Fairness Auditing with Multi-Agent Collaboration

Martijn de Vos[a, *], Akash Dhasade[a, *], Jade Garcia Bourrée[b, *], Anne-Marie Kermarrec[a], Erwan Le Merrer[b], Benoît Rottembourg[c] and Gilles Tredan[d]

[a]EPFL, Lausanne, Switzerland
[b]Inria, University of Rennes, Rennes, France
[c]Inria, Paris, France
[d]LAAS, CNRS, Toulouse, France
[*]Corresponding Authors. Email: martijn.devos@epfl.ch, akash.dhasade@epfl.ch, jade.garcia-bourree@inria.fr

Abstract. Existing work in fairness auditing assumes that each audit is performed independently. In this paper, we consider multiple agents working together, each auditing the same platform for different tasks. Agents have two levers: their collaboration strategy, with or without coordination beforehand, and their strategy for sampling appropriate data points. We theoretically compare the interplay of these levers. Our main findings are that *(i)* collaboration is generally beneficial for accurate audits, *(ii)* basic sampling methods often prove to be effective, and *(iii)* counter-intuitively, extensive coordination on queries often deteriorates audits accuracy as the number of agents increases. Experiments on three large datasets confirm our theoretical results. Our findings motivate collaboration during fairness audits of platforms that use ML models for decision-making.

1 Introduction

Machine learning (ML) models are becoming an integral part of many business and industrial processes, increasingly impacting various facets of our lives [20]. Such models are increasingly employed to drive decisions in high-stakes domains [4]. For example, many financial institutions use AI-driven models in which several attributes such as income, credit score, and employment history influence the decision to issue a particular loan [28]. These models are also used to automate the hiring process of certain companies, which would otherwise be labor-intensive [17, 21]. Because these models may significantly impact people's lives, their fairness and regulatory compliance have become increasingly important [33, 23].

Estimating the fairness of ML models is commonly done through algorithmic audits by regulators [32]. However, auditors are not granted unrestricted access to a ML model to protect trade secrets but instead send queries to the model and use the query responses (*i.e.*, a *black-box* interaction) to detect fairness violations. It is common to impose a cap on the queries that are sent to the black-box in order not to overload or interfere with the model being audited [31, 38, 34].

As of today, an auditor performs her audit tasks on each attribute of interest sequentially, one at a time, and independently of other auditors. For example, if she wants to audit a bank's ML model that predicts whether it is safe to issue a loan [16], she begins by auditing the fairness property of the *gender* attribute in a first step. In a subsequent step, she independently audits the fairness property on the *race* attribute. This procedure could result in a sub-efficient auditing scheme in terms of the amount of queries sent to the model. Instead, a coordinated –or *collaborative*– auditing scheme, in which information is shared between distinct audits, might have been more effective. In other words, there is an opportunity to mutualize queries, *i.e.*, through collaboration between the different agents of an auditor. While collaboration in a *single* audit task has recently been introduced in the community [37], to the best of our knowledge, collaborative auditing for multiple tasks has not yet been studied. Therefore, we pose the question: *can an auditor benefit from collaboration among individual audit tasks, e.g., by strategically constructing and sharing queries and responses*?

We answer this question by studying collaboration strategies for independent auditing agents. Specifically, their common goal is to enhance the efficiency and accuracy of auditing a target model for one of the most studied fairness estimation tasks: *demographic parity (DP)* [38, 35, 33]. To this end, we introduce and analyze two realistic forms of multi-agent collaboration. In the *a-posteriori* collaboration, agents share their queries and the responses they receive. In the *a-priori* collaboration, in addition, agents coordinate beforehand on their queries to maximize the information that can be gathered. Besides the types of collaboration, auditors need to wisely choose the strategy for querying data points for estimating DP. These strategies are the sampling methods that address the model's input space that are suitable to the audit black-box setup. Thus, the scientific challenge of this paper is to analyze the relevant combinations of collaboration strategies and sampling methods.

Contributions. This work makes the following contributions:

1. We propose a multi-agent setup in which agents collaboratively perform fairness audits of ML models (Section 4). The collaborations are driven using coordinated sampling methods on sensitive attributes under audit and by sharing query responses.
2. We provide a theoretical analysis of the effectiveness of the a-priori and a-posteriori collaboration strategies and their interplay with different sampling methods (Section 5). First, we show that collaboration is generally beneficial for audit accuracy compared to conducting independent audits. Second, we derive that the advantages of sampling strategies vanish when the number of auditors increases for a-posteriori collaboration. Third, we show that, surprisingly, performing extensive coordination on queries when the number of agents increases sometimes *hurts* audit accuracy.
3. Using three real-world large datasets (Folktables, German Credit, and Propublica), commonly used in fairness studies, we empirically confirm our main theoretical findings (Section 6).

This work is the first to explore the nuances and effectiveness of collaboration between different fairness audit tasks of black-box ML models. In summary, we find that collaboration among agents is a successful setup for increasing query efficiency and detecting biases.

Related Work. We refer the reader to surveys such as [33, 6, 27] for a general introduction to fairness. While the predominant focus within this domain lies on the fair learning, existing work extends to various subjects, including auditing black-box settings [1, 25].

The exploration of fairness in multi-agent systems using game-theoretic frameworks has received relatively little attention [9]. Previous work on fairness in collaborative frameworks, like FAIR [37], emphasizes data sharing and fairness between agents, but it does not align with the objective of estimating fairness in specific black-box models. FAIR focuses on fair collaboration between agents with different devices for scientific discovery, while we focuse on agents of equal importance studying the fairness of a common algorithm.

Beyond technical considerations, legal dimensions also influence the fairness landscape. Traditional legal protocols often constrain inter-agency collaborations for assessing legal compliance, with exceptions such as the recent precedent set by [7]. [8] examines the European Commission's implementation of the Digital Markets Act [15] in March 2024 and offers valuable insights and recommendations for optimizing compliance mechanisms and resolving Big Tech platform investigations effectively, along with recommendations for collaborative processes.

In conclusion, we expect our work to facilitate effective collaboration between different agencies, streamlining regulatory efforts and enhancing law enforcement.

2 Background: Sampling

Let us consider an auditor that wants to know the average value μ of a particular numerical $\{x_i\}_{1\leq i\leq N}$ value in a population of size N, *i.e.*, $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$. In situations where this auditor can only afford consulting a fraction $R < N$ of such values, she has to resort to an *estimator* $\hat{\mu}$. The Bienaymé-Chebyshev inequality states that $\hat{\mu}$, a random variable, has finite variance Var, the probability that $\hat{\mu}$ deviates from its mean by more than a standard deviation is bounded by Var. We hence use variances to characterize empirical errors.

Besides analyzing $\hat{\mu}$ for different sampling methods, this section also highlights the sampling variance of $\hat{\mu}$, which is important as a large sampling variance can lead to inaccurate estimates.

This work explores three sampling methods that auditors use to construct queries: *(i)* uniform sampling, *(ii)* stratified sampling, and *(iii)* Neyman sampling [3, 26]. While uniform sampling is a natural, well-studied, and easy-to-implement choice, stratified sampling offers a fairer method that an auditor can use in a black-box setting. Neyman sampling will serve as an upper bound, as it is the most accurate for an "omniscient" auditor (that would have white-box-like information on the problem).

2.1 Uniform Sampling

The most straightforward approach to estimate μ is to select R members of the population randomly. We refer to the sampled members as $x_1, \dots x_R$. With uniform sampling, the allocation of queries mirrors the real-world distribution of the population. The estimate of μ with uniform sampling, referred to as $\hat{\mu}_{uniform}$, is the mean of x_i over the sampled set: $\hat{\mu}_{uniform} = \frac{1}{R}\sum_{i=1}^{R} x_i$.

Sampling Variance. The sample variance associated with uniform sampling is: $Var(\hat{\mu})_{uniform} = \frac{1}{R}\sum_{i=1}^{R}(x_i - \hat{\mu})^2$.

2.2 Stratified Sampling

Uniform sampling often falls short of providing the most accurate estimator, especially in heterogeneous populations. Stratified sampling involves dividing the heterogeneous population into n subgroups, called *strata*, such that subgroups are non-overlapping and homogeneous with respect to μ. Once the population is segmented into strata, one selects R_j samples from each subgroup j. All the samples drawn from each stratum constitute a stratified sample of total size $R = \sum_{j=1}^{n} R_j$. Stratified sampling can enhance precision by ensuring that all subgroups are adequately represented, making it a more effective method when dealing with diverse groups [3, 22].

