Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?

Anonymous Author(s)
Affiliation
Address
email

Abstract

As the scope of machine learning broadens, we observe a recurring theme of *algorithmic monoculture*: the same systems, or systems that share components (e.g., datasets, models), are deployed by multiple decision-makers. While sharing offers advantages like amortizing effort, it also has risks. We introduce and formalize one such risk, *outcome homogenization*: the extent to which particular individuals or groups experience the same outcomes across different deployments. If the same individuals or groups exclusively experience undesirable outcomes, this may institutionalize systemic exclusion and reinscribe social hierarchy. We relate algorithmic monoculture and outcome homogenization by proposing the *component sharing hypothesis*: if algorithmic systems are increasingly built on the same data or models, then they will increasingly homogenize outcomes. We test this hypothesis on algorithmic fairness benchmarks based on the US Census, demonstrating that increased data-sharing exacerbates homogenization, especially for small datasets. Further, given the current regime in AI of foundation models, i.e., pretrained models that can be adapted to myriad downstream tasks, we test whether model-sharing homogenizes outcomes across tasks. We observe mixed results: we find that for both vision and language settings, the specific methods for adapting a foundation model significantly influence the degree of outcome homogenization. We also identify societal challenges that inhibit the measurement, diagnosis, and rectification of outcome homogenization in deployed machine learning systems.

1 Introduction

Machine learning is built on strong traditions of sharing: we share datasets (e.g., ImageNet), models (e.g., BERT), libraries (e.g., PyTorch), optimizers (e.g., Adam), evaluations (e.g., SuperGLUE) and much more. This ethos of sharing serves the field well: we are able to repeatedly capitalize on the effort required to build high-quality assets (e.g., ImageNet has supported thousands of researchers in computer vision), and improvements to these assets have sweeping benefits (e.g., BERT raised all boats in NLP). Yet does sharing also have risks? Could this central tenet of the field lead to undesirable outcomes?

We observe that certain forms of sharing can be reinterpreted as monoculture: Kleinberg and Raghavan [2021] define *algorithmic monoculture* as the state "in which many decision-makers all rely on the same algorithm." In parts of society where algorithmic systems are ubiquitous, we see trends towards such monoculture [Moore and Tambini, 2018, Engler, 2021]. Monocultures often pose serious risks: Kleinberg and Raghavan [2021] show monoculture is suboptimal for decision-makers when their decisions are interconnected, as when they compete to hire job candidates. Should we think of our practices of sharing assets in ML as monoculture and, if so, what harms should we worry about?

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

We investigate this question by proposing one potential risk we call *outcome homogenization*, i.e., the phenomenon of individuals (or groups) exclusively receiving negative outcomes from *all* ML models they interact with. We view outcome homogenization as an important class of *systemic* harms that arise when we study social *systems*, i.e., harms that require observing how individuals are treated by many decision-makers. In §2, we conceptually motivate outcome homogenization in the context of algorithmic hiring. In §3, we introduce the first mathematical formalism for outcome homogenization: we measure homogenization as the observed probability of systemic failure normalized by the base rate.

To link the practice of sharing in ML with the proposed harm of homogenization, we pose and test the *component sharing hypothesis*: algorithmic systems built using the same underlying components, such as training data and machine learning models, will systematically exclude the same individuals or groups. We see component sharing as a specific form of algorithmic monoculture, which generalizes the definition in Kleinberg and Raghavan [2021] from decision-makers deploying the *same* system to deploying *similar* systems in terms of how they are constructed. We investigate how two types of shared components — training data and foundation models — contribute to homogeneous outcomes.

In §4, we demonstrate that data-sharing reliably homogenizes outcomes for individuals and for racial groups, especially for small training datasets involving US census data. In §5, we discuss how the rise of foundation models [Bommasani et al., 2021], i.e., pretrained models that can be adapted to myriad downstream tasks, could yield unprecedented homogenization. Based on experiments with foundation models for vision (CLIP) and language (RoBERTa), to our surprise, we find the use of foundation models does not always exacerbate outcome homogenization. Instead, we find the specific mechanism for adapting the foundation model to the downstream task significantly influences homogenization: for example, linear probing consistently leads to more homogeneous outcomes than finetuning for both modalities. Through these experiments, it is clear that the relationship between sharing and homogenization is not fully explained by our hypothesis, but that there is some evidence that sharing homogenizes outcomes. To advance the study of homogenization in practice, where systemic harms are most consequential, we conclude by identifying key challenges for diagnosing, measuring, and rectifying homogenization in society (§6).

2 Outcome Homogenization in Resume Screening

To illustrate outcome homogenization and its potential causes, we will use the example of algorithmic resume screening. Companies use resumes to screen job applicants, choosing which candidates to interview and which to reject. Maximum homogenization occurs when every company makes the same decision about each candidate, such that each lucky candidate is interviewed by all companies and each unlucky candidate by no companies. We say that the unlucky candidates who receive no interviews experience a *systemic failure*.

What factors might homogenize outcomes? Historically, hiring managers at each company decided who to interview and often agreed in their decisions. This agreement can be attributed to multiple sources: first, if the needs of each company were identical, then managers at different companies may be incentivized to interview the same candidates, thereby homogenizing outcomes. Second, if hiring managers' choices are influenced by the same social biases, they will mistakenly reject the same people, thereby homogenizing their errors. Bias in resume screening is well-documented and remains significant [Jowell and Prescott-Clarke, 1970, Bertrand and Mullainathan, 2004, Kline et al., 2021, *inter alia*].

However, neither explanation implies that systemic failures are inevitable. Since companies have different needs and resumes are imperfect predictors of success in role, the "best" candidates will likely differ across companies. Further, bias is not uniform across companies: Kline et al. [2021] find that 21% of firms were responsible for 46% of the racial contact gap. Even if decisions are influenced by the same group-level biases, different companies may choose different individual members of the advantaged and disadvantaged groups. Variance in company needs, in prevalence of bias, and in individual hiring manager preferences all make it more likely that different resumes survive the screening stage at different companies, ensuring some diversity in resume screening outcomes.

How do these dynamics change with the introduction of automated decision-making? Most large companies now use automated resume screening software to parse resumes and decide which applicants advance. As a stylized example, if every company deploys the *same* deterministic system and has the same hiring criteria, then outcomes will be necessarily homogeneous: individuals will either receive interviews at every company or be rejected by all of them (i.e., systemic failure). This example is not far from reality: a few major vendors dominate the marketplace for algorithmic resume screening with 700 companies, including over 30% of Fortune 100 companies, relying on Hirevue [Hirevue, 2021]. This practice of different companies deploying the same system is defined as *algorithmic monoculture* by Kleinberg and Raghavan [2021].

More generally, different companies may instead deploy *similar*, but non-identical, systems. We expand the definition of algorithmic monoculture to encapsulate this broader setting, which is also alluded to in Kleinberg and Raghavan [2021]. Engler [2021] describe this as the reality for college enrollment management algorithms, writing "there are a relatively small number (between five and 10) of prominent vendors in the enrollment management algorithm market, . . . their process and analytics are markedly similar. Since their processes seem relatively consistent, the outcomes might be as well — potentially leading to consistently good results for students who match the historical expectations of colleges, and consistently poor results for students who don't".

