Algorithmic Targeting of Social Policies: Fairness, Accuracy, and Distributed Governance

Alejandro Noriega-Campero
Prosperia Labs
Cambridge, MA
anc@prosperia.ai

Bernardo Garcia-Bulle
MIT
Cambridge, MA
bernard0@mit.edu

Luis Fernando Cantu
ITAM
Mexico City, Mexico
lcantudi@itam.mx

Michiel A. Bakker
MIT
Cambridge, MA
bakker@mit.edu

Luis Tejerina
IADB
Washington, DC
luist@iadb.org

Alex Pentland
MIT
Cambridge, MA
pentland@mit.edu

ABSTRACT

Targeted social policies are the main strategy for poverty alleviation across the developing world. These include targeted cash transfers (CTs), as well as targeted subsidies in health, education, housing, energy, childcare, and others. Due to the scale, diversity, and widespread relevance of targeted social policies like CTs, the algorithmic rules that decide who is eligible to benefit from them—and who is not—are among the most important algorithms operating in the world today. Here we report on a year-long engagement towards improving social targeting systems in a couple of developing countries. We demonstrate that a shift towards the use of AI methods in poverty-based targeting can substantially increase accuracy, extending the coverage of the poor by nearly a million people in two countries, without increasing expenditure. However, we also show that, absent explicit parity constraints, both status quo and AI-based systems induce disparities across population subgroups. Moreover, based on qualitative interviews with local social institutions, we find a lack of consensus on normative standards for prioritization and fairness criteria. Hence, we close by proposing a decision-support platform for distributed governance, which enables a diversity of institutions to customize the use of AI-based insights into their targeting decisions.

CCS CONCEPTS

• **Applied computing → Law, social and behavioral sciences**;
• **Computing methodologies → Artificial intelligence**;

KEYWORDS

AI for social good, algorithmic fairness, targeted social programs, proxy means tests, cash transfers

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3375784

ACM Reference Format:
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 *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3351095.3375784

1 WORLDWIDE RELEVANCE OF TARGETED SOCIAL POLICIES — AND THEIR UNDERLYING ALGORITHMS

Algorithmic decision-making systems (ADS) have become increasingly ubiquitous—e.g., in criminal justice [30], medical diagnosis and treatment [29], human resource management [8], social work [18], credit [22], and insurance [32]. Although there is widespread excitement for the potential societal benefit that this type of technology can bring, there is also commensurate concern about how it can deepen social inequalities and systematize discrimination [38, 40]. Consequently, substantial work has surged in recent years, on conducting fairness audits of deployed systems, as well as on defining and optimizing for algorithmic fairness. Notably, however, the vast majority is focused on developed-world contexts: online domains such as targeted advertising [48], search engines [19], and facial recognition [4]; and offline domains such as criminal justice [3, 12], child maltreatment [13], and predictive policing [46].

Algorithmic targeting of social policies. The present work focuses on targeted social development policies, i.e.: policies that promote social development, and target only a subset of the population—most commonly, populations in poverty. Since two decades ago, algorithmic rules underlie the targeting decisions of a large fraction of social policies in the developing world, e.g., poverty prediction algorithms that precondition eligibility to cash transfer programs [16, 20, 23]. Here, we argue that the algorithms that directly influence or determine critical decisions regarding who benefits from targeted social policies—and who doesn't—at a large scale and across the global south, are among the algorithms of paramount importance operating in the world today.

Diversity. Targeted social policies constitute a major vehicle for fighting poverty and redistributing wealth across the developing world [2, 20]. In most countries, targeted programs run by governments and NGOs proliferate, touching every corner of social

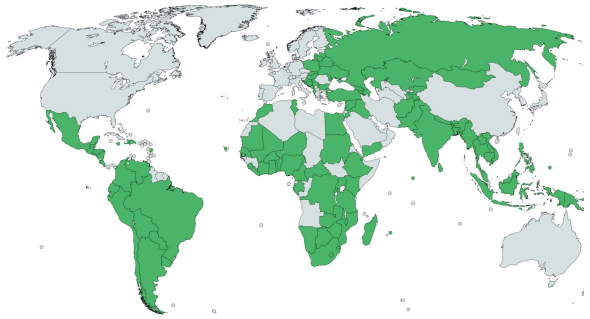


Figure 1: Global presence of national targeted cash transfer programs [2].

development: from cash transfer programs, to scholarship programs, subsidized health care systems, targeted housing and energy subsidies, targeted childcare, food security, retirement pension programs, targeted microloans, and others. Moreover, international aid agencies such as USAID require their funded institutions to guide the selection of beneficiaries based on poverty assessment tools [7], in an effort to increase effectiveness and prevent the misuse of funds.

Scale and Spread. To gauge the scale and global spread of targeted policies, consider the case of cash transfer programs (CTs). CTs provide a financial stipend to families in poverty, and are often conditional, i.e., requiring beneficiaries to comply with "co-responsibilities", such as keeping their children in school, and attending regular medical appointments [14, 16]. More than 110 countries worldwide have implemented national CTs (see a map in Figure 1) [2], playing a central role in the countries' poverty alleviation and wealth redistribution strategies. Only in Mexico and Brazil, for example, the national CTs combined reach more than 80 million people (roughly 25% of the population) [23].

The present work is motivated by the scale, diversity, and widespread relevance of targeted social policies like CTs. It reports on comprehensive engagements with two Latin American countries, towards the development and adoption of more fair, accurate, and transparent targeting systems.

2 THE ACCURACY EDGE: LARGE-SCALE IMPACT OF AI-BASED TARGETING

2.1 Algorithmic Targeting of Social Policies

Most social policies, like CTs, are targeted at the poor. However, in the developing world, reliable income data is typically not available and costly to procure, because the majority of people work in informal economic markets [26]. Hence, targeting most commonly relies on algorithms that estimate households' poverty based on observable and less costly proxy data, such as education levels, demographics, and their assets and services [7, 16, 20, 23, 27]. In most countries, these algorithms are trained to estimate poverty based on large, periodic, statistically representative household surveys, which collect both the proxy features and the income ground truth. In their daily operations, institutions cannot collect ground truth income data directly from potential beneficiaries, due to the cost of eliciting trustworthy income data at such massive scale, as well as candidates' strong incentives for under-reporting [16, 20, 23].

In practice, predictive algorithms are imperfect, leading to targeting errors. In Latin America and other regions, it is estimated that targeting systems incur more than 25% exclusion errors and inclusion errors (undercoverage and leakage) [23, 34]. However, the methods used during the past two decades for estimating poverty have relied on econometric approaches that are not optimized for out-of-sample prediction, and cannot leverage the many non-linear relationships that are typically found in high-dimensional data. Hence, we hypothesize that substantial accuracy gains could come from the use of modern computational and statistical methods borrowed from the field of artificial intelligence.