We consider in this work *disproportionate stratified sampling* which is a particular type of stratified sampling and in which an equal amount of the query budget is allocated to each stratum. Specifically, with n strata, $R_j = R/n$ of the budget is spent on each stratum j. This sampling strategy ensures that each stratum of a given attribute is sampled using the same number of queries.

Sampling Variance. $\hat{\mu}_{stratified}$ is given by a weighted average of the estimators by stratum sizes: $\hat{\mu}_{stratified} = \sum_{j=1}^{n} p_j\hat{\mu}_j$, with p_j being the probability to be in the stratum j and $\hat{\mu}_j$ being the estimator of μ in this stratum with R_j samples.

It is possible to show that the variance under stratified sampling is [26]: $Var(\hat{\mu})_{stratified} = \sum_{j=1}^{n} p_j^2\left(\frac{1}{R_j} - \frac{1}{p_j N}\right)Var(\hat{\mu}_j)^2$, with $Var(\hat{\mu}_j) = \frac{1}{R_j}\sum_{i=1}^{R_j}(x_{ji} - \hat{\mu}_j)^2$ being the sample standard deviation of stratum j.

While disproportionate stratified sampling offers advantages in terms of representation, it might not be the optimal strategy. In particular, in situations where strata are heavily unbalanced w.r.t. their sizes, the resulting strata may be of uneven interest to an auditor, and a more nuanced sampling strategy might have been preferable.

2.3 Neyman Sampling

Neyman sampling is the optimal sampling strategy defined as the stratified sampling strategy in which the allocation of queries among strata yields the most precise estimate of μ [26]. It minimizes the variance in the estimation process: $(R_1^*, \dots, R_n^*) = argmin\ (Var(\hat{\mu})_{stratified})$. Since Neyman sampling is a specific instance of stratified sampling, its mean and variance estimates are identically constructed: $\hat{\mu}_{Neyman} = \sum_{j=1}^{n} p_j\hat{\mu}_j$ and $Var(\mu)_{Neyman} = \sum_{j=1}^{n} p_j^2\left(\frac{1}{R_j} - \frac{1}{p_j N}\right)Var(\mu_j)^2$.

However, to compute the Neyman sampling allocation, one has to know the standard deviation values of each stratum beforehand. This assumption is unlikely to be met in practice, as an auditor knowing those values could directly derive μ values. Nonetheless, despite its lack of realism, Neyman sampling serves as an optimal baseline to compare practical approaches against it and reveal the impact of auditor's missing knowledge on the precision.

Discussion. The exploration of sampling methodologies reveals a spectrum of approaches, each with strengths and considerations. Uniform sampling is simple to implement but may not accurately represent unbalanced populations. Disproportionate stratified sampling ensures fair representation but can also be limited in unbalanced populations. Neyman sampling, although impractical to implement, is considered optimal as it balances stratum size and intra-stratum variance to minimize estimation variance and maximize precision. In this paper, we emphasize the importance of choosing a sampling method tailored to population characteristics for reliable results and explore the interaction between these methods and collaborative strategies.

3 Problem Statement: Collaborative Auditing

To provide a structured framework for our investigation, we formalize the audit process and delineate its key components.

Auditing Fairness. We investigate a *black-box* algorithm $\mathcal{A} : \mathcal{X} \mapsto \{0, 1\}$ (*e.g.*, a ML model) deployed within a platform setting. The input space $\mathcal{X}$ encompasses a set of m protected, binary attributes denoted as $X_1, X_2, \ldots, X_m$. Most of the work on fairness in ML deals with binary classification problems, and we also adopt this scenario for the sake of consistency [5, 1, 35]. While, most fairness metrics for binary targets can be generalized to support multi-class classification [11], this is beyond the scope of this paper.

In our context, m distinct agents $A_1, \ldots, A_m$ interact with the same black-box algorithm $\mathcal{A}$ to scrutinize fairness attributes associated with specific attributes. Each agent A_i concentrates on assessing a distinct protected attribute X_i. For example, in a loan attribution application [16], some agent A_1 could audit a fairness property of the *gender* attribute (X_1) and A_2 could audit a fairness property on the *race* attribute (X_2). Each attribute X_i induces two groups in X: the favored and unfavored groups. We call a *stratum* each intersection of these $2m$ groups, *i.e.*, with m agents, there are 2^m strata.

In line with related work, we assume that an auditor can issue a fixed total number of B queries, a common preamble in auditing [38]. B is called the *query budget*. We assume that each agent can send R queries to $\mathcal{A}$ where $R = B/m$ is the per-agent query budget.

Independence Assumption. The attributes $X_1, X_2, \ldots, X_m$ are considered *protected* due to their sensitive nature and potential implications for fairness. In line with other works [24, 36, 29], we assume that these attributes are *independent* of one another, *i.e.*, the value of one attribute does not influence or depend on the values of other ones. In theoretical analysis, the assumption of independence among protected attributes is crucial for clarity and insights. However, real-world scenarios may involve interdependencies among these attributes, which is a topic for future investigation.

Fairness Metric. The notion of fairness has been operationalized in several different metrics such as disparate impact, demographic parity, or equalized odds. We refer the reader to the recent survey of Pessach and Shmueli [33] for a comprehensive overview of the topic.

Demographic parity (DP), mathematically denoted as D, has emerged as a critical fairness metric for modeling social equality [4, 30, 38]. An algorithm satisfies DP if, on average, it produces the same predictions across different protected groups. Despite its simplicity, DP is a key measure of fairness, especially in high-stakes areas such as finance and recruitment, where decisions can significantly impact individuals' lives. EU regulations also recognize DP as a metric for detecting bias in algorithmic decision-making [13].

Definition 3.1. The demographic parity D_i with respect to a binary protected attribute X_i captures the impact of X_i on a prediction Y: $D_i = \mathbb{P}(Y = 1|X_i = 1) - \mathbb{P}(Y = 1|X_i = 0)$.

Agent A_i can use her R queries to estimate DP on protected attribute X_i by querying any $x \in \mathcal{X}$. Let $\hat{D}_i$ be a DP estimator:

$$\hat{D}_i = \hat{\mathbb{P}}(Y = 1|X_i = 1) - \hat{\mathbb{P}}(Y = 1|X_i = 0). \quad (1)$$

For brevity, we will write $\hat{Y}_i = \hat{\mathbb{P}}(Y = 1|X_i = 1)$ and $\hat{Y}_{\bar{i}} = \hat{\mathbb{P}}(Y = 1|X_i = 0)$.

According to the Digital Services Act [14], if $D_i = 0 \pm 0.2$, then $\mathcal{A}$ respects demographic parity on protected attribute X_i. While demographic parity could, in addition, be estimated relatively to each stratum (as in intersectional fairness [18]), we focus in this paper on attribute-level DP. DP is a group level fairness metric according to the classification proposed by Pessach and Shmueli [33], which is defined in opposition to individual-level fairness metrics.

While our work focuses on DP for concreteness, we argue that any group-level metric requires an auditor to sample members of different target groups to estimate the group property of interest and has to deal with the convergence of empirical estimators. Our work can hence be easily transposed to other group metrics. For instance, disparate impact uses the ratio of group-level estimates rather than their difference for DP (see eq. 1). Hence, while those two estimators will have different variances, the strategy of minimizing the variance of each subgroup is the same.

Objective of an Auditor. The goal of an auditor is to audit DP as accurately as possible, *i.e.*, by obtaining a DP estimate that is close to the ground truth DP, which is the value that can be computed provided one has access to the whole dataset. An auditor is typically interested estimating the DP for different attributes; in this paper, each such attribute estimation is embodied and performed by an *agent*. We consider the estimations of all attributes as equally valuable to the auditor: the auditor's goal is to have *the average of DP variances, i.e., the difference between her estimated DPs and the actual DPs, as small as possible*. Note that the choice of a regular average as the metric is arbitrary, acknowledging that alternative metrics exist; for instance, Xu et al. [37] proposes a metric ensuring fairness between agents. However, in our context, where all agents share the common goal of auditing and are controlled by a single auditor, this paper strives for the lowest averaged DP variance.

Finally, we assume that agents are homogeneous: they all use the same sampling and collaboration strategy. We leave the analysis of collaboration among heterogeneous agents to future work.

4 Multi-Agent Collaboration

To estimate DP, an agent relies on *(i)* a sampling method that influences the way she constructs her queries, and *(ii)* a collaboration strategy with other agents. We first introduce two collaboration strategies and then derive the sampling variances of these strategies.

4.1 Collaboration Strategies

We introduce two natural collaboration strategies in a black-box audit context and a baseline approach that does not involve collaboration.