Component Sharing Hypothesis. In this work, we study systems that are related in how they are constructed, akin to what is described by Engler [2021]. We pose the **component sharing hypothesis** that relates such algorithmic monoculture with outcome homogenization: *If deployed algorithmic systems share components, outcome homogenization will increase (i.e., there will be more systemic failures)*. In this work, we empirically test this hypothesis for two prominent forms of component sharing: (i) the sharing of training data in training all deployed systems (§4) and (ii) the sharing of the same foundation model for building all deployed systems (§5).

3 Formalizing Outcome Homogenization

While prior work [Kleinberg and Raghavan, 2021, Creel and Hellman, 2021, Bommasani et al., 2021] alludes to outcome homogenization, here we provide the first mathematical formalism of outcome homogenization.[1] In line with our running example of resume screening, we formalize outcome homogenization for individuals in terms of *systemic failures* (i.e., every algorithmic system fails for an individual). We then generalize to the group setting, where groups are systemically excluded rather than individuals, with a discussion of how these metrics relate to established fairness, robustness, and accuracy metrics (§3.4).

3.1 Formalizing Outcome Homogenization for Individuals

Notation. Since we define outcome homogenization as a systemic phenomenon, we consider a social *system* $\{h^i\}_{i=1}^k$ where every individual j interacts with the k deployed models $h^1,\dots,h^k$. Specifically, an individual $j \in [N]$ will submit features x_j^i (e.g., their resume) as input to company $i \in [k]$ and receive an outcome $h^i(x_j^i) = \hat{y}_j^i$ (e.g., an interview). Let D^i be the empirical distribution of inputs x^i for company i.[2]

To define a notion of failure, let $I^i(x_j^i)$ indicate if $\hat{y}_j^i$ is a negative outcome, i.e., individual j experiences a negative outcome from model h^i. The failure rate for model h^i is

$$\texttt{fail}(h^i) \triangleq \mathop{\mathbb{E}}_{x^i \sim D^i} I^i(x^i) = \Pr_{x^i \sim D^i}\left[I^i(x^i) = 1\right]. \tag{1}$$

Experimentally, we consider classification errors as failures (i.e., $I^i(x^i) \triangleq \mathbb{I}\left[h^i(x^i) \neq y^i\right]$), but other negative outcomes (e.g., rejections from hiring or educational opportunities; $I^i(x^i) \triangleq \mathbb{I}\left[h^i(x^i) = -1\right]$) also can be studied under our framework.

Systemic failures for individuals. If an individual exclusively experiences failure, we say they experience *systemic failure*. The *observed rate of systemic failure* S is

$$S \triangleq \mathop{\mathbb{E}}_{j}\left[\prod_i I^i(x_j^i)\right] = \Pr_j\left[I^1(x_j^1) = 1 \wedge \cdots \wedge I^k(x_j^k) = 1\right]. \tag{2}$$

[1]The formal model of Kleinberg and Raghavan [2021] is related, but substantially distinct. Concretely, their formalism considers harms experienced by decision-makers, whereas we center decision-subjects.

[2]Note that our framework is general: we permit the deployed models to be for different tasks and for the individual's inputs to not be the same, though in our resume screening example all the models perform the same task and applicants often submit the same resume to different companies.

In other words, it is the fraction of individuals who are failed by every model.

Homogenization metric for individuals. S quantifies homogeneous outcomes but is difficult to compare across real-world systems with different underlying accuracies: S will in general be higher for less accurate systems independent of a *specific* tendency to pick on the same person. While we may sometimes want to combine accuracy and outcome homogenization into an overall measure of utility or social welfare, which S implicitly does, we focus on a *relative* measure of homogenization that disentangles accuracy from homogenization. In particular, we are interested in outcome homogenization even, and perhaps especially, in systems that are highly accurate.

As a result, we measure individual-level outcome homogenization for a social system $\left\{h^i\right\}_{i=1}^k$ by normalizing the *observed* rate of systemic failure by the *expected* rate of systemic failure.

$$H^{\textbf{individual}}(h^1,\ldots,h^k) \triangleq \frac{S}{\prod_{i=1}^{k} \texttt{fail}(h^i)} = \frac{\mathbb{E}_j\left[\prod_i I^i(x_j^i)\right]}{\prod_i \left[\mathbb{E}_j I^i(x_j^i)\right]} \tag{3}$$

This measure can be interpreted as the ratio between (i) the probability that an individual experiences systemic failure and (ii) the probability that randomly sampled outputs for each model are all misclassified. That is, the measure captures how the rate of systemic failure changes when we attend to the structure of individuals.

3.2 Formalizing Outcome Homogenization for Groups

In addition to our individual-level metric, we also define a group-level metric for outcome homogenization. While our individual-level metric individualizes harm, complementing work on group-level harms like bias and inequity, we may also want to identify when broader social groups (e.g., Black women) are systemically excluded. This is especially relevant when we do not have access to individual-level information (e.g., due to privacy concerns; see §6) or when individuals are not shared across data distributions (e.g., hiring in different states).

Notation. For each input x^i, we denote the associated group as $G(x^i) \in \mathcal{G}$. Group identity may correspond with the data producer (e.g., the age of a user querying a search engine) or the data subject (e.g., the race of an individual subject to face recognition). Let D_g^i indicate restricting the data distribution to inputs to group g, i.e., $\forall x^i \in D_g^i, G(x^i) = g$. The *group failure rate* $\texttt{fail}_g(h^i)$ is

$$\texttt{fail}_g(h^i) \triangleq \mathbb{E}_{x^i \sim D_g^i} I^i(x^i). \tag{4}$$

Homogenization metric for groups. To measure group-level outcome homogenization for a social system $\left\{h^i\right\}_{i=1}^k$, we modify our individual-level metric to compute a weighted average over groups in place of a simple average over individuals.

$$H_G^{\textbf{group}}(h^1,\ldots,h^k) = \frac{\frac{1}{\sum_g W(g)} \sum_g \left[W(g) \prod_i \texttt{fail}_g(h^i)\right]}{\prod_{i=1}^{k} \texttt{fail}(h^i)} \tag{5}$$

Weight functions. We consider three weight functions $W : \mathcal{G} \to \mathbb{R}$ (full definitions in §A.1):

Average ($H_{\textbf{avg}}$) Each group g receives a weight proportional to its *frequency* across all deployments.

Uniform ($H_{\textbf{unif}}$) Each group g receives equal weight.

Worst ($H_{\textbf{worst}}$) The group with the highest rate of systemic failure receives a weight of 1 and all other groups receive a weight of 0.

We introduce these weight functions to clarify that, much like having both individual-level metrics and group-level metrics, we may want to weight groups differently in different circumstances. For example, weighting by frequency may provide a useful overall measurement of homogenization but obscure systemic exclusion experienced by minority groups or specifically the worst-off group.

3.3 Understanding our metrics

As a ratio of probabilities, our metrics take values in $[0,\infty)$ where 0 indicates no systemic failures, 1 indicates the observed rate matches the expected rate, and values greater than 1 indicate some degree of outcome homogenization. In the individual setting, we assume each individual generates exactly one input per deployment, which may not hold in practice (e.g., people may submit multiple resumes or not apply to every company). We appropriately generalize our individual-level metric to address this in Appendix A. Further, in the group setting, we recover the individual-level metric using the **uniform** weighting (or the **average** weighting) if each individual's inputs are treated as belonging to their own group.