Related work. Closest in spirit to the work presented in this section, is the work of Mcbride et. al 2016 [34]. The authors there show that out-of-sample estimation can reduce exclusion errors, but increase inclusion errors, compared to in-sample linear methods traditionally used in poverty prediction tools. However, they found no consistent advantage from using random forest classifiers over linear methods. Our results in this section partially contradict results in [34], showing that AI-based predictive approaches substantially outperform the status quo method, in both exclusion and inclusion errors. The partial discrepancy is most likely due to the fact that [34] works on small samples ($n \in [1800, 11280]$), does not perform feature engineering, does not conduct meta-parameter search to control overfitting, and does not compute the entire exclusion-inclusion error curves when comparing targeting approaches.

2.2 Empirical Evaluation

Household surveys data. The present study uses publicly-available data from two countries: Costa Rica and Colombia, which represent populations of medium (47M) and smaller (5M) size, characteristic of Latin American countries. In particular, we use data from household surveys that are collected annually by the countries' national statistical offices; the same data on which status quo targeting systems are trained and evaluated [23, 27]. These surveys constitute one of the most important information instruments in the countries, based on which the main poverty, prices, labor, and other socioeconomic indices are computed [15, 25]. The surveys are carefully designed to be statistically representative of the population (i.e., unbiased), and collect both ground truth income data, as well as relevant household information, such as socio-demographics, living conditions, education, assets, and services (information & communication services, utilities, sanitary, etc.) [15, 25].

The ground truth income in these surveys is widely considered the best income data available in the countries—as compared to income from census data and other surveys—due to the robustness of the sampling methodology, extensiveness of the questionnaire applied to each household, professionality of the field workers, and the lack of incentives to under-report [49]. Table 1 presents basic statistics of the datasets for each country.

Algorithms

We aim at assessing potential advantages of modern statistical and computational methods compared to the status quo.

	Poverty ratio	Years	Total Sample Size	Population Size
Costa Rica	22%	2015-18	22 k	4.9 M
Colombia	28%	2016-17	462 k	47.6 M

Table 1: Household survey data statistics.

Status quo methods. The status quo predictors used for income-based targeting—used over the past two decades, and still prevailing today—stem mainly from econometric methods [7, 23], more suitable for causal inference than out-of-sample prediction. Quantile linear regressions (QLR) are the most common [31], as it's often found empirically that these outperform other methods such as regular linear regressions and various matching estimators. In what follows, we refer to QLR as the *status quo* methodology.

AI-based methods. We implement a number of algorithms based on the AI paradigm of machine learning. In particular, the methodological elements that could increase the predictive performance compared to the status quo are: 1) feature engineering, 2) regularization, and 3) better approximator models.

Discussing the performance of different AI-based predictors is beyond the scope of this work. Here we describe only the best-performing methodology (See Section A. in the Supplementary Material (SM) for further methodological details.):

- **Feature engineering.** We preprocessed the survey data to generate three types of features. First, *expert features*, crafted by human experts, such as the ratio of people over rooms in the households, and the age of the head of the households. Second, *statistical features*, including means, modes and entropies for all individual-level variables of household members, such as age, gender, and education. Lastly, *deep features*, generated by a recursive neural network that condenses information of the individual-level features into a one-dimensional encoding—a technique akin to the AI subfield of multiple instance learning (MIL) [5].
- **Predictive algorithm.** The best-performing algorithm was a gradient boosting classifier, trained on the three feature sets concatenated. We used k-fold cross-validation to determine the model's meta-parameters and control overfitting.

In what follows, we refer to the above combination of feature engineering and predictive algorithm as the *AI-based* method.

Accuracy metrics

The two key metrics used by institutions in the social sector to assess the quality of their targeting are exclusion and inclusion errors [20, 23]. The ***exclusion error*** measures the % of poor households incorrectly classified as non-poor, denoted by $\epsilon_{\text{exc}} = \frac{\text{FN}}{\text{TP+FN}}$; while the ***inclusion error*** measures the % of non-poor households incorrectly classified as poor, denoted by $\epsilon_{\text{inc}} = \frac{\text{FP}}{\text{TP+FP}}$. TP, FP, TN, FN correspond to true positives, false positives, true negatives, and false negatives. In what follows, we compare the accuracy performance of alternative targeting systems based on these measures.

Exclusion-Inclusion Curve. Targeting rules are composed of two elements: a poverty score and an *acceptance threshold* above/below which candidates are accepted. Exclusion and inclusion errors are dependent on the acceptance threshold applied. Hence, for comparison of poverty prediction methodologies, we compute the entire ***exclusion-inclusion curve*** (EIC), which maps the full space of targeting rules that a predictive methodology enables: from the universal program, to all sizes of targeted programs, and down to the non-existent program (similar to ROC curves). Thereby, EICs map the fundamental trade-off between exclusion and inclusion errors. Figures 2a and 2b present instances of EICs.

Lastly, all error measures presented in this work are computed out-of-sample, and 95% confidence intervals are computed non-parametrically by means of bootstrapped resampling.