No Collaboration (baseline). To quantify the effectiveness of our collaboration strategies, we consider a non-collaborative setting where each agent queries the black-box algorithm by itself and independently of the others. This baseline is also used in [37] and represented in the left part of Figure 1.

***a-posteriori* Collaboration.** The most natural collaboration scheme involves sharing queries and their responses among agents. In this approach, each agent independently queries the black-box algorithm and then shares the queries and resulting responses with the other agents so that all agents can access a pooled set of queries. This is also visualized in the middle part of Figure 1. Even though a query-response pair obtained by agent i might not have been optimized to satisfy the interest of agent j, j might still be able to leverage the information to obtain a more precise DP estimation on attribute X_j.

***a-priori* Collaboration.** We introduce a second collaborative strategy, aiming at a preliminary coordination of agents. With a-priori collaboration, agents collectively implement a sampling strategy, accounting for the estimation tasks of the other participating agents. This is also visualized in the right part of Figure 1. More specifically,

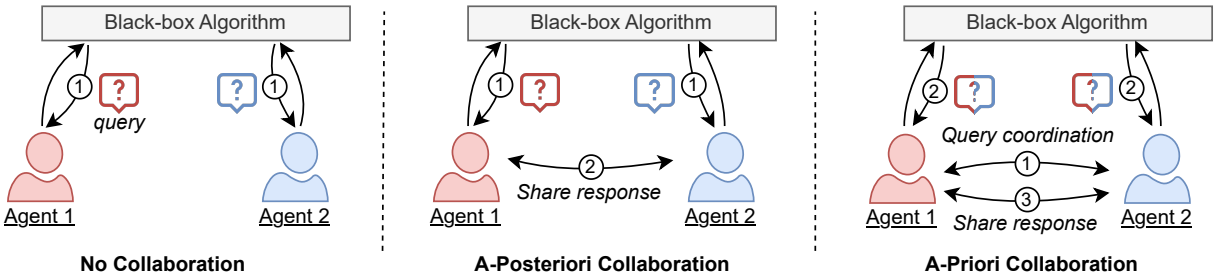

Figure 1: Possible collaboration strategies of an auditor with her two agents: no collaboration (left, baseline), a-posteriori collaboration where agents share queries and responses (middle), and a-priori collaboration where agents also coordinate on queries to be sent (right).

all agents divide the input space into 2^m strata and agree on the sampling strategy of each stratum. This coordination allows for a more integrated and strategic approach to querying, intending to enhance the overall effectiveness of the audit.

Note that a-priori and a-posteriori collaboration strategies cover the possible options in our black-box audit setting. Specifically, there are only two ways to collaborate in a bulk query-response-based approach: before or after sending the queries. a-priori collaboration has coordination before sending the queries, and a-posteriori shares queries and their responses after they have been sent to the platform.

4.2 *From Collaborations to Estimation Variances*

The objective of an auditor is to use a budget of R per-agent queries to derive an accurate estimator $\hat{D}_i$ of the demographic parity D_i of their respective protected attribute X_i. Note that $\hat{D}_i$ is a random variable. Analyzing its distribution is key to understanding the characteristics of the different estimators that result from the interplay of sampling and collaborations employed. This section thus analyzes the distribution of $\hat{D}_i$ for our collaboration strategies.

As seen in Eq. (1), the demographic parity, $\hat{D}_i$, is determined by comparing the two empirical probabilities $\hat{Y}_i$ and $\hat{Y}_{\bar{i}}$. Each empirical probability is calculated as the proportion of positive or negative responses collected from queries within the group of interest, *e.g.*, $\hat{Y}_i = \frac{|X_i=1,Y=1|}{|X_i=1|}$. The empirical probability $\hat{Y}_{\bar{i}}$ is defined by replacing $X_i = 1$ by $X_i = 0$ in the previous equation.

As the probabilities Y_i and $Y_{\bar{i}}$ can be calculated as the average number of positive answers on the stratum verifying $X_i = 1$, respectively $X_i = 0$, our study inherently accommodates the sampling methodologies presented in Section 2.

By linearity of the variance and independence among the protected attributes, the variance of the demographic parity D_i is equal to the sum of the variances of Y_i and $Y_{\bar{i}}$: $Var(D_i) = Var(Y_i) + Var(Y_{\bar{i}})$.

No Collaboration. In the absence of collaboration, agents prioritize their individual attributes regardless of other agents' estimations. More concretely, an agent i splits the input space into two distinct strata based on their binary attribute X_i. These strata represent instances where the attribute is favored ($X_i = 1$) and unfavored ($X_i = 0$). Within each stratum, homogeneity is assumed with respect to the characteristic of interest (Y), leading to the computation of average positive responses denoted as Y_i or $Y_{\bar{i}}$. Consequently, the variance of the DP metric on attribute X_i, denoted as D_i, in a setting without collaboration can be expressed as:

$$Var(\hat{D}_i)_{nocollab.} = \frac{1}{R_i}\sum_{j=1}^{R_i}(x_j - \hat{Y}_i)^2 + \frac{1}{R_{\bar{i}}}\sum_{j=1}^{R_{\bar{i}}}(x_j - \hat{Y}_{\bar{i}})^2. \quad (2)$$

Each agent's ability to measure D_i is constrained by the choice of sample sizes R_i and $R_{\bar{i}}$. The used sampling strategies determines the values of R_i and $R_{\bar{i}}$. All expressions of the total budget R_i and $R_{\bar{i}}$ for different collaboration strategies and sampling methods are summarized in Table 1 in Appendix A. Under uniform sampling, R_i is typically set equal to $p_i R$, representing the probability of observing $X_i = 1$. Conversely, in stratified sampling, a uniform allocation strategy assigns $R_i = R/2$, ensuring an equal budget allocation across all strata. However, Neyman sampling introduces a more nuanced approach, wherein R_i and $R_{\bar{i}}$ are strategically determined to minimize the expression defined in Eq. (2). Depending on the extent to which the division into two strata is relevant, Neyman sampling can, for example, give a distribution close to uniform, close to disproportionate sampling or something in-between [26].

***a-posteriori* Collaboration.** Under a-posteriori collaboration, agents face a similar situation regarding the variance of D_i as when not collaborating.

$$Var(\hat{D}_i)_{a-posteriori} = \frac{1}{R_i}\sum_{j=1}^{R_i}(x_j - \hat{Y}_i)^2 + \frac{1}{R_{\bar{i}}}\sum_{j=1}^{R_{\bar{i}}}(x_j - \hat{Y}_{\bar{i}})^2. \quad (3)$$

We note that R_i and $R_{\bar{i}}$ are expressed differently for the non-collaborative and a-posteriori collaborative settings. The total budget R_i allocated to each stratum is the sum of the budget allocated by the agent i on the stratum and the total budget allocated by the other agents on the stratum. Thus, as each agent homogeneously considers its own attribute in its strata, an agent $j \neq i$ uniformly samples the attribute X_i. The budget on stratum $X_i = 1$ is $p_i(m-1)R$ added to the budget allocated by agent i on this specific stratum.

In the case of stratified sampling, agent i spends half the budget ($R/2$) on each stratum. In total, R_i is thus equal to $R/2 + p_i(m-1)R$. Similarly, with Neyman sampling, the agent i spends the optimal budget R_i^* (obtained from Eq. (3)) on stratum $X_i = 1$. In total, $R_i = R_i^* + p_i(m-1)R$. While for uniform sampling, agent i samples the strata uniformly as the other agents, resulting in $R_i = p_i m R$.

***a-priori* Collaboration.** For a-priori collaboration with uniform sampling, $Var(\hat{D}_i)$ is the same as for a-posteriori collaboration with uniform sampling, as these situations are equivalent. We now analyze a-priori collaboration with stratified or Neyman sampling. With a-priori collaboration, each agent considers the $n = 2^m$ strata, encompassing all possible combinations of the protected attributes. Unlike in a-posteriori collaboration, where agents treat their strata as homogeneous and uniformly sample them, in this scenario, Y_i and $Y_{\bar{i}}$ are stratified, resulting in variances distinct from those observed in previous situations. Specifically, we can express this variance as follows:

$$Var(\hat{D}_i)_{a-priori} = \sum_{j=1}^{n} p_j^2 \left(\frac{1}{R_j} - \frac{1}{p_j N}\right) Var(\hat{D}_j)^2. \quad (4)$$

For stratified sampling, the budget on each stratum is $R_j = B/2^m$ because all agents spend an equal amount of requests on each stratum. For Neyman sampling, we have $R_j = R_j^*$ with $(R_1^*, \ldots, R_n^*) = argmin\ (Var(\mu)_{a-priori})$.