3.4 Relationship with related concepts

Given that we introduce (several) metrics, we want to be cognizant of how they relate to existing metrics for related concepts (e.g., accuracy, fairness, robustness). This speaks to the convergent and divergent validity of our metrics [Campbell and Fiske, 1959, Messick, 1987, Jacobs and Wallach, 2021], i.e., whether they are adequately correlated with metrics of similar constructs and adequately uncorrelated with metrics of dissimilar constructs. Here, we discuss theoretical relationships, whereas in §5.2 we look at the empirical correlations.

Accuracy. We design our metrics to minimize correlation with accuracy to separate homogenization effects explicable by the accuracy of the underlying models from those that go beyond what is predicted by accuracy. While not theoretically guaranteed, we demonstrate in Table 1 that our metrics are consistently uncorrelated, or very weakly correlated, with accuracy.

Fairness and Robustness. Beyond accuracy, outcome homogenization is closely related to fairness and robustness. However, we emphasize that outcome homogenization is fundamentally about correlated outcomes for social *systems*, whereas almost all robustness or fairness metrics are defined for a single model. Recent work [Zhao and Chen, 2019, D'Amour et al., 2020, Wang et al., 2021] has initiated the study of fairness in multi-task learning, however these works focus on favorable overall trade-offs across tasks as opposed to systemic modes of failure. Conversely, our metrics cease to be interesting (e.g., $H^{\textbf{individual}}$ is always 1) in the single-model setting as systemic failures are single-model failures.

At a more fine-grained level, algorithmic fairness metrics [e.g., Dwork et al., 2012, Hardt et al., 2016, Corbett-Davies and Goel, 2018], as well as fairness metrics from other fields like the Gini coefficient, emphasize *discrepancies between individuals/groups*. In contrast, our metrics do not (explicitly) center these differences: we are interested in the observed rate of systemic failures across all deployments (and whether this exceeds the expected rate). Performance differences across individuals/groups are not *sufficient* for outcome homogenization: if the performance disparities for each deployment do not align across deployments, then the observed rate of systemic failure may not exceed the expected rate. For robustness metrics, our metric $H_{\textbf{worst}}$ in the worst-case setting closely resembles the metrics studied in work on worst-group robustness [e.g., Sagawa* et al., 2020]. In particular, when there is only one deployment, our metric recovers the standard worst-group accuracy normalized by the overall accuracy.

3.5 Alternative metrics

In Appendix A, we more extensively discuss desiderata for our metric, alternatives we considered, and how we arrived at the metrics we present in the main paper. With that said, we also note conditions where we may instead favor alternatives, as well as connections to familiar quantities like the covariance, Pearson correlation, and (pointwise) mutual information in the binary setting ($k=2$).

4 Data-Sharing Experiments

Having stated our mathematical formalism and metrics for outcome homogenization, we test if sharing training data leads to outcome homogenization. We choose to look at the US Census, which has been well-studied in work on algorithmic fairness and ensures we have individual-level outcomes.

Data. We work with the **ACS PUMS** data[3] introduced by Ding et al. [2021], which contains US Census survey data recording 286 features (e.g., self-reported race and sex, occupation, average hours worked per week) for 3.6 million individuals. Ding et al. [2021] construct several classification tasks using

[3] Full reproducibility details are provided in §B.1.

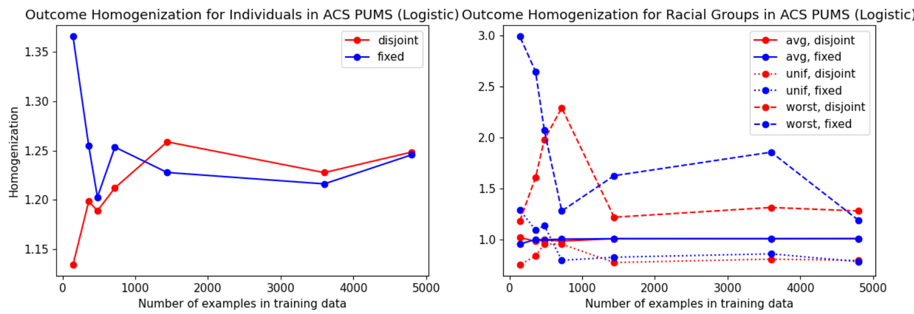

Figure 1: Results for data-sharing experiments showing homogenization (y) as a function of training dataset size (x). Training across tasks on the same data (**fixed**) yields more homogeneous outcomes than on non-identical but identically distributed data (**disjoint**), especially for small datasets.

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). We work with three such tasks: **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).

Individuals and Groups. For each individual, we have their true label for each of the three tasks, enabling us to measure individual-level homogenization. In the group setting, we measure how outcomes homogenize when grouping by the 9 self-identified racial categories (e.g., American Indian, Asian, Black/African American, White, two or more races).

Experimental Design. To test if data-sharing influences outcome homogenization, we formulate a controlled comparison by specifying two settings for the training data: **fixed** and **disjoint**. In the fixed setting, we sample n points without replacement from the entire **ACS PUMS** training dataset and set the training data for every task to this set. In the disjoint setting, we sample $3n$ points without replacement and randomly partition them across the three tasks. In other words, in the **fixed** setting, the three models are trained on data corresponding to exact same individuals, whereas in the **disjoint** setting, no pair of models is trained on data corresponding to the same individual. Having specified the training data, we train logistic regression models for each of three tasks following Ding et al. [2021]. To account for randomness, we report results averaged over 25 trials of the experiment (i.e., 5 samples of training data and 5 training runs per sample for every value of n we consider).

Results and Analysis. In Figure 1, we present our results for this controlled comparison.[4] Across all of our homogenization metrics, we see clear trends (for both individuals and racial groups) that the **fixed** setting yields more homogeneous outcomes than the **disjoint** setting (i.e., blue line is above the corresponding red line). This provides clear evidence for our hypothesis: the use of the same training data leads to greater outcome homogenization than the use of different (but identically distributed) training data. Interestingly, our results also suggest that homogenization can be improved by randomly subsampling a dataset. Given a fixed dataset D, Figure 1 (**left**) suggests that homogenization is better (lower) if each of the three applications randomly subsamples a third of D rather than using the entire dataset D. However, subsampling can reduce accuracy, which poses a tradeoff we discuss in §6. Further, we see that these effects wane as the amount of training data increases (i.e., gap between blue and red lines from left to right). This comes as no surprise: the two distributions converge due to concentration of measure, hence the distinction between the training data in the **fixed** and in the **disjoint** settings is increasingly small for larger amounts of data.

Picking on the same person. Further, we contrast the degree of homogenization in the individual and group settings. (Recall the **average** and **uniform** metrics are the group-level analogues of our individual-level metric.) Outcomes are consistently more homogeneous at the individual-group than for racial groups. This has significant ramifications for many works on algorithmic fairness, which only consider social groups (e.g., race): these works may miss systemic failures for particular individuals that are obscured at the group-level [cf. Kearns et al., 2018, Hashimoto et al., 2018]. Even

[4] We provide results on two additional datasets and for three additional model families (gradient boosted trees, SVMs, neural networks) in Appendix C, which reinforce the findings here.

intersectional approaches may not suffice to surface these systemic failures, unless each intersectional group comprises a single individual.