2.3 Results

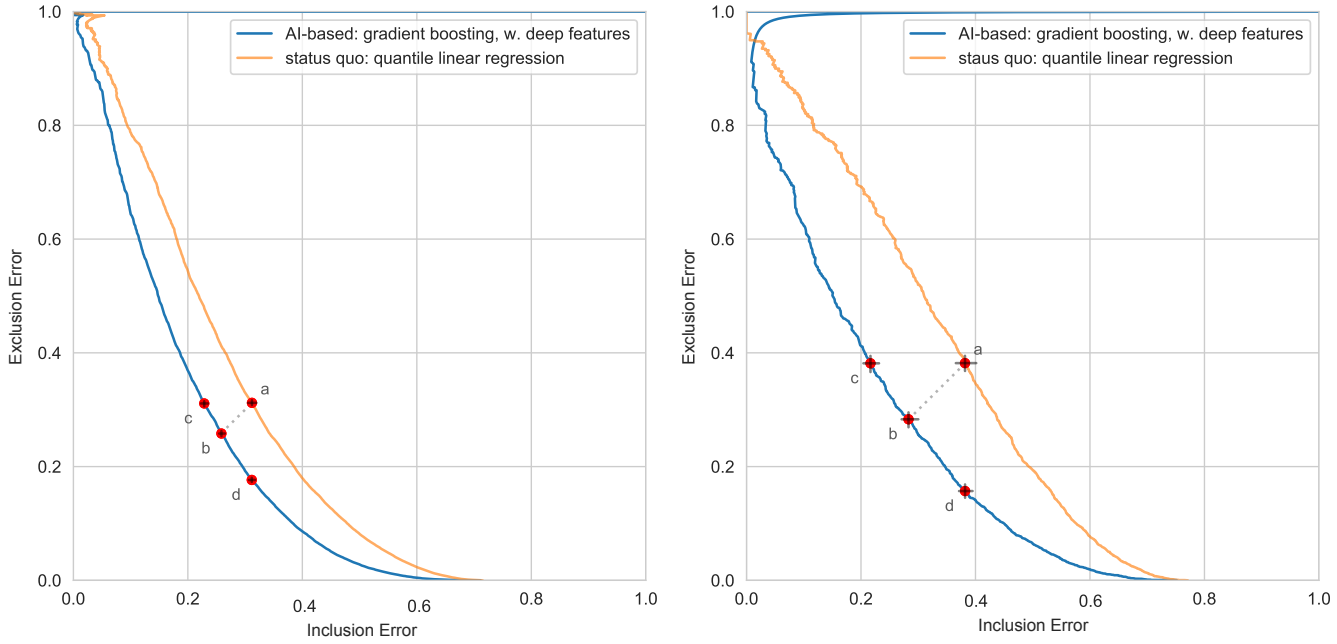
Figures 2a and 2b compare the accuracy performance of the AI-based methodology versus the status quo. For each method, the *exclusion-inclusion curve* is plotted, mapping the entire set of targeting rules, and the trade-off between exclusion and inclusion errors. The upper left corner corresponds to programs with no beneficiaries (100% exclusion and 0% inclusion error). The lower right corresponds to universal programs (0% exclusion and maximum inclusion error). In between, solutions correspond to all targeted programs ranging in sizes from zero to the population size.

In particular, point $\boldsymbol{a}$ on the status quo curve corresponds to the targeting rule with an acceptance threshold t_a that accepts a number of beneficiaries equal to the country's poverty rate. It is also the acceptance threshold that balances exclusion and inclusion errors (i.e., $\epsilon^{\text{sq}}_{\text{exc}}(t_a) = \epsilon^{\text{sq}}_{\text{inc}}(t_a)$). Similarly, point $\boldsymbol{b}$ on the AI-based curve corresponds to the threshold t_b that accepts a number of beneficiaries equal to the country's poverty rate, and which balances the errors $\epsilon^{\text{ai}}_{exc}(t_b) = \epsilon^{\text{ai}}_{inc}(t_b)$. Let $s(t)$ denote the size of the accepted population of a targeting rule with threshold t. Because $s^{\text{sq}}(t_a) = s^{\text{ai}}(t_b)$, the difference between points a and b provides a rather meaningful assessment, as it compares the performance of the two methods when both are constrained by a constant budget equal to the amount of poor in the population.

Points $\boldsymbol{c}$ and $\boldsymbol{d}$ on the AI-based curve correspond to solutions with either equal exclusion errors ($\epsilon^{\text{ai}}_{exc}(t_c) = \epsilon^{\text{sq}}_{exc}(t_a)$), or equal inclusion errors ($\epsilon^{\text{ai}}_{inc}(t_d) = \epsilon^{\text{sq}}_{inc}(t_a)$), in comparison to point a on the status quo curve. Finally, if we want to compare performances irrespective of any particular threshold level, but averaged across all, we compare the area under the inclusion-exclusion curve (AUEIC).[1]

Figure 2a and Table 2 present the performance comparison for Colombia. It is shown that the AI-based method dominates the status quo by a wide margin, and along the entire exclusion-inclusion error curve. This means that for any given budget, a social program will incur less exclusion error and less inclusion error if it targets based on the AI method. In particular, comparing points $\boldsymbol{a}$ and $\boldsymbol{b}$, a social program or policy accepting an amount of candidates equal to the country's poor would reduce its exclusion errors by 19.7%,

[1]This metric is equal to the area under the precision-recall curve, also known as average precision.



(a) Colombia. **(b) Costa Rica**

Figure 2: Exclusion versus inclusion error curve. Grey bars around the red points indicate 95% confidence intervals, computed non-parametrically via bootstrap resampling.

and its inclusion errors by 19.8% (Table 2), if it switches to the AI-based method. Most importantly, this reduction in errors means that 828,940 people in poverty, previously misclassified, would now be correctly included in the country's set of social protection policies.

Figure 2b and Table 2 present analogous results for Costa Rica, where a social program or policy accepting an amount of candidates equal to the country's poor population would reduce both its exclusion and inclusion errors by 25.9% (Table 2) if it switches to the AI-based method. The reduction in errors would mean that 110,377 people in poverty, previously misclassified, would be now correctly included in the country's set of social protection policies.

Figure 6 and Section A. in the SM report a detailed decomposition of the sources of increased accuracy, showing that both feature engineering and model non-linearity contributed to the substantial improvement of the AI-based over the status quo.

Beyond accuracy, relevant additional considerations for model choice include *fairness*, *robustness*, and *explainability*. Section 3 studies subgroup disparities in detail; Section D. in the SM tests the inter-temporal stability of the AI-based, and Section C. in the SM provides a model-agnostic layer of explainability, applicable to both AI-based and status quo methodologies.

3 DISPARITIES IN PREDICTIVE PERFORMANCE ACROSS SUBGROUPS

Section 2 shows that AI-based targeting systems can be substantially more accurate, enabling public institutions and NGOs to increase coverage of their target population while maintaining budget constant, or reduce the budget while maintaining coverage constant.

However, these results are aggregates across the population as a whole, which could hide systematic biases with respect to minorities or other population subgroups. In this section we study and compare the performance of AI-based vs. status quo targeting systems, computing error rates disaggregated by subgroups defined by relevant household characteristics, such as urban/rural, geographic region, gender of the head of the family, and family type.

Relevance. To the best of our knowledge, this is the first time that algorithmic rules of this sort are audited for potential exclusion disparities across population subgroups. As argued in Section 1, potential disparities could have a systematic impact on the lives of millions of socially disadvantaged households across the globe.

Contributions. In summary, 1) we show that AI-based targeting reduces the errors local to every population subgroup studied in both countries, rendering the transition to AI methods uncontroversial; 2) we find substantial disparities in exclusion errors across population subgroups, in both methods; and 3) the error disparities across subgroups are substantially narrowed by the shift from the status quo methods to the AI-based.