As the right-hand side of Eq. 4 does not depend on i, the variance $Var(\hat{D}_i)_{a-priori}$ is the same for all agents i.

Having established variances for each agent auditing its own attribute, we are now ready to compute the aggregate DP variance across all agents of the auditor. We recall that our overarching goal is to minimize the average variance in practical applications (see Section 3). To globally assess the interest of collaboration, we rely on the average DP variance realized by agents:

Definition 4.1. Average DP variance. According to Section 3, for a set I of collaborative agents where each agent (A_j) audits a demographic parity D_i with variance $Var(\hat{D}_i)$, the average one is:

$$Var(\hat{D}) = \frac{1}{m}\sum_{i=1}^{m} Var(\hat{D}_i). \tag{5}$$

5 The Dynamics of Collaboration

We now outline the foundational principles behind the interplay between collaboration strategies and sampling methods. Building upon our previous derivations, our main result comprises three theoretical outcomes that provide guidelines for collaborative fairness audits.

In the following, $\sum_j (x_j - \hat{Y}_i)^2$ will be denoted as σ_i^2 for brevity as each sample x_j has an equal variance. This shorthand simplifies expressions for clarity in mathematical formulas.

5.1 When Collaboration is Advantageous

We introduce two theorems that describe when collaboration leads to a more accurate audit accuracy.

Theorem 5.1. ***Except for stratified sampling under a-priori collaboration, a-posteriori and a-priori collaboration leads to more accurate results****. Apart from one situation (see Theorem 5.3), collaboration is always beneficial and is an effective approach to increase the accuracy of fairness audits, i.e.* $Var(\hat{D})_{collab} \leq Var(\hat{D})_{nocollab.}$.

Below we provide the proof of this result for the a-posteriori collaboration with stratified sampling.

Proof. As seen in Section 4, the variance of the average DP estimation $Var(\hat{D})_{a-posteriori}$ in this setting can be written as:

$$Var(\hat{D})_{a-posteriori} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{\frac{R}{2} + p_i(m-1)R} + \frac{\sigma_{\bar{i}}^2}{\frac{R}{2} + p_{\bar{i}}(m-1)R}\right).$$

Since $\forall i \in I, (m-1)p_i R > 0$ and $(m-1)p_{\bar{i}} R > 0$, the previous equation leads to the following inequality:

$$Var(\hat{D})_{a-posteriori} \leq \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{\frac{R}{2}} + \frac{\sigma_{\bar{i}}^2}{\frac{R}{2}}\right)}_{:=Var(\hat{D})_{nocollab.}}.$$

Combining Eq. 3 and Eq. 5, the right-hand side of the above inequality is exactly the definition of the variance of $\hat{D}$ without collaboration and stratified sampling (Eq. 2).

Thus, we have just proven that a-posteriori collaboration with stratified sampling is *always* beneficial. In Appendix B, we prove that all combinations of collaboration strategies and sampling methods (with the only exception of a-priori collaboration with stratified sampling) are beneficial over no collaborations. We also show in Appendices B.1.1 and B.2.1 that for all collaborative strategies with uniform sampling, the variance on $\hat{D}$ linearly decreases with the number of collaborating agents.

This linear reduction is also similarly observed for a-priori collaboration with Neyman sampling (Appendix B.2.3).

Conclusions. Except in the case of a-priori collaboration with stratified sampling, the gains from collaboration increases with the number of collaborating agents. The gains from collaboration can even be linear on m. It is therefore recommended that agents collaborate.

Theorem 5.2. ***Under a-posteriori collaboration, stratified and Neyman sampling methods are asymptotically equivalent to uniform sampling****. The advantages of more advanced sampling methods vanishes with the increasing number of agents under a-posteriori collaboration:* $Var(\hat{D})_{stratified} \underset{m\to+\infty}{\sim} Var(\hat{D})_{uniform}$ *and* $Var(\hat{D})_{Neyman} \underset{m\to+\infty}{\sim} Var(\hat{D})_{uniform}$.

Proof. We consider a-posteriori collaboration. Under stratified sampling, the agent i splits equally her budget on the two strata: $R/2$ for $X_i = 1$ and $R/2$ for $X_i = 0$. This distribution does not depend on m. The total budget on these strata with a-posteriori collaboration with m agents is $R_i = R/2 + (m-1)p_i R$. If $m \to +\infty$ then $R_i \sim m p_i R$ (and the same thing replacing i by $\bar{i}$). Thus:

$$Var(\hat{D})_{stratified} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{\frac{R}{2} + p_i(m-1)R} + \frac{\sigma_{\bar{i}}^2}{\frac{R}{2} + p_{\bar{i}}(m-1)R}\right) \underset{m\to+\infty}{\sim} \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{p_i m R} + \frac{\sigma_{\bar{i}}^2}{p_{\bar{i}} m R}\right)}_{:=Var(\hat{D})_{uniform}},$$

which is exactly $Var(\hat{D})_{a-posteriori}$ with uniform sampling. Therefore, when using a-posteriori collaboration with a large number of agents, each agent can simply adopt uniform sampling for its queries. The proof for a-posteriori collaboration with Neyman sampling follows similarly, where $R/2$ is replaced by R_i^* (Appendix C).

Conclusions. We know that without missing knowledge on $\mathcal{A}$, the best sampling method is Neyman sampling. As the variance with Neyman sampling converges to that of uniform, the benefits of extra-information on $\mathcal{A}$ vanishes with the increasing number of agents under a-posteriori collaboration. Thus, if the number of agents collaborating is large in a-posteriori collaboration, each agent can simply do uniform sampling.

5.2 When Collaboration is Disadvantageous

We now highlight potential issues in the collaboration. This section focuses on a-priori collaboration with stratified sampling, which intuitively seems like a desirable candidate in practical settings. This is because of its coordinated nature, with the fairest sampling method that is compatible with black-box assumptions.

The forthcoming theorem leverages the following observation:

Observation 1. ***For any collaboration with*** m ***agents, there is a stratum among the existing*** 2^m ***that represents at least*** $1/(2m)$ ***of the population.*** *It is mathematically expressed as:*

$$\exists j^*, 1 \leq j^* \leq 2^m, p_{j^*} \geq \frac{1}{2m}.$$

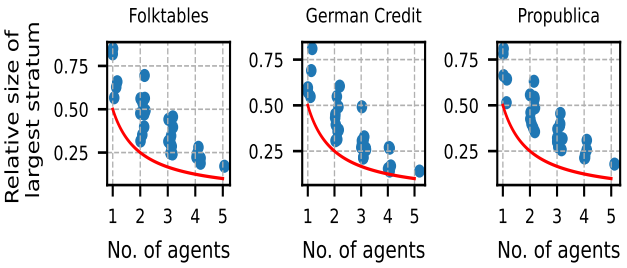

Figure 2: The relative size of the largest stratum for all possible m auditor configurations and three datasets. The red curve is $y = \frac{1}{2x}$.

This observation formalizes that in collaborations involving m agents, there will typically be certain subgroups or strata that represent a significant portion of the overall population. As the number of agents m increases, the number of potential collaboration configurations grows exponentially, reaching 2^m. Yet, as m increases, strata become unbalanced with some stratum being consistently larger than the average 2^{-m} fraction of the dataset. We verify this also within the three classical datasets used for our experiments with binarized attributes in Section 6, and even with non-binary attributes (Appendix E.2). Figure 2 shows the relative size of a largest stratum for each possible assignment of m agents to the five protected attributes considered in each dataset. Any point below the red line would violate Observation 1. Hence, all of the $\sum_{m=1}^{5} \binom{5}{m} = 31$ configurations tested for each dataset confirm Observation 1.

Theorem 5.3. ***The a-priori collaboration can be disadvantageous.*** *The variance on the estimator using stratified a-priori increases with the number of agents.*

If $\forall m > 0, \exists j^*, 1 \leq j^* \leq 2^m, p_{j^*} \geq \frac{1}{2m}$*,(i.e. Observation 1 holds) then* $Var(\hat{D})^{stratified}_{a-priori} \underset{m\to\infty}{\to} +\infty$.

Proof. With a-priori collaboration and stratified sampling, the variance of DP is $Var(\hat{D}) = \frac{1}{m}\sum_{j=1}^{2^m} p_j^2 \left(\frac{2^m}{B} - \frac{1}{p_j N}\right) Var(\hat{D}_j)^2$.

