Algorithmic Decision-Making, Fairness, and the Distribution of Impact: Application to Refugee Matching in Sweden

Kirk Bansak* Linna Martén†

July 2021

Abstract

The literature on algorithmic fairness has predominantly focused on differences in predictive performance, in and of itself. This perspective obscures or ignores several key dimensions of algorithmic decision-making in the real-world, especially in public policy settings, including (a) the directness with which predictions themselves dictate the final decisions, (b) the fairness or equity involved in status quo decision-making in the absence of algorithmic support, and (c) the possible impacts across other outcomes aside from that being predicted. This paper provides an alternative approach to evaluating the fairness of algorithmic decision-making on the basis of the distribution of causal impact. To do so, the approach formalizes the algorithmic assignment procedure within the context of the potential outcomes causal framework. Further, it is flexible in allowing for the consideration of outcomes of different types (continuous or discrete), impact on multiple outcomes of interest, any number of policy options to which units can be assigned, and various ways in which predictions map to actual decisions.

To illustrate the approach, as well as the limits of the conventional algorithmic fairness perspective, the paper further presents an in-depth application: the geographic assignment of refugees in Sweden, where refugee employment outcomes are a high priority for resettlement authorities. Using real-world data on refugees in Sweden to perform retrospective counterfactual impact evaluations, the paper assesses the fairness implications if refugees were algorithmically assigned to labor market regions, in particular focusing on fairness across gender. As is shown, the framework allows for a multifaceted evaluation of the distribution of causal impact. Specifically, in addition to considering the algorithmic target outcome (i.e. the outcome used in building the underlying predictive models that inform the decision-making procedure), the framework also allows for evaluation of unintended cross-outcome impacts with respect to other outcomes (e.g. skill development) and their implications for fairness.

*Assistant Professor, Department of Political Science, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, United States. E-mail: kbansak@ucsd.edu

†Researcher, Institute for Social Research, Stockholm University, Universitetsvägen 10 F, 10691 Stockholm, Sweden. E-mail: linna.marten@sofi.su.se

I. Overview

Algorithmic decision-making—the use of predictive algorithms and their outputs as inputs into a decision-making process—has spread across a variety of domains in both the private and public spheres, such as criminal justice (Desmarais et al., 2021; Kleinberg et al., 2018a), health care (Bates et al., 2014; Obermeyer et al., 2019), and hiring decisions (Chalfin et al., 2016). Equally pervasive have become concerns among scholars and policymakers alike that the use of algorithms as decision-making aids could reproduce, exacerbate, or create inequities that cut across race, gender, and other socioeconomic dimensions (for an overview, see Kleinberg et al., 2018b). Accordingly, a literature on "algorithmic fairness" has rapidly emerged over the past several years, defining metrics and laying out guidance for assessing the resulting or anticipated fairness implications of predictive algorithms as decision-making aids (for an overview of different fairness definitions that have been proposed, see Verma and Rubin, 2018).

The most predominant conceptions and definitions of algorithmic fairness have limited their focus on differential *predictive* performance across groups, and typically with respect to a single outcome of interest.[1] This perspective, however, obscures or ignores several key dimensions along which algorithmic decision-making contexts can vary:

1. The directness with which predictive outputs translate into decisions, a function of whether: (i) the output is considered alongside additional information and/or filtered through additional procedures, which may be human-led or automated, prior to the determination of ultimate decisions, and (ii) decisions made for one unit can affect the decisions made for other units via constraints or other problem features.

2. The fairness or equity of the procedures under the reference or status quo decision-making process, in the absence of algorithmic assistance, which may range from the ideal of perfect evaluation (e.g. a human labeler being able to perfectly classify images) to highly imperfect and potentially biased judgments (e.g. judges appraising the risk posed by criminal defendants in order to make bail or release decisions).

3. The number of outcomes of interest and the number of policy decisions to which a unit can be assigned, with the simplest case being a single outcome of interest (e.g. recidivism) and a binary policy assignment choice (e.g. hold in jail or release).

Scholarship on algorithmic fairness has tended to paint a picture of the simplest case of algorithmic decision-making. In focusing predominantly on predictive behavior, this view designates the predictive output in and of itself as what we should ultimately care about, as if predictions equate with decisions and impact. Further, in considering this algorithmic performance without reference to the status quo, this view engages in nirvana-fallacy reasoning. And finally, the vast majority of scholarship in this area has focused on the

[1] For instance, one of the most high-profile examples is the possibility of unequal false positive rates when predicting recidivism for criminal defendants from different racial groups.

simple case of a binary policy assignment choice and the results for a single outcome of interest, hiding the complications posed by the possible existence of multiple outcomes of interest and/or multiple policy assignment options for understanding fairness in real-world algorithmic decision-making contexts.

This paper argues for an alternative approach to evaluating the fairness of algorithmic decision-making procedures that views fairness from the lens of distributive causal effects and facilitates assessment via retrospective causal impact evaluation. To do so, the paper formalizes the algorithmic decision-making procedure within the context of the potential outcomes causal framework, and considers the fairness of the distribution of impact in a way that accommodates a wider range of scenarios and use cases than other approaches have. While some previously proposed definitions and conceptions of algorithmic fairness have incorporated causality, they have focused on the causal relationships between the predictors and outcome variables (e.g. by allowing and/or excluding specific edges in a causal graph; Kusner et al., 2017; Kilbertus et al., 2018; Nabi and Shpitser, 2018) or built directly upon statistical definitions within causally defined strata of the target population (e.g. Imai and Jiang, 2020). In other words, they focus on specific causal relationships that are explicitly encoded (or not encoded) into the predictive model(s). In contrast, the focus here is counterfactual reasoning from the perspective of impact and understanding fairness to be a question of impact, which is achieved by explicitly evaluating the counterfactual distribution of outcomes under an algorithmic procedure relative to the status quo (or some other alternative). In doing so, the proposed framework is also agnostic to (and hence flexible with respect to) the manner in which predictions lead to decisions, the type of outcomes (continuous or discrete), the number of policy options, and whether there are other outcomes of interest beyond the target outcome that serves as the prediction model response. In other words, it is a framework that can be flexibly employed across a more comprehensive variety of contexts, and is geared toward allowing for evaluations of fairness that are intuitive, actionable, and facilitate policy action rather than setting up a framework that aspires unrealistically toward "transcendental justice" (Sen, 2009).

To highlight these dynamics and illustrate the approach, the paper considers the case of the assignment of refugees to geographic regions within host countries. This is a recently growing area of algorithmic decision-making research, as national resettlement programs seek to improve the post-arrival outcomes (e.g. employment) of ever-increasing flows of refugees and asylum seekers (e.g. Bansak et al., 2018; Mousa, 2018; Andersson et al., 2018; Gölz and Procaccia, 2019; Olberg and Seuken, 2019; Acharya et al., 2021; Ahani et al., 2021). The basic idea behind the algorithmic assignment of refugees and asylum-seekers is to, on the basis of prior data and predictive analytics, match families to those locations within a host country where they are most likely to thrive, according to specific metrics of integration success. This is an algorithmic decision-making case with various complexities, including the existence of various constraints that make families' location assignments rivalrous with one another, an imperfect status quo decision-making procedure in which the location assignments are far from optimal, and a context in which there are many outcomes of interest (e.g. employment,

education, health) and many policy options (locations to be assigned to).

Using real-world data on refugees in Sweden, where refugee employment outcomes are a high priority for resettlement authorities, the paper performs retrospective counterfactual impact evaluations that assess the fairness implications if refugees were algorithmically assigned to labor market regions. These results are also compared to evaluations on the basis of conventional algorithmic fairness definitions. In this application, the focus is on fairness across gender. Previous research has documented a large and persistent employment gap between natives and refugees, particularly among women (e.g. Bratsberg et al., 2014; Schultz-Nielsen, 2017). Refugees' low employment is viewed as a major obstacle to their integration. With this in mind, several countries have designed specific programs to mitigate existing gender differences (Albrecht et al., 2021). The Swedish government has also provided targeted funding to the Public Employment Service and the Adult Education Association in order for them to improve their services and increase the employment rate for newly arrived women (Prop., 2017).

The methods employed follow on work in policy learning/optimization and retrospective impact assessment (e.g. Hubbard and Van der Laan, 2008; Zhang et al., 2015; Samii et al., 2016; Jung et al., 2020; Athey and Wager, 2021), and the leveraging of machine learning techniques to undertake causal inference by more flexibly and quasi-parametrically estimating counterfactual response surfaces (cf. Hill, 2011; Zhao and Hastie, 2021; Athey, 2018; Van der Laan and Rose, 2011). In addition, this paper fits into a smaller subset of the algorithmic fairness literature that has sought to establish connections between notions of fairness and social welfare (Kasy and Abebe, 2021; Hu and Chen, 2020; Mullainathan, 2018; Heidari et al., 2019; Kleinberg et al., 2018b; Corbett-Davies et al., 2017), but it provides several contributions beyond this existing literature.

