# *Beyond Reasonable Doubt: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds*

# Beyond Reasonable Doubt: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds

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

Duy Patrick Tu
Massachusetts Institute of Technology
MIT-IBM Watson AI Lab
Cambridge, MA, USA
patrick2@mit.edu

Krishna P. Gummadi
Max Planck Institute for Software Systems
Saarbrücken, Germany
gummadi@mpi-sws.org

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

Kush R. Varshney
IBM Research
Yorktown Heights, NY, USA
krvarshn@us.ibm.com

Adrian Weller
University of Cambridge
Cambridge, United Kingdom
The Alan Turing Institute
London, United Kingdom
aw665@cam.ac.uk

## ABSTRACT

Prior work on fairness in machine learning has focused on settings where all the information needed about each individual is readily available. However, in many applications, further information may be acquired at a cost. For example, when assessing a customer's creditworthiness, a bank initially has access to a limited set of information but progressively improves the assessment by acquiring additional information before making a final decision. In such settings, we posit that a fair decision maker may want to ensure that decisions for all individuals are made with similar expected error rate, even if the features acquired for the individuals are different. We show that a set of carefully chosen confidence thresholds can not only effectively redistribute an information budget according to each individual's needs, but also serve to address individual and group fairness concerns simultaneously. Finally, using two public datasets, we confirm the effectiveness of our methods and investigate the limitations.

## CCS CONCEPTS

• **Applied computing** → *Law, social and behavioral sciences*; • **Computing methodologies** → *Machine learning*; • **Social and professional topics** → **Computing / technology policy**.

## KEYWORDS

fairness, individual fairness, active feature acquisition

**ACM Reference Format:**
Michiel A. Bakker, Duy Patrick Tu, Krishna P. Gummadi, Alex 'Sandy' Pentland, Kush R. Varshney, and Adrian Weller. 2021. Beyond Reasonable Doubt: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462575



*AIES '21, May 19–21, 2021, Virtual Event, USA.*

ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462575

## 1 INTRODUCTION

The machine learning community has proposed myriad definitions for fairness [37], that can be broadly categorized in two groups. (1) *Group* or *statistical* definitions of fairness focus on balancing classification errors across protected population subgroups (based on attributes like race or gender), towards achieving equal error rates (*overall accuracy equality*), equal false-positive rates (*predictive equality*), equal false-negative rates (*equal opportunity*), or both (*equal odds*). Although these notions can be easily verified, they fail to give meaningful guarantees to individuals. (2) *Individual* notions of fairness, on the other hand, provide individual-level guarantees, as opposed to enforcing parity of some quantity that is averaged over a group. For example, Dwork et al. [12] require that 'similar individuals should be treated similarly', using a predefined distance function to measure similarity between individuals. Individual fairness definitions have stronger semantics but existing proposals have proven difficult to implement in practice [9].

Methods for improving fairness have focused predominantly on predictive models that make or support decisions when all data is readily available. In such a setting, the model makes a classification decision for each subject based on the same set of features that is observed at training time. In practice, however, there are many scenarios where some or all features are missing initially and features can be acquired during the decision making process, possibly at a cost [24]. As our running example, we consider a set of applicants applying for a job. If the position is temporary, a hiring manager might make the final decision only based on the set of resumes. By contrast, when the applicants are instead considered for a more permanent position, a hiring manager will want to invite a subset of the applicants for an assessment or interview to gather additional information before making a more informed hiring decision. Taking all candidates through the full process is

prohibitively expensive and time-consuming, hence, for each candidate, an important decision should be made as to which information should be gathered and how much information is required. This problem, in which a decision maker sequentially selects a subset of costly features for each individual, is called *active feature acquisition* (AFA) and is relevant in a wide range of domains such as credit assessment, criminal justice, medical diagnosis, advertising and recruiting [15, 24, 25, 27, 35].

There is a large literature concerned with different AFA strategies that, given a set of already selected features, predict the most cost-effective feature to select next [21, 24, 28, 35]. However, there is a second important part of the problem that has received comparatively little attention: *how should we split a limited information budget across individuals?* Most work on AFA simply assumes an equal information budget for each individual such that the feature acquisition process stops when this budget is reached [24, 28]. Though this approach does treat everyone equally in terms of information budget, it can yield outcomes that are unfair in terms of the quality of the decision.

Consider again the example of the hiring manager. While going through a stack of resumes, they might observe a pool of candidates that can broadly be categorized in two groups: some candidates have more familiar, traditional backgrounds while other candidates have taken unconventional paths or have unfamiliar schools or employers on their resume. Neither group is, on average, more qualified for the job but note that, at least in the context of the "training data", decisions made about the latter group of candidates will have lower confidence and are thus more likely to be erroneous. Collecting more information for each candidate, for example through assessments or interviews, will probably not close this gap if we keep splitting the information budget equally and collect the same amount of information for all candidates. As a result, good but unconventional candidates are more likely to be rejected and, perhaps even worse, a conservative hiring strategy will not even consider the unfamiliar set of candidates. Similarly, many U.S. households lack a formal credit history, making it difficult to apply for loans at traditional financial institutions [5]. However, the absence of a credit history does not mean that households are less creditworthy *per se*, it might only require more effort to assess how creditworthy these individuals really are.

### 1.1 Our contributions

We propose a more equitable strategy for deciding how to spend a limited feature budget across individuals, inspired by the legal concept of "Beyond a reasonable doubt". Beyond a reasonable doubt is a legal standard of proof required to validate a criminal conviction [19]. Simply put, if the jury (or judge) perceives that the probability a defendant committed a crime is equal or greater than their interpretation of beyond reasonable doubt, they will decide to convict [22]. The "burden of proof" in this case rests not on the defendant, who is "innocent until proven guilty", but on the prosecution, who is tasked with finding sufficient evidence to support the innocence or guilt of the accused.

Building on this legal foundation, we argue that for fair decision making, the "burden of uncertainty" should not be put on the decision subject but on the decision maker. In the example of hiring, a hiring manager should thus request more information from the unfamiliar set of applicants so that the eventual hiring decision can be made with equal confidence for each applicant. Similarly, in a fair financial system, institutions would be required to first close the information gap between loan applicants by collecting more information before making a decision. Hence, to make fair decisions, a decision maker should continue the collection of new information until every individual can be guaranteed a decision that is equally likely to be erroneous. We thus want to guarantee an equal error rate in expectation, an individual notion of fairness which we call *individual error parity*.

We will formalize the notion of individual error parity and show how, in contrast to most existing proposals for individual fairness, this notion has a natural connection to group fairness – in particular, how it implies overall accuracy equality across groups, and in some cases also equal odds. Subsequently, we show how "confidence thresholds" as stopping criteria in AFA are a natural mechanism to guide information collection and mitigate individual error disparity. In particular, we derive a set of thresholds, determining when to stop collecting features for each individual, which ensure that we only classify an individual's outcome once we have acquired a sufficient number of features to reach a predefined expected error rate. Because the expected error rate will be the same across individuals, we attain error parity for calibrated probabilistic classifiers across groups and, in expectation, across individuals within these groups.

