Welfare-based Fairness through Optimization

Violet (Xinying) Chen and J. N. Hooker
Carnegie Mellon University
xinyingc@andrew.cmu.edu, jh38@andrew.cmu.edu

July 2021

Abstract

We propose optimization as a general paradigm for formalizing welfare-based fairness in AI systems. We argue that optimization models allow formulation of a wide range of fairness criteria as social welfare functions, while enabling AI to take advantage of highly advanced solution technology. In particular, we highlight that social welfare optimization supports a broad perspective on fairness motivated by general distributive justice considerations. We illustrate this advantage by reviewing a collection of social welfare functions that capture various concepts of equity. Most of these functions have tractable optimization formulations that can be efficiently solved by state-of-the-art methods. To further demonstrate the potentials of social welfare optimization in AI, we show how to integrate optimization with rule-based AI and machine learning, and outline research directions to explore for practical implementation of integrated methods.

1 Introduction

Artificial intelligence is increasingly used not only to solve problems, but to recommend action decisions that range from awarding mortgage loans to granting parole. The prospect of making decisions immediately raises the question of ethics and fairness. If ethical norms are to be incorporated into artificial decision making, these norms must somehow be automated or formalized. The leading approaches to this challenge include

- *value alignment*, which strives to train or modify AI systems to reflect human ethical values automatically, e.g. Allen et al. [2005], Russell [2019], Gabriel [2020];
- *logical formulations* of ethical and fairness principles that attempt to represent them precisely enough to govern a rule-based AI system, e.g. Bringsjord et al. [2006], Lindner et al. [2020], Hooker and Kim [2018]; and

- *statistical fairness metrics* that aim to ensure that benefits are allocated equitably in the decision process, e.g. Dwork et al. [2012], Mehrabi et al. [2019], Chouldechova and Roth [2020].

Each of these approaches can be useful in a suitable context. We wish to propose, however, an alternate framework for formalizing ethics and fairness that has received less attention:

- *optimization*, which allows one to achieve equity or fairness by maximizing a *social welfare function*.

Welfare economics has long used social welfare functions (SWFs) as a tool to measure the desirability of a given distribution of benefits and harms. A SWF is a function of the utility levels allocated to affected parties, where utility reflects a party's gain or loss as a consequence of the decisions of interest. Using a SWF motivates explicit consideration of the downstream outcomes of fairness and equity criteria. In contrast to leading notions of AI fairness that focus on eliminating disparity between groups, SWFs allow a broader perspective that emphasizes fairness in the welfare impacts of decisions.

AI research is beginning to recognize the importance of a welfare perspective on fairness (e.g., Corbett-Davies and Goel [2018], Hu and Chen [2020]), due in part to its potential for aligning fairness concepts with social well-being. Despite this rising attention, there is no general framework for incorporating welfare-based fairness into AI systems. In this paper, we utilize social welfare optimization as the core component of one possible framework. This framework allows one to take advantage of the flexibility of SWFs to represent a wide range of fairness and equity concepts, as well as to harness powerful optimization solvers. Optimization methods are of course already employed in AI to train neural networks, calibrate machine learning models, and the like. Our proposal is to bring fairness under the optimization umbrella.

We begin below by stating some specific advantages of social welfare optimization as a paradigm for implementing equity and fairness in AI. We then state the general optimization problem and its potential for solution by advanced mathematical programming software. Following this, we introduce as a running example the mortgage loan processing problem that is often discussed in an AI fairness context, and we review some previous work on social welfare optimization in both the operations research and AI communities. We then examine several SWFs to illustrate how they can capture a variety of fairness concepts. We indicate how they correspond to mathematical programming models and assess their suitability for the mortgage problem in particular. We conclude by outlining a general framework and research program for social welfare optimization as a basis for formalizing fairness in AI.

2 Advantages of Optimization

The optimization of social welfare functions offers several advantages as a framework for incorporating fairness into AI.

- Social welfare functions provide a *broader perspective on fairness* than can be achieved by focusing exclusively on bias and concepts of parity across groups. They not only have the flexibility to represent a wide range of fairness concepts, but they encourage modelers to take into account the overall welfare of those affected. While AI-based decision making already strives to maximize predictive accuracy, a welfare perspective allows it to consider explicitly the more general benefits that accurate predictions can deliver, as well as whether the benefits are distributed justly.

- Social welfare functions allow one to *balance equity and efficiency* in a principled way. Where equity is an issue, there is often a desire for efficiency as well. A social welfare approach obliges one to consider how equity and utilitarian goals should be represented and balanced when one chooses the function to be maximized. One can of course maximize efficiency subject to a constraint on some measure of inequity, but this provides no principled way of regulating the trade-off between the two.

- Optimization models allow one to harness *powerful optimization methods*, which have been developed and refined over a period of 80 years or more. A wide variety of social welfare functions can be formulated for solution by highly advanced linear, nonlinear, and mixed integer programming solvers. We provide examples in Section 6.

- Optimization models offer enormous flexibility to *include constraints on the problem*. Decisions are normally made in the context of resource constraints or other limitations on possible options. These can be represented as constraints in the optimization problem, as nearly all state-of-the-art optimization methods are designed for constrained optimization. Also, a complex social welfare function can often be simplified by adding constraints to the optimization problem, resulting in a problem that is easier to solve.

3 The Basic Optimization Problem

The general problem of maximizing social welfare can be stated

$$\max_{\boldsymbol{x}} \left\{ W\big(\boldsymbol{U}(\boldsymbol{x})\big) \;\middle|\; \boldsymbol{x} \in S_{\boldsymbol{x}} \right\} \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a vector of resources distributed across stakeholders $1, \ldots, n$, and $S_{\boldsymbol{x}}$ is the set of feasible values of $\boldsymbol{x}$ permitted by resource limits and other constraints. $\boldsymbol{U} = (U_1, \ldots, U_n)$ is a vector of *utility functions*, where $U_i(\boldsymbol{x})$ defines the utility experienced by stakeholder i as a result of the resource distribution $\boldsymbol{x}$. We can normally write $U_i(\boldsymbol{x})$ as $U_i(x_i)$, since a stakeholder's utility typically depends only on the resources allotted to that stakeholder. Finally, $W(\boldsymbol{u})$ is a *social welfare function* that measures the desirability of a vector $\boldsymbol{u}$ of utilities. Problem (1) maximizes social welfare over all feasible resource allocations.