First, it considers these dimensions of fairness through a rigorous causal impact formalization, and further highlights how this perspective of fairness can be assessed intuitively via restrospective causal impact evaluations. The framework presented connects most closely with the work of Kasy and Abebe (2021), who also consider the relationship between fairness and impact within the causal inference tradition. However, the operationalization of assessing fairness via influence functions in Kasy and Abebe (2021) may provide barriers to usage in practical policy contexts, where intuitive presentation and results are necessary to allow for policy setting and decision-making. In line with that imperative, the approach presented in this paper evaluates fairness and the distribution of causal effects within a particular decision-making system via retrospective causal impact evaluation (Samii et al., 2016; Hubbard and Van der Laan, 2008), which yield counterfactual results that are easy and intuitive to interpret (e.g. what would have happened in the presence of algorithmic decision-making vs. in its absence).[2] In doing so, this paper also presents concrete metrics

[2]Conceptually, this approach connects closely with the concept of "disparate benefit" discussed informally in Kleinberg et al. (2018b). Other studies that similarly consider the causal impact of algorithmic decision-making relative to status quo procedures, but not with reference to conceptions of fairness, include Bansak et al. (2018); Kleinberg et al. (2018a). Also relevant is Obermeyer et al. (2019), who evaluate the fairness

that could serve as guidelines for evaluating fairness from the perspective of the distribution of causal impact.

Second, in both the formalization and its application to refugee assignment, this paper presents an approach that allows use cases to deviate substantially and in multiple ways from the prototypical context upon which the vast majority of the algorithmic fairness literature (including that connecting fairness to social welfare) has been built. Indeed, conceptualizations, concrete definitions, and illustrations of algorithmic fairness have predominantly revolved around binary screening use cases where predictions or classification are mapped onto a simple decision between two options—such as to grant bail or not, to give a loan or not, to hire or not, to provide additional health resources or not. By allowing for greater complexities as described above (e.g. more options and constraints), and illustrating such complexities via the refugee matching application, this paper is able to provide greater richness in highlighting the limitations of conventional definitions of algorithmic fairness as well as the flexibility of the proposed approach. Indeed, these deviations from the simplest case challenge not only the usefulness, but in some cases even prima facie relevance of the conventional definitions.

Third, this paper also introduces notions of cross-outcome distributive impact, which are defined and accommodated within the presented causal impact framework. The possibility of diffuse effects across many outcomes is an important, but generally neglected dimension of algorithmic decision-making. Typically, there is a single target outcome, which is the outcome that serves as the dependent or response variable when building the corresponding predictive model(s). Notions of algorithmic fairness are then defined with respect to that target variable, and consideration of other outcome variables would require new predictive models with these other variables as the response. In practice, however, even if there are additional outcome variables of interest, there will still be a target variable (or potentially an index of variables) that is chosen to build the predictive model that feeds into the decision-making. In other words, building additional predictive models for other outcomes in order to assess fairness has no real-world bearing if those other predictive models are not to be used to guide decision-making. Hence, the question of interest is: given predictive models built on the basis of the target outcome, what is the impact of decisions guided by those same predictive models on other non-target outcomes of interest. The framework presented in this paper naturally handles this consideration.[3]

of high-risk health care management algorithms using the statistical fairness notion of calibration, and then separately (and informally) consider counterfactual impacts but do not formalize or connect the latter to fairness; and Corbett-Davies et al. (2017), who highlight the tradeoffs between satisfying statistical notions of fairness and the efficacy of an algorithmic decision-making system (in their use case of pre-trial algorithms, in terms of delivering public safety).

[3]This consideration of cross-outcome impacts is distinct from an algorithmic bias phenomenon termed "label choice bias" (Obermeyer et al., 2019). In the case of label choice bias, there is a known outcome variable of interest at a conceptual level (e.g. health), but the choice of the specific measure to use to operationalize that variable (e.g. health spending vs. number of comorbidities) may be unclear and can lead to unintended bias depending on the socioeconomic processes that underlie the alternative measures. In contrast, the case of cross-outcome impact considered here focuses on the possibility of multiple outcome variables of

The allocation of public services across democratic countries has become more and more dependent on algorithmic assistance.[4] Given the clear implications for justice surrounding the distribution of public resources, evaluating the fairness of these tools is essential. Yet in many of the cases where algorithmic decision-making systems have been deployed, they are developed by private companies with little transparency, and the public has only been able to scrutinize them after seeing evidence of skewed outputs or unintended effects, which can engender deep mistrust with such systems and the policymakers who sign off on them. Hence, trying to establish and standardize best practices and intuitive metrics for understanding the expected impacts of algorithmic tools not only during but also *prior* to their deployment should be a key priority to help guide policymakers. And while the study of algorithmic fairness originated in computer science, the growing use of algorithmic decision-making in public policy and other realms of society has led to the increasing engagement of economists and political scientists with this area of scholarship (e.g. Bansak et al., 2018; Imai and Jiang, 2020; Yang and Roberts, 2021; Kennedy et al., 2021; Rambachan et al., 2020). Social scientists have contributed key perspectives in the past on the design of various types of institutions that allocate opportunities and resources. Similarly, they should also be able to make important contributions to research on algorithmic decision-making and fairness, especially given their disciplines' unique blend of expertise in various methods, social choice theory, political and moral philosophy, and practical policy issues.

II. Fairness and the Distribution of Causal Impact

To fix ideas, the presentation of the following formalization will focus on the specific context of refugee assignment. However, it establishes a general formulation for any context where there is a finite set L of different policy options that units can be assigned to, and some procedure for making those assignments on the basis of anticipated (predicted) responses to or outcomes from those assignments. It is more general than the most common formulation in the statistical literature on policy learning/evaluation, in which a unit is mapped to a policy as a function of only its own characteristics, thereby not allowing for constraints or rivalrous policy decisions. In addition, it also nests as a special case the most common scenario in the algorithmic fairness literature, where there is a binary, screening decision to be made (i.e. $|L| = 2$) with few constraints on how decisions are made for each individual (i.e. the constraints and family structure laid out in the refugee matching application could be eliminated).

interest that are conceptually distinct (e.g. employment and education), and how building an algorithmic decision-making system around one could lead to unintended consequences with respect to the other.

[4] In Sweden, for instance, these tools have been used to match students to schools and decide what kind of labor market training different types of unemployed individuals receive. In the United States, public policy areas in which such tools have been used include criminal justice, immigration, child welfare, education, law enforcement, fire safety, health care, housing, and public benefits.

A. *Preliminaries*

Let S denote a set of refugee families, indexed by $i = 1, ..., n$, that have arrived and need to be geographically placed within a host country. Further, let individual refugees be indexed by $j = 1, ..., m$. For each family, let V_i denote the set of refugees belonging to that family, such that $j' \in V_{i'}$ if and only if refugee j' belongs to family i'. Further, let L denote a set of locations in the host country, where each location $k \in L$ has capacity for q_k families. For simplicity, assume the number of location slots and families are exactly equal: $\sum_{k \in L} q_k = |S|$.

For each individual refugee j, let Y_j denote a post-arrival outcome of interest (e.g. employment), $A_j \in L$ denote the refugee's actual assigned location, and X_j a vector of pre-arrival background characteristics. Denote the expected outcome for refugees with characteristics x assigned to location k as:

$$\mu(k, x) = E[Y_j | A_j = k, X_j = x]$$

On the basis of past data, one can then use statistical learning methods to estimate and approximate $\mu(k, x)$, resulting in predictive models $\hat{\mu}(k, x)$. These predictive models can then be applied to refugees in S, yielding an estimate of the expected outcome for each refugee j if assigned to each location k, denoted by $\hat{\mu}_{jk}$. Use of the outputs from these models to determine (or help inform) the location assignments for newly arriving refugees—which has previously been proposed and implemented in various countries (Bansak et al., 2018; Gölz and Procaccia, 2019; Acharya et al., 2021; Ahani et al., 2021)—would then constitute an algorithmic decision-making system.

B. *Conventional Algorithmic Fairness*

Conventional algorithmic fairness definitions, typically referred to as "statistical" definitions of fairness (Verma and Rubin, 2018; Chouldechova and Roth, 2020), focus on the observed performance of the predictive component in and of itself, not the final decision-making (i.e. assignment) component. That is, the evaluation made by these algorithmic fairness definitions take the following general form:

$$f(\hat{\mu}_{jA_j}, Y_j | G_j = g) \overset{?}{\approx} f(\hat{\mu}_{jA_j}, Y_j | G_j = g')$$

where f denotes a fairness metric function and G_j denotes a salient group variable (e.g. gender) that includes at least one segment of the overall population that is disadvantaged and perhaps legally protected. The underlying imperative behind algorithmic fairness definitions of this form is to achieve equality in the function f between advantaged and disadvantaged/protected groups. As mentioned earlier, in the case of refugee assignment in Sweden, the most salient dimension to employ as the variable G_j is gender. In any particular context, the decision on which attribute(s) to focus on is guided by context-specific ethical, legal, and political considerations.