5 Model-Sharing Experiments

Having found that data-sharing appears to exacerbate outcome homogenization, especially for small datasets, we now turn to model-sharing. Specifically, we test how sharing pretrained models, or *foundation models* [Bommasani et al., 2021], affects outcome homogenization. Bommasani et al. [2021] define foundation models as "models trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks". These models have had a sweeping impact on the AI research community, most notably in NLP, and are increasingly central to deploying ML at both startups (e.g., Hugging Face, Anthropic, Inflection) and established technology companies (e.g., Google, Microsoft, Meta). Sharing is endemic to foundation models: to justify their immense resource requirements, the models must be used repeatedly for these costs to amortize favorably. In the extreme, if an entire domain like NLP comes to build almost all downstream systems on one or a few foundation models, then any biases or idiosyncrasies of these models that pervasively manifest downstream could potentially yield unprecedented systemic failures and outcome homogenization [Bommasani et al., 2021, Fishman and Hancox-Li, 2022]. We see initial evidence for such algorithmic monoculture: BERT was downloaded 10 million times in the past month alone and GPT-3 enables hundreds of deployed apps.[5] Consequently, we believe it is especially timely to understand if, and to what extent, outcomes get homogenized as these models become entrenched as infrastructure.

5.1 Experiments

Data. To test how foundation models influence homogenization, we run experiments for both vision and language data.[6] On the vision side, we work with the **CelebA** dataset [Liu et al., 2015] 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]. On the language side, we use four standard English text classification datasets following Gururangan et al. [2019]: **IMDB** [Maas et al., 2011], **AGNews** [Zhang et al., 2015], **Yahoo** [Chang et al., 2008], and **HateSpeech18** [de Gibert et al., 2018].

Individuals and Groups. Since the vision tasks are all based on CelebA, we have individual-level information. However, since the language tasks involve entirely different data (e.g., movie reviews vs. news articles), there is no (shared) individual-level information. At the group-level, for vision we use annotations for *hair color* and for whether the individual has a *beard*, whereas for language we automatically group inputs by *binary gender*.

Experimental Design. To test if model-sharing influences outcome homogenization, we contrast setting with differing degrees of model-sharing. In the vision experiments, we produce task-specific models for each task by either (i) training from **scratch** on CelebA data, (ii) linearly **probing** by fitting a linear classifier on frozen CLIP [Radford et al., 2021] features, or (iii) **finetuning** CLIP. To ensure meaningful comparisons, the models trained from scratch shared the same ViT architecture [Dosovitskiy et al., 2021] used in CLIP but with weights initialized randomly.

In the language experiments, we further hone in on the specific *adaptation method* used to adapt the foundation model (specifically RoBERTa-base [Liu et al., 2019]) to each task. We consider (i) linear **probing**, (ii) **finetuning**, and (iii) **BitFiT** [Ben Zaken et al., 2022], which is a recent *lightweight finetuning* method in NLP that involves freezing all the RoBERTa weights except the bias parameters which are updated as in finetuning. Consequently, BitFit is an intermediary between probing and finetuning, which has been shown to achieve similar accuracy as finetuning while updating very few of the pretrained parameters. For both vision and for language, all models are trained for the same number of epochs and we repeat each experiment for 5 random seeds per adaptation method.

Hypotheses. Much like data-sharing, model-sharing is graded and is not binary: different downstream systems can share varying degrees of underlying models. By design, our experimental design

[5] `https://huggingface.co/bert-base-uncased` as of May 2022.

[6] Full reproducibility details for vision are in §B.2; for language are in §B.3.

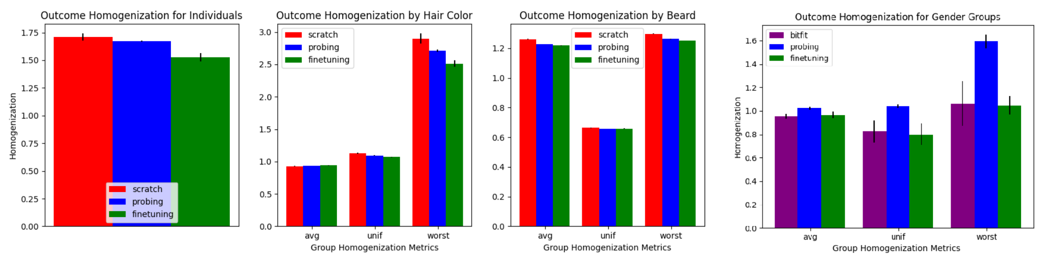

Figure 2: Results for model-sharing experiments showing homogenization as a function of training method for vision (**left three**) and language (**rightmost**).
Vision: **scratch** is the most homogeneous, then **probing**, then **finetuning**.
Language: **probing** is the most homogeneous; **finetuning** and **BitFit** are similarly homogeneous.

suggests a continuum in sharing: first, downstream system either can share a foundation model or not (**scratch**). Second, among methods that involve foundation models, all methods initialize the weights using the pretrained weights but differ in which parameters remain the same *after* adaptation is completed: **finetuning** changes all the parameters, **BitFit** only changes the bias parameters, and **probing** changes none of the parameters. As a result, overall, we can rank methods from most to least sharing as (i) **probing**, (ii) **BitFit**, (iii) **finetuning**, (iv) **scratch**, which leads us to predict the degree of homogenization will also follow this ranking under our component-sharing hypothesis.

Results. In Figure 2 (**left**), across all vision settings, we surprisingly find that **scratch** is the most homogeneous, i.e., more homogeneous than either approach involving shared foundation models. This is the opposite of what we hypothesized: we posit that this may indicate model sharing is not the key explanatory variable for outcome homogenization here, but instead it is a more complex form of data sharing. Specifically, we conjecture that since the **scratch** models are only trained on **CelebA** data, whereas the others also are trained on the much larger WebImageText via the CLIP foundation model, this may mean that the models based on CLIP are effectively regularized from learning idiosyncrasies of **CelebA** that the **scratch** models acquire. This may more generally suggest that a more correct hypothesis around data sharing should factor in the relationship (e.g., distribution shift) between the training data and the evaluation data for each model. Additionally, we find **probing** is consistently more homogeneous than **finetuning**, which aligns with our hypothesis. Finally, akin to the census results (§4), we once again find that outcome homogenization is significantly higher for individuals than for groups (comparing to $H_{\textbf{avg}}$ and $H_{\textbf{unif}}$).

In Figure 2 (**right**), across all language settings, we find the ordering of homogenization matches what our hypothesis predicts. Specifically, we find **BitFit** and **finetuning** achieve similar levels of homogeneity, even though **BitFit** updates 0.08% of the parameters full finetuning does (i.e., the number of shared parameters for BitFit is more like probing than finetuning), suggesting the number of shared parameters is not the right lens for understanding model sharing. More broadly, these results do suggest parameter-sharing effects may contribute to outcome homogenization within the foundation model regime, but comparisons between foundation models and no foundation models may be more complex to explain.