In practice, it is often convenient to model the utility functions $\boldsymbol{U}$ using constraints, because this results in problems better suited for optimization solvers. One therefore writes (1) as

$$\max_{\boldsymbol{x},\boldsymbol{u}} \left\{ W'(\boldsymbol{u}) \;\middle|\; (\boldsymbol{x},\boldsymbol{u}) \in S_{\boldsymbol{xu}} \right\} \tag{2}$$

where $\boldsymbol{u}$ is a vector of utilities, and $S_{\boldsymbol{xu}}$ is defined so that $(\boldsymbol{x},\boldsymbol{u}) \in S_{\boldsymbol{xu}}$ implies $\boldsymbol{x} \in S_{\boldsymbol{x}}$ and $\boldsymbol{u} = \boldsymbol{U}(\boldsymbol{x})$. The function W' is a possibly simplified version of W that yields an equivalent optimization problem due to constraints defining $S_{\boldsymbol{xu}}$.

To simplify exposition, we assume that the original problem constraints that define $S_{\boldsymbol{x}}$ consist of (or can be approximated by) a system of linear inequalities and equations. Thus, for example, when we say that (1) is a linear programming (LP) problem for a given W, we mean that (1) can be formulated as an LP problem (2) when $S_{\boldsymbol{x}}$ is defined by a linear system. The linearity assumption actually allows a great deal of modeling flexibility, because $S_{\boldsymbol{xu}}$ can be approximated by linear constraints whenever $S_{\boldsymbol{x}}$ is convex and $\boldsymbol{U}(\boldsymbol{x})$ is a concave function of $\boldsymbol{x}$. The latter occurs in the common situation where $\boldsymbol{U}$ is linear or represents decreasing returns to scale.

All of the SWFs considered here can be formulated as linear, nonlinear, or mixed integer programming problems for which advanced solution technology exists. An LP model optimizes a linear function over continuous variables, subject to linear inequality constraints. The problem is *extremely* well solved. Nonlinear programming (NLP) models optimize a nonlinear function over continuous variables, subject to linear or nonlinear inequality constraints. All the NLP models considered here are relatively easy to solve. Mixed integer/linear programming (MILP) models are LP problems except that some variables must take integer values. They are combinatorial in nature, but state-of-the-art software frequently solves industrial instances with thousands of discrete variables.

If some of the original problem variables x_i are discrete, an otherwise LP problem becomes an MILP problem, and an NLP problem becomes a mixed integer/nonlinear programming (MINLP) problem. The latter can be quite hard to solve. An MILP problem of course remains an MILP problem.

4 Example: Mortgage Loans

We use mortgage loan processing as a running example, as it is a much-discussed application of AI-based decision making. Issues of fairness arise when an AI system is more likely to deny loans to members of certain groups, perhaps reflecting minority status or gender. A frequently used remedy is to apply statistical bias metrics to detect the problem and adjust the decision algorithms in an attempt to solve it.

Yet bias is only one element of a broader decision-making context. For one thing, there is a clear utilitarian imperative. The reason for automating mortgage decisions in the first place is to predict more accurately who will default, because defaults are costly for the bank and devastating to home

buyers. The desire for accurate prediction is, at root, a desire to maximize utility. Furthermore, bias is regarded as unfair in large part because it reduces the welfare of a segment of society that is already disadvantaged. An aversion to bias is, to a great degree, grounded in a desire for distributive justice in general. All this suggests that loan decisions should be designed to achieve what we really want: efficiency and distributive justice, rather than focusing exclusively on predictive accuracy and group parity.

The social welfare function W in (1) should be selected to balance efficiency and equity in a suitable fashion; we consider some candidate SWFs in Section 6. The stakeholders $1, \ldots, n$ might include the loan applicants, the bank, the bank's stockholders, and the community at large. The utility function $\boldsymbol{U}$ converts a given set of loan decisions $\boldsymbol{x} = (x_1, \ldots, x_n)$ to a vector of expected utilities $\boldsymbol{u} = (u_1, \ldots, u_n) = \boldsymbol{U}(\boldsymbol{x})$ that the stakeholders experience as a result. Since granting a loan is a yes-or-no decision, we can define x_i to be a binary variable with $x_i = 1$ if applicant i receives a loan (we fix $x_i = 0$ if i is a stakeholder other than an applicant). The utility measure $u_i = U_i(x_i)$ for applicant i could depend on the applicant's financial situation as well as the amount of the loan, as for example when the marginal value of a loan dollar is greater for an applicant who is less well-off. The SWF can reflect a preference for granting loans to disadvantaged applicants even when they have a somewhat higher probability of default, so as to ensure a more just distribution of utility. This could have the effect of avoiding bias against minority groups, but as part of a more comprehensive assessment of social welfare.