For a comprehensive overview of the various algorithmic fairness definitions that fall into this general form, see Verma and Rubin (2018). In the case of a binary outcome variable,

some of the most common and high-profile fairness considerations revolve around false positive/negative rates. Of particular interest if the binary outcome variable is a good/beneficial outcome (e.g. employment) is the "equal opportunity" fairness metric (equivalent to false negative error rate balance), for which $f(\hat{\mu}_{jA_j}, Y_j|G_j = g)$ corresponds to the within-group true positive rate $Pr(\mathbf{1}(\hat{\mu}_{jA_j} > t)|Y_j = 1, G_j = g)$, where t denotes the classification threshold (Hardt et al., 2016).[5] Note that this formulation is not unique to the refugee matching context and covers the general case where there is a finite set of different policies (L) that units can be assigned to, and models that output predicted responses to those policies ($\hat{\mu}_{jk}$), for any unit j and any policy $k \in L$. In the prototypical context where a binary screening decision needs to be made (e.g. grant bail/loan/assistance or not), $|L| = 2$. Often times in that simple case, the predictions will focus on the expected responses to only one of the two policy options (e.g. what would happen if bail is granted, or if a loan is given), in which case $\hat{\mu}$ is no longer subscripted with k, yielding the definition of equal opportunity as commonly presented:

$$Pr(\mathbf{1}(\hat{\mu}_j > t)|Y_j = 1, G_j = g) = Pr(\mathbf{1}(\hat{\mu}_j > t)|Y_j = 1, G_j = g')$$

Definitions of statistical fairness such as this, in which algorithmic fairness is considered as a function statistical measures of performance in and of themselves, have attained significant prominence in public discourses (e.g. Angwin et al., 2016) as well as academic research (e.g. Obermeyer et al., 2019). Accordingly, such definitions have been increasingly proposed and employed as standard reporting metrics to be included in AI/algorithm audits (e.g. Mitchell et al. (2019); see also Brown et al. (2021) and Raji et al. (2020)).

B.1. Shortcomings of Statistical Notions of Algorithmic Fairness

One common line of critique levied at statistical definitions of fairness is that there are many and that, as an unavoidable mathematical reality, they can generally not be all achieved simultaneously (Kleinberg et al., 2017; Chouldechova, 2017). However, putting this issue aside (or assuming that agreement can be made on which of the many competing definitions to prioritize), there are additional problems that span both the philosophical and the practical.

First, the existing statistical notions of fairness privilege an absolutist view in which fairness is either achieved or not, neglecting the fact that there can be a continuum of relative fairness. As such, these notions fall into what Sen (2009) describes as a transcendental theory of justice that envisions perfectly just or fair social arrangements, and in doing so, obstructs decision-making and policy setting in the real-world, where perfection is impossible.

Second, and relatedly, by evaluating only the performance behavior of the predictive models themselves, the existing statistical notions of fairness ignore the inherent and unavoidable

[5] In cases where the binary outcome variable corresponds to a bad/undesirable outcome, false positive error rate balance (also known as the "predictive equality" fairness metric) has received more attention (Chouldechova, 2017). This has particularly been the case in the area of criminal justice, in which many jurisdictions have deployed algorithmic tools that predict whether or not criminal defendants will re-offend or fail to appear at their court date, which then aid judges in deciding whether or not to release defendants from jail. For instance, see Angwin et al. (2016).

imperfection involved in decision-making in many areas, including any area of policy that involves predicting people's future outcomes. Indeed, the status quo and other alternatives to algorithmic support are, in such cases, characterized by their own degree of unfairness as a result. In such cases, even if there exist independent or agreed upon criteria for a fair set of results or outcomes, it is not necessarily possible for any real-world procedure to accomplish this ideal—what Rawls (2020) referred to as "imperfect procedural justice."

And third, also relatedly, these existing statistical notions of fairness are more concerned with fairness of the algorithmic instrument or rule itself than fairness of the resulting outcomes or realizations of their use. Indeed, note from the general form of statistical notions of algorithmic fairness—$f(\hat{\mu}_{jA_j}, Y_j|G_j = g) \overset{?}{\approx} f(\hat{\mu}_{jA_j}, Y_j|G_j = g')$—that they are a function of predictions and observed outcomes, and hence completely neglect the manner in which predictions are connected to decisions. Along these lines, several previous studies have shown how applying constraints to predictive algorithms to achieve statistical notions of fairness can result in worsened outcomes for both advantaged and disadvantaged groups (Kasy and Abebe, 2021; Hu and Chen, 2020; Mullainathan, 2018). These studies have made such observations in the prototypical context of screening decisions, where predictions on binary outcomes are viewed as leading (more or less) directly to a decision to grant or withhold a benefit/privilege.

The distinction—and possible tradeoffs—between prediction and decision become even more stark when one considers the broader classes of problem to which algorithmic decision-making support can be applied. In the refugee matching case, for instance, there are (a) many options, rather than just two; (b) the decision involves a choice between substitutable options of the same class (locations), rather than wholesale granting or withholding of a benefit/privilege, where those options may have different utility for different people; and (c) decision options are governed by binding constraints, such that the options are rivalrous and at the immediate decision point, each person's decision is also affected by decisions applied to others. In this substantially more complicated case, it is unclear if standard statistical notions of fairness are even conceptually relevant.

All of these considerations raise the question: What are we actually trying to achieve with the deployment of algorithmic decision-making aids and systems, and how much do evaluations of conventional statistical notions of algorithmic fairness aid in this? One might argue that the simplest answer to the first question is that we are trying to improve lives. If so, and if we are beginning with a starting point that is imperfect and unfair, and also if it is impossible to produce an eventuality that is perfect, then a new procedure that could replace the old should not necessarily be evaluated from a absolutist, isolated, transcendentalist fairness perspective, but rather should be considered from the perspective of whether it will either produce outcomes that are more fair and/or produce a fair distribution of the benefits that are created.[6]

[6]This fits into the comparative perspective on justice for which Sen (2009) advocates. This perspective also fits into principles of practice in discrimination case law, in which legal findings of discrimination on the basis of disparate impact require the plaintiff to establish not simply that the defendant's actions have

Such is the case with the use of algorithmic decision-making in public policy spheres, where inevitable and inherent uncertainty surrounding human behavior means perfection will never be achieved, and where status quo decision-making processes are always marked by some degree of imperfection and unfairness. This can be contrasted with more objective classification tasks that are pervasive in commercial applications of algorithmic systems, and around which much of the thinking on algorithmic bias and fairness has revolved, in which the objects being labeled or classified could be done so perfectly because they do not involve predicting the future (e.g. classifying images, emails, etc. as belonging to a particular category). For these latter cases, statistical notions of fairness may capture everything that is necessary to know, because (a) the prediction directly and wholly renders the decision, and (b) the relative reference point may actually be perfect and hence perfectly fair (albeit costly or manual) decision-making. But any sort of understanding of fairness that is appropriate with respect to this latter situation should not be expected to be appropriate for understanding fairness in the former situation.

C. Algorithmic Assignment

As already described above, we can define $\hat{\mu}(k, x)$ as the models that predict the target outcome for any individual with characteristics x if assigned to location (or, more broadly, policy option) k, and then denote the estimate of the expected outcome for a particular individual j if assigned to location k as $\hat{\mu}_{jk}$. Yet this has still not defined the precise procedure or mechanism by which individuals are actually assigned to locations—that is, the mapping of predictions to actual decisions on which policy option each individual gets.

First, consider the class of procedures Π that assign the refugees to locations subject to the location capacity constraints q_k defined above and the requirement that all refugees in a family are assigned to the same location. Formally, let Π denote the class of procedures that produce a feasible matching, $S \to L$. Define feasibility such that $|\Phi^{-1}(k)| = q_k$ for any $k \in L$ and $\Phi \in \Pi$. Let $\bar{\phi}_i$ denote the location assignment of family i and ϕ_j the location assignment of refugee j under Φ. Hence, for any refugee j', if $j' \in V_{i'}$, then $\phi_{j'} = \bar{\phi}_{i'}$.

We can then refer to an "algorithmic assignment procedure" as a procedure that uses estimates $\hat{\mu}_{jk}$ (i.e. predictions) to produce such a feasible matching. Let $\Phi_1 \in \Pi$ denote a particular algorithmic assignment procedure, and ϕ_{1j} the location that would be (or would have been) assigned to refugee j under this procedure. Further, let $\Phi_0 \in \Pi$ and ϕ_{0j} denote the same under some status quo decision-making procedure.

In the present case of refugee matching, the following is one possibility for the algorithmic assignment procedure:

$$\Phi_1 = \arg\max_{\Phi \in \Pi} \sum_{j=1}^{m} \hat{\mu}_{j\phi_j}$$

This procedure assigns refugees to locations such that the highest expected outcome is

disparate effects across groups but also that the goals of those actions, if deemed to have a legitimate business purpose, could be achieved via a less discriminatory alternative (Supreme Court, 2014).

achieved subject to the constraints represented in Π (i.e. location capacity constraints and the requirement that families are kept together). Variants of this procedure can also be formulated to take into account additional practical or administrative constraints. As one example, stochastic programming methods can be integrated to facilitate dynamic (i.e. sequential, one-by-one) assignment of families over time (for the technical details, see Bansak, 2020). The present paper, however, will abstract away some of these optimization details and simply consider Φ_1 to be a feasible procedure to fulfill the maximization objective (exactly or in expectation).