Our approach to fairness differs from the typical approaches to mitigating unfairness such as data pre-processing [6, 14], constraining or regularizing the learning process [3, 38], or post-processing predictions [17]. However, we suggest that in many settings our approach is intuitive. In the standard machine learning paradigm, in which a decision-maker collects the exact same information for each individual, individuals from groups that are underrepresented in the training data will naturally face more erroneous decisions. Redistributing the amount of information we collect across individuals is therefore an intuitive way to guarantee equitable decision-making. Our approach trades off inequality (the set of features used is personalized in our framework and thus varies across individuals) for equity (each of the individuals are classified with an equal expected error rate).

## 2 ADDITIONAL RELATED WORK

### 2.1 Fairness in machine learning

Most recent work on fairness in machine learning focuses on *group fairness* notions, fixing a number of protected demographic subgroups and requiring parity of some statistical measure across these subgroups. One such group fairness notion is *demographic parity*, requiring parity of the raw positive classification rate across group. However, demographic parity can lower a model's accuracy, especially when the base rate across groups differ [17]. To tackle this, one can instead consider parity in error rates. *Overall accuracy equality* is achieved when the total classification error is the same across protected subgroups [4]. When either false-positives or false-negatives are desirable, one can consider equal false-positive rates (*predictive equality*) or false-negative rates (*equal opportunity*), while *equal odds* requires parity in both false-positive and

false-negative rates [17]. We refer to Verma and Rubin [37] for an overview of definitions.

In contrast, *individual fairness* notions define unfairness at the level of a single individual. An early formalization of individual fairness was proposed by Dwork et al. [12], defining fairness as the smoothness of the classification function by requiring that 'similar individuals should be treated similarly'. The authors introduce a framework to maximize accuracy subject to a fairness constraint that binds on pairs of individuals. Their framework has two drawbacks: it requires a predefined distance function that determines how similar two individuals are and the optimization does not produce an inductive rule that generalizes to previously unseen data [12, 20].

Given the limitations of individual and group fairness notions, is there a way to get the 'best of both worlds'? Kearns et al. [23] and Hébert-Johnson et al. [18] ask for group fairness definitions to hold on an exponential class of groups instead of a small number of subgroups. Although promising, the approach still inherits the weaknesses of group fairness at a smaller scale [8]. Sharifi-Malvajerdi et al. [34] propose *average individual fairness*, computing the statistical error rate not across individuals but across different classification tasks an individual is subjected to during a period of time. Our work is similar in the sense that is aimed at error rate parity at the individual level. However, where they define rate as an expectation over multiple classification tasks over time, we define rate as the expected error rate for a single task.

To measure individual unfairness with respect to the expected error rate, we build on a fairness definition introduced by Speicher et al. [36]. Using inequality indices from economics, they formulate a measure of inequity of the distribution of a benefit over the population. This naturally captures notions of both individual and group fairness. Our approach focuses on a measure of individual expected error rate which we call 'risk' instead of 'benefit', such that by minimizing this, we mitigate both individual and group unfairness.

## 2.2 Prediction-time active feature-value acquisition

Different from *active learning* where labels are queried to improve a model, prediction-time active feature-value acquisition (AFA) describes the problem where costly feature-values of a test instance are unknown and are acquired sequentially. The different but related problem of training-time active feature-value acquisition is concerned with which feature values must be acquired for model improvement.

An AFA system consists of three components: 1) a classifier that can handle partially observed feature sets, 2) a strategy for determining which feature to select next based on the features that are already collected, and 3) a stopping criterion for determining when to stop acquiring more features and make a final prediction. First, there are different ways that classifiers handle partial features sets. Generative models handle missing features naturally by first integrating out missing values. However, for tabular datasets as we consider here, we found the best performance using distribution-based imputation for random forests in which the possible assignments of missing values are weighted proportionally [33].

Second, to determine which feature to select next, we need a method that estimates the cost-effectiveness of each of the unselected features based on the features we have already selected. For simplicity, we use a method that maximizes the expected utility of a feature, where the utility function is based on the expected increase in the absolute difference between the estimated class probabilities of the two most likely classes [21]. However, we emphasize that our framework is agnostic to the specific model and acquisition strategy that is used. Hence, simpler strategies such as population-level feature selection methods (e.g. LASSO) would yield similar results.

Third, to determine when to stop selecting additional features, most prior work assumes some given feature budget per individual such that the decision maker is tasked only with selecting the most cost-effective features within that budget [24]. The work that is most similar to ours develops an optimization framework that is used to find an information budget for each population subgroup such that an AFA classifier achieves parity in false-positive or false-negative rates [29]. Notably, by using the information budget as an additional degree of freedom during optimization, they show that several statistical (group) notions of fairness can be achieved in an AFA setting. Where their method accommodates different information budgets according to the needs of each population subgroup, our method adapts the information budget dynamically to the needs of each individual. Our work thus provides a novel method aimed at individual fairness which, in turn, gives rise to group fairness.

Finally, a more recent method considers all three AFA components jointly in one reinforcement learning framework, trading off the cost of each feature with the expected decrease in loss [35]. Bakker et al. [2] extend this framework by adding an adversarial loss to force the agent to acquire feature sets from which one can only predict the label but not the sensitive attribute. In this way, they guarantee demographic parity, a group fairness measure, over the protected subgroups split by the sensitive attribute. This work is the first that to address and improve individual fairness in an AFA setting.

# 3 DATA COLLECTION AND BROADER IMPACTS

The confidence thresholds introduced in this work help mitigate individual error disparity across individuals. However, we note that careful consideration should be given to the potential privacy implications on individuals. If an individual belongs to a population subgroup for which the classifier faces more uncertainty (for example because of underrepresentation in the training data), more features will be collected to reach the same expected error rate. In practice, this could result in an increased privacy burden on minority communities who are often already victim of over-policing [13]. Although this is an important concern that needs careful consideration when applying confidence thresholds in practice, we note that some form of data collection is often unavoidable.

Disparities in error rates are often caused by skewed sample sizes or a selection of features that are more predictive for some population subgroups than others [7]. Although typical approaches like restricting the model class, pre-processing the data or post-processing predictions can improve fairness without collecting

additional data, these strategies do not directly address the underlying causes but instead achieve fairness by harming predictive accuracy [10]. In high-stakes domains like criminal justice, health care diagnosis, loan approval, or hiring, the reduction in accuracy that is introduced by these methods is hard to justify and can even be considered unethical. Chen et al. [7] highlight this issue and propose collecting additional training examples to close the disparity gap between groups that are naturally over- and underpresented in the data. Though effective, this approach only improves group-level error rate disparities, measured across protected groups, and thus fails to give any individual-level guarantees. Moreover, this approach inevitably puts a increased privacy burden on the group for which more data is collected. And, in contrast to the prediction-time data collection method that we propose, this burden is not on the individuals that are subject to the decision, but on the group as a whole as more members of the group will have to contribute their data for training.[1]

Even though a trade-off between privacy and fairness (or between accuracy and fairness) seems inevitable, there are measures that could alleviate some of these concerns in practice. Institutional controls and privacy laws such as HIPAA and FERPA can stipulate how information should be maintained and limit the access of data to properly authorized individuals [11]. A more cryptographic approach could be to not release the data to decision maker but instead have a third-party trusted entity that stores and processes the data, and subsequently communicates only the decision to the decision maker. Other approaches that build on this paradigm include privacy preserving computational techniques such as fully homomorphic encryption and secure multiparty computation [1, 39]. Note that these approaches can only mitigate some of the privacy risks and do not directly address the issue of potentially excessive data collection.