This framework can be applied to an AI-based decision-making context in several ways. Machine learning can estimate the probability p_i of default for a given applicant i, based on available data, and that estimate would feed into the expected utility $\bar{u}_i$ that results from granting the loan. In particular, we would have $\bar{u}_i = p_i u_i^1 + (1 - p_i) u_i^0$, where u_i^1 is the utility that results if i repays the loan and u_i^0 if i defaults. If $\bar{v}_i$ is applicant i's utility without a loan, we have $U_i(x_i) = \bar{v}_i + (\bar{u}_i - \bar{v}_i)x_i$. When confronted with a batch of loan decisions, the bank could maximize $W(\boldsymbol{U}(\boldsymbol{x}))$ subject to a constraint $\sum_i c_i x_i \leq B$ on the funds available (where c_i is the requested loan amount) and perhaps other constraints. Another option is for the bank to solve the optimization problem in advance, before particular applicants are considered. It would maximize $W(\boldsymbol{U}(\boldsymbol{x}))$ over a set of hypothetical applicants i corresponding to various financial profiles, again using ML-based default probabilities as input. In this case, the utility $U_i(x_i)$ that accrues to a potential applicant type would depend in part on the estimated number of applicants in the population that have the corresponding profile. Then when someone with financial profile i applies for a loan, the bank would award the loan if $x_i = 1$ in the optimal solution of the welfare maximization problem. We will later refer to these two options as examples of *post-processing* integration of machine learning and social welfare optimization, because the fairness element is injected after the learning phase.

In-processing integration can be achieved by incorporating social welfare into the actual training of the machine learning system, which the bank could use to predict whether a loan application should be approved. The training

algorithm would maximize a SWF rather than predictive accuracy—with the understanding that predictive accuracy is a major determinant of social welfare. Using this approach, the output of the ML system for a particular applicant would already reflect general welfare and fairness concerns.

5 Previous Work

Social welfare optimization is already fairly well established in the operations research literature, and it is beginning to attract interest in the AI community. Our proposal is that AI expand these initial efforts into a general research program for formulating fairness. We review here some of the previous work in both literatures.

An excellent survey of equity models used in operations research is provided by Karsu and Morton [2015]. We mention a few examples that combine equity and efficiency. Bandwidth allocation in telecommunication networks is a popular application studied in early works on fair resource allocation (Luss [1999], Ogryczak and Śliwiński [2002], Ogryczak et al. [2008]). For problems in this domain, a standard setup is to interpret bandwidth as utility and define a SWF that is consistent with a Rawlsian maximin criterion. The corresponding optimization problem seeks equitable allocations that optimize the worst performance among activities or services that compete for bandwidth. Project assignment is another application where fairness is often relevant, as the involved stakeholders may have different preferences over projects. For instance, Chiarandini et al. [2019] work with a real-life decision to assign projects to university students. They use student rankings of projects as utilities and study a variety of SWFs that capture different fairness-efficiency balancing principles. Fair optimization has also received attention in humanitarian operations. Eisenhandler and Tzur [2019] study an important logistical challenge in food bank operations, food pickup and distribution. They design a routing resource allocation model to seek both fair allocation of food to different agencies and efficient delivery of as much food as possible. The utilities of agencies are measured by the amount of food delivered. An SWF is selected to combine utility and the Gini coefficient. Mostajabdaveh et al. [2019] consider a disaster preparation task of selecting shelter locations and assigning neighborhoods to shelters. They choose a SWF that combines the Gini coefficient with neighborhood utilities based on the travel distances to their assigned shelter.

Recent AI research has developed efficient algorithms that take fairness into account. This effort is not directly comparable to our proposal in that it develops algorithms to solve specific problems that have a fairness component, rather than formulating optimization models that can be submitted to state-of-the-art software. Algorithmic design tasks are often associated with fair matching decisions, such as kidney exchange McElfresh and Dickerson [2018], paper-reviewer assignment in peer review Stelmakh et al. [2019], or online decision procedures for a complex situation such as ridesharing Nanda et al. [2020].

Fair machine learning is a rapidly growing field in recent years. Fair ML

methods in literature can be categorized as pre-, in-, or post-processing, which respectively seek fairness by modifying standard ML methods before, during, or after the training phase. The majority of fair ML methods seek to eliminate bias and discrimination in standard ML models, via fairness notions that measure certain type of disparity in the generated predictions. Many of these methods rely on optimization in the fairness-seeking components. Pre-processing methods can use optimization models to find the best data modifications to the training data to prevent bias and disparity (see e.g. Zemel et al. [2013], Calmon et al. [2017]). Similarly, post-processing methods can use optimization models to determine the optimal tuning rules to adjust the predictions generated from the trained model to seek fairness (see e.g. Hardt et al. [2016], Alabdulmohsin [2020]). Moreover, fairness through optimization fits naturally into in-processing methods, which modify standard ML models by adding fairness constraints or including fairness components in objective function (see e.g. Zafar et al. [2019], Olfat and Aswani [2018], Donini et al. [2018].

Different from this dominant statistical view of fairness, an emerging research thread advocates welfare-based fairness in ML to seek better compatibility between fair ML and distributive justice. This is in line with our proposal of using social welfare functions to capture a broader perspective on fairness. We next discuss a few representative papers in this thread, and review their chosen utility and social welfare definitions. Heidari et al. [2018] consider a standard supervised learning setting with true labels $\{y_i\}$ and predicted labels $\{\hat{y}_i\}$. They define the utility function as a function of $y_i, \hat{y}_i$, and the specific format is chosen to reflect whether i is risk averse, neutral or seeking, and how close the predicted outcome $\hat{y}_i$ is to i's desirable outcome. They then define a utilitarian sum of these individual utilities as the social welfare measure, and propose to add a constraint on this social welfare value to standard ML models as an in-processing fair ML approach. Hu and Chen [2020] study a similar utility definition without the risk component in a classification setup. They evaluate the overall welfare associated with classification decisions through comparing a vector of welfare values, which measure the utilitarian welfare by group. Also in a classification setting, Corbett-Davies and Goel [2018] suppose each group has fixed benefits and costs associated with classification outcomes, and these values are used as parameters in the utility functions. A group's utility aggregates the benefits and costs that individuals of the group incur from their classification outcomes. A more refined view of utility is studied in Heidari et al. [2019]: they partition one's actual utility into an effort-based component and an advantage component. Utilizing this partition, they group individuals by effort-based utilities and propose a fairness measure equivalent to the expected advantage utility of the worst-off group.