In the general case, an algorithmic assignment procedure can also incorporate additional institutional, administrative, or other constraints or features. For instance, in the case of complex health care management (Obermeyer et al., 2019), Φ_1 combines predictions of health risk with multiple threshold rules. In the case of pretrial bail decisions, Φ_1 involves human decisions with discretionary incorporation of the predictions. In cases where human discretion is involved, formalizing Φ_1 would require the modeling of human judgment on the basis of past behavior. Further, in the more general case, the various feasibility constraints that are laid out above (in particular, limits or quotas on the number of individuals that can be assigned to any particular policy option) may or may not pertain, and other constraints could be incorporated into the definition of Π.

D. Fairness of Algorithmic Decision-Making and Distribution of Causal Impact

Now, following the potential outcomes framework of Neyman (1923) and Rubin (1974), let $Y_j(k)$ denote the potential outcome for refugee j if assigned to location k, for all $k \in L$. The relationship between the observed outcome and the potential outcomes for any refugee j is such that $Y_j = Y_j(A_j)$. Further, define the following:

$$\theta(k, x) = E[Y_j(k)|X_j = x]$$

The counterfactual outcome of the algorithmic assignment procedure for the refugees in the cohort S of families can then on average be defined as:

$$\frac{1}{m}\sum_{j=1}^{m}\theta(\phi_{1j}, X_j)$$

Under conditional independence, $Y_j(k) \perp\!\!\!\perp A_j \mid X_j$, and overlap, $P_k(x) \equiv P(A_j = k|X_j = x) > 0 \ \ \forall \ \ k, x$, this is identified by:

$$\frac{1}{m}\sum_{j=1}^{m}\mu(\phi_{1j}, X_j)$$

where $\mu(k, x)$ is as already defined earlier. In the case of centralized refugee assignment, these assumptions are easily justifiable, as refugees are assigned to locations by case officers either on a quasi-random basis or on the basis of X_j.

The causal impact of algorithmic assignment relative to the status quo can thus be defined as $\frac{1}{m}\sum_{j=1}^{m}\theta(\phi_{1j}, X_j) - \theta(\phi_{0j}, X_j)$, which can be estimated via $\frac{1}{m}\sum_{j=1}^{m}\hat{\mu}(\phi_{1j}, X_j) - \hat{\mu}(\phi_{0j}, X_j)$.

D.1. Benefits of this Perspective

By explicitly incorporating (a) the mapping of predictions to decisions and (b) the counterfactual impacts of decisions on outcomes, this approach allows for a comprehensive evaluation of the actual impacts of algorithmic decision-making procedures. Furthermore, it is flexible to incorporate any number of policy options, both discrete and continuous variables, and consideration of additional outcomes beyond that targeted in prediction (as will be developed in more detail below). The following subsections highlight how this framework can then be employed to assess the fairness of the distribution of an algorithmic decision-making procedure's impact.

D.2. No Group Harm

By design, for the algorithmic assignment procedure to be achieving the baseline goal of making average improvements above the status quo, the quantity $\frac{1}{m}\sum_{j=1}^{m}\theta(\phi_{1j}, X_j) - \theta(\phi_{0j}, X_j)$ should be positive.[7] More generally, however, for some subgroup $G_j = g$, define the following subgroup average outcome under some feasible assignment procedure Φ_*:

$$\tau_{\Phi_*}(g) = \frac{1}{\sum_{j=1}^{m}\mathbf{1}(G_j = g)}\sum_{j=1}^{m}\theta(\phi_{*j}, X_j)\cdot\mathbf{1}(G_j = g)$$

A first, necessary criterion for a fair distribution of the causal impact would require that no group is harmed:

$$\tau_{\Phi_1}(g) - \tau_{\Phi_0}(g) > 0$$
$$\forall\ g$$

D.3. Balanced Distribution of Causal Impact

Assuming this simple requirement is met, the fairness of the causal impact distribution would also be evaluated by comparisons across group. Specifically, for any two groups g' and g'', the evenness of the distribution of causal impact can be evaluated (along with analog estimators) via:

$$f(\tau_{\Phi_1}(g'), \tau_{\Phi_0}(g')) \overset{?}{\approx} f(\tau_{\Phi_1}(g''), \tau_{\Phi_0}(g''))$$

where f denotes a causal effect function. Two natural choices would evaluate absolute or relative measures (depending on which is more appropriate given the objectives and outcome of interest) of causal effect: $f(a, b) = a - b$ and $f(a, b) = \frac{a}{b}$, respectively.

In practice, this would mean to evaluate these comparisons with an appropriate amount of tolerance for disparity since establishing exact equality would be virtually impossible. In the case of comparing absolute causal effects (i.e. comparing $\tau_{\Phi_1}(g') - \tau_{\Phi_0}(g')$ against $\tau_{\Phi_1}(g'') - \tau_{\Phi_0}(g'')$), a reasonable approach would be to evaluate their relative size by measuring

[7] This, of course, assumes without loss of generality that the target outcome is a positive/beneficial outcome.

the deviation of their ratio from 1:

$$1 - \frac{min\Big([\tau_{\Phi_1}(g') - \tau_{\Phi_0}(g')], [\tau_{\Phi_1}(g'') - \tau_{\Phi_0}(g'')]\Big)}{max\Big([\tau_{\Phi_1}(g') - \tau_{\Phi_0}(g')], [\tau_{\Phi_1}(g'') - \tau_{\Phi_0}(g'')]\Big)} \overset{?}{<} \epsilon_a$$

where ϵ_a represents the tolerance. The value of ϵ_a would need to be contextually determined by stakeholders in the policy process. However, there may be relevant guidelines and precedents that have been set in other realms, such as in the area of discrimination law, that could then be applied or modified here. For instance, the "four-fifths rule" is often employed as a guideline for determining adverse impact in hiring or other selection procedures—a distinct use case, but not far afield of the present discussion of the distribution of impact. Adapting this guideline here would mean setting the tolerance to $\epsilon_a \equiv \frac{1}{5}$.

In the case of comparing relative causal effects (i.e. comparing $\frac{\tau_{\Phi_1}(g')}{\tau_{\Phi_0}(g')}$ against $\frac{\tau_{\Phi_1}(g'')}{\tau_{\Phi_0}(g'')}$), since the two components are already expressed in relative terms, a simple difference could be evaluated:

$$\left| \frac{\tau_{\Phi_1}(g')}{\tau_{\Phi_0}(g')} - \frac{\tau_{\Phi_1}(g'')}{\tau_{\Phi_0}(g'')} \right| \overset{?}{<} \epsilon_r$$

where ϵ_r again represents the amount of tolerance.[8]

Note again that this formulation covers the more general case where there is a finite set of different policies (L) that units can be assigned to, an algorithm that outputs predicted responses to those policies ($\hat{\mu}_{jk}$, for any unit j and any policy $k \in L$), and a procedure Φ_1 that makes assignments on the basis of those predictions (along with any other constraints that need to be accounted for), as well as the status quo decision-making process Φ_0.

As can be seen, in contrast to the conventional algorithmic fairness perspective, this conception of fairness is focused on evaluating the actual impact of the algorithmic decision-making, and making that evaluation relative to the status quo state of affairs.

D.4. Distribution of Cross-Outcome Impacts

Now consider an additional outcome, denoted by $\tilde{Y}_j$, along with potential outcomes $\tilde{Y}_j(k)$. For instance, the outcome Y may be the primary outcome targeted for optimization in the algorithmic assignment process (e.g. employment), while $\tilde{Y}$ may be an additional outcome not explicitly used by the algorithm but still of some importance to policymakers (e.g. post-arrival education or instruction). Define $\tilde{\theta}(k, x)$ and $\tilde{\mu}(k, x)$ as before but with respect to $\tilde{Y}_j(k)$ and $\tilde{Y}_j$, and under analogous assumptions as before, $\tilde{\theta}(k, x) = \tilde{\mu}(k, x)$, the latter of which can also be estimated using the data—call those estimated models $\hat{\tilde{\mu}}(k, x)$.

The following subgroup average outcome for $\tilde{Y}$, under the algorithmic assignment procedure Φ_1 optimizing for Y, can thus be denoted by:

$$\tilde{\tau}_{\Phi_1}(g^*) = \frac{1}{\sum_{j=1}^{m} \mathbf{1}(G_j = g^*)} \sum_{j=1}^{m} \tilde{\theta}(\phi_{1j}, X_j) \cdot \mathbf{1}(G_j = g^*)$$

[8] Setting $\epsilon_2 \equiv \frac{1}{5}$ could potentially be motivated by the four-fifths rule in this case as well.

The cross-outcome fairness of algorithmic decision-making—i.e. collateral impact with respect to a separate outcome—can thus be evaluated (along with analog estimators) via:

$$f(\tilde{\tau}_{\Phi_1}(g'), \tilde{\tau}_{\Phi_0}(g')) \stackrel{?}{=} f(\tilde{\tau}_{\Phi_1}(g''), \tilde{\tau}_{\Phi_0}(g''))$$

As above, a first consideration would be the desire for no cross-outcome group harm:

$$\tilde{\tau}_{\Phi_1}(g) - \tilde{\tau}_{\Phi_0}(g) > 0$$

$$\forall\ g$$

Beyond this, evaluations of the fair distribution of cross-outcome impact could proceed using the same (relative and absolute) causal effect functions and approaches presented in the previous section.