3.1 Methodology

We extend the accuracy analysis in Section 2 to the subgroup level. For every subgroup and each method (i.e., status quo and AI-based), exclusion errors were computed by setting acceptance thresholds equal to the poverty rate (base rate) of each subgroup. This choice of thresholds is analogous to the analysis in Section 2 and emulates the most prevalent practice, where social institutions set acceptance thresholds that admit a number of beneficiaries equal to the poverty

	Reduction in Area Under the Curve (AUEIC)	Reduction in Exclusion Error @constant budget $\frac{\epsilon^{sq}_{exc}(t_a) - \epsilon^{ai}_{exc}(t_b)}{\epsilon^{sq}_{exc}(t_a)}$	Reduction in Inclusion Error @constant budget $\frac{\epsilon^{sq}_{inc}(t_a) - \epsilon^{ai}_{inc}(t_b)}{\epsilon^{sq}_{inc}(t_a)}$	Increase in Poor Population Covered @constant budget
Colombia	26.4%	17.3%	17.1%	728,830
Costa Rica	37.1%	25.9%	25.9%	110,377

Table 2: Comparative Accuracy Results and their Impact on Coverage of the Poor.

rate [23]. For example, Figure 3a shows the exclusion error rates of urban/rural subgroups in Colombia under the status quo method.

We then compare the performance across methodologies by computing the relative improvement of the AI-based method over the status quo. For example, if the status quo and AI-based systems yield exclusion errors of 25% and 20%, respectively, for a given subgroup, then we report that the AI-based achieves a 20% reduction.

Beyond assessing the subgroup-level benefits of transitioning to the AI system, we also assess disparities among subgroups that exist *within* each method. We define the ***subgroup disparity*** metric σ_A as the standard deviation of the exclusion errors across subgroups, i.e., $\sigma_A = \sqrt{\sum_{g \in A} (e_g - \bar{e})^2}$, where A is a segmentation attribute such as *family size*, and $g \in A$ are the subgroups that it defines.

3.2 Results

Subgroup-specific improvements.

The right column in Figure 3 shows the subgroup-specific improvements in exclusion errors achieved by the AI-based over the status quo, for two selected examples of country-attribute pairs. It is observed that the AI-based method reduced errors in all cases, with a minimum improvement of 10% for rural households in Colombia (figures 3c and 3f).

Tables 3 and 4 provide an overview of the relative reduction in group-specific errors for all combinations of the two countries and five segmentation attributes: *urban/rural*, *gender of the head*, *with children*, *family size*, and *region*. Overall, we find that the AI-based method reduced the errors for each subgroup, across the five segmentation attributes and two countries. The minimum improvement was 5.7% for region 12 in Colombia, whereas a maximum improvement of 36.4% was attained for unipersonal households in Costa Rica (also shown in Figure 3f). All subgroup improvement results were significant at 95% confidence. Tables 5 and 6 in SM B. show full results in absolute terms for all country-attribute pairs.

Disparities across subgroups.

We are interested in knowing: a) whether there are significant disparities in predictive performance across subgroups; and b) whether these disparities are more salient for one method or another.

The left and center columns of Figure 3 show that substantial imbalances do exist in predictive performance across subgroups. For example, figures 3a-b show that poor Colombian households in urban areas are more likely to be misclassified by either algorithm. Similarly, figures 3d-e show that poor Costa Rican unipersonal households are significantly more likely to be misclassified than households with two members or more, by either method.

However, we also observe that the AI-based method has a strong effect not only in reducing errors overall, but also on balancing error disparities across subgroups. Figure 4 shows the subgroup disparities σ_A for the status quo method and the AI-based method. It is shown that σ_A disparities were reduced in both countries and for most segmentation attributes. Notable examples are those of Costa Rica, where disparities with respect to the *with children* and *family size* attributes were cut in half, and disparity with respect to *urban/rural* was reduced to nearly zero.

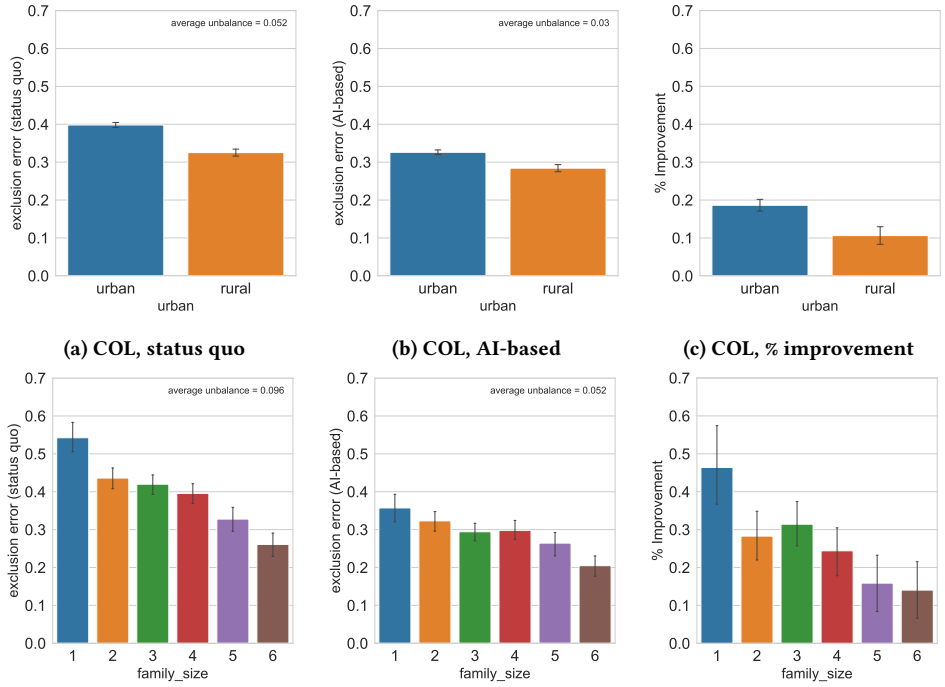
3.3 Section Conclusions

The above results provide three key results. First, AI-based targeting systems reduced predictive errors for every subgroup, according to the five segmentation attributes studied, in both countries. Because all subgroups result benefited, the AI-based Pareto-dominates the status quo at the subgroup level. This result is rather positive, as it renders uncontroversial among subgroups a decision to transition from the status quo to the AI-based targeting system.

Second, substantial disparities were found across population subgroups for both methods. These disparities were dependent on the particularities of each country and sensitive attributes, hence supporting the case for requiring empirical evaluations and fairness audits to understand subgroup-level performance prior to deploying algorithmic targeting systems. Lastly, although both methods entailed disparities in prediction errors across population subgroups, the AI-based method had a strong effect in balancing errors and narrowing disparities compared to the status quo.