6 A Sampling of Social Welfare Functions

We briefly review a collection of SWFs to illustrate how they can embody various conceptions of equity. For each, we indicate the type of optimization model

it yields, and whether it is appropriate for our running example of mortgage loan processing. We classify the SWFs as pure fairness metrics, functions that combine fairness and efficiency, and statistical fairness metrics.

6.1 Pure fairness measures

Social welfare functions that measure fairness alone, without an element of efficiency, are of two basic types: inequality metrics and fairness for the disadvantaged.

Inequality metrics abound in the economics literature. Some simple ones are represented by the following SWFs (which negate the inequality measure):

$$W(\boldsymbol{u}) = \begin{cases} -(1/\bar{u})(u_{\max} - u_{\min}) & \text{for the } \textit{relative range} \\ -(1/\bar{u}) \sum_i |u_i - \bar{u}| & \text{for the } \textit{relative mean deviation} \\ -(1/\bar{u}) \Big[(1/n) \sum_i (u_i - \bar{u})^2 \Big]^{\frac{1}{2}} & \text{for the } \textit{coefficient of variation} \end{cases}$$

There is also the well-known *Gini coefficient*, which is proportional to the area between the Lorenz curve and a diagonal line representing perfect equality. It corresponds to the SWF

$$W(\boldsymbol{u}) = 1 - \frac{1}{2\bar{u}n^2} \sum_{i,j} |u_i - u_j|$$

Although these SWFs are nonlinear, all but the coefficient of variation have LP models. The coefficient of variation has a convex quadratic programming model with linear constraints, for which there are very efficient specialized solvers.

Other fairness-based SWFs are concerned with the lot of the disadvantaged. The *Hoover index* measures the fraction of total utility that would have to be transferred from the richer half of the population to the poorer half to achieve perfect equality. The SWF is

$$W(\boldsymbol{u}) = -\frac{1}{2n\bar{u}} \sum_i |u_i - \bar{u}|$$

The Hoover index is proportional to the relative mean deviation and can therefore be optimized using the same LP model.

The *McLoone index* compares the total utility of individuals at or below the median utility to the utility they would enjoy if all were brought up to the median utility. The index is 1 if nobody's utility is strictly below the median and approaches 0 if there is a long lower tail. The SWF is

$$W(\boldsymbol{u}) = \frac{1}{|I(\boldsymbol{u})|\tilde{u}} \sum_{i \in I(\boldsymbol{u})} u_i$$

where $\tilde{u}$ is the median of utilities in $\boldsymbol{u}$ and $I(\boldsymbol{u})$ is the set of indices of utilities at or below the median. The McLoone index can be optimized in an MILP model.

The Hoover and McLoone indices measure only the relative welfare of disadvantaged parties, and not their absolute welfare. The *maximin* criterion addresses both. It is based on the Difference Principle of John Rawls, which states that inequality should exist only to the extent it is necessary to improve the lot of the worst-off (Rawls [1999], Freeman [2003], Richardson and Weithman [1999]). It can be plausibly extended to a lexicographic maximum principle. The SWF is simply

$$W(\boldsymbol{u}) = \min_i \{u_i\}$$

and has an LP model.

Purely fairness-oriented SWFs can be used when equity is truly the only issue of concern. In particular, they are unsuitable for the mortgage problem, where overall utility is a prime consideration.

6.2 Combining fairness and efficiency

Several SWFs combine equity and efficiency, sometimes with a parameter that regulates the relative importance of each. Perhaps the best known is *alpha fairness*, for which the SWF is

$$W_\alpha(\boldsymbol{u}) = \begin{cases} \dfrac{1}{1-\alpha} \displaystyle\sum_i u_i^{1-\alpha} & \text{for } \alpha \geq 0,\ \alpha \neq 1 \\ \displaystyle\sum_i \log(u_i) & \text{for } \alpha = 1 \end{cases}$$

Larger values of α imply a greater emphasis on equity, with $\alpha = 0$ corresponding to a pure utilitarian criterion $\sum_i u_i$, and $\alpha = \infty$ to a pure maximin criterion. An important special case is $\alpha = 1$, which corresponds to *proportional fairness*, also known as the *Nash bargaining solution*. It is widely used in telecommunications and other engineering applications. Both proportional fairness and alpha fairness have been given axiomatic and bargaining justifications (Nash [1950], Harsanyi [1977], Rubinstein [1982], Binmore et al. [1986], Lan et al. [2010]). The alpha fairness SWF is irreducibly nonlinear, but because it is concave for all α, it can be maximized with reasonable efficiency by NLP methods.

Alpha fairness is conceptually a reasonable choice for the mortgage problem, because the bank can obtain any desired balance between utility and fairness by adjusting α. While it is difficult to justify to stakeholders any particular choice for the value of α, a perceived bias against minorities can always be addressed by increasing α. On the other hand, the presence of 0–1 variables x_i produces an MINLP model, which can be hard to solve. Thus alpha fairness may be practical only for problems with at most a few hundred applicants.

The *Kalai-Smorodinsky* (K–S) bargaining solution, proposed as an alternative to the Nash bargaining solution, minimizes each person's relative concession. That is, it provides everyone the largest possible utility relative to the maximum

one could obtain if other players are disregarded, subject to the condition that all persons receive the same fraction β of their maximum. In addition to the bargaining justification of Kalai and Smorodinsky [1975], this approach has been defended by Thompson [1994] and is implied by the contractarian philosophy of Gautier [1983]. The SWF can be formulated

$$W(\boldsymbol{u}) = \begin{cases} \sum_i u_i, & \text{if } \boldsymbol{u} = \beta \boldsymbol{u}^{\max} \text{ for some } \beta \text{ with } 0 \leq \beta \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

where $u_i^{\max} = \max_{(\boldsymbol{x},\boldsymbol{u}) \in S_{\mathbf{xu}}} u_i$ for each i. It can be optimized by maximizing β subject to $\boldsymbol{u} = \beta \boldsymbol{u}^{\max}$ and $\beta \leq 1$, an easy LP problem.

