Fair and Interpretable Algorithmic Hiring Using Evolutionary Many-Objective Optimization

Michael Geden, Joshua Andrews
Modern Hire, Delafield, Wisconsin
{michael.geden, joshua.andrews}@modernhire.com

Abstract

Hiring is a high-stakes decision-making process that balances the joint objectives of being fair and accurately selecting the top candidates. The industry standard method employs subject-matter experts to manually generate hiring algorithms; however, this method is resource intensive and finds sub-optimal solutions. Despite the recognized need for algorithmic hiring solutions to address these limitations, no reported method currently supports optimizing predictive objectives while complying to legal fairness standards. We present the novel application of Evolutionary Many-Objective Optimization (EMOO) methods to create the first fair, interpretable, and legally compliant algorithmic hiring approach. Using a proposed novel application of Dirichlet-based genetic operators for improved search, we compare state-of-the-art EMOO models (NSGA-III, SPEA2-SDE, bi-goal evolution) to expert solutions, verifying our results across three real world datasets across diverse organizational positions. Experimental results demonstrate the proposed EMOO models outperform human experts, consistently generate fairer hiring algorithms, and can provide additional lift when removing constraints required for human analysis.

Introduction

Hiring is a high-stakes practice that influences people's livelihoods across the globe. It is an important and challenging socio-technical problem relevant to industrial, academic, and governmental organizations (Raghavan, Barocas, Kleinberg, and Levy 2020; Shute et al. 2016). The hiring process encompasses the joint objectives of providing a fair, legally compliant process and selecting candidates that are best fit for the position. Governments vary in their requirements for fairness in hiring; however, many conceptualize fairness as the absence of at least two types of discrimination: disparate treatment and adverse impact (e.g., USA, EU, ILO). Disparate treatment represents differential treatment of protected groups, such as the use of quotas, while adverse impact represents decisions that have a disproportionate impact on minority groups, such as unequal hiring rates.

The industry standard method, rational weighting, involves subject-matter experts balancing predictive accuracy and fairness objectives by manually generating algorithms using heuristics and data (SIOP 2018). Rational weighting addresses fairness requirements, but it is resource-intensive, finds sub-optimal solutions, encodes human biases, and cannot guarantee the legal requirement that a fairer and equally predictive hiring algorithm does not exist. These limitations have led to the growing interest in methods that learn hiring algorithms solely from the data—algorithmic hiring. Algorithmic hiring carries transformative potential for reducing bias while enhancing reproducibility, predictive accuracy, and scalability for larger feature sets (Raub 2018; Houser 2019). Despite the significant promise of algorithmic hiring, and the claims of many organizations, no reported method currently supports optimizing predictive objectives while adhering to legal fairness standards. This has led to the continued reliance on rational weighting (Raghavan et al. 2020; Sánchez-Monedero, Dencik, and Edwards 2020).

Many existing solutions eschew fairness requirements to solely focus on predictive accuracy, leading to algorithms that are not legal for practical implementation (Hemamou et al. 2019; Meijer et al. 2020). Alternatively, methods have been proposed to reduce relationships between predictors and minority classes; however, these methods do not account for adverse impact (Calmon et al. 2017; He, Burghardt, and Lerman 2020). Multi-objective optimization (MOO) models have demonstrated promising results when simplifying protected groups into a binary minority variable, yet performance was not compared to competitive baselines, such as rational weighting (De Cortes, Sackett, and Lieven 2011; Song, Wee, and Newman 2017). MOO models produce an approximate pareto front—set of non-dominated solutions—which can incorporate client preferences and

support human-in-the-loop decision-making. Nonetheless, real hiring problems require fairness objectives for multiple demographic classes, rapidly increasing the number of objectives, which can degrade MOO performance (Deb and Jain 2013). Evolutionary many-objective optimization (EMOO) models address MOO's objective size limitation through modifying dominance relationships, using reference directions for exploration, and reducing the objective space (Li et al. 2015; Ishibuchi, Tsukamoto, and Nojima 2008).