D.5. Remarks

Note that a proper evaluation of the distribution of cross-outcome impacts requires the full procedure detailed above, beginning with estimation of the models $\hat{\tilde{\mu}}(k, x)$. One might think that cross-outcome impacts could be inferred simply through the impact evaluation on the target outcome combined with a simple analysis of the relationship between target and additional outcomes. For instance, if the no group harm criterion is met for the target outcome, and the target outcome and additional outcome are positively (negatively) correlated, a naive conclusion would be that this implies that no cross-outcome group harm is met (violated). This conclusion is incorrect. First, note that a positive observed correlation between the target and additional outcomes means that $cor(Y_j, \tilde{Y}_j) = cor(Y_j(\phi_{0j}), \tilde{Y}_j(\phi_{0j})) > 0$. There is, of course, nothing from this that guarantees the same sign for a correlation between these two outcomes given counterfactual, algorithmic assignments. In other words, $cor(Y_j(\phi_{0j}), \tilde{Y}_j(\phi_{0j})) > 0$ does not imply $cor(Y_j(\phi_{1j}), \tilde{Y}_j(\phi_{1j})) > 0$. And further, even if $cor(Y_j(\phi_{1j}), \tilde{Y}_j(\phi_{1j})) > 0$, it could still be the case that optimizing with respect to the target outcome could harm with respect to the additional outcome if values of $\tilde{Y}_j(\phi_{1j}))$ are for some reason systematically lower than values of $\tilde{Y}_j(\phi_{0j}))$.[9]

Why does the observed correlation between the target and additional outcomes $cor(Y_j, \tilde{Y}_j) = cor(Y_j(\phi_{0j}), \tilde{Y}_j(\phi_{0j}))$ provide such limited traction by itself for understanding cross-outcome impacts? The fundamental issue is that individuals will have different locations (or in more general terms, policy options) assigned to them under Φ_1 versus Φ_0, and correlations between the target and additional outcomes could be heterogeneous across locations. Further, note that assignments under Φ_1 (and perhaps also, but not necessarily, under Φ_0) are guided on

[9]Though note that there are situations in which, for theoretical or functional reasons, optimizing with respect to one outcome should be expected to also improve other, closely related outcomes if achieving success on the former outcome can lead to or even cause success on the latter. For instance, optimizing on near-term employment is likely to also lead to improvements on longer-term employment since employment in the near term sets a foundation for having employment later as well, whether in the same job or in another job that was enabled by the first.

the basis of individuals' values X_j. For any particular stratum of individuals $X_j = x$, their within-stratum correlation between the target and additional outcomes may vary and even switch signs across different locations. In one location, it may be the case that employment and education move together for a particular stratum. This could be, for instance, because there are employment opportunities in particular industries that are well-suited to that stratum, and employment in such industries leads to training or educational opportunities (it could also be the case that the location is independently good for both education and employment for that stratum). In contrast, in some other location but for the same stratum, it may be that there are strong employment opportunities but without corresponding educational opportunities. Calculations of $cor(Y_j, \tilde{Y}_j)$, and even within-stratum calculations of $cor(Y_j, \tilde{Y}_j)$, will not provide sufficient insight on these aspects.

III. Application

A. Institutional Details and Data

The majority of refugees coming to Sweden apply for asylum upon arrival.[10] After submitting their application, they are offered temporary housing by the Swedish Migration Agency (SMA) but also have the option of finding housing on their own. The share receiving housing assistance varies over time, and reached 54% in 2018 (SMA, 2019). Once the application has been reviewed and approved, a process that can take up to a year, those who are still unable to find housing on their own are assigned to municipalities all over the country.

In the past, almost all municipalities (97%) signed voluntary agreements with the SMA outlining the number of assigned refugees they had to accept in a given year. The assignment process was reformed in March 2016 when all municipalities were required to accept a given yearly quota of assigned refugees. The quota is based on the local labor market, the population size, the previous number of newly arrived refugees in the municipality, available housing, and the number of asylum seekers currently living in the municipality. Importantly, refugees are not able to influence which municipality they get assigned to.[11] The SMA assigns refugees to the municipalities continuously during the year, and individuals are assigned to the municipality that currently has the lowest fulfillment degree of their quota, creating a quasi-random assignment process. Table A2 in the appendix also confirms that individuals' background characteristics are unrelated to characteristics of the municipality they are assigned to, suggesting that unobservable characteristics are also unlikely to be correlated with the type of municipality to which an individual is assigned. Refugees are only given one assignment offer, and have to find housing on their own if they choose to decline.

After receiving a residence permit most refugees also participate in a two year long labor market program organized by the Swedish Public Employment Service (PES). Most individuals participate full time in the program, but it is also possible to only participate

[10] Sweden also accepts 5000 UNHCR refugees each year, but they are given a residence permit before arrival.

[11] Exceptions are made for individuals who suffer from health issues that require care that is only available at a few specific hospitals.

part time for a longer duration than two years. At the beginning of the program, caseworkers at the PES set up an individualized activity plan depending on the refugees' education and labor market experience. The program always includes language training known as Swedish for Immigrants (SFI) as well as an orientation class about the Swedish society. SFI classes are divided into four levels A-D, and individuals with longer education typically start at a higher level. It is also possible to study, do internships, and participate in labor market training classes as a program activity. During this period individuals receive an allowance that is conditional on their participation in the program. In order to increase incentives to find work as soon as possible, the allowance is not reduced if an individual finds employment but is still able to participate in the program.[12] In 2018, around 40% of the individuals who left the program (and belonged to the labor force) were employed 90 days later (Gäfvert et al., 2018).

This paper uses individual-level administrative register data covering all refugee assignments that occurred in 2016 and 2017. The data set was compiled by Statistics Sweden and contains information about individuals' demographic background characteristics and their asylum case, such as the application date, the type of residence permit, the assignment date, and municipality. This data set was merged with yearly information about income, educational participation and performance, and location of residence. We restrict our sample to adults (aged 19-60) who applied for asylum in Sweden (i.e. we exclude resettled refugees). Table A1 provides summary statistics for the 21,520 individuals in our dataset. Almost 40% of the refugees are women, and most of the sample arrived from Syria, Afghanistan, Eritrea, and Iraq. All outcome variables are measured in the assignment year, and we see that roughly 10% found employment already in their arrival year, although the employment rate for men (13%) was considerably higher than for women (5%).

B. Evaluation Procedure

In order to assess the potential impact of algorithmic assignment of refugees in Sweden, along with its implications for fairness, we perform a retrospective counterfactual analysis. We compare the outcomes of the 2017 cohort (adult refugees who were geographically assigned in 2017) given their actual (status quo) assignments, to estimates of their expected outcomes had they instead been assigned via algorithmic assignment in order to optimize employment. The outcome of interest Y targeted in the algorithmic assignment procedure Φ_1 is a binary variable measuring whether or not an individual found employment within the year that they arrived, and the geographic basis for the assignments (i.e. the locations in L) are 41 local labor market regions in Sweden.[13]

[12] For additional information about the program, see the evaluation by Joona et al. (2016)

[13] Local labor markets are areas defined as being self sufficient when it comes to labor demand and supply, and are based on commuting patterns. The smallest possible unit is a municipality, but typically they consist of several adjacent municipalities. The 290 Swedish municipalities have been classified as 70 different labor market regions by Statistics Sweden. In order to observe enough individuals in each location we have re-grouped the smallest units with their neighbors, giving us a total of 41 local labor markets.

To mimic a real-world implementation of algorithmic assignment in order to lend the counterfactual analysis as much external validity as possible, several procedures and constraints are employed. First, since in a real-world implementation models built with past data are used to make predictions and guide the assignment of new arrivals, the estimated models $\hat{\mu}(k, x)$ are based on the 2016 cohort. Once trained, these models are then applied to the 2017 cohort to generate estimates of their expected success in each location and, subsequently, guide their counterfactual algorithmic assignment. Second, under the counterfactual algorithmic assignment, real-world geographic constraints are employed such that the same number of families that were actually assigned to each local labor market under the status quo are also assigned via the algorithmic process. Third, of course, families are also kept together under the counterfactual algorithmic assignment. And finally, whereas the algorithmic assignment Φ_1 was previously formalized as the optimal assignment as applied to all refugees at once in the relevant set (i.e. solution to a single, static linear assignment problem), in reality assignments would need to occur over time dynamically or in batches; hence, in the counterfactual algorithmic assignment evaluated here, optimal assignments are applied sequentially to two-week batches of refugee families based on their actual assignment dates.

In addition to evaluating the impact of algorithmic assignment on the target employment outcome Y, cross-outcome impact is also evaluated. The additional outcome $\tilde{Y}$ that is considered is a key skills-based measures of integration. Specifically, $\tilde{Y}$ is a binary indicator for the completion of at least one Swedish for Immigrants (SFI) class. As described earlier, SFI is a national free Swedish language learning program, and completion of these classes is highly valued by the Swedish resettlement program as an important factor for successful integration.