The K–S criterion cannot be used for the mortgage problem, because the role of 0–1 variables in the problem almost ensure that the optimization model will be infeasible. Since $u_i^{\max} = \bar{u}_i^1$, we must have $U_i(x_i) = \beta u_i^1$ for all i. But $U_i(x_i) = \bar{v}_i + (\bar{u}_i - \bar{v}_i)x_i$, which means that there must be a ratio β that, for each i, is equal to either $\bar{v}_i/u_i^1$ or $\bar{u}_i/u_i^1$ (which correspond to setting $x_i = 0$ or $x_i = 1$, respectively). It is very unlikely that the problem data will have this property.

Williams and Cookson [2000] suggest two *threshold* criteria for combining maximin and utilitarian objectives in a 2-person context. One uses maximin until the cost of fairness becomes too great, whereupon it switches to utilitarianism, and the other does the opposite. Hooker and Williams [2012] generalize the former to n persons by proposing the following SWF:

$$W_\Delta(\boldsymbol{u}) = (n-1)\Delta + \sum_{i=1}^{n} \max\left\{u_i - \Delta, u_{min}\right\}$$

where $u_{\min} = \min_i\{u_i\}$. The parameter Δ regulates the equity/efficiency tradeoff in a way that may be easier to interpret in practice than the α parameter: parties whose utility is within Δ of the lowest utility receive special priority. Thus the disadvantaged are favored, and Δ defines who is disadvantaged. As with the α parameter, $\Delta = 0$ corresponds to a purely utilitarian criterion and $\Delta = \infty$ to a maximin criterion. Hooker and Williams provide an MILP model of the SWF and show that it is sharp (i.e., its continuous relaxation describes the convex hull of its feasible set). Partly for this reason, they found that the model solves rapidly in computational tests.

This threshold approach is a reasonable choice for the mortgage problem. Since the problem has discrete variables regardless of the SWF used, the MILP-based threshold formulation adds relatively little complexity to the problem. In addition, loan officers can specify in a meaningful way when an applicant is to be considered disadvantaged, by selecting an appropriate value of Δ.

One weakness of the model is that the actual utility levels of disadvantaged parties other than the very worst-off have no effect on the measurement of social welfare, as long as those utilities are within Δ of the lowest. As a result, the socially optimal solution may not be as sensitive to equity as one might desire. Chen and Hooker [2020a,b] address this issue by *combining utilitarianism with a*

leximax rather than a maximin criterion. A leximax (lexicographic maximum) solution is found by first maximizing the lowest utility, then while holding it fixed, maximizing the second lowest utility, and so forth. Chen and Hooker combine leximax and utilitarian criteria by maximizing a sequence of threshold SWFs that have tractable MILP models. Their approach may yield more satisfactory solutions of the mortgage problem.

6.3 Statistical bias metrics

While we argue that bias metrics afford an overly narrow perspective on fairness, they nonetheless can be expressed as SWFs if desired. The utility vector $\boldsymbol{u}$ becomes simply a binary vector in which $u_i = 1$ if individual i is selected for some benefit, and $u_i = 0$ otherwise. In the mortgage example, the benefit is a mortgage loan. We set constant $a_i = 1$ when person i actually qualifies for selection (as for example when person i in the mortgage training set repaid the loan), and $a_i = 0$ otherwise. Two groups are compared, respectively indexed by N and N'. One is a protected group, such as a minority subpopulation, and the other consists of the rest of the population.

For example, *demographic parity* has the SWF

$$W(\boldsymbol{u}) = 1 - \left| \frac{1}{|N|} \sum_{i \in N} u_i - \frac{1}{|N'|} \sum_{i \in N'} u_i \right|$$

Equalized odds can be measured in two ways, one of which is *equality of opportunity*:

$$W(\boldsymbol{u}) = 1 - \left| \frac{\sum_{i \in N} a_i u_i}{\sum_{i \in N} a_i} - \frac{\sum_{i \in N'} a_i u_i}{\sum_{i \in N'} a_i} \right|$$

Another SWF represents *accuracy parity*:

$$W(\boldsymbol{u}) = 1 - \left| \frac{1}{|N|} \sum_{i \in N} \big(a_i u_i + (1 - a_i)(1 - u_i)\big) - \frac{1}{|N'|} \sum_{i \in N'} \big(a_i u_i + (1 - a_i)(1 - u_i)\big) \right|$$

and still another *predictive rate parity*:

$$W(\boldsymbol{u}) = 1 - \left| \frac{\sum_{i \in N} a_i u_i}{\sum_{i \in N} u_i} - \frac{\sum_{i \in N'} a_i u_i}{\sum_{i \in N'} u_i} \right|$$

The computational challenge varies widely across the various bias-oriented SWFs. The first three SWFs above give rise to linear models (which become MILP models due to the 0–1 restriction on u_i), while the last produces an extremely difficult nonconvex MINLP model.

Bias measures are inappropriate as social welfare objectives for the mortgage problem, because they take no account of efficiency. One can, of course, maximize predictive accuracy subject to constraints on the amount of bias, but this has a number of drawbacks:

- As previously argued, it provides a very limited perspective on the utility actually created by decisions. Indeed, the utility vector consists only of 0–1 choices.
- There is no consensus on which bias measure is suitable in a given context, if any. Bias measures were developed by statisticians to measure predictive accuracy, not to assess fairness.
- There is no principle for balancing equity and efficiency. If equity is one of the *objectives*, it should be part of the *objective function*. The choice of that function obliges one to justify the equity/efficiency trade-off mechanism in a transparent manner.
- Bias measurement forces one to identify *a priori* which individuals in a training set should be selected for benefits (as indicated by a_i). In a social welfare approach, no prior decisions of this kind are necessary.
- Bias measurement forces one to designate "protected groups" (as indicated by the index set N). There is no clear principle for selecting which groups should be protected, unless one is content simply to recognize those mandated by law.