Finally, under one interpretation, our proposed method for achieving individual error parity actually has a *positive* impact on the privacy of individuals. While it may appear unreasonable to require more information for some people, and we emphasize the need for transparent discussion and debate with stakeholders given important concerns about selective surveillance, we are in fact only collecting the smallest possible set of features to reach a desirable level of confidence — in contrast to methods that necessitate all features to be collected before making a prediction. Therefore we believe it follows the 'need-to-know' or 'data minimization' principle expressed in Article 5(1)(c) of the EU's General Data Protection Regulation (GDPR) which provides that personal data shall be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed".[2] Though GDPR does not clearly define "adequate, relevant and limited", the UK regulator, the ICO, stipulates that "you should identify the minimum amount of personal data you need to fulfil your purpose" and that "you may need to consider this separately for each individual".[3] We argue that confidence thresholds present a natural way to operationalize data minimization. When assessing creditworthiness, for example, a bank or regulator might set the acceptable error rate to 1%. In practice, the bank would thus collect information up to reaching this expected error rate after which any further collection could be considered excessive.

[1] A controversial example of a training-time data collection effort to mitigate bias made the news recently when a Google contractor targeted dark-skinned homeless people in Atlanta to gather more facial data in an attempt to improve Google's facial recognition algorithm, https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/.

[2] https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04

[3] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation/

## 4 PROBLEM SETUP

Let $(\mathbf{x}^{(i)}, y^{(i)}) \sim P$ be an individual $i$ in $P$ represented by a $d$-dimensional feature set $\mathbf{x}^{(i)}$ and a binary label $y^{(i)} \in \{0, 1\}$. In the AFA setting, we acquire the features in sequential order starting with the empty set $O_0^{(i)} := \emptyset$ at time $t = 0$.[4] At every later timestep $t$ we choose a feature from the unselected set of features, $j_t^{(i)} \subseteq \{1, \ldots, d\} \setminus O_{t-1}^{(i)}$ and examine the value of $j_t^{(i)}$ at a cost $c_j$. After each new acquisition step, the classifier will have access to features $O_t^{(i)} := j_t^{(i)} \cup O_{t-1}^{(i)}$. We keep acquiring features up to time $T^{(i)}$ when we meet a stopping criterion or collect the last available feature. At that point, we will classify $\mathbf{x}^{(i)}$ using only the set of features in $O_T^{(i)}$, $\{x_j\}_{j \in O_T^{(i)}}$. The cost vector $\mathbf{c}$ is equal for every individual in $P$ and can represent different types of costs such as monetary or privacy costs. The final set of features for individual $i$, $O_T^{(i)}$, is determined by two factors. First, we compute the expected utility of each feature to determine, for each individual and at each timestep, which feature should be selected to balance costs and expected increase in accuracy [21]. Second, to mitigate unfairness, we introduce a set of confidence thresholds that determine when to stop selecting features.

## 5 INDIVIDUAL ERROR PARITY

For classification, let $h : \mathbb{R}^d \to [0, 1]$ be a binary classifier that determines, given a partial feature set $O_t^{(i)}$, the probability that individual $i$ belongs to the positive class, $h(O_t^{(i)}) = P\left[y = 1 \mid \{x_j\}_{j \in O_t^{(i)}}\right]$. We omit the index $i$ and denote $\{x_j\}_{j \in O_t^{(i)}}$ as $O_t^{(i)}$ when it is clear from context. Similarly, $1 - h(O_t)$ corresponds to the probability that individual $i$ belongs to the negative class.

Definition 1. *At prediction time, each decision carries a finite risk $r_{err}(O_T^{(i)})$, defined as the expectation that the prediction affecting individual $i$ is erroneous given a probabilistic classifier $h$ and a set of features $O_T^{(i)}$.*

For a well-calibrated probabilistic classifier $h$, we can derive the following equation for this risk

$$\begin{aligned} r_{err}(O_T) &= \mathbb{E}_{(\mathbf{x},y)\sim P}\left[|h(O_T) - y|\right] \\ &= \int_0^1 p\, \mathrm{P}_{(\mathbf{x},y)\sim P}\left[h(O_\mathrm{T}) = p \mid y = 0\right] \mathrm{P}_{(\mathbf{x},y)\sim P}\left[y = 0\right] \\ &\quad + (1-p)\, \mathrm{P}_{(\mathbf{x},y)\sim P}\left[h(O_\mathrm{T}) = p \mid y = 1\right] \mathrm{P}_{(\mathbf{x},y)\sim P}\left[y = 1\right] dp, \end{aligned}$$

[4] We assume an empty set at the start but note that a decision maker could instead have access to an initial set of features that is the same for all individuals $O_0^{(i)}$. This does not change the further feature acquisition process.

where we drop the index $(i)$ for brevity. We can now use Bayes rule to find

$$r_{err}(O_T) = \int_0^1 \big(p\,\mathrm{P}_{(\mathbf{x},y)\sim P}\left[y=0 \mid h(O_\mathrm{T})=p\right] + (1-p)\,\mathrm{P}_{(\mathbf{x},y)\sim P}\left[y=1 \mid h(O_\mathrm{T})=p\right]\big)\,\mathrm{P}_{(\mathbf{x},y)\sim P}\left[h(O_\mathrm{T})=p\right]dp.$$

Substituting $\mathrm{P}_{(\mathbf{x},y)\sim P}\left[y=0 \mid h(O_\mathrm{T})=p\right] = 1-p$ and $\mathrm{P}_{(\mathbf{x},y)\sim P}\left[y=1 \mid h(O_\mathrm{T})=p\right] = p$ results in[5]

$$\begin{aligned} r_{err}(O_t) &= \int_0^1 2\left(p-p^2\right)\mathrm{P}_{(\mathbf{x},y)\sim P}\left[h(O_\mathrm{T})=p\right]dp \\ &= 2\,\mathbb{E}_{(\mathbf{x},y)\sim P}\left[h(O_\mathrm{T}) - h(O_\mathrm{T})^2\right]. \end{aligned} \tag{1}$$

Using the risk $r_{err}$ we can now define a corresponding individual fairness definition:

Definition 2. *Individual error parity across a population $P$ where each individual $i$ in $P$ is represented by a set of features $O_T$ requires $r_{err}(O_T^{(i)}) = r_{err}(O_T^{(j)}) \quad \forall i,j \in P$ .*

In practice, however, exact risk equality is hard to enforce so we follow Speicher et al. [36] and measure the degree of inequality in risk among individuals within a population using half the squared coefficient of variation.[6] This inequality measure is part of a family of inequality indices called the generalized entropy indices.