It corresponds to Eq. (4) for $R_i = B/2^m$ summed of all agents. As we are interested in show that the variance of the estimator increases, we lower bound it: $Var(\hat{D}) > \frac{1}{m}\sum_{j=1}^{2^m} p_j^2 \left(\frac{2^m}{B} - 1\right) Var(\hat{D}_j)^2$. In particular, under Observation 1, there is a stratum among the existing 2^m that represents at least $1/(2m)$ of the population. That means that $\exists j^*, 1 \leq i \leq 2^m, p_j \geq \frac{1}{2m}$. The sum of the variances is at least greater than the variance of each of its components: $Var(\hat{D}) > \frac{2^{m-2}}{Bm^3} Var(\hat{D}_{j^*})^2$. In the majority of cases, the decisions of $\mathcal{A}$ on this stratum are not always the same, so $Var(\hat{D}_{j^*}) \neq 0$. Thus, $\frac{2^{m-2}}{Bm^3} Var(\hat{D}_{j^*})^2 \underset{m\to\infty}{\to} +\infty$ and so does $Var(\hat{D})$.

Conclusions. Under Observation 1, Theorem 5.3 demonstrates that the estimator variance tends to go to infinity with a-priori collaboration and stratified sampling. Although this may appear as an appealing strategy, it turns out to be detrimental to the general audit accuracy in realistic settings where strata are severely unbalanced. More precisely, the strategy of allocating a constant number of samples per stratum turns out detrimental as m grows since the number of samples allocated to the majority stratum decreases exponentially with m whereas its size (and hence its contribution to the overall estimation variance) decreases only proportionally to m. While our theoretical results characterizes the asymptotic behaviour as m grows, our experimental results (see Section 6) already show this behaviour for low values of m.

Counter-intuitively, we thus find that a-priori collaboration is not the best strategy to consider. Its variance with stratified sampling increases with the number of agents and its combination with Neyman sampling is impossible in practice. Besides, a-priori collaboration with uniform sampling is equivalent to a-posteriori collaboration with uniform sampling by definition. On the contrary, a-posteriori collaboration exhibits advantages with all sampling methods. We also showed that the advantages of advanced sampling methods vanish when the number of agents is large. So when agents collaborate, if there are many, they have every interest in using a-posteriori collaboration with uniform sampling.

6 Experiments and Results

To empirically evaluate and understand collaboration with real-world datasets, we implement a simulation framework. We leverage three datasets: German Credit [19], Propublica [2] and Folktables [12] (full description is deferred to Appendix D). We consider five attributes on each dataset where each attribute is binary by default or binarized by following a certain scheme (see Appendix D). The labels for the prediction task in each dataset are also binary. To simulate black-box models, we adopt a unique strategy by treating dataset labels as responses from the ML model, avoiding the traditional need to train specific models for each audit task. This approach views datasets as extensive passive sampling sets of the target model, eliminating the need to select a training algorithm and an ML model from diverse choices. We open source all source code and documentation [10].

Setup. We consider combinations of the three sampling methods described in Section 2 and the collaboration strategies described in Section 4. The Folktables dataset is composed of 5 916 565 samples, German Credit of 1000 and Propublica of 6172 samples. We run each experiment for 300 repetitions. These repetitions yield a good balance between accuracy and computational efficiency. Our experiments report the average DP error, whose minimization is the objective of an auditor (see Section 3).

6.1 Impact of Collaboration with Two Agents

We simulate different collaboration strategies with stratified sampling for all three datasets and observe the average DP error for different per-agent query budgets R. The query budgets are varied depending on the dataset, ranging from 100 to 1000 for the Folktables and the Propublica while for the German Credit dataset, we vary from 50 to 250 given its small size.

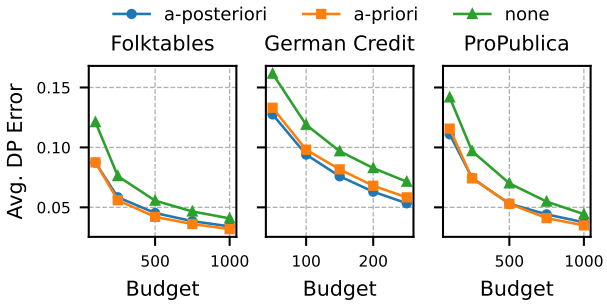

Figure 3: 2-agent collaboration with stratified sampling. The budget ranges are relative to the size of the dataset being studied. We observe that collaboration (a-posteriori and a-priori) can significantly improve DP error. This is in line with Theorem 5.1.

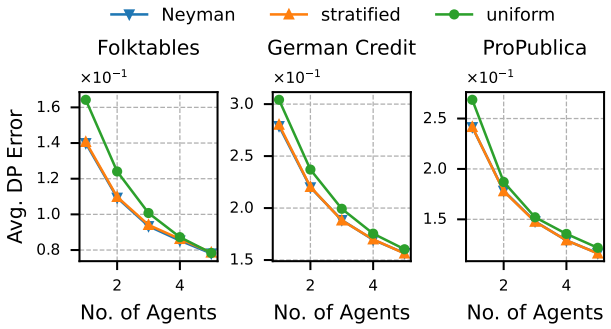

Figure 4: Different sampling methods with a-posteriori collaboration. The more agents collaborate, the more all sampling methods tend to converge. This is in line with Theorem 5.2.

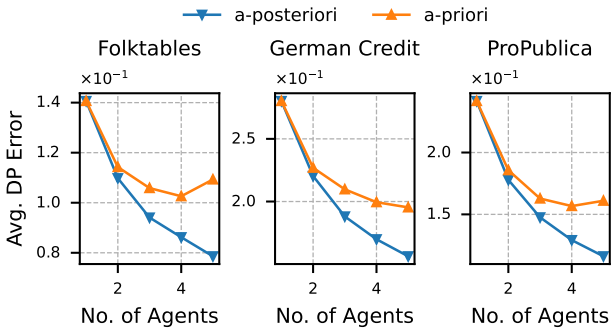

Figure 5: Different collaborative strategies with stratified sampling. We observe that as more agents collaborate, the a-priori strategy can be disadvantageous. This is in line with Theorem 5.3.

For this experiment, two agents audit a particular attribute in each dataset; we consider gender and marital status for the Folktables dataset, age and gender for the German Credit dataset and lastly, gender and African-American origin for the Propublica dataset. These results are shown in Figure 3 where each column represents a different dataset. Our first observation is that for all configurations, the average DP error decreases as R increases since individual DP estimations become closer to the ground-truth value. Secondly, we observe that a-posteriori or a-priori collaboration always decreases the average DP error compared to when not collaborating and therefore *it is always beneficial to collaborate*.

This behaviour is consistent across all three datasets, providing strong empirical evidence for Theorem 5.1. The average DP error reduces anywhere between 17.4% to 24.6% by collaboration. Furthermore, we observe that both the collaborative strategies have similar performance in this two-agent setting. We further extend this comparison in Section 6.3 while considering more agents.

6.2 *Performance of Different Sampling Methods with Multi-agent Collaboration*

This experiment aims to observe how the average DP error changes for different sampling methods as we increase the level of collaboration. We consider the uncoordinated collaboration *i.e.*, the a-posteriori strategy in this section and analyse the coordinated collaboration *i.e.*, a-priori strategy in the following section. For each dataset, we vary the number of collaborating agents (m) from 1 (no collaboration) to 5. For each specific value of m, we report the average over all $\binom{5}{m}$ possible combinations of m collaborating agents. For example, there are 10 combinations of protected attributes when $m = 2$. Each combination is still run several times with different random seeds, as previously described. We set the budget $R = 500$ for the Folktables dataset, $R = 100$ for the German Credit dataset and $R = 250$ for the Propublica dataset.

Figure 4 displays the results of all sampling methods across different datasets, with a column per dataset. We note that the average DP error for all methods decreases as collaboration increases, reinforcing Theorem 5.1. Stratified and Neyman sampling consistently shows lower error rates than uniform sampling when m is small, highlighting the advantages of stratification. However, as m increases, these advantages diminish, and the performance gap with uniform sampling methods narrows significantly. This convergence in performance, anticipated in Theorem 5.2 as $m \to +\infty$, is observed empirically even at moderate values of m. For example, at $m = 5$, the performance of uniform and stratified sampling is alike for the Folktables dataset and closely matched for the other datasets. This empirical evidence confirms Theorem 5.2. Lastly, we also note that while Neyman sampling performs optimally, the empirical difference of average DP error with stratified sampling is very low (< 0.001).