In this paper, we present the first fair and legally compliant algorithmic hiring method using a diverse set of state-of-the-art EMOO models (i.e., NSGA-III, SPEA2-SDE, bi-goal evolution). Additionally, we propose the use of Dirichlet-based genetic operators for constrained EMOO problems as opposed to traditional repair approaches. We evaluate models on three real world datasets along diverse positions at Fortune 500 companies, comparing to the industry standard rational weighting baseline created by professional Industrial/Organizational (I/O) Psychologists. Empirical results demonstrate that EMOO substantially outperforms human experts across all datasets. Additionally, when constraints required for human experts are removed, EMOO can further improve performance.

Related Work

Hiring

Hiring algorithms transform candidate data gathered from selection systems (e.g., situational judgement, work history) into a ranked composite score with the joint goals of selecting high performing candidates and maximizing fairness across protected attributes (e.g., gender, race; Pyburn, Ployhart, and Karvitz 2008). The legal requirements for fair employment decisions vary by country; nevertheless, there is broad consensus that algorithms must meet at least two requirements: they cannot include protected attributes (disparate treatment) and they cannot disproportionally exclude minority groups (adverse impact). Additional properties may be required by region, such as the job relevance of features (i.e., face validity) and the absence of differential prediction and measurement bias (US EEOC Uniform Guidelines; EEOC UG). Government organizations and AI experts have also stressed the need for interpretability in algorithms applied to high-stakes decision-making, which includes hiring (Rudin 2019; EU GDPR; Schumann et al. 2020; Goodman and Flaxman 2016).

The de facto method for industry applications is *rational weighting*, where subject matter experts manually select a weighted linear combination, which sums to one, using spot checking, heuristics, and the available literature to predict performance while ensuring fairness (Ployhart and Holtz 2008; SIOP 2018). Rational weighting is interpretable, easy to perform, and can guarantee that most fairness requirements are met; however, it is also time consuming, expensive, hard to replicate, and inefficient at optimization. Another simple method is to equally weight all standardized predictors, known as *unit weighting* (Einhorn and Hogarth 1975); however, this approach is inefficient for predictive accuracy and cannot guarantee fairness. Authors have also proposed linear regression (O'Neill and Steel 2018; Meijer et al. 2020) to improve predictive validity, but it similarly fails to account for adverse impact.

Fair Machine Learning

Fair machine learning is a critical issue for the ethical deployment of AI/ML and encompasses a diverse set of definitions and approaches depending upon the target domain (Friedler et al. 2019). Algorithmic approaches hold transformative potential for fair and predictive hiring (Houser 2019; Raub 2018; Schumann et al. 2020); although, their success is contingent upon the adoption of fair ML methods to meet social, ethical, and legal fairness requirements.

Preprocessing methods attempt to prevent models from learning subtle patterns of discrimination by striping predictors of their relationship to a protected class (Feldman et al. 2015; Calmon et al. 2017; He, Burghardt, and Lerman 2020). A key limitation of preprocessing methods for hiring is that they typically focus exclusively on feature sets and do not control for adverse impact. An alternative strategy is post-processing of predictors, but these methods often require direct use of protected classes, which would constitute disparate treatment (Zehlike et al. 2017; Hardt, Price, and Srebro 2016; Dwork et al. 2018). Regularization methods incorporate fairness objectives directly into their loss function using hyperparameters to weight each objective's relative importance (Fould et al. 2019; Zemel et al. 2013). They provide efficient optimization at the cost of sensitivity to fairness/utility trade-offs from the selected objective weights and poor performance for concave pareto fronts (Boyd and Bandenberghe 2004). Regularization methods can be interpreted as an aggregation approach to multi-objective optimization (MOO; Li et al. 2015). Some research has been conducted demonstrating the potential of other simple evolutionary methods for hiring systems with promising results (De Cortes, Sackett, and Lieven 2011; Song, Wee, and Newman 2017; De Corte, Sackett, and Lievens 2010). These studies suffered from two primary limitations: (1) they did not include competitive baselines and (2) they simplified protected classes into a binary variable (majority, minority) rather than considering all represented minority groups.