Definition 3. *The amount of individual unfairness within a population $P$ is defined as*

$$\mathcal{E}_P(\mathbf{r}_{err}) = \frac{1}{2|P|}\sum_{i\in P}\left[\left(\frac{r_{err}^{(i)}}{\bar{r}_{err,P}}\right)^2 - 1\right],$$

*where $r_{err}^{(i)}$ is the risk received by individual $i \in P$ and $\bar{r}_{err,P} = \frac{1}{N}\sum_{i\in P}^{N} r_{err}^{(i)}$ is the average risk across all individuals in $P$. Perfect individual fairness, when the risk is equal for all individuals, corresponds to $\mathcal{E}_P(\mathbf{r}_{err}) = 0$.*

Importantly, many of the axioms that are satisfied by generalized entropy indices are also appealing properties for fairness measures:

- *zero-normalization* – the inequality is zero when every individual receives the same level of risk.
- *anonymity* – the inequality depends only on the risk of the individuals and not on other characteristics.
- *population invariance* – the inequality is independent of the size of the population.
- *transfer principle* – transferring risk high-risk to a low-risk individual decreases inequality.

## 5.1 Connection to group fairness definitions

We connect equal error parity to four corresponding group fairness definitions: overall accuracy equality (equal error rates), predictive equality (equal false-positive rates), equal opportunity (equal false-negative rates), and equal odds (equal false-positive and false-negative rates). To make this connection, we first assume a set of disjoint subgroups $G_a$ in our population $P$ with $a \in \mathcal{A}$, which, for example, represent subgroups split by race or gender. Generally, these subgroups can have different base rates $\mu_a$, defined as the probability of belonging to the positive class $\mu_a = P[y=1 \mid A=a]$. To compute the errors of our probabilistic classifier with respect to the groups, we use the generalized rates introduces by Pleiss et al. [32].

Definition 4. *The generalized false-positive rate for a classifier $h$ computed over a group $G_a$ is $c_{fp,a}(h) = \mathbb{E}_{(\mathbf{x},y)\sim G_a}[h(O_T) \mid y=0]$. The generalized false-negative rate is $c_{fn,a}(h) = \mathbb{E}_{(\mathbf{x},y)\sim G_a}[1 - h(O_T) \mid y=1]$. The generalized error rate is equivalent to the $L_1$ loss $c_{err,a}(h) = \mathbb{E}_{(\mathbf{x},y)\sim G_a}[|y - h(O_T)|]$.*

We use probabilistic classifiers and thus generalized rates because it helps decision makers interpret the predictions. If, for an individual, the confidence of the classifier is low, it will be reflected in the final probability and a decision maker can thus choose to reject the decision or collect more information. Moreover, if the classifier would output binary predictions $h \in \{0,1\}$ instead of probabilities, these rates would simply represent standard false-positive rates, false-negative rates, and the zero-one loss. Similarly, we use generalized notions of equal accuracy, equal opportunity, and predictive equality for probabilistic classifiers:

Definition 5. *Equal accuracy for a set of probabilistic classifiers $h_1$ and $h_2$ for groups $G_1$ and $G_2$ requires $c_{err,1}(h) = c_{err,2}(h)$. Similarly, predictive equality requires $c_{fp,1}(h) = c_{fp,2}(h)$, equal opportunity $c_{fn,1}(h) = c_{fn,2}(h)$ and equal odds $c_{fp,1}(h) = c_{fp,2}(h)$ and $c_{fn,1}(h) = c_{fn,2}(h)$.*

Exact equality is hard to enforce in practice so we study the degree to which these constraints are violated: $|c_{fp,1} - c_{fp,2}|$, $|c_{fn,1} - c_{fn,2}|$, $|c_{err,1} - c_{err,2}|$. Further, for probabilistic classifiers, we require the probabilities to be calibrated, $P_{(\mathbf{x},y)\sim G_a}[y=1 \mid h(O_t)=p] = p$.

We can make two important connections between individual error parity and these group fairness definitions:

Proposition 1. *Individual error parity across a population $P$ necessarily yields equal accuracy across groups $G_a \subseteq P$.*

Proof. When individual error parity is satisfied, the risk for each member of the population will be equal to the risk of each other member of the population, irrespective of group membership (see Definition 2). If we assume that this risk has some finite value $r_{err} = \beta_{err}$ then the expected error for each group in the population will also be equal to this risk $\beta_{err}$. Hence, each group will have an equal error rate, resulting in equal accuracy across groups

$$c_{err,a}(h(O_\mathrm{T})) = \mathbb{E}_{\mathrm{G_a}}\left[r_{err}(h(O_\mathrm{T}))\right] = \beta_{err} \quad \forall a \in \mathcal{A}. \tag{2}$$

where we write $\mathbb{E}_{(\mathbf{x},y)\sim G_a}$ as $\mathbb{E}_{G_a}$ for brevity. □

Proposition 2. *Individual error parity implies equal false-positive and false-negative rates (equal odds) across all groups that have an equal base rate $\mu_a$.*

[5] Here, we assume perfect calibration. The derivation for approximate calibration can be found in the Appendix.

[6] [36] measures inequality in terms of 'benefits' but in our case, it makes more sense to share expected errors or 'risk' fairly.

Proof. Similar to Proposition 1, we assume an equal risk across the population of $r_{err} = \beta_{err}$. We can then derive from Definition 4

$$\begin{aligned} c_{fp,a}(h) &= \mathbb{E}_{G_a}\left[h(O_T) \mid y = 0\right] \\ &= \int_0^1 p \; \mathrm{P}_{G_a}\left[h(O_T) = p \mid y = 0\right] dp \\ &= \int_0^1 p \; \frac{1 - \mathrm{P}_{G_a}\left[y = 1 \mid h(O_T) = p\right]}{1 - \mathrm{P}_{G_a}\left[y = 1\right]} \mathrm{P}_{G_a}\left[h(O_T) = p\right] dp. \end{aligned}$$

Again, using $\mathrm{P}_{G_a}\left[y = 1 \mid h(O_T) = p\right] = p$ and $\mathrm{P}_{G_a}\left[y = 1\right] = \mu_a$, we can rewrite this as

$$\begin{aligned} c_{fp,a}(h) &= \frac{1}{1-\mu_a} \int_0^1 p(1-p) \, \mathrm{P}_{G_a}\left[h(O_T) = p\right] dp \\ &= \frac{1}{1-\mu_a} \mathbb{E}_{G_a}\left[h(O_T) - h(O_T)^2\right] \\ &= \frac{1}{2(1-\mu_a)} \mathbb{E}_{G_a}\left[r_{err}(h(O_T))\right]. \end{aligned} \tag{3}$$

Following the same steps for the false-negative rate, we find

$$\begin{aligned} c_{fn,a}(h) &= \frac{1}{\mu_a} \mathbb{E}_{G_a}\left[h(O_T) - h(O_T)^2\right] \\ &= \frac{1}{2\mu_a} \mathbb{E}_{G_a}\left[r_{err}(h(O_T))\right]. \end{aligned} \tag{4}$$

If the base rate $\mu_a$ is equal across groups and $r_{err}^{(i)} = \beta_{err} \quad \forall i \in P$,

$$c_{fp,a}(h(O_T)) = \frac{\beta_{err}}{2(1-\mu_a)} \quad \forall a \in \mathcal{A},$$

$$c_{fn,a}(h(O_T)) = \frac{\beta_{err}}{\mu_a} \quad \forall a \in \mathcal{A},$$

which yields equal-false positive and equal false-negative rates across groups. If equal odds is satisfied, predictive equality and equal opportunity must also be satisfied. □

In contrast to $c_{err}$, $c_{fp}$ and $c_{fn}$ now depend on the group-specific base rate, and individual error parity therefore only implies equal odds (equal false-positive and equal false-negative rates) when the base rates are the same across group. However, an error rate disparity can still be present when the base rates are different if unfairness is caused only by a difference in variance across groups, for example, because of different sample sizes $|G_a|$ or because of a difference in group-conditional feature variance $\mathrm{Var}(\mathbf{x} \mid a)$. In that case the base rates across group will be equal, leading to attaining individual error parity, as well as equal odds.