4 FAIR TARGETING AND DISTRIBUTED GOVERNANCE VIA INTERACTIVE DECISION SUPPORT

Sections 2 and 3 demonstrate that the AI-based targeting system provides superior accuracy, both at the global and subgroups levels; and that it reduces disparities among population subgroups when compared to status quo methods. Unfortunately, these results do not entail that the AI targeting system is fair in absolute terms. On the contrary, Figure 4 shows that substantial performance disparities exist, demanding careful reflection on algorithmic fairness, and



(a) COL, status quo (b) COL, AI-based (c) COL, % improvement

(d) CR, satus quo (e) CR, AI-based (f) CR, % improvement

Figure 3: Subgroup-level performance for two examples of country-attribute pairs. All subgroups substantially benefited from switching to the AI-based method. Error bars denote 95% confidence intervals.

	Urban		Gender of Head		Children		Family Size					
Subgroup	urban	rural	male	female	with	without	1	2	3	4	5	6
Reduct. %	18.09%	12.62%	16.09%	18.95%	17.88%	13.56%	15.87%	11.85%	14.83%	9.61%	16.33%	19.44%
	Region											
Subgroup	1	2	3	4	5	6	7	8	9	10	11	12
Reduct. %	19.7%	21.65%	14.49%	14%	16.43%	17.64%	11.17%	15.54%	16.81%	13.64%	18.48%	5.74%
Subgroup	13	14	15	16	17	18	19	20	21	22	23	24
Reduct. %	22.57%	18.37%	15.43%	11.68%	14.98%	11.35%	20.59%	16.87%	19.1%	14.66%	18.52%	15.14%

Table 3: The AI-based method reduced exclusion errors for each population subgroup studied in Colombia.

potential fairness mechanisms to be implemented, prior to the operational deployment of any of these systems[2].

However, given the centralized-decentralized architecture of real-world social targeting systems (Subsection 4.1), and existing heterogeneous preferences over fairness and prioritization criteria (Subsection 4.2), a major challenge arises: how to foster fairness of a multiplicity of targeting rules, all using a common algorithmic input, but each working under heterogeneous preferences about what fair targeting should be?

[2]Note, of course, that the status quo has been deployed in multiple countries for more than two decades [7, 23].

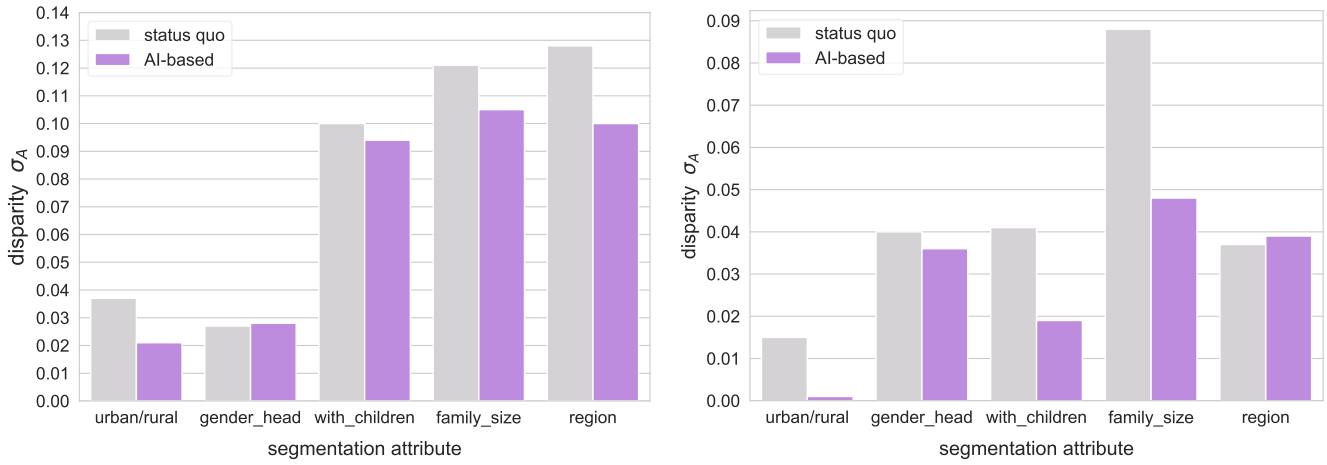
Subsection 4.3 describes a proposed decision-support platform for fostering fair targeting and distributed governance, currently being piloted with several social institutions in both countries.

4.1 The Centralized/Decentralized Architecture of Real-World Social Targeting

In many countries, like Costa Rica, Colombia, Panama, Mexico, and Dominican Republic, a unified poverty index is constructed by the central government, establishing a common methodology for assessing the poverty status of households in the country [23, 27].

	Urban		Gender of Head		Children		Family Size					
Subgroup	urban	rural	male	female	with	without	1	2	3	4	5	6
Reduct. %	24.24%	29.98%	24.13%	27.24%	25.28%	30.54%	34.13%	25.92%	30%	24.56%	19.51%	21.46%
	Region											
Subgroup	1	2	3	4	5	6	-	-	-	-	-	-
Reduct. %	22.79%	24.7%	32.78%	23.7%	36.39%	27.25%	-	-	-	-	-	-

Table 4: The AI-based method reduced exclusion errors for each population subgroup studied in Costa Rica.



(a) Colombia. **(b) Costa Rica**

Figure 4: Subgroup disparities for the status quo versus AI-based methods, according to five segmentation attributes. The AI-based method substantially reduces disparities across most segmentation attributes in both countries.

Similarly, agencies like USAID establish a per-country poverty assessment tool [7]. These central indices are then consumed in a distributed manner by a wide diversity of social programs and NGOs, serving as core and unified criterion for prioritizing and selecting beneficiaries in the country. The use of the central methodology is most often mandated by law, in an effort to focus public funds on the population segments most in need, as well as avoiding manipulation of public funds for political or private interests [7, 9].

However, the centralization of prioritization and selection criteria creates strong tensions across the diversity of institutions mandated to consume the central poverty index. These institutions include everything from cash transfer programs, to subsidized healthcare systems, subsidized energy, housing, and child care services, scholarship programs, pensions to non-formal workers, support to micro-entrepreneurs, and others. They also include a range of geographic scopes, encompassing the national, regional, and local levels of government.

Hence, targeting decisions are the compound product of a central agency issuing the guiding criterion for prioritization, and a diversity of social programs consuming that index, applying acceptance thresholds based on it, and most often complementing it with additional criteria like multidimensional poverty, and positive/negative preferences across population subgroups.