5.2 Correlations between Metrics

Since we introduce several metrics, we measure the correlations between our metrics. Further, we measure correlations with accuracy (specifically, the expected rate of systemic failure) to test if homogenization is disentangled from accuracy. Since outcome homogenization is related to fairness, we also measure the correlation between our metrics and a standard group fairness metric. Fairness metrics are generally defined for a single model h, whereas we study entire systems $\{h^i\}_{i=1}^k$. We extend the fairness definition used by Khani et al. [2019] as the variance in the systemic failure rates across groups.

$$\text{Fairness}_G(h^1,\ldots,h^k) \triangleq \operatorname*{Var}_g\left[\prod_i \texttt{fail}_g(h^i)\right] \tag{6}$$

Results. In Table 1, we report the pairwise correlation between metric pairs, based on the models we trained in §5.1. These correlations are for 45 systems (3 methods $\times$ 3 groupings $\times$ 5 random seeds) of 2 models for vision and 15 systems of 4 models for language. For vision, our metrics are highly correlated

Table 1: The correlation between our metrics and other metrics. Correlations are (Pearson R^2, Spearman ρ) with * indicating significance at $p = 0.05$ and *italics* indicating $p = 0.001$.

	Vision $H_{\mathbf{avg}}$	Vision $H_{\mathbf{unif}}$	Vision $H_{\mathbf{worst}}$	Vision Accuracy	Vision Fairness	Language $H_{\mathbf{avg}}$	Language $H_{\mathbf{unif}}$	Language $H_{\mathbf{worst}}$	Language Accuracy	Language Fairness
$H_{\mathbf{avg}}$	-	(*0.87*, *0.93*)	(0.0, *0.96*)	(0.0, 0.09*)	(0.0, *0.8*)	-	(*0.22*, *-0.47*)	(*0.11*, *0.56*)	(0.06*, -0.22*)	(0.02, 0.09)
$H_{\mathbf{unif}}$	(*0.87*, *0.93*)	-	(0.0, *0.96*)	(0.0, -0.02)	(0.0, *0.74*)	(*0.22*, *-0.47*)	-	(*0.63*, *-0.53*)	(0.0, 0.19*)	(0.0, -0.01)
$H_{\mathbf{worst}}$	(0.0, *0.96*)	(0.0, *0.96*)	-	(*0.05*, 0.1*)	(*1.0*, 0.82)	(*0.11*, *0.56*)	(*0.63*, *-0.53*)	-	(0.02, 0.13)	(*0.13*, *0.47*)

with each other, whereas for language, $H_{\mathbf{unif}}$ patterns quite differently (columns 1-3, 6-8). This is to be expected in that the vision groups (e.g., hair colors) all share similar frequencies, whereas the female group is significantly rarer in the language datasets. For both language and vision, we find that our metrics are generally not correlated, or perhaps weakly correlated, with accuracy as we intended (columns 4, 9). With respect to fairness, our worst-case metric $H_{\mathbf{worst}}$ is strongly correlated for both modalities, but for the other two metrics we see no linear correlations and only monotone correlations for the vision experiments (columns 5, 10). This is in line with our broader expectations that fairness and outcome homogenization are indeed related (especially for the worst-performing group), but that given they are distinct theoretical constructs, they should not always be correlated [Campbell and Fiske, 1959].

5.3 Discussion

Based on our experimental findings for both data-sharing and model-sharing, we provide considerable evidence that sharing in various forms does contribute to homogenization. However, we also provide nuance with our findings, indicating that the simple notion of sharing will need to be refined to properly capture the full effects observed in terms of homogenization. This is particularly relevant when the findings in §4 and §5 are contrasted with each other, given model-sharing in the foundation model regime also implies immense data-sharing in the form of the pretraining data (as mediated by the pretrained initialization). We emphasize that the regimes we study are considerably different: low-dimensional tabular data with simple model families in the data-sharing experiments vs. high-dimensional images/text with large and deep neural networks in the model-sharing, so it is not surprising the results do not fully coincide. Most broadly, we believe a more complete explanation will require factoring in the nature of the various data distributions and the associated distribution shifts (e.g., between pretraining and adaptation) at play. What we believe is clear, however, is that our findings provide an empirical basis to build on our conceptual arguments that sharing in ML can increase homogenization, and that this motivates exploration into whether these effects manifest in deployed high-stakes systems like those used for algorithmic hiring.

6 Societal Considerations and Challenges

To situate our work in a broader social context, we identify and discuss core challenges in **diagnosing**, **measuring**, and **rectifying** outcome homogenization.

Diagnosis. To diagnose homogenization as a byproduct of monoculture requires knowing which companies rely on the same vendor, dataset, or foundation model. How specific algorithmic systems are constructed is often so opaque that determining the responsibility of a shared component for outcome homogenization is nigh impossible. However, if a high outcome homogenization score was measured across deployments, the measurement itself could justify a request for increased transparency and access. We believe this could be a useful practical mechanism for balancing tensions in auditing to provide conditional access to auditors.

Measurement. Measuring homogenization only requires black box access to algorithmic systems, which is often achievable in practice [see Buolamwini and Gebru, 2018, Raji and Buolamwini, 2019, Metaxa et al., 2021]. However, identifying individual-level effects requires observing outcomes for the same individual across deployments. Due to privacy constraints, correlating individuals across datasets from different companies might be challenging or impossible. Such constraints incentivize the study of homogenization for groups as a more accessible concept, as in §3.2.

Rectification. Even once outcome homogenization is identified, organizations deploying systems responsible for homogeneous outcomes may not be incentivized to reduce it. When outcome homogenization does not align with each organization's interests (e.g., maximizing accuracy), it may be necessary to introduce regulation, policy, or other compliance mechanisms to encourage organizations to alter their outcomes. Navigating this trade-off is further complicated if the harms of homogeneous outcomes take time to observe or accrue. More optimistically, Kleinberg and Raghavan [2021] show that in some instances, there may not be a trade-off between policies optimal for each organization and those that diversify outcomes for societal benefit.

7 Limitations and Conclusion

We have introduced, formalized, and measured outcome homogenization as a systemic harm that may arise from practices of sharing in ML. Our work has an important limitation: our measure cannot accommodate every setting and may be brittle given that systemic failures are often rare. Furthermore, direct optimization of our measure may not lead to desirable outcomes, potentially contributing to ethics-washing. The interpretation of our measure must be **contextual**: the implications and severity of homogeneous outcomes must be situated in a broader societal context.

Outcome homogenization is essential to comprehensively characterizing algorithmic harm at scale, especially given the trend towards algorithmic monoculture. Without scrutiny, the harms of homogenization may become insidiously entrenched. Consequently, we believe further study and early intervention is necessary to mitigate and prevent such harms in society.