# 6 CONFIDENCE THRESHOLDS

Intuitively, the stopping criteria should be chosen such that we collect more features for individuals and groups for which the model is less certain. By stopping later, we have more predictive power, and thus decrease the expected error rate. In this section, we present confidence thresholds as a way to achieve approximate individual error parity and will derive corresponding upper and lower thresholds $\alpha_u$ and $\alpha_l$. The upper threshold corresponds to predicting $\hat{y} = 1$ with probability $\alpha_u$ while the lower threshold corresponds to predicting $\hat{y} = 0$ with probability $1 - \alpha_l$. We reach these thresholds by sequentially adding features one-by-one, slowly increasing confidence (moving the probabilistic estimate towards either of the thresholds). We stop collecting features at time step $T$ when the probability reaches either one of the thresholds, $h(O_T) \geq \alpha_u$ or $h(O_T) \leq \alpha_l$.

## 6.1 Individual error parity and overall accuracy equality

Our aim is to find the thresholds $\alpha_u$ and $\alpha_l$ that ensure $r_{err}^{(i)} = \beta_{err} \quad \forall i \in P$. When the risk is equal across all individuals we obtain individual error parity and thus also equal accuracy across groups. During the feature selection process, each individual reaches either the upper threshold or the lower threshold first. We consider potential overshooting later in this section but, for now, assume that we can stop the process at exactly the upper or lower threshold such that

$$h(O_T) = \alpha_u \quad \text{or} \quad h(O_T) = \alpha_l.$$

We then use Equation (1) to find, respectively, for each individual

$$r_{err} = 2(\alpha_u - \alpha_u^2) \quad \text{or} \quad r_{err} = 2(\alpha_l - \alpha_l^2).$$

To ensure $r_{err} = 2\left(h(O_T) - h(O_T)^2\right) = \beta_{err}$ in both cases, the solution for the confidence thresholds follows as

$$\alpha_u = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 2\beta_{err}}, \quad \alpha_l = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 2\beta_{err}}. \tag{5}$$

If, for every individual, we acquire features one-by-one until we reach either of these thresholds, we attain a generalized error rate $\beta_{err}$ in expectation for each individual, and across groups, achieving overall accuracy equality. Importantly, these thresholds are independent of the group label $a$ and thus guarantee group and individual fairness with respect to arbitrary subgroups, even when the subgroup labels are unknown at training and prediction time.

## 6.2 Equal false-positive or false-negative rates

When the desired measure of group fairness is equal false-positive rates (predictive equality) or equal false-negative rates (equal opportunity), the thresholds derived for individual error parity will only suffice if each group has an equal base rate $\mu_a$ (see Proposition 2). In contrast, choosing the same thresholds for individuals in groups with different base rates will result in different false-positive and false-negative rates across these groups. To address this, we can derive a new set of group-specific thresholds that account for these differences in base rate. Following Equations (3) and (4), we find

$$\begin{aligned} c_{fp}(h_a) &= \frac{1}{1-\mu_a} \mathbb{E}_{G_a}\left[h_a(O_T) - h_a(O_T)^2\right] \\ c_{fn}(h_a) &= \frac{1}{\mu_a} \mathbb{E}_{G_a}\left[h_a(O_T) - h_a(O_T)^2\right]. \end{aligned}$$

Moreover, in line with Definition 1, we can define a false-positive and false-negative risk $r_{fp}$ and $r_{fn}$, corresponding to the probability that a prediction for an individual in group $G_a$ is a false-positive or a false-negative

$$\begin{aligned} r_{fp}(h(O_T)) &= \frac{1}{1-\mu_a} \mathbb{E}_{G_a}\left[(h(O_T) - h(O_T)^2\right], \\ r_{fn}(h(O_T)) &= \frac{1}{\mu_a} \mathbb{E}_{G_a}\left[h(O_T) - h(O_T)^2\right]. \end{aligned}$$

Following a similar derivation as for equal error rates, we define a target false-positive rate $\beta_{fp}$ to find the stopping criteria for each

group such that $r_{fp}(h(O_T^{(i)})) = \beta_{fp}, \forall i \in P$. In turn, this implies $c_{fp}(h(O_T)) = \mathbb{E}_{G_a}\left[r_{fp}(h(O_T))\right] = \beta_{fp}, \forall a \in \mathcal{A}$ and thus equal false-positive rates across group. To ensure $\beta_{fp}$ for individuals that stop at either one of the thresholds, we find a set of stopping criteria for equalizing false-positive rates, analogously to those for $c_{err}$,

$$\alpha_u = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\beta_{fp}(1-\mu_a)}, \tag{6}$$

$$\alpha_l = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\beta_{fp}(1-\mu_a)}. \tag{7}$$

For false-negative rates we find a similar set of stopping criteria for a target false-negative rate $\beta_{fn}$,

$$\alpha_u = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\beta_{fn}\mu_a}, \quad \alpha_l = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\beta_{fn}\mu_a}. \tag{8}$$

Even though equal false-positive or false-negative rates can be achieved using these thresholds at a group level, this approach could raise some concerns when used in practice. In contrast to the thresholds for $c_{err}$, the thresholds for $c_{fp}$ and $c_{fn}$ now depend on the group-specific base rate, precluding individual error parity across the full population and requiring different confidence thresholds across groups which could be perceived as unfair. Nonetheless, within each subgroup, we still set the same thresholds and we can thus achieve within-group individual error parity. Though not equivalent to population-wide individual error parity, it does mitigate potential intra-group unfairness. The latter is a concern when one imposes group fairness constraint with typical methods like constraint optimization or post-processing predictions (see Kearns et al. [23] for examples). In that case, a solution could be group fair but simultaneously exhibit strong error rate differences within groups, a problem that within-group individual error parity would address.

### 6.3 Sources of residual unfairness

If the probabilities for each individual are calibrated and each individual's probabilistic estimate $h(O_T)$ stops exactly at the upper or lower threshold, we can guarantee exact individual error parity. In practice, however, there will still be some residual unfairness. First, in real-world datasets, there will be a non-zero classification error even when all features are collected such that, for some individuals, we will not reach thresholds close to 0 or 1 even with unlimited budget for feature acquisition. Not reaching the threshold will lead to individual error disparity, as part of the estimates will fall between the thresholds ($\alpha_l \leq h(O_T) \leq \alpha_u$) and potentially also to group unfairness when one subgroup has a disproportionate number of members that do not reach the thresholds. Hence, decision makers should ensure there are sufficient predictive features for each individual to reach the desired threshold. Fortunately, AFA allows you to have much more diverse and personalized feature sets as features are only acquired if they are expected to be predictive for an individual.