4.2 Heterogeneous Preferences Over Fairness and Prioritization Criteria

The hybrid centralized/decentralized architecture creates tensions over normative preferences regarding what fairness and prioritization criteria should guide targeting rules.

Tensions Over Fairness Criteria

There are at least three parity-based fairness criteria discussed in the algorithmic fairness literature applicable to the context of selecting beneficiaries of social programs:

- **Demographic parity**, which requires the amount of positively classified individuals of any given group to be proportional to the size that group in the population [21]. For example, a 50/50 gender composition in a president's cabinet.

- **Threshold parity**, which defines fair classification as applying the same acceptance threshold to all individuals in the population [11].
- **Error rate parity**, which defines fair classification as equal classification error rates across relevant population subgroups. Here in particular, we consider the special case of *equal opportunity*, which requires only parity in exclusion error—or equivalently, parity in coverage of the positive class—, because the context of targeting social benefits is a canonical example of positive interventions [21].

It is well known that these parity criteria are most often incompatible and trade-off with one another [30]. Moreover, there is no consensus in the academic literature about the appropriateness of one over the rest. Regarding real-world practice, we find from official documentation and qualitative interviews that a complex mix of parity criterion is involved in allocating resources and selecting beneficiaries. For example, while several official documentation stresses the prevalence of threshold parity criteria [9, 47], interviewees mentioned that geographic and other subgroup quotas also often play a role in budget allocation negotiations.

Positive discrimination. Even more fundamental, there is no consensus about parity *per se*, regardless of its definition, being the *a priori* desirable fairness property for social targeting systems. On the contrary, positive discrimination prevails. Institutions targeting social programs commonly prioritize population subgroups that are considered socially disadvantaged, guided by a paradigm of positive discrimination and affirmative action, rather than parity. Examples of these include victims of violence and forced migration in Colombia [36, 41], and households with members with disabilities in Costa Rica [24, 35].

Hence, it is not clear that one should, for example, attempt to recalibrate a centrally-defined poverty score to balance exclusion errors—selecting relevant subgroups and fairness criteria *a priori*—unless there is certainty that such balance conforms to the policy goals and ethical frameworks of all institutions that will build targeting rules based on it.

Tensions Over Prioritized Population Subgroups

In addition to tensions over fairness criteria, the diverse social institutions within countries often demand tailored and more complex prioritization criteria, beyond monolithic income poverty scores [24, 41]. Many struggle to accommodate additional criteria that respond to their policy goals, and to the particular characteristics of their social programs and target population.

For example, while cash transfer programs focus mainly on household poverty [6, 23, 28], programs intended to broaden access to student loans also use standardized educational test scores as complimentary prioritization criteria [10]. Similarly, subsidized healthcare systems need to incorporate the quality of alternative healthcare services accessible to potential beneficiaries, as stated in official program operating rules [37]. From a set of 14 interviews conducted as part of this project with managers and field workers of local social institutions, we found several additional criteria used in regional-level programs. For example, in programs providing pensions to poor elderly citizens, prioritization criteria often include the amount of time (months) that applicants have waited in the admissions queue, their age, as well as the severity of any medical condition they may suffer.

In a similar vein, social programs often attempt to incorporate contextualized preferences over discrete population subgroups, which vary substantially according to programs' policy goals, resources, and target population. For example, interviewees from local institutions mentioned the creation of *socioeconomic profiles*—such as single mother households, elderly households with any disability, underage orphans or orphans under 25 still in university. These *socioeconomic profiles* are then used in periodic program-specific resource allocation meetings, where group-specific acceptance thresholds are discussed and set for each profile until exhausting the period's budget. Likewise, other commonly prioritized subgroups include indigenous populations, population with disabilities, and victims of violence and forced migration [35, 36].

4.3 Interactive Decision Support Platform for Fair Targeting and Distributed Governance

It is in this context of centralized/decentralized decision making, and heterogeneous preferences, that we must reflect on what fair targeting is, and how we can design systems that support it.

Towards it, we developed an interactive decision support tool with the following design goals, addressing the opportunities described in sections 1-2, and challenges described in sections 3-4:

(1) **Accuracy:** Empower social institutions to leverage highly accurate AI-based predictors on which to build their targeting rules, without requiring expertise in machine learning.
(2) **Distributed Governance:** Enable decentralization of three design choices fundamental to the design of fair targeting rules: (a) the choice of relevant population subgroups; (b) the choice of fairness criteria—type of parity criteria, or conversely positive discrimination—; and (c) the choice of prioritization criteria (e.g., income or multidimensional poverty).
(3) **Awareness:** Educate stakeholders in understanding their available space of targeting rules, and underlying tradeoffs.
(4) **Inclusiveness and transparency:** Enable wider discussions among stakeholders over the space of options. Increase the transparency and auditability of deployed targeting rules, and their rationales.

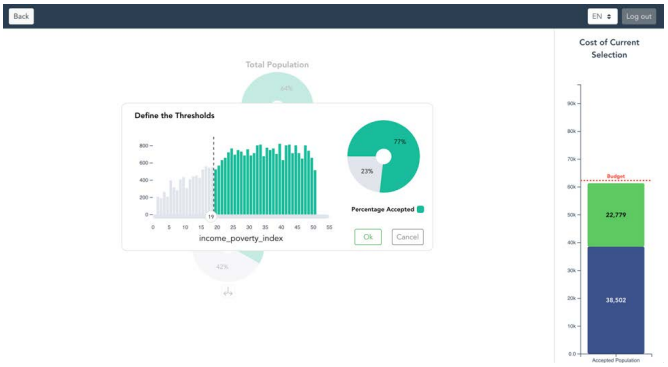
Consider the working example of a nation-wide cash transfer program in a Latin American country. With the start of the fiscal year, the institution's prioritization committee has to select which of the 100,000 applicants will take one of the 65,000 available spots this year. The committee session starts, and the decision support tool is projected on screen. First, the platform asks to specify the:

- **Prioritization criteria:** the core prioritization criteria to use, e.g., the centrally-defined income poverty score.
- **Segmentation attributes:** the categorical attributes by which they might choose to segment the population, e.g., urban/rural.
- **Budget:** the budget for the period, i.e., 65,000 available spots;
- **Hard filters:** any hard exclusion criteria to apply, according to the target population of the program—e.g., age requirement for a pension program.