6.3 *Performance of different collaborative strategies with multi-agent collaboration*

In this section we thoroughly examine the coordinated collaboration strategy *i.e.*, a-priori strategy in comparison to the uncoordinated collaboration strategy *i.e.*, a-posteriori. We keep the setup same as Section 6.2 and observe the average DP error when increasing the number of collaborating agents from $m = 1$ to $m = 5$. Figure 5 depicts our results for stratified sampling. We include the results with Neyman sampling in Figure 6 (Appendix E). We observe that, under stratification, the error of a-priori shows a decreasing trend as m grows from 2 to 4 but then either reduces very little or exhibits an increasing trend at full collaboration ($m = 5$). This is particularly evident for the Folktables dataset. While Theorem 5.3 shows that the estimator variance tends to infinity as $m \to +\infty$, we observe the detrimental effects of a-priori collaboration with stratified sampling even for $m = 5$ agents. On the other hand, the error of a-posteriori under stratified sampling always decreases with increasing collaboration. Thus contrary to expectations, *extensive coordination using the a-priori approach can be disadvantageous*, whereas simpler, uncoordinated collaboration consistently proves beneficial.

7 Conclusions and Future Work

Decision-making algorithms and models are now widespread online and often lack transparency in their operation. Multiple regulatory bodies are willing to conduct efficient fairness audits. However, agents can only estimate fairness attributes since they have a hard cap on the number of queries they can issue. This paper shows that collaboration among previously independent audit tasks can yield a substantial gain in accuracy under fixed query budgets. We observed and analyzed an interesting case where prior agent coordination on the queries causes worse outcomes than a non-coordinated collaboration strategy. We also show that, in practice, the latter strategy performs nearly as well as the (infeasible) optimal strategy, underlining the relevance of that proposed collaboration strategy.

Future work directions include exploring the intersectional fairness and review a-priori collaboration in this context [18]. Additionally, collaboration using active approaches, like adaptive sampling, could yield efficient and accurate audits. However, this may entail higher synchronization costs.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, 2018.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[3] P. K. Arieska and N. Herdiani. Margin of error between simple random sampling and stratified sampling. In *Proc. of International Conference Technopreneur and Education*, 2018.

[4] M. Buyl and T. De Bie. Inherent limitations of ai fairness. *Communications of the ACM*, 2024.

[5] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 2022.

[6] S. Caton and C. Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020.

[7] Court of justice of the European Union. Judgment of the court in case c-252/21, press release no. 113/23, 2023. URL https://alinea-avocats.eu/fr/actualite/une-autorite-nationale-de-concurrence-est-competente-pour-constater-une-violation-du-rgpd-dans-le-cadre-de-lexamen-dun-abus-de-position-dominante/.

[8] J. Crémer, D. Dinielli, P. Heidhues, G. Kimmelman, G. Monti, R. Podszun, M. Schnitzer, F. Scott Morton, and A. De Streel. Enforcing the digital markets act: institutional choices, compliance, and antitrust. *Journal of Antitrust Enforcement*, 2023.

[9] S. De Jong, K. Tuyls, and K. Verbeeck. Fairness in multi-agent systems. *The Knowledge Engineering Review*, 2008.

[10] M. de Vos, A. Dhasade, J. Garcia Bourrée, A.-M. Kermarrec, E. Le Merrer, B. Rottembourg, and G. Tredan. Code and data for "fairness auditing with multi-agent collaboration", 2024. URL https://github.com/sacs-epfl/fairness-audits-with-collaboration.

[11] C. Denis, R. Elie, M. Hebiri, and F. Hu. Fairness guarantees in multiclass classification with demographic parity. *Journal of Machine Learning Research*, 2024.

[12] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 2021.

[13] EU. Auditing the quality of datasets used in algorithmic decision-making systems, 2022. URL https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf.

[14] European Commission. Proposal for a regulation of the european parliament and of the council on a single market for digital services (digital services act) and amending directive 2000/31/ec., 2020. URL https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-parliament-and-council-single-market-digital-services-digital-services.

[15] European Parliament and Council of the European Union. Regulation (eu) 2022/1925 of the european parliament and of the council of 14 september 2022 on contestable and fair markets in the digital sector and amending directives (eu) 2019/1937 and (eu) 2020/1828 (digital markets act), 2022. URL https://eur-lex.europa.eu/eli/reg/2022/1925/oj.

[16] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proc. of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015.

[17] V. Giang. The potential hidden bias in automated hiring systems. *Fast Company*, 2018.

[18] U. Gohar and L. Cheng. A survey on intersectional fairness in machine learning: notions, mitigation, and challenges. In *Proc. of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.

[19] H. Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994.

[20] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proc. of CHI conference on human factors in computing systems*, 2019.

[21] A. L. Hunkenschroer and C. Luetge. Ethics of ai-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, 2022.

[22] M. Keramat and R. Kielbasa. A study of stratified sampling in variance reduction techniques for parametric yield estimation. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 1998.

[23] E. Le Merrer, R. Pons, and G. Trédan. Algorithmic audits of algorithms, and the law. *AI and Ethics*, 2023.

[24] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 2017.

[25] P. Maneriker, C. Burley, and S. Parthasarathy. Online fairness auditing through iterative refinement. In *Proc. of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

[26] O. O. Mathew, A. F. Sola, B. H. Oladiran, and A. A. Amos. Efficiency of neyman allocation procedure over other allocation procedures in stratified random sampling. *American Journal of Theoretical and Applied Statistics*, 2013.

[27] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys*, 2021.

[28] D. Mhlanga. Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International journal of financial studies*, 2021.

[29] C. Mougan, L. State, A. Ferrara, S. Ruggieri, and S. Staab. Beyond demographic parity: Redefining equal treatment. *arXiv preprint arXiv:2303.08040*, 2023.

[30] H. Mouzannar, M. I. Ohannessian, and N. Srebro. From fair decision making to social equality. In *Proc. of the Conference on Fairness, Accountability, and Transparency*, 2019.

[31] W. Neiswanger, K. A. Wang, and S. Ermon. Bayesian algorithm execution: Estimating computable properties of black-box functions using mutual information. In *International Conference on Machine Learning*, 2021.

[32] A. Ng. Can auditing eliminate bias from algorithms?, Feb 2021. URL https://themarkup.org/the-breakdown/2021/02/23/can-auditing-eliminate-bias-from-algorithms. Accessed: 2023-10-10.

[33] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys*, 2022.

[34] B. Rastegarpanah, K. Gummadi, and M. Crovella. Auditing black-box prediction models for data minimization compliance. *Advances in Neural Information Processing Systems*, 2021.

[35] H. Singh and R. Chunara. Measures of disparity and their efficient estimation. In *Proc. of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023.

[36] E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 2014.

[37] X. Xu, Z. Wu, A. Verma, C. S. Foo, and B. K. H. Low. Fair: Fair collaborative active learning with individual rationality for scientific discovery. In *Proc. of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[38] T. Yan and C. Zhang. Active fairness auditing. In *International Conference on Machine Learning*, 2022.

A Sampling Methods

In Section 4, the variance formulas have been driven of all sampling methods (uniform, stratified and Neyman) with all the collaboration strategies (no collab., a-posteriori and a-priori). In particular, Equations 2 3 and 4 depend on the number of queries on each subpopulation. The following table sums up the total queries allocation R_i for each subpopulation $X_i = 1$ with m agents. All budget $R_{\bar{i}}$ are the complementary proportions of the total budget (R for no collab., B for a-posteriori and a-priori collaboration).

Table 1: Number of queries R_i in the subpopulation $X_i = 1$ (left table) and $R_{\bar{i}}$ in the subpopulation $X_i = 0$ (right table). The notation R_i^* stands for $R_i^* = argmin\ \left(Var(\hat{DP}_i)\right)$.

	R_i		
	uniform	**stratified**	**Neyman**
no collab.	P_iR	$R/2$	R_i^*
a-posteriori	mP_iR	$R/2 + (m-1)P_iR$	$R_i^* + (m-1)P_iR$
a-priori	mP_iR	$Rm/2$	$m\ R_i^*$

	$R_{\bar{i}}$
no collab.	$R - R_i$
a-posteriori	$B - R_i$
a-priori	$B - R_i$

B Proofs of Theorem 5.1

This section provides the proof of Theorem 5.1(in Section 5) and some complementary results in all cases.

Theorem 5.1. ***Except for stratified sampling under a-priori collaboration, a-posteriori and a-priori collaboration leads to more accurate results***. *Apart from one situation (see Theorem 5.3), collaboration is always beneficial and is an effective approach to increase the accuracy of fairness audits, i.e.* $Var(\hat{D})_{collab} \leq Var(\hat{D})_{nocollab.}$.

B.1 a-posteriori collaboration

Proof. As seen in Section 4, the variance of the average DP estimation $Var(\hat{D})$ with any sampling method and collaboration can be written as:

$$Var(\hat{D}) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{R_i} + \frac{\sigma_{\bar{i}}^2}{R_{\bar{i}}}\right)$$

We refer to Table 1 for the budget expressions R_i depending on the sampling methods and collaboration.