References

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22), 2021. ISSN 0027-8424. doi: 10.1073/pnas.2018340118. URL `https://www.pnas.org/content/118/22/e2018340118`. (Cited on 1, 2, 3, 10)

Martin Moore and Damian Tambini, editors. *Digital Dominance: The Power of Google, Facebook, Amazon and Apple*. Oxford University Press, New York, NY, 05 2018. ISBN 9780190845131. (Cited on 1)

Alex Engler. Enrollment algorithms are contributing to the crises of higher education. report, The Brookings Institution, September 2021. URL `https://www.brookings.edu/research/enrollment-algorithms-are-contributing-to-the-crises-of-higher-education/`. (Cited on 1, 3)

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, D. Card, Rodrigo Castellon, Niladri S. Chatterji, Annie Chen, Kathleen Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jackson K. Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021. URL `https://crfm.stanford.edu/assets/report.pdf`. (Cited on 2, 3, 7)

Roger Jowell and Patricia Prescott-Clarke. Racial discrimination and white-collar workers in britain. *Race*, 11(4):397–417, April 1970. doi: 10.1177/030639687001100401. URL `https://doi.org/10.1177/030639687001100401`. (Cited on 2)

Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, August 2004. doi: 10.1257/0002828042002561. URL https://doi.org/10.1257/0002828042002561. (Cited on 2)

Patrick Kline, Evan Rose, and Christopher Walters. Systemic discrimination among large u.s. employers. Technical report, National Bureau of Economic Research, July 2021. URL https://doi.org/10.3386/w29053. (Cited on 2)

Hirevue. Hirevue customers conduct over 1 million video interviews in just 30 days, Oct 2021. URL https://www.hirevue.com/press-release/hirevue-customers-conduct-over-1-million-video-interviews-in-just-30-days. (Cited on 3)

Kathleen Creel and Deborah Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision making systems. *Virginia Public Law and Legal Theory Research Paper No. 2021-13*, 2021. URL https://ssrn.com/abstract=3786377. (Cited on 3)

Donald T. Campbell and Donald W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81, 1959. URL https://pubmed.ncbi.nlm.nih.gov/13634291/. (Cited on 5, 9)

Samuel Messick. Validity. *ETS Research Report Series*, 1987(2):i–208, 1987. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2330-8516.1987.tb00244.x. (Cited on 5, 18)

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, FAccT '21, New York, NY, USA, 2021. Association for Computing Machinery. URL https://arxiv.org/abs/1912.05511. (Cited on 5, 18)

Chen Zhao and Feng Chen. Rank-based multi-task learning for fair regression. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 916–925, 2019. doi: 10.1109/ICDM.2019.00102. (Cited on 5)

Alexander D'Amour, Katherine A. Heller, Dan I. Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin G. Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *ArXiv*, abs/2011.03395, 2020. (Cited on 5)

Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H. Chi. *Understanding and Improving Fairness-Accuracy Trade-Offs in Multi-Task Learning*, page 1748–1757. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383325. URL https://doi.org/10.1145/3447548.3467326. (Cited on 5, 7, 21, 23, 24)

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255. (Cited on 5)

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016. (Cited on 5)

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *ArXiv*, abs/1808.00023, 2018. (Cited on 5)

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS. (Cited on 5, 7, 21)

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=bYi_2708mKK. (Cited on 5, 6, 20, 23, 24)

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kearns18a.html. (Cited on 6)

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *ICML*, 2018. (Cited on 6)

Nic Fishman and Leif Hancox-Li. Should attention be all we need? the epistemic and ethical implications of unification in machine learning. In *Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency*, FAccT '22, New York, NY, USA, 2022. Association for Computing Machinery. URL https://arxiv.org/abs/2205.08377. (Cited on 7)

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. (Cited on 7, 21, 22)

Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 196–205, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445883. URL https://doi.org/10.1145/3442188.3445883. (Cited on 7, 21)

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A. Smith. Variational pretraining for semi-supervised text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5880–5894, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1590. URL https://aclanthology.org/P19-1590. (Cited on 7, 22)

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015. (Cited on 7)

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf. (Cited on 7)

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 830–835. AAAI Press, 2008. ISBN 9781577353683. (Cited on 7)

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5102. URL https://www.aclweb.org/anthology/W18-5102. (Cited on 7)

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. (Cited on 7, 21)

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy. (Cited on 7, 21)

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. (Cited on 7, 22)

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.acl-short.1`. (Cited on 7)

Fereshte Khani, Aditi Raghunathan, and Percy Liang. Maximum weighted loss discrepancy. *ArXiv*, abs/1906.03518, 2019. (Cited on 8)

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. URL `https://proceedings.mlr.press/v81/buolamwini18a.html`. (Cited on 9, 21)

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 429–435, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314244. URL `https://doi.org/10.1145/3306618.3314244`. (Cited on 9)

Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4):272–344, 2021. ISSN 1551-3955. doi: 10.1561/1100000083. URL `http://dx.doi.org/10.1561/1100000083`. (Cited on 9)

Jane Loevinger. Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3):635–694, 1957. doi: 10.2466/pr0.1957.3.3.635. URL `https://doi.org/10.2466/pr0.1957.3.3.635`. (Cited on 18)

D.J. Hand. *Measurement: A Very Short Introduction*. Very short introductions. Oxford University Press, 2016. ISBN 9780198779568. URL `https://books.google.com/books?id=QBIBDQAAQBAJ`. (Cited on 18)

L. E. Dubins and E. H. Spanier. How to cut a cake fairly. *The American Mathematical Monthly*, 68 (1):1–17, 1961. ISSN 00029890, 19300972. URL `http://www.jstor.org/stable/2311357`. (Cited on 19)

Monika Henzinger, Charlotte Peale, Omer Reingold, and Judy Hanwen Shen. Leximax approximations and representative cohort selection. *ArXiv*, abs/2205.01157, 2022. (Cited on 19)

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL `http://jmlr.org/papers/v12/pedregosa11a.html`. (Cited on 20, 24)

Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546, 2021. doi: 10.1109/WACV48630.2021.00158. (Cited on 21)

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *ArXiv*, abs/2110.01963, 2021. (Cited on 21)

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas

Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-demo.21. URL https://aclanthology.org/2021.emnlp-demo.21. (Cited on 22)

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6. (Cited on 22)

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.556. URL https://aclanthology.org/2020.emnlp-main.556. (Cited on 22)

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. What's in a name? Reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4187–4195, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1424. URL https://aclanthology.org/N19-1424. (Cited on 22)

Konstantinos Tzioumis. Demographic aspects of first names. *Scientific Data*, 5, 2018. (Cited on 22)

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1720347115. URL https://www.pnas.org/content/115/16/E3635. (Cited on 22, 23)

Yang Trista Cao and Hal Daumé III. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.418. URL https://aclanthology.org/2020.acl-main.418. (Cited on 22)

Maria Antoniak and David Mimno. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.148. URL https://aclanthology.org/2021.acl-long.148. (Cited on 23)

Yu Zhang and Qiang Yang. A survey on multi-task learning. *ArXiv*, abs/1707.08114, 2017. (Cited on 23)

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12, 05 2022. doi: 10.1002/widm.1452. (Cited on 23)

L.F. Wightman, H. Ramsey, and Law School Admission Council. *LSAC National Longitudinal Bar Passage Study*. LSAC research report series. Law School Admission Council, 1998. URL https://books.google.com/books?id=dtA7AQAAIAAJ. (Cited on 23, 24)

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml. (Cited on 23)

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1827–1858, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533232. URL https://doi.org/10.1145/3531146.3533232. (Cited on 27)

Charles T. Marx, Flavio Du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. (Cited on 27)

L. Breiman. Statistical modeling: The two cultures. *Quality Engineering*, 48:81–82, 2001. (Cited on 27)

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

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 $p(g) \triangleq \prod_{i=1}^{k} p^i(g)$.