7 Welfare-based Fairness: A General Framework

In practical applications, we can rely solely on a fully specified optimization model to determine optimal decisions with respect to the selected notion of fairness and social welfare; in fact, this is the standard approach in the optimization literature we have reviewed. We discuss in the mortgage loan example that specifying the model might require estimation and prediction tools such as machine learning models. Therefore, a more flexible and realistic use case is to integrate the optimization-for-fairness paradigm with general AI methods.

Drawing motivation from both settings, we formalize a general framework for designing AI systems with welfare-based fairness guarantees. Through this framework, we hope to streamline the process of making fair decisions to attain the desirable welfare outcomes, and distinguish the role played by social welfare optimization from the other components. This framework consists of three steps, which we explain in detail as follows.

7.1 Step 1: Specify decision problem

We begin by specifying the needed components of the decision problem. This step is critical for the success of later steps as it ensures we have a precise understanding of the problem scope and context. We highlight some key components that commonly exist in problem instances. Note that additional factors may be needed in specific problems.

- Task: the task refers to the decision actions in question. To specify the task is to describe the downstream actions and identify the resources to allocate. In our running example, the bank's task is to decide whether to grant loans to applicants.

- Stakeholders: stakeholders are individuals or groups directly or indirectly affected by the decisions, namely, they are the utility recipients in the problem. When a decision involves a wide variety of stakeholders that is impractical to all be considered, it is important to select stakeholders in a compatible manner with the chosen task and goals. For instance, in the running example, loan applicants and the bank are necessary stakeholders, and optional choices include stockholders and the community at large.

- Goals: we characterize the desirable outcomes as goals of the decision problem. A high-level structure is to define separate fairness and welfare goals. As an example, when allocating mortgage loans, the bank's fairness goals may include prioritize applicants in need to improve their access to opportunity; its welfare goals may include guarantee a sufficiently low loan default rate to benefit the bank and its stockholders, and seek an efficient use of loan funds to benefit the local community. These goals will later serve as the guiding principles for defining the social welfare function $W(\boldsymbol{u})$ to be used as the objective function.

- Constraints: these are restrictions in the problem context that limit which actions are feasible, namely, we specify constraints to define the domain $S_{\mathbf{x}}$. A main source of restriction is the scarcity of resources, for example, the bank is subject to a budget constraint. In addition, the decision contexts may impose constraints on actions, for instance, the loan allocated to an applicant should not exceed the requested amount.

7.2 Step 2: Define utility and social welfare functions

With a clear problem statement, we continue to define utility functions and social welfare functions. Both definitions need to be compatible with the decision task and goals. The utility function is a function of the decision actions: the utility value indicates the degree of preference a party has for its assigned outcome, namely, when i is better off under $\mathbf{x}$ in comparison to $\hat{\boldsymbol{x}}$, we should have $U_i(\mathbf{x}) > U_i(\hat{\boldsymbol{x}})$. Depending on the problem context, the utility function may be dependent on a single component, such as, assign a fixed positive utility for receiving a positive classification and zero otherwise. Alternatively, the utility value can aggregate multiple components relevant to an individual's well-being, such as, in the mortgage example, we can define utility to incorporate negative cost and positive wealth impacts from the loan decision.

A social welfare function evaluates the desirability of an outcome $\boldsymbol{x}$ via its corresponding utility distribution $\boldsymbol{u}$. The selected SWF should capture the decision goals in a way that a greater social welfare value indicates improvement

in achieving the goals. As we discuss in Section 6, the mortgage problem requires a SWF combining fairness and efficiency.

7.3 Step 3: Develop decision models

In the previous steps, we identify the components needed to formulate the social welfare optimization problem (2). We next distinguish two types of information context that call for different schemes for developing optimization-based decision models.

Full Information: Integration with Rule-based AI

The first context occurs when we have the full information necessary to state the social welfare optimization problem. Specifically, we know all the parameters needed in the utility function, constraints and the social welfare function. This information may be available from past data, or provided by experts utilizing their domain knowledge. For instance, in the mortgage example, the bank may hire experts to evaluate parameter values based on their expertise. With a fully defined optimization model, we can directly solve for the optimal solution and make decisions based on the obtained solution. Recall from our literature review that this case has been broadly studied in the optimization literature.

More broadly, the full information context is suitable for integration by means of rule-based AI, which utilizes a set of rules to encode knowledge relevant to the decision and to produce pre-defined outcomes. Rule-based systems are increasingly recognized for their capacity to support principled and transparent AI in various application domains. For instance, Brandom [2018] observes the trend in autonomous vehicle industry whereby "companies have shifted to rule-based AI, an older technique that lets engineers hard-code specific behaviors or logic into an otherwise self-directed system." Moreover, Kim et al. [2021] demonstrate that ethical principles can be precisely represented as rules to include in an AI system. In fact, they suggest that a rule-based formulation is necessary for making ethical decisions.

We highlight two possible schemes to implement the integration of rule-based AI and optimization. The first method is to use the optimization problem to guide the selection of rules to encode into the AI system, then rely on the rule-based system to make decisions. To illustrate in the mortgage example, we suppose the bank pre-defines several classes of applicants and wishes to specify rules on whether to approve loans from each class. The bank can determine these rules using an optimization model which contains a 0-1 variable to denote the loan approval status (yes or no) for each class, and optimizes a social welfare objective function defined with historical data about the loan applications, decisions and default outcomes for all classes. We then use the solved optimal solutions to state the decisions rules for each class. Such a rule-based system is straightforward to use: for a new loan applicant, the bank would first identify which class the applicant belongs to, then approve the loan if the corresponding rule for the class says so and reject otherwise.

Alternatively, in an AI rule base, we can include rules that provide instructions for formulating the optimization problem and for choosing actions based on the optimal solution. This is consistent with the proposal from Bringsjord et al. [2006] that one could constrain AI systems with ethical principles formalized as logic statements, such as if-then statements. For example, the bank may consider rules that require applicants with certain features to receive reasonable prioritization, and these rules can be captured as constraints or incorporated into the objective function in the optimization model. Furthermore, when making the final loan decisions, the bank may define rules about implementing the allocation solution obtained from the optimization problem.