Second, we achieve individual error parity among individuals that stopped when the probabilistic estimates after stopping are exactly $h(O_T) = \alpha_u$ or $h(O_T) = \alpha_l$. In practice, however, each estimate progresses in discrete jumps when a new feature is acquired and thus find that estimates 'overshoot' the thresholds $h(O_T) \geq \alpha_u$ or $h(O_T) \leq \alpha_l$. We refer to the Appendix for a detailed analysis of the residual unfairness due to residual error and overshooting.

Finally, we have assumed access to a well-calibrated classifier when defining our notion of risk. However, the standard notion of calibration is a property that strictly holds only on average across all predictions made by the classifier. The probabilities and the corresponding risks could therefore still be inaccurate with respect to a structured subgroup of individuals that is unknown at training time and hence not taken into account explicitly when calibrating. We analyze the effect of miscalibration on group fairness in the appendix but encourage future work analyzing the effect of miscalibration on individual error parity, as well as work that combines our method with individual calibration approaches that have recently been developed [18, 40].

### 6.4 Choosing a target error rate

The optimal strategy for picking a target error rate, and thus for picking a set of thresholds, will be strongly context dependent. In the context of the law, Kaplan [22] developed a decision theoretic framework for setting the probabilistic threshold that represents the standard of proof. Focusing on the avoidance of errors, they argue that a decision should be made to convict if the expected disutility of a decision to acquit is greater than the expected disutility of a decision to convict $PD_g > (1-P)D_i$ where $P$ is the probability that a defendant is guilty, $D_g$ the disutility of acquitting a guilty person and $D_i$ the disutility of convicting an innocent person. Extending Kaplan's framework for confidence thresholds, we can pick a target error rate by balancing the cost of a potential error with the cost spent on features. A bank could, for example, compute the average costs of making an erroneous loan decision and find the corresponding feature budget to minimize the overall costs. Alternatively, we could see scenarios, for example in criminal justice, where the target error rates are instead set by a regulatory body.

## 7 EXPERIMENTS

In this section, we show that our method mitigates individual and group unfairness while choosing the confidence thresholds to achieve individual error parity and overall accuracy equality, equal false-positive rates or equal false-negative rates. Experimental analysis of the residual unfairness and results for a second dataset can be found in the Appendix.

We measure the residual individual error disparity using half the squared coefficient of variation, $\mathcal{E}(\mathbf{r})$, where we compute the inequality across individuals using the risk functions $r_{err} = 2(h(O_T) - h(O_T)^2)$, $r_{fp} = \frac{1}{1-\mu_a}(h(O_T) - h(O_T)^2)$, and $r_{fn} = \frac{1}{\mu_a}(h(O_T) - h(O_T)^2)$. Group-level unfairness between two groups $G_1$ and $G_2$ is measured using the absolute difference in generalized error $|\Delta c_{err}| = |c_{err,1}) - c_{err,2}|$, $|\Delta c_{fp}| = |c_{fp,1} - c_{fp,2}|$ or $|\Delta c_{fn}| = |c_{fn,1} - c_{fn,2}|$ . As this is the first work concerned with individual fairness in this setting, we compare the accuracy-unfairness trade-offs in each experiment against an 'equal budget' benchmark. To compute this benchmark, we use the exact same model and feature acquisition strategy but, instead of using our stopping criteria to redistribute budget, we now equally distribute the feature budget across individuals. A second 'all features' benchmark represents the point on the equal budget accuracy-unfairness trade-off when all features are queried for all individuals.

**Table 1: Overview of the datasets. Accuracy is computed on a dataset-level using the full feature set, while $\mu$ is the dataset-level base-rate $P(y = 1)$. For each subgroup $n_a$ is the relative number of individuals and $\mu_a$ the base rate.**

| Dataset | | | | | Subgroup$_1$ | | | Subgroup$_0$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Name | N$_{samples}$ | N$_{feat}$ | Acc | $\mu$ | Label$_1$ | n$_1$ | $\mu_1$ | Label$_0$ | n$_0$ | $\mu_0$ |
| Mexican poverty | 70,305 | 99 | 75.9% | 35.5% | Urban | 63.6% | 34.9% | Rural | 36.4% | 36.6% |
| Adult income | 48,842 | 98 | 85.1% | 23.9% | White | 85.4% | 25.4% | Non-white | 14.6% | 15.3% |

## 7.1 Implementation

In addition to the confidence-based stopping criteria, implementation requires two more elements: a model and a feature acquisition strategy. We refer to the Appendix for further implementation details but provide a summary in this section. First, we need a model that allows us to estimate $P(y \mid O_t)$ for arbitrary feature subsets. While this is easier with generative models like Naive Bayes, we use distribution-based imputation in random forest as a random forest model has superior predictive performance (accuracy of 75.9% using the full feature set on the Mexican Poverty dataset set versus 73.4% for Naive Bayes) [33]. We use the full feature set at training-time but note that there are methods for training with partial feature sets [16, 24].

Second, we implement a feature acquisition strategy to estimate which next feature should be selected based on the expected utility for of each feature given the current feature set $O_t$ [21]. In contrast to population-level feature selection methods like LASSO, this strategy allows us to pick a next feature on an individual level. However, we note that confidence thresholds are agnostic to the model and the feature acquisition strategy.

The cost for each feature $c_j$ can be different and can represent for example monetary or privacy costs. To make the results more interpretable, we choose the costs to be the same for each feature $c_j = 1$. Changing these costs to make them more realistic will only lead to a different ordering of features and will not further impact the results. Finally, to account for overshooting, we use a validation set to learn a mapping function that maps the confidence thresholds to target error rates $\beta_{err}$, $\beta_{fp}$, and $\beta_{fn}$. We sweep a range of $\alpha_u$ (and $\alpha_l = 1 - \alpha_u$), while measuring the empirical error rates for individuals that reached either of the thresholds. At prediction time, when we are given a target rate $\beta$, we use the inverse of this function to find the optimal set of thresholds.

## 7.2 Datasets

An overview of the datasets is given in Table 1 while preprocessing details can be found in the Appendix. The Mexican Poverty dataset is a household survey used in [29] for fair feature selection, motivated by a real-world example of fair distribution of social programs where there are insufficient resources to gather all information for each individual [30]. The dataset contains a series of household-level features and binary poverty levels for prediction. The Adult income dataset is a standard benchmark for fair classification [26], comprising demographic and occupational attributes, with the goal to classify a person's income as above $50,000 or not. We use 5-fold cross-validation with random 60%/20%/20% train/validation/test splits.