$$H_{\textbf{avg}}(h^1,\ldots,h^k) \triangleq \frac{\frac{1}{\sum_{g\in\mathcal{G}} p(g)} \sum_{g\in\mathcal{G}} p(g)\texttt{err}_g(h^1)\times\cdots\times\texttt{err}_g(h^k)}{\texttt{err}(h^1)\times\cdots\times\texttt{err}(h^k)} \tag{7}$$

$$H_{\textbf{unif}}(h^1,\ldots,h^k) \triangleq \frac{\frac{1}{|\mathcal{G}|} \sum_{g\in\mathcal{G}} \texttt{err}_g(h^1)\times\cdots\times\texttt{err}_g(h^k)}{\texttt{err}(h^1)\times\cdots\times\texttt{err}(h^k)} \tag{8}$$

$$H_{\textbf{worst}}(h^1,\ldots,h^k) \triangleq \frac{\max_{g\in\mathcal{G}} \texttt{err}_g(h^1)\times\cdots\times\texttt{err}_g(h^k)}{\texttt{err}(h^1)\times\cdots\times\texttt{err}(h^k)} \tag{9}$$

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:

$$H^{\textbf{individual}}(h^1,...,h^k) \triangleq \frac{\mathbb{E}_j\left[\prod_{i \in c_j} I^i(x^i_j)\right]}{\mathbb{E}_j\left[\prod_{i \in c_j} \texttt{fail}(h^i)\right]} \tag{10}$$

Notably, when $\forall j$, $c_j = [k]$, the denominator simplifies to $\prod_{i \in [k]} \texttt{fail}(h^i)$, 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 $\log$.

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

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

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

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*.

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

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

In particular, when dealing with accurate models that are homogeneous (i.e., $\mathbb{E}\left[I^1I^2\right]>>\mathbb{E}\left[I^1\right]\mathbb{E}\left[I^2\right]$), 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.

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

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

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

$$\tag{19}$$

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 $z=\frac{\prod_{i=1}^k \texttt{fail}(h^i)}{\texttt{fail}(h^{\text{best}})}$ 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.[7] 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 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

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

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.[8] 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].[9] 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.[10] 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

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

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

[10] 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.

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 $y_j^1,\ldots,y_j^k$. 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.[11] 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`[12] and we use an 80/20 train-test split like Wang et al. [2021]. All other experimental conditions match what we describe for the data-sharing experiments in §4 with further details given in §B.1.

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

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

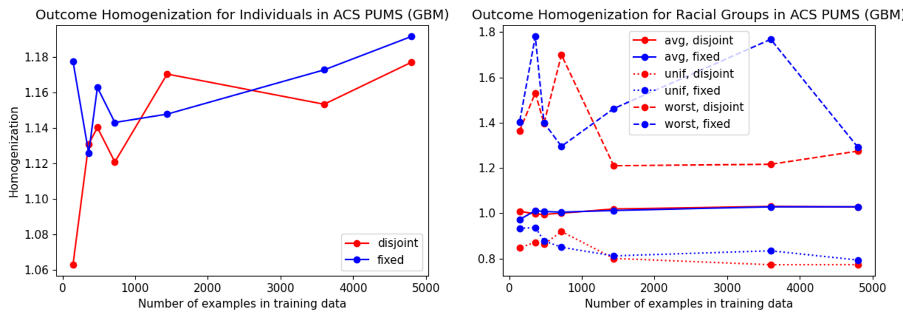

Figure 3: Results for data-sharing experiments on **ACS PUMS** with gradient boosted classifiers showing homogenization (y) as a function of training dataset size (x). Training across tasks on the same data (**fixed**) yields more homogeneous outcomes than on non-identical but identically distributed data (**disjoint**), especially for small datasets.

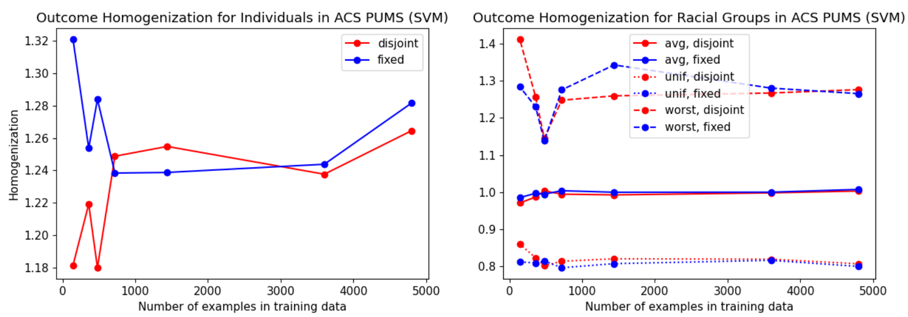

Figure 4: Results for data-sharing experiments on **ACS PUMS** with support vector machines showing homogenization (y) as a function of training dataset size (x). Training across tasks on the same data (**fixed**) yields more homogeneous outcomes than on non-identical but identically distributed data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

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].[13] The data is accessed through the `tempeh`[14] 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),[15] 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 3–5, for **LSAC** in Figures 6–8, and for **GC** in Figures 9–12. 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 contrasting individual-level measures with the appropriate racial group-level measures.[16] However, we do note the absolute scale of homogenization is quite different across the datasets: we note that this is

[13] We filter any students where only one of the outcomes is reported.
[14] https://github.com/microsoft/tempeh
[15] Also considered by Ding et al. [2021] in their work with **ACS PUMS**.
[16] **ACS PUMS** is the only dataset where we have race metadata.

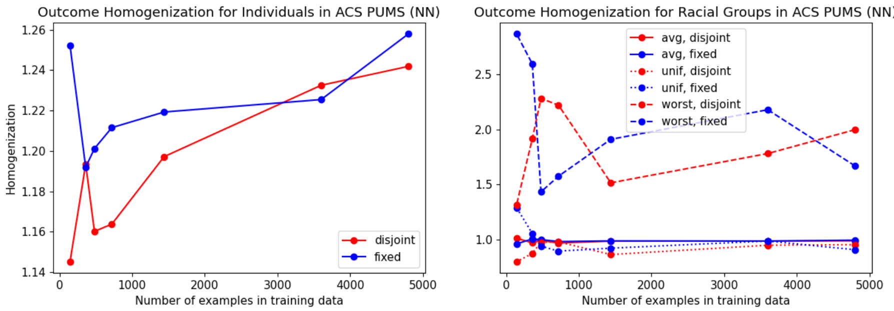

Figure 5: Results for data-sharing experiments on **ACS PUMS** with neural network classifiers showing homogenization (y) as a function of training dataset size (x). Training across tasks on the same data (**fixed**) yields more homogeneous outcomes than on non-identical but identically distributed data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

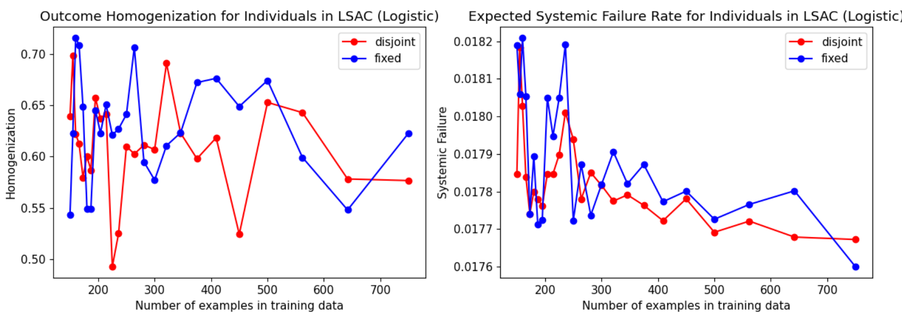

Figure 6: Results for data-sharing experiments on **LSAC** with logistic regression classifiers showing homogenization (**left**) and expected systemic failure rate ($\prod_{i \in [k]} \texttt{fail}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