Partial Information: Integration with Machine Learning

A second type of context arises when a limited amount of information is required to formulate the relevant optimization problem. The partial information case motivates a natural integration of optimization with machine learning. In particular, we focus on supervised learning methods that train predictive models from labelled data. Suppose a training data set is $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i$ is the feature vector and y_i is the true label, then a supervised learning method trains a predictor function h with the accuracy in the predicted labels $\{f(\mathbf{x}_i)\}$ as the primary goal. The ML literature has studied a large number of formats for h, ranging from a simple functional form in logistic regression and support vector machine to more complex structures like decision tree and neural network. Optimization, as a technique, is broadly used to train ML models, but our emphasis is to integrate optimization as the fairness-seeking strategy.

We propose two integration approaches that differ in the role played by ML. The first one follows a post-processing view that uses ML solely for estimating unknown parameters. To be more specific, after formulating the optimization model in the first two steps, we identify the unknown parameters in the formulation. We then consider each set of parameters separately, and choose an appropriate ML method to train a predictor function for the parameter values from available data. For instance, to predict the probability p_i of loan default, the bank may train a neural network based predictor using historical data containing past loan applicants' profiles and their default records. The predicted probabilities then serve as input to the optimization problem underlying the bank's loan decisions. It is notable that all supervised learning methods are suitable for such post-processing integration, and the decision maker has the flexibility to choose the ML methods fitting for the problem context and computational requirement.

The other approach is in-processing integration where social welfare optimization is directly embedded into a machine learning model. We can consider this approach as a type of in-processing fair ML method, and the key distinction with the majority of literature is that we encode fairness in a social welfare function. More precisely, we implement the integration by modifying the standard accuracy objective in a training model with a social welfare function. The SWF captures the desirable fairness and efficiency criteria, and can contain the

accuracy objective as an efficiency-related component. To further formalize, we use $L(h, \mathcal{D})$ to denote a loss function that evaluates the prediction error on the training data, then a ML model is trained to seek a loss minimizing predictor $h^* \in \text{argmin}_h L(h, \mathcal{D})$. Since h has a pre-selected form, e.g. regression function, neural network, with unknown parameters, the training problem is solving for parameter values. To incorporate social welfare considerations, we modify the standard loss function to include a suitable social welfare measure W. A simple example is $h^* \in \text{argmin}_h \lambda L(h, \mathcal{D}) - (1-\lambda) W(h, \mathcal{D})$, where the objective function allows the accuracy of prediction (measured via L) to contribute to welfare. Since the choice of W clearly affects the complexity of the learning model, successful in-processing integration requires the ability to design W into a format that can be handled effectively in machine learning.

As a final remark, we briefly discuss the integration potentials with two other core machine learning methods, unsupervised learning and reinforcement learning (RL). Fairness has been studied in both methods, but the progress is much more limited compared to fair supervised learning. Within unsupervised learning, we focus on clustering methods. We can easily apply post-processing integration to clustering methods and utilize the trained clusters as input to specify the optimization problem. For instance, in the loan example, the bank can use clustering algorithms to decide a categorization of financial profiles that will play a role in the optimization formulation. Recent works in fair clustering, e.g.Abraham et al. [2019], Deepak and Abraham [2020], have explored an in-processing strategy to extend K-means clustering to include fairness considerations by adding a fairness component to the usual K-means objective function. This indicates the potentials of in-processing integration, that is, we can define social welfare based fairness component to modify the usual clustering objective functions. In reinforcement learning, the goal is to search for a reward-maximizing policy in a dynamic environment that is typically modelled as a Markov Decision Process. Defining and achieving fairness in RL is more challenging due to the sequential and dynamic structure. Weng [2019], Siddique et al. [2020] propose a novel framework for fair multi-objective reinforcement learning based on welfare optimization. The key component of their proposal is to replace the standard reward objective with a particular social welfare function on the reward distribution. This exactly captures the perspective of in-processing integration, hence demonstrates the potentials of social welfare optimization for seeking fairness in RL.

8 Discussion and Conclusion

We formalize a general framework for using optimization to incorporate welfare-based fairness into AI applications. The framework provides a guideline for formulating a decision task into a social welfare optimization problem. In particular, we illustrate how optimization can be integrated with rule-based AI systems and machine learning models. By expanding the fairness problem to the optimization of social welfare functions, one can achieve a broader perspective

on fairness that are driven by the well-beings of stakeholders and characterize the broader fairness concepts in a principled way. Optimization models also provide the flexibility of adding constraints on resources and other problem elements, while harnessing the power of highly advanced optimization solvers.

We conclude the paper by outlining a research program to explore some key questions related to the framework.

- There is a wide gap between the presented general formalization of integration strategies and practical implementations of integrated methods. For integration with rule-based AI, one important direction is to investigate how to build ethics-sensitive rule bases to fit into different social welfare optimization scenarios. Previous works on formulating ethics principles into rules, e.g. Bringsjord et al. [2006], Kim et al. [2021], may provide guidance for this direction. For integration with machine learning, future research could explore the in-processing perspective and study how to define social welfare functions to use as the objective in machine learning models. The modified objective functions need to have a format that can be efficiently trained, and the trained models need to provide the desirable fairness and welfare guarantees.

- Although optimization solvers have been developed over decades, not all classes of optimization models are readily solvable by state-of-the-art software. Among all classes, linear programming and convex programming problems can be considered tractable up to reasonably large sizes, but non-convex formulations including some mixed integer programming problems are more restricted. For practical use of social welfare optimization models, one may need to apply available computational strategies or design problem-specific heuristics to speed up solving the optimization problems.