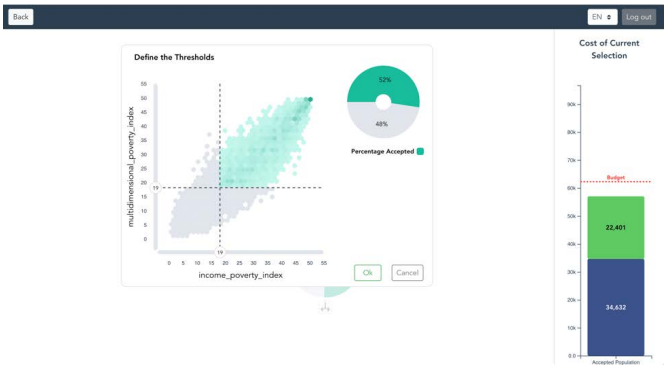
After defining the above key elements, the user moves to the main interactive tool for designing targeting rules. The interface

(a) Initial state: total population without segmentation.

(b) Segmentation by *urban/rural*.



(c) one-dimensional prioritization criteria.



(d) two-dimensional prioritization criteria.

(e) Segmentation by *family type*.

(f) Segmentation by *vulnerability*.

(g) Coverage parity.

(h) Positive discrimination.

Figure 5: Interactive decision-support platform.

is composed of three elements: **1) the population tree**, which displays—for each population segment—the acceptance thresholds set, and the percentage of coverage that the thresholds entail (figures 5a and 5b). **2) the thresholds adjustment window** (figures 5c and 5d), and **3) the costs stacked bar**, which displays the segmented costs versus the institution's budget (all figures 5a-h).

The key functions supported by the platform are:

a) Thresholds adjustment. Users can observe the distribution of the population or any segment in terms of the prioritization variable, and how the threshold divides the distribution in accepted/non-accepted (Figure 5c). Next, they can adjust the threshold, interactively observing the impact that threshold shifts have on: 1) the % of coverage, which corresponds to 1 − *exclusion error* (pie chart), and 2) the associated costs (stacked bar on the right).

b) Population Segmentation. Users can segment the population in subgroups by clicking the "split node" button on any leaf node and selecting a segmentation attribute (e.g., family type; see figures 5b, 5e, and 5f). The platform initially assigns an equal threshold to all subgroups in a split. Users can then adjust subgroup-specific thresholds as in (a), and appreciate interactively the percentage coverage and costs of all groups. Hence, it forces users to reflect on fundamental trade-offs between coverage and cost, and among population subgroups; yet providing flexibility for defining relevant segments, and priorities among them. For example, in a country the law may require to give preference to groups such as: indigenous households, households with disabilities, displaced by armed conflict, and victims of domestic violence. In that case, the user may explore shifting thresholds to reach 100% coverage of each vulnerable group, but will then be forced to acknowledge the reductions in coverage of non-vulnerable groups needed to keep the program in budget (Figure 5h).

c) Balance for parity. Users can easily implement different parity criteria across subgroups, by clicking "balance button" on any parent node and selecting one of three parity criteria to implement: coverage parity (which implies exclusion error parity), thresholds parity, or demographic parity. The platform automatically finds the combination of thresholds that achieves the specified parity (while maintaining constant the cost incurred by the parent node). Figures 5f, 5g, and 5h exemplify the cases of threshold parity, coverage parity, and positive discrimination.

a) Dual prioritization criteria. Institutions often require to complement income poverty scores with additional indices. Use cases include scholarship programs considering both economic need and academic performance [9, 17]; pension programs considering income poverty and the amount of months applicants have waited in the admissions queue; loan programs considering income poverty and repayment risk; and the relevant use case of multidimensional poverty indices [9, 43]. For dual criteria, the platform maps the joint distribution of the population (or segment), and enables the user to explore combinations of the two thresholds (Figure 5d). Thereby, institutions can reflect on emphasizing either criteria, in accordance with their policy goals, and with proper awareness of the coverage and cost implications.

CONCLUSIONS

The present work presents a comprehensive study focused on a relevant domain: the targeting of social policies. We demonstrate that a shift towards the use of AI methods in poverty-based targeting can substantially increase accuracy, both globally and with respect to each population subgroup studied. This improvement enables an increase in coverage of the poor of nearly a million people in the two countries studied, without increasing their social budgets. However, it is also shown that both the status quo and AI-based systems suffer from significant performance disparities across population subgroups. Hence, we close by proposing an interactive decision support platform that empowers social institutions to design fair and accurate targeting rules tailored to their needs, under a distributed governance framework and enhanced transparency.

Despite the advantages here shown, potential adopters of AI-based approaches to social targeting should be mindful of at least two potential disadvantages. First, the shift from linear to non-linear models loses the interpretability of linear coefficients. Yet, this disadvantage is mitigated by the construction of a model-agnostic layer of explainability [33, 44], such as the feature sensitivity maps presented in Section C. of the SM. These make AI algorithms more transparent and allow for accountability during the decision-making process. Second, deployed systems should be tested for inter-temporal robustness, and monitored for relevant changes in the statistical distributions underlying the prediction task. In the case of the two countries here studied, distributions were stable and one-year inter-temporal accuracy was virtually equal to same-year cross-validated accuracy (see SM D.).

We are currently conducting pilots with multiple social institutions in Costa Rica, Colombia, and Panama, on the use of the decision-support platform described in Section 4. Future work may report on the multi-stakeholder preferences over prioritization and fairness criteria elicited during pilots, as well as on the dynamics towards or away from consensus.

Finally, the improvement of poverty-estimation models, and the eligibility rules derived from them, are only two of several high-value opportunities in the use of AI approaches for improving the targeting of social policies. Exciting applications currently in development include: a) The use of transfer and multi-task learning for leveraging multi-country data towards improved models, particularly in small countries with small data samples. b) The integration of remote sensing and household survey data into comprehensive targeting systems that inform both, the deployment of field workers on the ground (prospection), and the prioritization of applicants (selection), under a *proactive targeting* paradigm that optimizes information cost, accuracy, and fairness [39]. And c) the use of AI approaches to reduce survey measurement errors, by assisting field workers on the ground in targeting their information verification efforts.

We hope the work here presented sparks attention towards these high-value opportunities in the use of AI for social good, and ultimately contributes to lead governments, multinational organizations, and NGOs towards more effective and equitable targeting of social policies across the global south.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] World Bank. 2015. World Databank. *World Development Indicators* (2015).

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207* (2017).

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[5] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition* 77 (2018), 329–353.