To generate flexible and quasi-parametric models that estimate the corresponding conditional expectation functions, we use stochastic gradient boosted tree ensembles and implement the modeling and estimation on a location-by-location basis. We fit the models using cross-validation, specifically tuning the interaction depth and number of boosting iterations. Calibration analysis and other metrics are used to validate the models and assess fit.

C. Results

C.1. Statistical Fairness Evaluation

As described earlier, the equal opportunity metric (i.e. equal true positive rates) has been proposed as a suitable statistical fairness measure when predicting a binary outcome that is good/beneficial. Hence, given the target outcome of employment in the present context, we assess the true positive rates of the estimated models—i.e. $Pr(\mathbf{1}(\hat{\mu}_{jA_j} > t)|Y_j = 1, G_j = g)$, where G_j denotes gender and t is set to 0.5, the natural threshold for predicted probabilities—separately for male and female refugees and compare their values. The resulting values are 0.06 for male refugees but only 0.03 for female refugees.

That is, the true positive rate is twice as high for male refugees relative to female refugees,

seeming at first glance to represent an unfair disadvantage for female refugees, who are already disadvantaged to begin with. That being said, however, the fact that both true positive rates are so close to 0 is no surprise at all: for refugees of all types, the baseline employment rates in the first year are extremely low simply because finding employment quickly as a refugee in Sweden (and in any country for that matter) is exceedingly difficult. As a result, there are very few strata of refugees (i.e. values of the vector of background characteristics X_j) with predicted probabilities of employment greater than 0.5.

In addition, the baseline employment rate that is observed in reality is particularly low among female refugees. As has been discussed and formalized elsewhere (e.g. Corbett-Davies et al., 2017), unequal baseline rates in the outcome generally lead to a mathematical guarantee that false negative and positive rates will not be balanced, and hence the imbalance discussed above is unsurprising. Nonetheless, the critique against having imbalanced true positive rates (which is equivalent to having imbalanced false negative rates) is that even if this imbalance may be the result of real-world disparities, those disparities are often the result of biases and inequities that exist in society, and allowing for this to be encoded into algorithmic decision-making systems can thereby serve to reproduce or even exacerbate those biases in society. In the present case, then, the concern is that the lower true positive rate in the estimated models for women not only may be due to their lower employment rate but also may reinforce their lower employment rate if the models were deployed to guide decision-making. Is this the case, however? As already explained earlier, this cannot be known simply by evaluating conventional statistical definitions of fairness, and instead what is needed is an evaluation of the distribution of actual impact. The following section provides such an analysis, highlighting how the present concern over unequal true positive rates rates does not actually translate to a corresponding unequal impact.

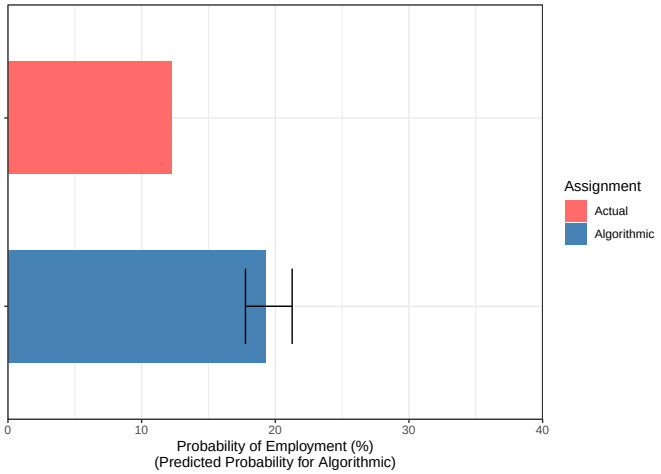
C.2. Target Outcome Impact Distribution

The figures below display the results of the retrospective counterfactual analysis, and hence serve as the basis for evaluating fairness from the perspective of the distribution of causal impact. In each figure, the red bars correspond to the actual (under status quo) average outcome achieved by the 2017 cohort, while the blue bars display the counterfactual average outcome expected under the algorithmic assignment, along with 95% confidence intervals computed via nonparametric bootstrap. Figures 1 and 2 highlight the impact of algorithmic assignment on employment, the target outcome used in the algorithmic assignment procedure, and how that impact varies across groups. As mentioned above, of particular interest for fairness considerations in the present context is the distribution of impact across gender, though results are also shown across age groups and educational levels.

As can be seen from Figure 1 substantial gains in employment are expected on average from algorithmic assignment. Under the status quo, the 2017 cohort achieved an employment rate of 12.3%, while under a counterfactual algorithmic assignment for this cohort, the estimated expected employment rate is 19.3%. However, are these average gains driven

disproportionately by one particular group or do they come at the expense of another? In particular, are male refugees further privileged under algorithmic assignment ahead of female refugees? Figure 2 allows for an assessment of this. For each group g, the red bars correspond to $\tau_{\Phi_0}(g)$, while the blue bars correspond to $\tau_{\Phi_1}(g)$. Quite clearly, the no group harm criterion is met across the board.

Figure 1: Overall Impact of Algorithmic Assignment on Employment



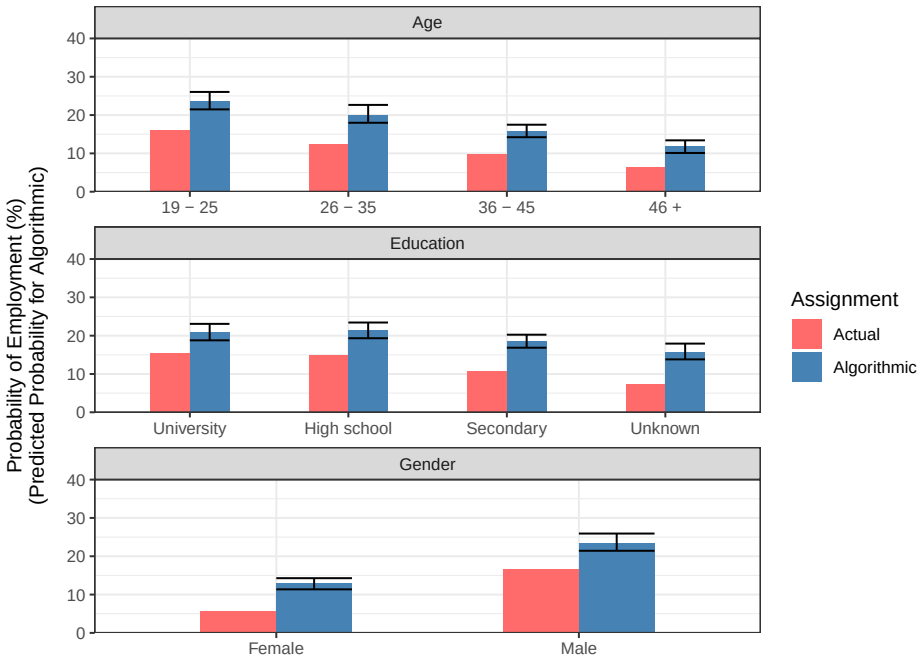
Focusing in more depth on gender, under the status quo, the employment rate was 5.5% for female refugees and 16.5% for male refugees, while under algorithmic assignment the expected employment is 12.9% and 23.4% respectively. That is, the algorithmic assignment would result in a gain of 7.4 percentage points for women and 6.9 percentage points for men. Comparing these gains in terms of absolute effects, it is immediately clear that they are very close in size. As described earlier, a more concrete and systematic comparison can be made by calculating the deviation from 1 of the ratio of these two effects and assessing whether the resulting value is deemed sufficiently small:

$$1 - \frac{min\Big([\tau_{\Phi_1}(g') - \tau_{\Phi_0}(g')], [\tau_{\Phi_1}(g'') - \tau_{\Phi_0}(g'')]\Big)}{max\Big([\tau_{\Phi_1}(g') - \tau_{\Phi_0}(g')], [\tau_{\Phi_1}(g'') - \tau_{\Phi_0}(g'')]\Big)} = 0.07$$

The resulting value is 0.07, which falls well below the value of 0.2 that could be motivated through precedent set by the "four-fifths rule. Further, also note that it is the disadvantaged group (women) that have the larger expected gain, which should further allay fairness concerns about these results.

The gains across gender could also be compared in terms of relative effects, taking into account the different baseline employment levels. For female refugees, the relative effect is $12.9/5.5 = 2.35$, while for male refugees, the relative effect is $23.4/16.5 = 1.42$. The absolute discrepancy between these two (0.93) is certainly nontrivial, owing to the fact that

Figure 2: Subgroup Impact of Algorithmic Assignment on Employment



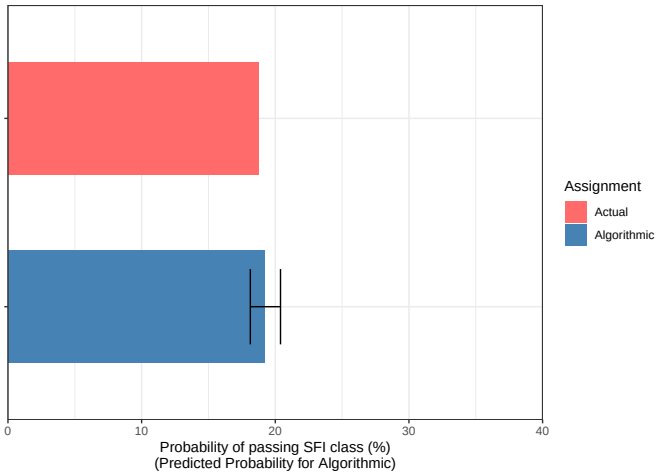
the absolute gains across gender are so similar and yet the baseline employment rates are so different. In this case, because it is the disadvantaged group (women) that are achieving the larger relative effect, and given the similarly sized absolute effects, this discrepancy in relative effects should not trigger any serious fairness concerns. The relevant policymakers and stakeholders could, of course, still decide that more equal relative effects across gender is desired, though note that this would imply increasing the gains for men and decreasing the gains for women (as will be described later).