- The social welfare functions we consider are of a static nature, that is, a SWF does not attempt to capture potential dynamics in the utilities. A SWF takes utility values as the input, and the function values characterize the associated static utility distributions. While such a static view is often sufficient and reasonable for a one-shot decision problem, a dynamic perspective may be required in sequential decision problems where decisions need to be made repeatedly and the selected actions have incremental impacts on the long term social welfare. Future research could explore how to extend the presented optimization based framework to fit a dynamic view of welfare and fairness. Although this is not a trivial task, there are many well-developed techniques to utilize, such as, stochastic optimization, Markov decision process, etc.

References

S. S. Abraham, S. S. Sundaram, et al. Fairness in clustering with multiple sensitive attributes. *arXiv preprint arXiv:1910.05113*, 2019.

I. Alabdulmohsin. Fair classification via unconstrained optimization. *arXiv preprint*, 2005.14621, 2020.

C. Allen, I. Smit, and W. Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7:149–155, 2005.

K. Binmore, A. Rubinstein, and A. Wolinsky. The Nash bargaining solution in economic modeling. *RAND Journal of Economics*, 17:176–188, 1986.

R. Brandom. Self-driving cars are headed toward an ai roadblock. *The Verge*, 2018.

S. Bringsjord, K. Arkoudas, and P. Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21:38–44, 2006.

F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

V. Chen and J. N. Hooker. A just approach balancing Rawlsian leximax fairness and utilitarianism. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 221–227, 2020a.

V. Chen and J. N. Hooker. Balancing fairness and efficiency in an optimization model. *arXiv preprint*, 2006.05963, 2020b.

M. Chiarandini, R. Fagerberg, and S. Gualandi. Handling preferences in student-project allocation. *Annals of Operations Research*, 275(1):39–78, 2019.

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

S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

P. Deepak and S. S. Abraham. Representativity fairness in clustering. In *WebSci*, pages 202–211, 2020.

M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *Symposium on Innovations in Theoretical Computer Science (ITCS)*, pages 214–226, 2012.

O. Eisenhandler and M. Tzur. The humanitarian pickup and distribution problem. *Operations Research*, 67:10–32, 2019.

S. Freeman, editor. *The Cambridge Companion to Rawls*. Cambridge University Press, 2003.

I. Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30:411–437, 2020.

D. Gautier. *Morals by Agreement*. Oxford University Press, 1983.

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

J. C. Harsanyi. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, 1977.

H. Heidari, C. Ferrari, K. Gummadi, and A. Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, pages 1265–1276, 2018.

H. Heidari, M. Loi, K. P. Gummadi, and A. Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 181–190, 2019.

J. N. Hooker and T. W. Kim. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 130–136, 2018.

J. N. Hooker and H. P. Williams. Combining equity and utilitarianism in a mathematical programming model. *Management Science*, 58:1682–1693, 2012.

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

E. Kalai and M. Smorodinsky. Other solutions to Nash's bargaining problem. *Econometrica*, 43:513–518, 1975.

O. Karsu and A. Morton. Inequity-averse optimization in operational research. *European Journal of Operational Research*, 245:343–359, 2015.

T. W. Kim, J. Hooker, and T. Donaldson. Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research*, 70:871–890, 2021.

T. Lan, D. Kao, M. Chiang, and A. Sabharwal. An axiomatic theory of fairness in network resource allocation. In *Conference on Information Communications (INFOCOM 2010)*, pages 1343–1351. IEEE, 2010.

F. Lindner, R. Mattmüller, and B. Nebel. Evaluation of the moral permissibility of action plans. *Artificial Intelligence*, 287, 2020.

H. Luss. On equitable resource allocation problems: A lexicographic minimax approach. *Operations Research*, 47(3):361–378, 1999.

C. McElfresh and J. Dickerson. Balancing lexicographic fairness and a utilitarian objective with application to kidney exchange. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 1161–1168, 2018.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint*, 1908.09635, 2019.

M. Mostajabdaveh, W. J. Gutjahr, and S. Salman. Inequity-averse shelter location for disaster preparedness. *IISE Transactions*, 51(8):809–829, 2019.

V. Nanda, P. Xu, K. A. Sankararaman, J. Dickerson, and A. Srinivasan. Balancing the tradeoff between profit and fairness in rideshare platforms during high-demand hours. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2210–2217, 2020.

J. Nash. The bargaining problem. *Econometrica*, 18:155–162, 1950.

W. Ogryczak and T. Śliwiński. On equitable approaches to resource allocation problems: The conditional minimax solutions. *Journal of Telecommunications and Information Technology*, pages 40–48, 2002.

W. Ogryczak, A. Wierzbicki, and M. Milewski. A multi-criteria approach to fair and efficient bandwidth allocation. *Omega*, 36(3):451–463, 2008.

M. Olfat and A. Aswani. Spectral algorithms for computing fair support vector machines. In *International Conference on Artificial Intelligence and Statistics*, pages 1933–1942, 2018.

J. Rawls. *A Theory of Justice* (revised). Harvard University Press (original edition 1971), 1999.

H. S. Richardson and P. J. Weithman, editors. *The Philosophy of Rawls* (5 volumes). Garland, 1999.

A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50: 97–109, 1982.

S. Russell. *Human Compatible: AI and the Problem of Control*. Bristol, UK: Allen Lane, 2019.

U. Siddique, P. Weng, and M. Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pages 8905–8915. PMLR, 2020.

I. Stelmakh, N. B. Shah, and A. Singh. Peerreview4all: Fair and accurate reviewer assignment in peer review. *Proceedings of Machine Learning Research*, 98:1–29, 2019.

W. Thompson. Cooperative models of bargaining. In R. J. Aumann and S. Hart, editors, *Handbook of Game Theory*, volume 2, pages 1237–1284. North-Holland, 1994.

P. Weng. Fairness in reinforcement learning. *arXiv preprint arXiv:1907.10323*, 2019.

A. Williams and R. Cookson. Equity in health. In A. Culyer and J. Newhouse, editors, *Handbook of Health Economics*. Elsevier, 2000.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.