[6] Simone Cecchini and Aldo Madariaga. 2011. Programas de Transferencias Condicionadas - Balance de la experiencia reciente en América Latina y el Caribe. https://repositorio.cepal.org/bitstream/handle/11362/27854/S2011032_es.pdf

[7] IRIS Center. 2009. Manual for the implementation of USAID poverty assessment tools. *povertytools. org/training_documents/Manuals/USAID_PAT_Manual_Eng. pdf, accessed* 1 (2009).

[8] Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. Productivity and selection of human capital with machine learning. *American Economic Review* 106, 5 (2016), 124–27.

[9] CONPES. 2016. Declaratoria de importancia estrategica del Sistema de Identificacion de Potenciales Beneficiarios (Sisben IV). https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3877.pdf

[10] Consejo nacional de política económica y social. República de Colombia. Departamento nacional de planeación. 2016. Declaración de importancia estratégica del sistema de identidicación de potenciales beneficiarios (sisbén iv). https://www.sisben.gov.co/Documents/Compes%20IV/6285-CONPES%203877.pdf

[11] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint arXiv:1808.00023* (2018).

[12] 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.

[13] Stephanie Cuccaro-Alamin, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. 2017. Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review* 79 (2017), 291–298.

[14] Alan De Brauw and John Hoddinott. 2011. Must conditional cash transfer programs be conditioned to be effective? The impact of conditioning transfers on school enrollment in Mexico. *Journal of Development Economics* 96, 2 (2011), 359–370.

[15] Departamento Administrativo Nacional de Estadística. 2018. *Pobreza monetaria en Colombia.* Technical Report. https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2018/bt_pobreza_monetaria_18.pdf

[16] Ariel Fiszbein and Norbert R Schady. 2009. *Conditional cash transfers: reducing present and future poverty.* The World Bank.

[17] Fodesaf. 2018. Ficha: Programa Becas Estudiantiles. https://www.fodesaf.go.cr/prog_soc_selectivos/programacion_anual/fichas_cronogramas/2018/fichas/Ficha%20descriptiva%20FONABE%202018.pdf

[18] Philip Gillingham. 2015. Predictive risk modelling to prevent child maltreatment and other adverse outcomes for service users: Inside the 'black box'of machine learning. *The British Journal of Social Work* 46, 4 (2015), 1044–1058.

[19] Eric Goldman. 2011. Revisiting search engine bias. *Wm. Mitchell L. Rev.* 38 (2011), 96.

[20] Rema Hanna and Benjamin A Olken. 2018. Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives* 32, 4 (2018), 201–26.

[21] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[22] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. 2007. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications* 33, 4 (2007), 847–856.

[23] Pablo Ibarrarán, Nadin Medellín, Ferdinando Regalia, Marco Stampini, Sandro Parodi, Luis Tejerina, Pedro Cueva, and Madiery Vásquez. 2017. How Conditional Cash Transfers Work. (2017).

[24] Instituto Mixto de Ayuda Social Subgerencia de Desarrollo Social Sistemas de Información Social. 2019. Informe del Programa Protección y Promoción Social (Del 01 de enero al 31 de diciembre de 2018). http://www.imas.go.cr/sites/default/files/docs/Informe%20PPPS%20anual%20del%202018%20VF%2015-02-2019.pdf

[25] Instituto Nacional de Estadística y Censos. 2018. Encuesta Nacional de Hogares Julio 2018: Resultados Generales. http://www.inec.go.cr/sites/default/files/documetos-biblioteca-virtual/enaho-2018.pdf

[26] International Labour Organization. 2018. *Women and men in the informal economy. A statistical picture.* International Labour Organization.

[27] Oleksiy Ivaschenko, Claudia Rodríguez, Marina Novikova, Carolina Romero, Thomas Bowen, and Linghui Zhu. 2018. *The State of Social Safety Nets.* The World Bank.

[28] Julia Johannsen, Luis Tejerina, and Amanda Glassman. 2009. Conditional cash transfers in Latin America: Problems and opportunities. (2009).

[29] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–95.

[30] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[31] Roger Koenker and Kevin F Hallock. 2001. Quantile regression. *Journal of economic perspectives* 15, 4 (2001), 143–156.

[32] Ilker Kose, Mehmet Gokturk, and Kemal Kilic. 2015. An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing* 36 (2015), 283 – 299. https://doi.org/10.1016/j.asoc.2015.07.018

[33] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[34] Linden McBride and Austin Nichols. 2016. Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review* 32, 3 (2016), 531–550.

[35] Ministerio de Educación. 2018. Programa Becas Estudiantiles. http://www.siteal.iipe.unesco.org/sites/default/files/sit_accion_files/siteal_costa_rica_0700.pdf

[36] Minsalud. 2019. Preguntas Frecuentes. https://www.minsalud.gov.co/Lists/FAQ/Tematica.aspx?View={9B7912F5-7A3D-40BF-8BF1-078A2FB53F97}&FilterField1=Tem_x00e1_tica&FilterValue1=Salud&FilterField2=Subtema&FilterValue2=R%C3%A9gimen%20Subsidiado

[37] Minsalud. 2019. Régimen subsidiado. https://www.minsalud.gov.co/proteccionsocial/Regimensubsidiado/Paginas/regimen-subsidiado.aspx

[38] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How search engines reinforce racism.* NYU Press.

[39] Alejandro Noriega-Campero, Michiel A Bakker, Bernardo Garcia-Bulle, and Alex'Sandy' Pentland. 2019. Active Fairness in Algorithmic Decision Making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 77–83.

[40] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books.

[41] Registraduría Nacional del Estado Civil. 2015. Resolución 11143 de 2015. https://www.registraduria.gov.co/IMG/pdf/R_RN_2015_11143.pdf

[42] Stephen A Rhoades. 1993. The herfindahl-hirschman index. *Fed. Res. Bull.* 79 (1993), 188.

[43] Rafael Perez Ribas, Guilherme Issamu Hirata, Fabio Veras Soares, et al. 2008. *Debating Targeting Methods for Cash Transfers: A Multidimensional Index vs. an Income Proxy for Paraguay? s Tekoporã Programme.* Technical Report. International Policy Centre for Inclusive Growth.

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).

[45] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[46] Andrew D Selbst. 2017. Disparate Impact in Big Data Policing. *Ga. L. Rev.* 52 (2017), 109.

[47] Sistema Costarricense de Información Jurídica. 2016. Reglamento a la Ley de Desarrollo Social y Asignaciones Familiares. http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?param1=NRTC&nValor1=1&nValor2=67607&nValor3=105961&strTipM=TC

[48] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10.

[49] United Nations. 2017. *Principles and Recommendations for Population and Housing Censuses, Revision 3.* 315 pages. https://doi.org/10.18356/bb3ea73e-en