B.1.1 uniform sampling

With uniform sampling, $R_i = mp_iR$ and $R_{\bar{i}} = m(1-p_i)R$. So

$$Var(\hat{D})^{uniform}_{a-posteriori} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{mp_iR} + \frac{\sigma_{\bar{i}}^2}{m(1-p_i)R}\right)$$

By factoring by $1/m$ we find $Var(\hat{D})^{uniform}_{nocollab.}$:

$$Var(\hat{D})^{uniform}_{a-posteriori} = \frac{1}{m}\Bigg(\underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{p_iR} + \frac{\sigma_{\bar{i}}^2}{(1-p_i)R}\right)}_{:=Var(\hat{D})^{uniform}_{nocollab.}}\Bigg)$$

Thus ($m > 1$), $Var(\hat{D})^{uniform}_{a-posteriori} < Var(\hat{D})^{uniform}_{nocollab.}$. We even demonstrate that using the uniform sampling, the variance of a-posteriori collaboration decreases by a factor m compared to no collab..

B.1.2 stratified sampling

It is the same proof as in Section 5.

With stratified sampling, $R_i = \frac{R}{2} + p_i(m-1)R$ and $R_{\bar{i}} = \frac{R}{2} + p_{\bar{i}}(m-1)R$. So

$$Var(\hat{D})^{stratified}_{a-posteriori} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{\frac{R}{2} + p_i(m-1)R} + \frac{\sigma_{\bar{i}}^2}{\frac{R}{2} + p_{\bar{i}}(m-1)R}\right)$$

Since $\forall i \in I, (m-1)p_iR > 0$ and $(m-1)p_{\bar{i}}R > 0$, the previous equation leads to the following inequality:

$$Var(\hat{D})^{stratified}_{a-posteriori} < \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{\frac{R}{2}} + \frac{\sigma_{\bar{i}}^2}{\frac{R}{2}}\right)}_{:=Var(\hat{D})^{stratified}_{nocollab.}}$$

Combining Equation (3) and Equation (5), the right-hand side of the above inequality is exactly the definition of $Var(\hat{D})^{stratified}_{nocollab.}$, or the variance of the average DP without collaboration with stratified sampling.

B.1.3 Neyman sampling

With Neyman sampling, $R_i = R_i^* + (m-1)p_iR$ and $R_{\bar{i}} = (R - R_i^*) + (m-1)p_iR$. We note The proof is identical to that of stratified sampling.

The variance of the average DP estimation $Var(\hat{D})^{Neyman}_{a-posteriori}$ can be written as:

$$Var(\hat{D})^{Neyman}_{a-posteriori} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{R_i^* + (m-1)p_iR} + \frac{\sigma_{\bar{i}}^2}{(R - R_i^*) + (m-1)p_iR}\right)$$

Since $\forall i \in I, (m-1)p_iR > 0$ and $(m-1)p_{\bar{i}}R > 0$, the previous equation leads to the following inequality:

$$Var(\hat{D})^{Neyman}_{a-posteriori} < \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{R_i^*} + \frac{\sigma_{\bar{i}}^2}{R - R_i^*}\right)}_{:=Var(\hat{D})^{Neyman}_{nocollab.}}$$

Combining Equation (3) and Equation (5), the right-hand side of the above inequality is exactly the definition of $Var(\hat{D})_{nocollab.}$, or the variance of the average DP without collaboration with stratified sampling.

B.1.4 Conclusion

Thus, a-posteriori collaboration is *always* beneficial whether with uniform, stratified or Neyman. We even demonstrate that using the uniform sampling, the variance of a-posteriori collaboration decreases by a factor m compared to no collab..

B.2 a-priori collaboration

Let us now move on to the case of a-priori collaboration. We treat the uniform sampling and Neyman sampling. The case of stratified sampling is a specific case leading to Theorem 5.3 which we will prove in Appendix B.1.

B.2.1 uniform sampling

As established in Section 4.2, the sampling variance of a-priori collaboration with uniform sampling, $Var(\hat{D})^{uniform}_{a-priori}$ is the same as for a-posteriori collaboration with uniform sampling, $Var(\hat{D})^{uniform}_{a-posteriori}$, as these situations are equivalent. The result proved in Appendix B.1.1 shows that $Var(\hat{D})^{uniform}_{a-priori} < Var(\hat{D})^{uniform}_{nocollab.}$. We even demonstrate that using the uniform sampling, the variance of a-priori collaboration decreases by a factor m compared to no collab..

B.2.2 stratified sampling

It is the specific case leading to Theorem 5.3.

B.2.3 Neyman sampling

The sampling variance with a-priori collaboration is defined in Equation (4) as:

$$Var(\hat{D}_i)_{a-priori} = \sum_{j=1}^{n} p_j^2\left(\frac{1}{R_j} - \frac{1}{p_jN}\right)\sigma_j^2$$

With Neyman sampling, $R_j = mR_j^*$. The variance of the average DP estimation $Var(\hat{D})^{Neyman}_{a-priori}$ can be written as:

$$Var(\hat{D}_i)^{Neyman}_{a-priori} = \sum_{j=1}^{2^m} p_j^2 \left(\frac{1}{mR_j^*} - \frac{1}{p_j N}\right)\sigma_j^2$$
$$\leq \frac{1}{m}\sum_{j=1}^{2^m}\frac{\sigma_j^2}{R_j^*}$$

The inequality is obtained by harsh bounds ($\forall j, p_j^2 \leq 1$ and $\frac{1}{p_j N} > 0$). It means that the sampling variance of a-priori collaboration with Neyman sampling is lower than the unweighted sum of the sampling variance of the 2^m stratas.

In the other hand, the sampling variance of no collab. with Neyman sampling is: $Var(\hat{D})_{nocollab.} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{R_i^*} + \frac{\sigma_{\bar{i}}^2}{R-R_i^*}\right)$ (Equation (2) in Equation (5) with R_i and $R_{\bar{i}}$ defined with Table 1). The sum can be splits on the 2^m strata considered in the collaboration (the intersection of all possible subpopulation). In that case, all strata are counted once per agent so m times in the global sum: $Var(\hat{D})_{nocollab.} = \frac{1}{m}\sum_{j=1}^{2^m} m\frac{\sigma_j^2}{R_j^*}$. We can thus write $Var(\hat{D})_{nocollab.} = mVar(\hat{D}_i)^{Neyman}_{a-priori}$ *i.e.* $Var(\hat{D}_i)^{Neyman}_{a-priori} \leq \frac{1}{m}Var(\hat{D})_{nocollab.}$.

We even demonstrate that using the Neyman sampling, the variance of a-priori collaboration decreases at least by a factor m compared to no collab..

B.2.4 Conclusion

a-priori collaboration is beneficial with uniform or Neyman. We even demonstrate that for those two sampling methods, the variance of a-priori collaboration decreases by a factor m compared to no collab..

C Proofs of Theorem 5.2

Theorem 5.2. ***Under a-posteriori collaboration, stratified and Neyman sampling methods are asymptotically equivalent to uniform sampling***. *The advantages of more advanced sampling methods vanishes with the increasing number of agents under a-posteriori collaboration:* $Var(\hat{D})_{stratified} \underset{m\to+\infty}{\sim} Var(\hat{D})_{uniform}$ *and* $Var(\hat{D})_{Neyman} \underset{m\to+\infty}{\sim} Var(\hat{D})_{uniform}$.

This theorem has been proven for a-posteriori collaboration with stratified sampling in Section 5. The proof is exactly the same for $Neyman$ sampling with a-posteriori collaboration by replacing $R/2$ by R_i^* in the proof:

Proof. We consider a-posteriori collaboration. Under Neyman sampling, agent i splits her budget on the two strata as follows: R_i^* for $X_i = 1$ and $B - R_i^*$ for $X_i = 0$. This distribution does not depend on m The total budget on these strata, with a-posteriori collaboration with m agents is $R_i = R_i^* + (m-1)p_i R$. If $m \to +\infty$ then $R_i \sim mp_i R$ (and the same thing replacing i by $\bar{i}$). Thus:

$$Var(\hat{D})^{Neyman}_{a-posteriori} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{R_i^* + p_i(m-1)R} + \frac{\sigma_{\bar{i}}^2}{R - R_i^* + p_{\bar{i}}(m-1)R}\right)$$
$$\underset{m\to+\infty}{\sim} \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sigma_i^2}{p_i mR} + \frac{\sigma_{\bar{i}}^2}{p_{\bar{i}} mR}\right)}_{:=Var(\hat{D})^{uniform}_{a-posteriori}}$$

which is exactly $Var(\hat{D})^{uniform}_{a-posteriori}$. Therefore, when using a-posteriori collaboration and the number of agents is large, each agent can adopt uniform sampling for its queries.