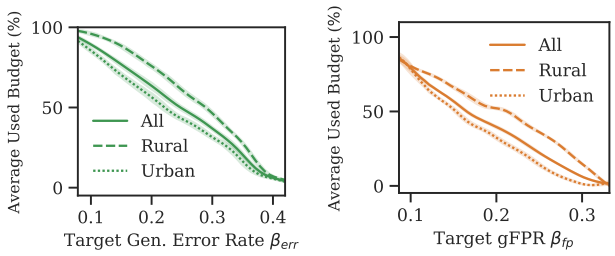
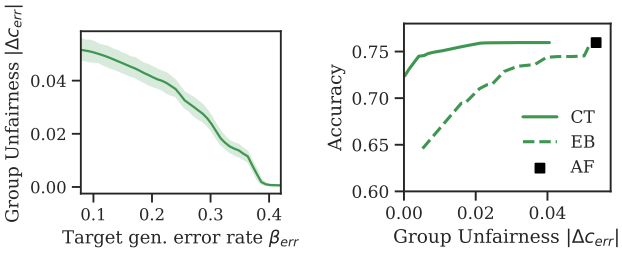



**Figure 1: Average used feature budget for the Rural and Urban subgroups, and across the full population ('All') in the Mexican Poverty dataset. We apply confidence thresholds for equalizing generalized error rates (left) and generalized false-positive rates (right). Curves and shaded regions are the average and 95% confidence intervals, computed using 5-fold cross validation. To mitigate group unfairness our method successfully assigns more budget to the subgroup for which the classifier is less certain.**

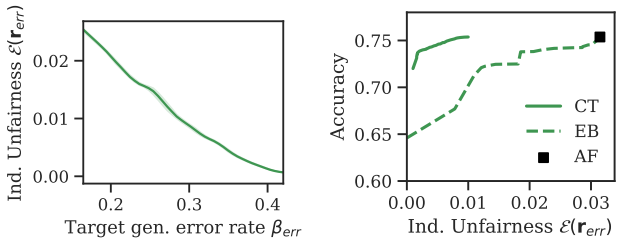
## 7.3 Individual error disparity and overall accuracy equality

We use the confidence thresholds derived in Equation (5) to mitigate individual error disparity and improve overall accuracy equality across groups. To ensure calibrated probabilities, we fit a sigmoid function to the classifier's probabilities using a validation set; a method known as Platt scaling [31]. Importantly, we calibrate across the entire population, effectively ignoring the underlying groups, to demonstrate that we can mitigate unfairness with respect to the generalized error rates without explicitly accounting for these subgroups. In the left panel in Figure 1, we show the average number of features collected per subgroup ('Rural' and 'Urban') and across the full population ('All') as the target error rate $\beta_{err}$ varies. Note that for larger values of $\beta_{err}$, when the acquisition process stops early as the confidence thresholds are closer to 0.5, only a few features are collected, while for smaller values of $\beta_{err}$ nearly all features are collected. Moreover, the average number of collected features differs across groups, demonstrating that our method assigns more budget to the group that is harder to classify without having knowledge of the underlying group structure (using the full feature set, the accuracy is 74.0% for the Rural subgroup and 76.9% for the Urban subgroup).

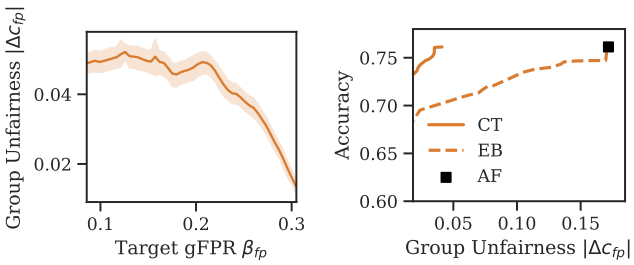
In practice, a decision maker would choose a single constant target error rate $\beta_{err}$. However, to investigate the behavior for different $\beta_{err}$, the left panels in Figures 2a and 2b show respectively the residual individual and group unfairness across the full range



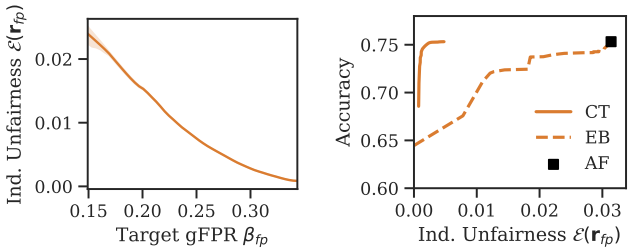

(a) Group unfairness for equalizing error rates.



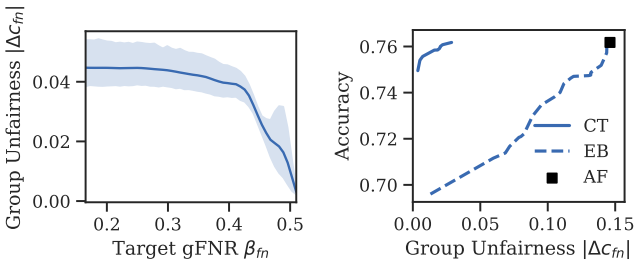

(b) Individual unfairness for equalizing error rates.



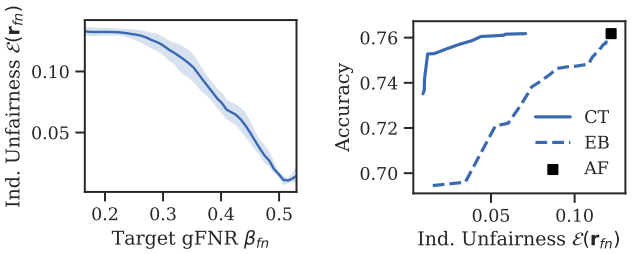

(c) Group unfairness for equalizing false-positive rates.




(d) Individual unfairness for equalizing false-positive rates.




(e) Group unfairness for equalizing gen. false-negative rates.




(f) Individual unfairness for equalizing gen. false-negative rates.

**Figure 2: Results for the Mexican Poverty dataset with group-level error disparity (Figures 2a, 2c and 2e) between the Urban ($a = 1$) and Rural ($a = 0$) subgroups and individual error disparity (Figures 2b, 2d and 2f) measured across the full population. The panel in the left of each subfigure measures the residual unfairness while sweeping a range of target rates $\beta_{err}$, $\beta_{fp}$ or $\beta_{fn}$. The curves and shaded confidence regions are the average and the 95% confidence intervals, computed using 5-fold cross-validation. The curves in the panel in the right of each subfigure are generated by sweeping $\beta_{err}$, $\beta_{fp}$ or $\beta_{fp}$, taking the average across the 5 folds, and computing the Pareto front. The accuracy-fairness trade-offs of our method (solid) Pareto dominate an equal budget (EB) baseline (dashed) as well as a baseline where we select all features (AF, black square).**

of target error rates $\beta_{err}$. For large values of $\beta_{err}$, when the confidence thresholds are close to 0.5, there are sufficient features for each individual to reach the thresholds and we effectively mitigate unfairness. For smaller values of $\beta_{err}$, when the information budget grows, there are an increasing number of individuals for which we exhaust all relevant features before we reach the confidence thresholds, limiting the effectiveness of the confidence thresholds. In deployment, a decision maker should thus avoid setting the target rate too low, or should abstain from making a final decision when the threshold is not reached and should instead reassess whether the feature set contains sufficient predictive features for each member of the population. Finally, in the right panels in Figures 2a and 2b, we observe that the accuracy-fairness trade-off of our method Pareto dominates the benchmark in terms of both individual error parity and overall accuracy equality.

