beyond our initial empirical demonstrations