C.3. Cross-Outcome Impact Distribution

Figures 3 and 4 focus on the unintended impact of algorithmic assignment on the additional outcome, completion of an SFI class, and its consequences for fairness across groups. The first apparent finding highlighted by Figure 3 is that optimizing the algorithmic assignment with respect to employment does not appear to have any meaningful effect on the overall average rate of SFI class completion. Under the counterfactual algorithmic assignment, the average rate is 19.2%, which is substantively close to and statistically indistinguishable from the baseline rate of 18.7% under the status quo.

However, turning to the cross-outcome distribution of impact at the group level, it appears from Figure 4 that the algorithmic assignment leads to opposite effects across gender. For female refugees the algorithmic assignment increases SFI class completion from 13.5% to 17.5% (a +4.0 percentage-point effect), while for male refugees the algorithmic assignment decreases SFI class completion from 22.0% to 20.2% (a -1.8 percentage-point effect),

Figure 3: Overall Unintended Impact of Algorithmic Assignment on SFI



with both of the effects being statistically significant. Clearly, this does not strictly meet the no cross-outcome group harm criterion, nor by extension any measure of cross-outcome impact balance. From that standpoint, the algorithmic assignment procedure in its current form leads to some undesirable behavior in this context, and hence modifications may be required (and will be discussed in the next section). At the same time, however, the sizes of these effects are substantively small, particularly the negative effect for men. The SFI class completion rate is also higher for men at baseline, and remains so under the counterfactual algorithmic assignment. And further, men are not the disadvantaged group. Thus, this small negative effect on SFI class completion for men may be tolerable in the larger picture—which includes impact on the core employment outcome of interest that is large, positive, and reasonably distributed across gender, along with a positive (albeit small) effect on SFI class completion for women.

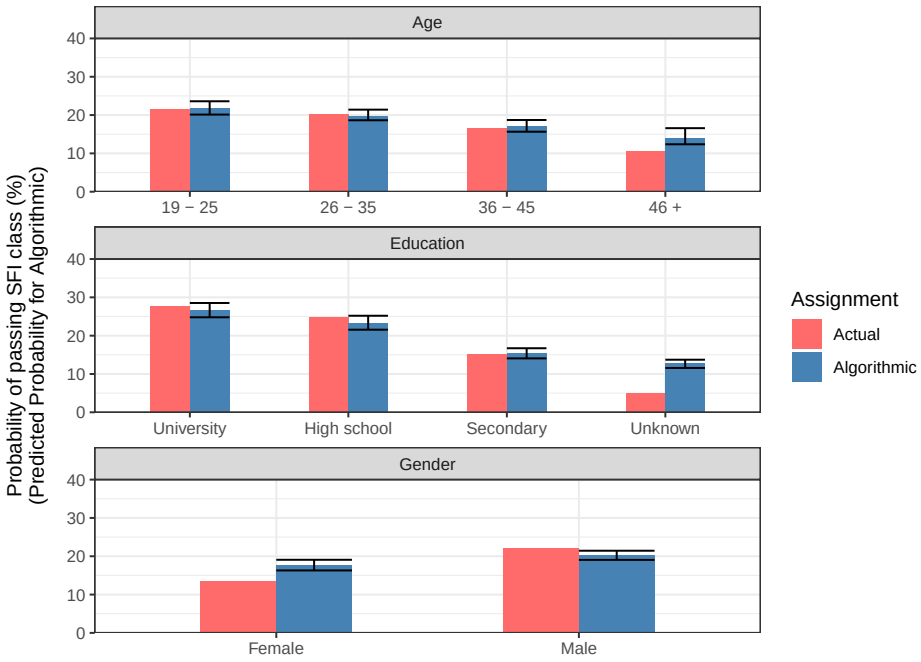
Whether or not such results on SFI (or any additional outcome of interest in any context) would be deemed tolerable would need to be determined by the careful judgment of the relevant policymakers and stakeholders. The framework presented in this paper for evaluating the distribution of impact provides a clear and intuitive set of results for thinking through these issues in a real-world context, as illustrated by the reasoned exposition in the previous paragraph.

IV. Discussion

A. Fairness Fixes

The framework presented above allows for an evaluation of the distribution of impact and the resulting fairness implications. The natural question that follows is what to do if the distribution of impact is deemed insufficiently fair across the relevant sub-groups. For instance, the results on employment above yield roughly equal gains in absolute terms across gender,

Figure 4: Subgroup Unintended Impact of Algorithmic Assignment on SFI



but the relative gains are much larger for women. Putting aside the fact that this disparity in the relative distribution of gains is less concerning from a fairness perspective than if the gains were larger for men, given that women are *a priori* disadvantaged, there may still be a desire to pursue a more equal distribution on these relative terms. There could also be other contexts or use cases in which the gains for the disadvantaged group are disproportionately smaller on relative and/or absolute terms. Are there steps that can be taken to address this issue and establish the desired fairness in the algorithmic decision-making procedure? The answer is yes. In particular, modifications should be made specifically to the algorithmic assignment procedure Φ_1.

Some of the previous literature on algorithmic fairness has argued for establishing fairness on the basis of modifying the predictive models themselves $\hat{\mu}(k, x)$. This perspective runs somewhat counter to core principles in statistics and estimation theory, where estimated models are meant to be the "best" (in some objective sense) approximations of the true underlying data-generating process that can be determined from the available empirical data. Instead, to satisfy certain conventional statistical definitions of fairness, interventions into this modeling process have been proposed, such as de-correlating sensitive variables from other variables in the data or even removing them entirely from the estimation process (e.g. Kusner et al., 2017). In doing so, this necessarily means that the resulting models will be worse approximations of the underlying data-generating process, though the justification is the possibility that the underlying data-generating process (i.e. what is happening in society) is biased and hence such modifications to the estimated models are a means of removing such

bias. However, is it necessarily the case that such modifications aimed at achieving some metric of fairness to be satisfied by predictive models, in and of themselves, will actually lead to greater fairness of impact in the real world? This turns out not to be the case, as such modifications have been shown in various situations to lead to worse outcomes, with the possibility of harming both advantaged and disadvantaged groups (Kleinberg et al., 2018b; Mullainathan, 2018).

Instead, rather than making modifications to $\hat{\mu}(k, x)$, in line with the underlying goal of achieving fairness of the distribution of actual impact, modifications should be made specifically to the algorithmic assignment procedure Φ_1. One straightforward modification to achieve distributive impact fairness with respect to the target outcome involves applying weights to the construction of Φ_1. That is, if the gains achieved under the algorithmic assignment procedure relative to the status quo are insufficiently balanced, more weight can be placed on the individuals belonging to the subgroup(s) whose gains are deemed too small. In the refugee matching case, for instance, the algorithmic assignment procedure would become the following:

$$\Phi_1 = \arg\max_{\Phi \in \Pi} \sum_{j=1}^{m} w_{G_j} \hat{\mu}_{j\phi_j}$$

where w_{G_j} denotes a positive weight to apply to the group to which individual j belongs. The counterfactual evaluations of the distribution of impact could then be iteratively re-computed on the basis of new weights for each group until the desired balance is met, and the resulting weighting rule can then be incorporated into the final version of Φ_1.

Note, however, that such a procedure would come at a cost. By deviating from equal weights across all individuals, this guarantees that the overall average gain will decrease, and if this performed to increase the gains for one group, this necessarily decreases the gains for another. In the application presented above, for instance, equalizing the relative gains across gender would necessarily require decreasing the gains among women.

Achieving impact distributional fairness with respect to both the target and additional outcomes would be more complicated, as there could prove to be unavoidable tradeoffs between the two depending upon the underlying data-generating processes. This would be a useful area for additional research. One approach that could be taken but would not be guaranteed to succeed would be to combine the different outcomes into a single index for use as the target outcome to be optimized upon, and then search for a weighting scheme that achieves a satisfactory distributional impact for both outcomes. A different possible approach would be to apply a more complex assignment procedure that explicitly considers both outcomes but does so separately. For instance, one could adapt the assignment mechanism developed in Acharya et al. (2021), which performs assignments of individuals to policy options (i.e. locations) on the basis of the individuals' preferences but subject to meeting a constraint with respect to average values for some outcome of interest. In the present case, preferences could be replaced by the target outcome values, and hence individuals would be assigned so as to optimize the target outcome but subject to constraints with respect to

the additional outcome (or outcomes). The mechanism would need to be further modified, however, to operate with respect to multiple sub-groups rather than simply on average.