## 7.4 Equal false-positive or false-negative rates

To mitigate the group and individual-level unfairness with respect to the expected false-positive and false-negative rates, we use the confidence thresholds derived in in Equations (6) to (8). Equalizing false-positive rates or false-negative rates necessitates access to the sensitive attribute because computing the confidence thresholds requires the group-specific base rates. As we now have access to the sensitive attribute, we calibrate the probabilities for each group separately, effectively creating separate classifiers for each group.

First, in the left frames of Figures 2c and 2d, we observe the residual unfairness with respect to the generalized false-positive rate and

in Figures 2e and 2f with respect to the generalized false-negative rate. Analogous to the results for equalizing error rates, we see the strongest reduction in unfairness for larger values of the target error rates $\beta_{fp}$ and $\beta_{fn}$. In that regime, the thresholds are closer to 0.5 and we thus have a sufficient number of features for each individual to reach the confidence thresholds. Finally, in the right frames of Figures 2c to 2f we show that, on a group and individual level, our confidence thresholds method Pareto dominates an equal budget baseline along the full accuracy-unfairness trade-off.

## 8 DISCUSSION AND CONCLUSION

We proposed individual error parity as an individual fairness notion in an AFA setting and related it to a set of commonly used group fairness notions. However, we argue that, more generally, even in settings that are not commonly seen a budget-constrained, measuring individual error disparity could help investigate how representative a chosen set of features is, and guide whether additional variables should be measured to foster more equitable decision making. We then introduced a framework for mitigating unfairness in this setting, addressing in a novel way both individual and group fairness concerns. The framework is straightforward and intuitive, and applies generally for any model that can handle partial feature sets and any feature acquisition strategy. The thresholds redistribute an information budget across individuals, allocating additional budget to those individuals for which the classifier faces most uncertainty.

On two public datasets, we demonstrated empirically that our method mitigates unfairness. Especially for larger target error rates, our framework strongly decreases disparities while for smaller rates we exhaust the relevant features before reaching the confidence thresholds. This issue also represents a limitation of the datasets which have not been collected for active feature acquisition, and features have thus been chosen to be cost-effective for the majority of individuals. In our framework, however, it is natural to add features that are relevant only to a handful of individuals. Hence, we encourage future work that investigates the applications of our framework to datasets and settings that meet this criterion. One interesting avenue for future research is to use individual error disparity as a guide for which variables should be measured during dataset creation. Rather than choosing features which are cost-effective for the majority of individuals, we imagine an iterative process in which individuals for which there are insufficient features are used to guide which additional variables should be measured.

We also encourage further research that investigates the implications on the privacy of individuals. Even though our method reduces error disparities and naturally follows the data minimization principle, it can actually create privacy disparities as for each individual a different set of features will be collected. Although we emphasize that some form of data collection is often unavoidable, we encourage future work in which the privacy burden of each feature is taken into account explicitly. A natural extension would then be to work towards a framework that holistically trades-off monetary costs for decision makers, privacy costs for decision subjects, and fairness. Finally, we encourage future work that furhter investigates the effects of miscalibration on individual error parity and mitigates these effects using methods that are aimed at more individual-level notions of calibration [18, 40].

## 9 ACKNOWLEDGMENTS

All authors would like to thank Humberto Riveron Valdes and Prasanna Sattigeri for useful discussions and contributions to earlier versions of this work. AW acknowledges support from a Turing AI Fellowship under grant EP/V025379/1, The Alan Turing Institute under EPSRC grant EP/N510129/1 and TU/B/000074, and the Leverhulme Trust via CFI.

## REFERENCES

[1] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. 2017. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security* 13, 5 (2017), 1333–1345.
[2] Michiel A Bakker, Duy Patrick Tu, Humberto Riverón Valdés, Krishna P Gummadi, Kush R Varshney, Adrian Weller, and Alex Pentland. 2019. DADI: Dynamic Discovery of Fair Information with Adversarial Reinforcement Learning. *arXiv preprint arXiv:1910.13983* (2019).
[3] Yahav Bechavod and Katrina Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* (2017).
[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.
[5] William Cai, Johann Gaebler, Nikhil Garg, and Sharad Goel. 2020. Fair Allocation through Selective Information Acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 22–28.
[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.
[7] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory?. In *Advances in Neural Information Processing Systems*. 3539.
[8] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. 134–148.
[9] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).
[10] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.
[11] John Davis and Osonde Osoba. 2016. Privacy Preservation in the Age of Big Data. *Available at SSRN 2944731* (2016).
[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.
[13] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
[14] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.
[15] Tianshi Gao and Daphne Koller. 2011. Active classification based on value of classifier. In *Advances in Neural Information Processing Systems*.
[16] Wenbo Gong, Sebastian Tschiatschek, Richard Turner, Sebastian Nowozin, and José Miguel Hernández-Lobato. 2019. Icebreaker: Element-wise Active Information Acquisition with Bayesian Deep Latent Gaussian Model. *arXiv preprint arXiv:1908.04537* (2019).
[17] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315.
[18] Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*. 1944–1953.
[19] Hock Lai Ho. 2015. The Legal Concept of Evidence. In *The Stanford Encyclopedia of Philosophy* (winter 2015 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.
[20] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333.
[21] Pallika Kanani and Prem Melville. 2008. Prediction-time active feature-value acquisition for cost-effective customer targeting. *Advances In Neural Information Processing Systems (NIPS)* (2008).

[22] John Kaplan. 1968. Decision Theory and the Factfinding Process. *Stanford Law Review* (1968), 1065–1092.

[23] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv* (2017).

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

[25] Yoad Lewenberg, Yoram Bachrach, Ulrich Paquet, and Jeffrey S Rosenschein. 2017. Knowing what to ask: A Bayesian active learning approach to the surveying problem. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[26] Moshe Lichman et al. 2013. UCI machine learning repository.

[27] Li-Ping Liu, Yang Yu, Yuan Jiang, and Zhi-Hua Zhou. 2008. TEFE: A time-efficient approach to feature extraction. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE.

[28] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. 2019. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *International Conference on Machine Learning*. 4234–4243.

[29] Alejandro Noriega-Campero, Michiel Bakker, Bernardo Garcia-Bulle, and Alex Pentland. 2019. Active Fairness in Algorithmic Decision Making. *Proceedings of AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society* (2019).

[30] Alejandro Noriega-Campero, Bernardo Garcia-Bulle, Luis Fernando Cantu, Michiel A Bakker, Luis Tejerina, and Alex Pentland. 2020. Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 241–251.

[31] John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

[32] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[33] Maytal Saar-Tsechansky and Foster Provost. 2007. Handling missing values when applying classification models. *Journal of machine learning research* 8, Jul (2007), 1623–1657.

[34] Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. 2019. Average Individual Fairness: Algorithms, Generalization and Experiments. In *Advances in Neural Information Processing Systems*. 8242–8251.

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

[36] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2239–2248.

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

[38] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics*. 962–970.

[39] Chuan Zhao, Shengnan Zhao, Minghao Zhao, Zhenxiang Chen, Chong-Zhi Gao, Hongwei Li, and Yu-an Tan. 2019. Secure multi-party computation: theory, practice and applications. *Information Sciences* 476 (2019), 357–372.

[40] Shengjia Zhao, Tengyu Ma, and Stefano Ermon. 2020. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*. PMLR, 11387–11397.