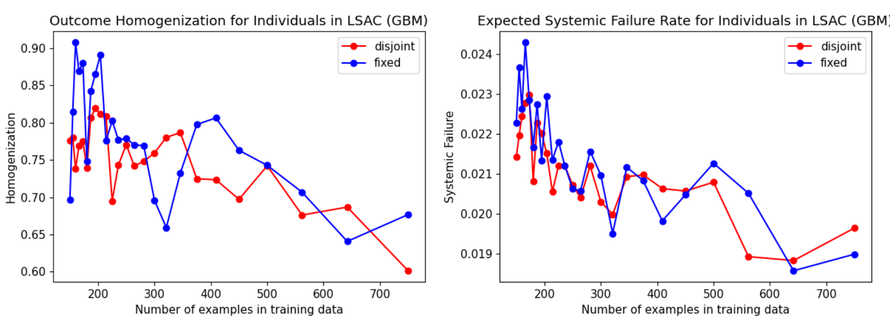

Figure 7: Results for data-sharing experiments on **LSAC** with gradient boosted classifiers showing homogenization (**left**) and expected systemic failure rate ($\prod_{i \in [k]} \texttt{fail}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

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.

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

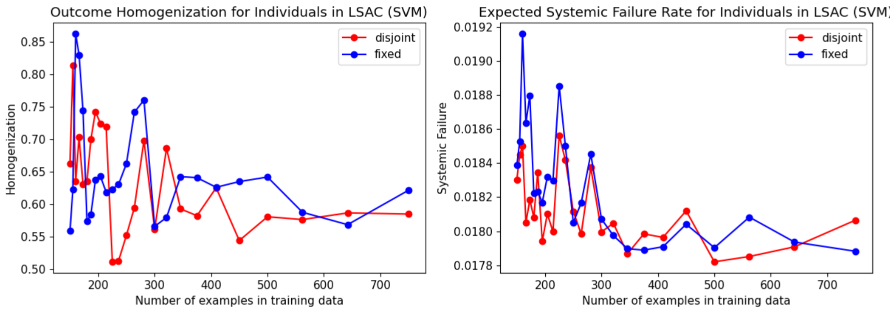

Figure 8: Results for data-sharing experiments on **LSAC** with support vector machines showing homogenization (**left**) and expected systemic failure rate ($\prod_{i \in [k]} \texttt{fail}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

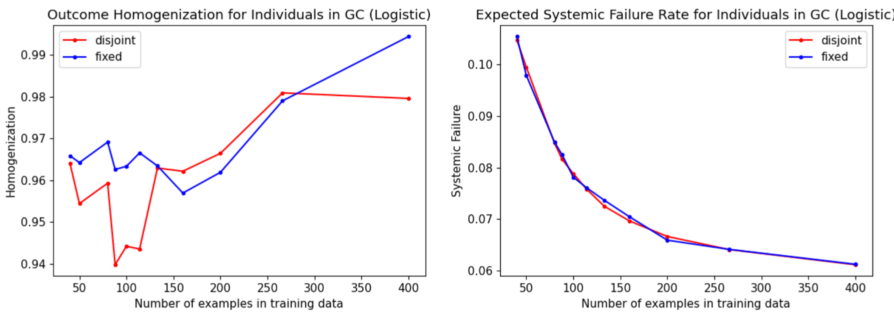

Figure 9: Results for data-sharing experiments on **GC** with logistic regression showing homogenization (**left**) and expected systemic failure rate ($\prod_{i \in [k]} \texttt{fail}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

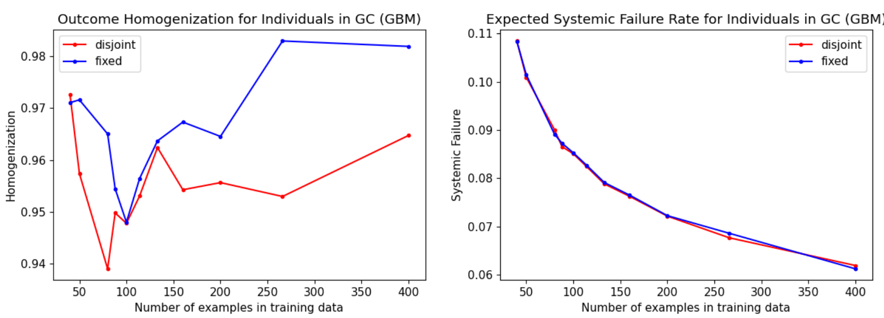

Figure 10: Results for data-sharing experiments on **GC** with gradient boosted classifiers showing homogenization (**left**) and expected systemic failure rate ($\prod_{i \in [k]} \texttt{fail}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

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 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

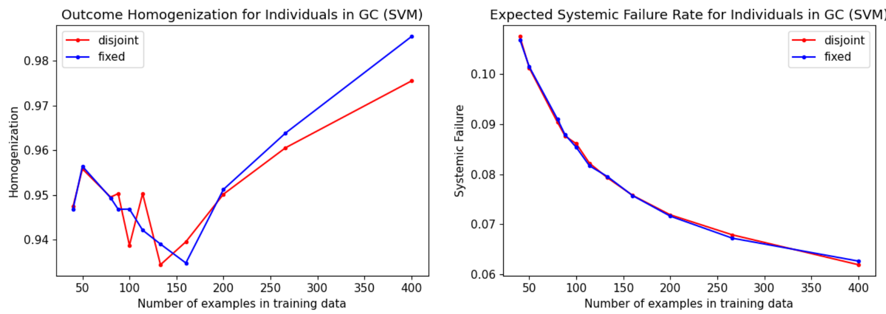

Figure 11: Results for data-sharing experiments on **GC** with support vector machines showing homogenization (**left**) and expected systemic failure rate ($\prod_{i\in[k]} \texttt{fail}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

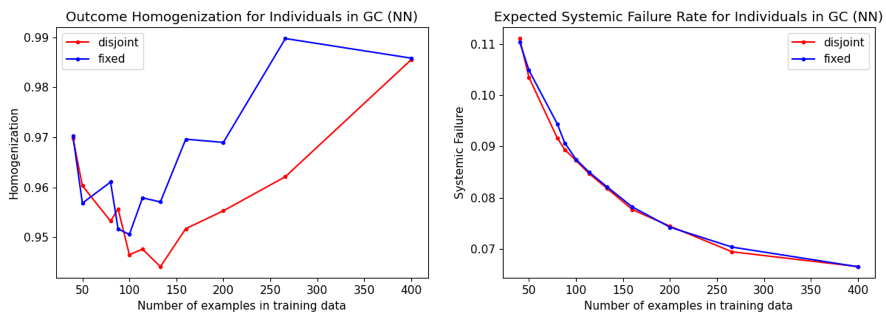

Figure 12: Results for data-sharing experiments on **GC** with neural network classifiers showing homogenization (**left**) and expected systemic failure rate ($\prod_{i\in[k]} \texttt{fail}(h^i)$; **right**), which is the denominator in homogenization, as a function of training dataset size (x).

greater homogenization. This demonstration is reminscent of work that demonstrates[17] 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**).

[17] 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