D Additional Notes on Experiment Setup

D.1 Description of the leveraged datasets

We conduct our study using three datasets: German Credit [19], Propublica [2] and Folktables [12]. In the German Credit dataset, the task involves predicting the creditworthiness of loan applicants. Within the Propublica dataset, we consider the recidivism risk task, predicting whether an individual will re-offend within two years after their initial criminal involvement. In the Folktables dataset, we consider the ACSPublicCoverage task that predicts whether low-income individuals, ineligible for Medicare, are covered by public health insurance. Attributes, such as age, gender, and demographic information, among others, are employed in these prediction tasks. While some attributes are inherently binary, we binarize others by grouping values. For instance, the marital status attribute in the Folktables dataset is binarized as 1 for married and 0 for other statuses (widowed, divorced, separated, or never married). In total, after binarization, we have five attributes corresponding to five auditing agents for each dataset. Lastly, the prediction labels for each of the above tasks are also binary. A comprehensive summary of the adopted datasets and the binarization strategy for each attribute can be found in Appendices D.2 to D.4.

In practice, the agents audit a platform through a black-box model. To simulate such an audit, we must train a model for each task to be audited later. In this work, we take a different approach and consider the labels in the dataset to be the response of the ML model when queried with the corresponding attributes. In other words, the datasets can be interpreted as a large passive sampling set of the target model (in our case, the real process that generated a given dataset). This strategy prevents the need of having to choose a training algorithm along with a ML model, among the diverse array of choices that exist.

D.2 Folktables dataset

In the Folktables dataset [12], we address the ACSPublicCoverage task, predicting whether a low-income individual without Medicare eligibility is covered by public health insurance. We consider the following five attributes for auditing: gender, marital status, age, nativity and mobility status. Their binarisation scheme is detailed in Table 2.

Table 2: Attributes in the Folktables ACSPublicCoverage task. The value to description mapping for the original values can be found in [12].

Attribute X_i	How was it binarized ?	$P(X_i = 1)$
SEX	Binary by default	0.43
NATIVITY	Binary by default	0.85
MIG	Class 1 is original value $\{1\}$ and 0 for original values $\{\text{N/A}, 2, 3\}$	0.82
AGEP	Class 1 is when $age \geq 25$ and 0 when $age < 25$	0.66
MAR	Class 1 is original value $\{1\}$ and 0 for original values $\{2, 3, 4, 5\}$	0.37

D.3 German Credit dataset

The task in the German Credit dataset involves predicting whether a given individual is a good or bad credit risk [19]. We chose the following five attributes for auditing: age, gender, marital status, whether the person has own telephone and employment status. Their binarisation scheme is detailed in Table 3.

Table 3: Attributes in the German Credit dataset. More information regarding the dataset can be found in [19].

Attribute X_i	How was it binarized ?	$P(X_i = 1)$
Own telephone	Class is 0 when original value is 'none' and 1 otherwise	0.40
Marital status	Class is 0 for original value 'single' and 1 otherwise	0.45
Gender	Class is 0 when original value is 'female' and 1 otherwise	0.69
Age	Class is 1 when $age > 25$ and 0 when $age \leq 25$	0.81
Employment status	Class is 1 for original values ≥ 4 and 0 otherwise	0.42

D.4 Propublica dataset

The recidivism risk task in the ProPublica dataset involves predicting whether an individual will re-offend within 2 years after their initial criminal involvement [2]. We consider the following five attributes for auditing: female, African-American origin, age below twenty five, misdemeanor and number of prior crimes. Their binarisation scheme is detailed in Table 4.

Table 4: Attributes in the Propublica dataset. More information regarding the dataset can be found in [2].

Attribute X_i	How was it binarized ?	$P(X_i = 1)$
Female	Binary by default	0.19
Misdemeanor	Binary by default	0.36
African-American	Binary by default	0.51
Age below twenty five	Binary by default	0.22
Number of prior crimes	Class is 1 if original value > 0 and 0 otherwise	0.66

E Additional Experiments

E.1 On Neyman sampling

In this section we include the results for a-priori and a-posteriori collaboration with Neyman sampling, expanding on results in Section 6.3. We observe that while a-priori strategy with stratified sampling can perform poorly (Figure 5), a-priori in combination with Neyman sampling always performs optimally as expected. However, Neyman sampling is infeasible in practice as we discussed in Section 2.3.

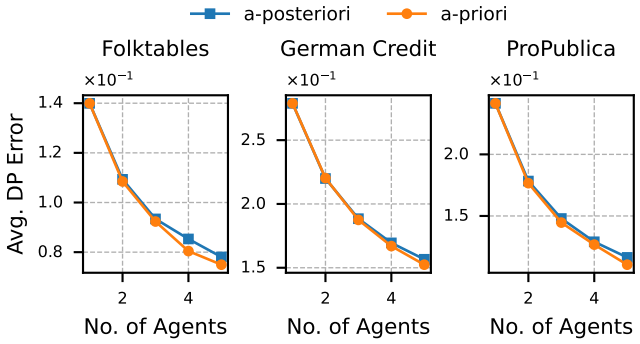

Figure 6: Different collaborative strategies with Neyman sampling. We observe that a-priori strategy with Neyman sampling performs optimally as expected. However, it remains infeasible in practice.

E.2 On multi-valued attribute setting

In this paper, we assume that agents audit only binary attributes. This assumption is not always verified in reality. As there is no clear consensus on how to define fairness for non-binary attributes, we considered demographic parity (DP) as the standard fairness metric and we binarized attributes in the experiments. However, to add some insight on non-binary attributes, we show in Figure 7 that Observation 1 also holds in multi-valued attribute settings.

For example, rather than binarizing the 'Employment status' attribute in the German Credit dataset, we retain the five original groups as distinct values for this multi-valued attribute. Similarly, the 'Age' attribute in the German Credit dataset is divided into three groups: $\{< 25, [25-50], > 50\}$. Tables 5 to 7 present all the multi-valued attributes in the Folktables, the German Credit and the Propublica datasets respectively. In Figure 7, we observe that even in the multi-valued attribute case, the largest stratum represents a significant portion of the overall population. Thus, a-priori collaborations may be disadvantageous in other fairness audits, even when the attributes are not binary.

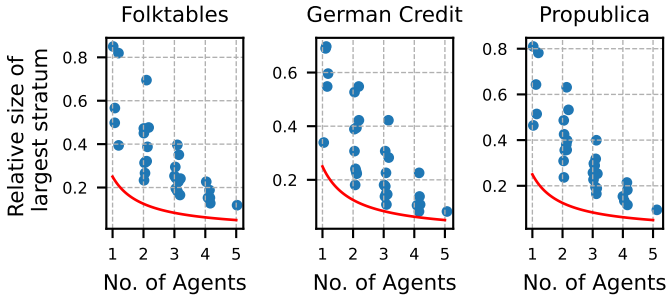

Figure 7: The relative size of the largest stratum for non-binary attributes in the three datasets. The regression curve is $y = \frac{1}{4x}$.

Table 5: Multi-valued attributes in the Folktables ACSPublicCoverage task.

Attribute X_i	Total classes and their relation to original attribute values
SEX	Binary by default
NATIVITY	Binary by default
MIG	3 classes corresponding to original values $\{1, 2, 3\}$
AGEP	3 classes corresponding to $age < 25$, $age \in [25, 50]$, $age > 50$
MAR	5 classes corresponding to original values $\{1, 2, 3, 4, 5\}$

Table 6: Multi-valued attributes in the German Credit dataset.

Attribute X_i	Total classes and their relation to original attribute values
Own telephone	Binary by default
Marital status	4 classes corresponding to {'single', 'div/dep/mar', 'div/sep' and 'mar/wid'}
Gender	Binary by default
Age	3 classes corresponding to $age < 25$, $age \in [25, 50]$, $age > 50$
Employment status	5 classes corresponding to the bins $< 1, [1, 4), [4, 7), \geq 7$ and 'unemployed'

Table 7: Multi-valued attributes in the Propublica dataset.

Attribute X_i	Total classes and their relation to original attribute values
Female	Binary by default
Misdemeanor	Binary by default
African-American	Binary by default
Age below twenty five	Binary by default
Number of prior crimes	4 classes corresponding to the bins $0, [1, 5], [6, 10], > 10$