B. Limitations

The primary limitation of the proposed framework is, of course, the need to estimate the results of counterfactual assignment realizations. Estimation of such causal effects requires strong assumptions that are often untestable and may need to be justified on the basis of subject matter expertise. It is for this reason that some scholars have proposed to focus on or reframe algorithmic assignment procedures into "prediction policy problems," where decisions can be made solely on the basis of predictions without needing to explicitly engage with causal phenomena (Kleinberg et al., 2015). For the various reasons described in this paper, however, fully understanding the distribution of ultimate impacts from algorithmic decision-making can only be achieved by taking causality seriously.

Fortunately, in many areas where algorithmic decision-making assistance is considered for deployment (or already being deployed), decisions are being applied to individuals on the basis of their observed characteristics X (and in some cases, perhaps even arbitrarily), without the possibility of self-selection. Such are situations in which the existing assignment procedure can be learned and hence precisely the cases where causal impact evaluation of the type proposed here is more feasible.[14] In contrast, reliable modeling of the counterfactual relationships at stake in many of the other causality-oriented definitions of algorithmic fairness (Kusner et al., 2017; Kilbertus et al., 2018; Nabi and Shpitser, 2018; Imai and Jiang, 2020) cannot be similarly motivated by knowledge of policy assignment mechanisms.

References

Acharya, A., Bansak, K., and Hainmueller, J. (2021). Combining outcome-based and preference-based matching: A constrained priority mechanism. *Political Analysis*, Forthcoming.

Ahani, N., Andersson, T., Martinello, A., Teytelboym, A., and Trapp, A. C. (2021). Placement optimization in refugee resettlement. *Operations Research*, Forthcoming.

Albrecht, C., Pérez, M. H., and Stitteneder, T. (2021). The integration challenges of female refugees and migrants: Where do we stand? In *CESifo Forum*, volume 22, pages 39–46.

Andersson, T., Ehlers, L., and Martinello, A. (2018). Dynamic refugee matching. Working Paper 2018:7, Lund University, Department of Economics.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica, May*, 23(2016):139–159.

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*, pages 507–547. University of Chicago Press.

[14] This is because in such cases, the conditional independence assumption $Y_j(k) \perp\!\!\!\perp A_j \mid X_j$ is defensible by design.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Bansak, K. (2020). A minimum-risk dynamic assignment mechanism along with approximations, extensions, and application to refugee matching. *arXiv preprint arXiv:2007.03069v2*.

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., and Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science*, 359(6373):325–329.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., and Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health affairs*, 33(7):1123–1131.

Bratsberg, B., Raaum, O., and Røed, K. (2014). Immigrants, labour market performance and social insurance. *The Economic Journal*, 124(580):F644–F683.

Brown, S., Davidovic, J., and Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, 8(1):2053951720983865.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., and Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5):124–27.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Chouldechova, A. and Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806.

Desmarais, S. L., Zottola, S. A., Duhart Clarke, S. E., and Lowder, E. M. (2021). Predictive validity of pretrial risk assessments: A systematic review of the literature. *Criminal Justice and Behavior*, 48(4):398–420.

Gäfvert, A., Ekström, B. M., and Vågen, A. (2018). Förbättra genomförandet av etableringsuppdraget. Technical report, the Swedish Public Employment Service.

Gölz, P. and Procaccia, A. D. (2019). Migration as submodular optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 549–556.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Heidari, H., Ferrari, C., Gummadi, K. P., and Krause, A. (2019). Fairness behind a veil of ignorance: a welfare analysis for automated decision making. *Advances in Neural Information Processing Systems 31*, pages 1265–1276.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hu, L. and Chen, Y. (2020). Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545.

Hubbard, A. E. and Van der Laan, M. J. (2008). Population intervention models in causal inference. *Biometrika*, 95(1):35–47.

Imai, K. and Jiang, Z. (2020). Principal fairness for human and algorithmic decision-making. *arXiv preprint arXiv:2005.10400*.

Joona, P. A., Lanninger, A. W., and Sundström, M. (2016). Reforming the integration of refugees: The swedish experience. Technical report, IZA Discussion Papers.

Jung, J., Shroff, R., Feller, A., and Goel, S. (2020). Bayesian sensitivity analysis for offline policy evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 64–70.

Kasy, M. and Abebe, R. (2021). Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586.

Kennedy, R., Waggoner, P., and Ward, M. (2021). Trust in public policy algorithms. *Journal of Politics*. Forthcoming.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2018). Avoiding discrimination through causal reasoning. In *31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*, pages 657–667. Curran Associates, Inc.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018a). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–95.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. R. (2018b). Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Kusner, M., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4069–4079.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Mousa, S. (2018). Boosting refugee outcomes: Evidence from policy, academia, and social innovation. *Academia, and Social Innovation (October 2, 2018)*.

Mullainathan, S. (2018). Algorithmic fairness and the social welfare function. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 1–1.

Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Olberg, N. and Seuken, S. (2019). Enabling trade-offs in machine learning-based matching for refugee resettlement. Technical report, University of Zurich.

Prop. (2017). Budgetpropositionen för 2018. Technical Report 1, the Swedish Ministry of Finance.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44.

Rambachan, A., Kleinberg, J., Ludwig, J., and Mullainathan, S. (2020). An economic perspective on algorithmic fairness. In *AEA Papers and Proceedings*, volume 110, pages 91–95.

Rawls, J. (2020). *A Theory of Justice*. Harvard University Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Samii, C., Paler, L., and Daly, S. Z. (2016). Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia. *Political Analysis*, pages 434–456.

Schultz-Nielsen, M. L. (2017). Labour market integration of refugees in denmark. *Nordic Economic Policy Review*, 7:55–90.

Sen, A. K. (2009). *The Idea of Justice*. Harvard University Press.

SMA (2019). Migrationsverkets Årsredovisning 2018. Technical report, The Swedish Migration Agency.

Supreme Court (2014). *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.* 135 S. Ct. 46, 576 U.S., 189 L. Ed. 2d 896.

Van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE.

Yang, E. and Roberts, M. E. (2021). Censorship of online encyclopedias: Implications for nlp models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 537–548.

Zhang, Y., Laber, E. B., Tsiatis, A., and Davidian, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, 71(4):895–904.

Zhao, Q. and Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281.

Appendix

A.3. Data

The data set was compiled by Statistics Sweden and contains individual level information collected from multiple authorities, such as the Migration Agency, the Tax Agency, and the Public Employment Service. We restrict our sample to adults (aged 19-60) who applied for asylum in Sweden (i.e. we exclude resettled refugees) and were given a residence permit based on refugee status or being in need of subsidiary protection. We focus on individuals who were assigned to a municipality 2016-2017, i.e. did not find housing on their own. We also exclude individuals who had previously been in Sweden (e.g. on a work permit) and drop individuals with missing information on their asylum application date or municipality assignment date.

Outcomes are measured at a yearly level, and in our case they refer to the year the individuals are assigned to municipalities. We look at two different outcomes:

1. Employment, defined as having any earnings
2. SFI, defined as having passed at least one SFI class

A.4. Tables

Table A1: Assigned refugees 2016-2017

	mean
Age	32.11
Woman	0.37
Married	0.56
Secondary school or less	0.43
High School	0.22
University	0.26
Family size	2.31
Syria	0.61
Eritrea	0.16
Afghanistan	0.06
Iraq	0.04
Work permit	0.50
Permanent residence permit	0.63
Any employment	0.10
Pass SFI	0.16
Observations	21520

Table A2: Balance test

	Unemployment share (non-EU)$_m$	ln(Population$_m$)	ln(co-nationals)$_m$
Woman	0.055	0.004	0.033
	(0.168)	(0.019)	(0.026)
Married	0.011	-0.040	-0.057
	(0.265)	(0.051)	(0.062)
25< Age <35	-0.178	0.044	0.024
	(0.221)	(0.043)	(0.060)
Age $\geq$ 35	0.030	0.029	0.055
	(0.273)	(0.042)	(0.060)
Secondary school or less	0.114	0.074	0.160
	(0.459)	(0.097)	(0.113)
High School	-0.357	0.103	0.154*
	(0.382)	(0.064)	(0.085)
University	-0.316	0.123	0.178
	(0.449)	(0.092)	(0.109)
Syria	0.189	-0.071	
	(0.303)	(0.045)	
Eritrea	-0.582	0.102	
	(0.592)	(0.165)	
Afghanistan	-0.195	-0.042	
	(0.527)	(0.064)	
Iraq	-0.041	-0.066	
	(0.503)	(0.057)	
Year FE	✓	✓	✓
Family size FE	✓	✓	✓
Country FE			✓
F-test	0.11	0.21	0.41
AdjR^2	0.01	0.01	0.20
Observations	21520	21520	21520

Note: Sample resticted to assigned adults 2016-2017. All outcomes are measured at the municipality level one year before assignment. Standard errors clustered at municipality. * p < 0.1, ** p < 0.05, *** p < 0.01.