Many-Objective Optimization

Including fairness objectives for all minority groups rapidly increases the number of objectives under consideration. The

balancing of many (>4) objectives becomes increasingly complex for several reasons: (1) the pareto space becomes increasingly large and difficult to represent, (2) non-dominated sorting degrades as most solutions belong to the pareto frontier, (3) genetic operators become inefficient at searching the wider pareto space, (4) evaluating diversity and performance metrics becomes computationally expensive (Deb and Jain, 2013; Ishibuchi et al. 2014; Ishibuchi, Tsukamoto, and Nojima 2008; Wagner, Beume, and Naujoks 2007). Evolutionary many-objective optimization (EMOO) algorithms attempt to circumvent these issues through modifying the definition of dominance, the objectives being optimized, and the search mechanisms to improve the exploration process and create smaller tiers of domination.

Methods such as geometric modifications of dominance relations and the use of rank methods can improve selection pressure toward the pareto front at the cost of solution diversity (Ishibuchi, Tsukamoto, and Nojima 2008; Kukkonen and Lampinen 2007). Aggregation and reduction approaches condense the problem space down to a small number of objectives (Wang et al. 2018; Saxena, et al. 2013), and therefore a multi-objective space, but at the cost of solution diversity (Li, Yang, Liu 2015; Yuan et al. 2016). Indicator based methods directly optimize on metrics representing the quality of a pareto-front, often using hypervolume as it is strictly monotonic to Pareto dominance (Bader and Zitzler 2011; Sun, Yen and Yi 2019). Indicator-based methods have demonstrated strong performance; however, they are computationally intensive, particularly as the number of objectives increases. Finally, some approaches alter the diversity maintenance function using reference directions, an explicit set of weight vectors, to select dominance between similar solutions and widen the search process (Deb and Jain 2013; Zhang and Li 2007).

Evolutionary Many-Objective Algorithmic Hiring

In this section, we formally introduce the problem of hiring from a pool of candidates with the joint goals of selecting future top performers and adhering to legal fairness requirements. Then, we outline the application of evolutionary many-objective optimization to the hiring problem. Finally, we introduce Dirichlet-based genetic operators to flexibly account for our interpretability constraints.

Problem Overview

Consider a hiring system composed of N applicants and P available positions. The hiring system is composed of L sets of Q_l interrelated questions $X_{q,l}$. Each set represents a different activity relevant to the job, such as situational judgement or mathematical ability. Questions within each activity are aggregated to create activity composite scores $\bar{X}_l$. Hiring algorithms serve to combine activity scores into a singular composite θ_i for ranking and selecting P candidates.

Interpretability is critical in effective and fair algorithmic decision-making (Bigu and Cernea 2019), leading us to incorporate the following model constraints. First, we limit hiring algorithms to weighted linear combinations of features. Second, we constrain feature weights to sum to one, mimicking traditional grade scoring, as the primary goal is ranking candidates. Finally, we enforce a rounding increment ψ for predictor weights at 0.05 or 0.01 to provide clear communication to lay audiences, such as organizational stakeholders and applicants themselves. More formally:

$$0 \le w_l \le 1 : \psi \,|w_l, \mathrm{l} \in L$$
$$\psi \in \{0.01, 0.05\}$$
$$\sum_{l=1}^{L} w_l = 1$$
$$\sum_{l=1}^{L} w_l \bar{X}_{i,l} = \theta_i$$

Objectives

Many-objective algorithmic hiring methods generate a set of A approximately pareto-optimal hiring algorithms that jointly optimize J job performance and F fairness objectives.

$$A = \{W_x \in \Omega_{J+F} : \nexists \mathrm{W_y} \in \Omega_{J+F}, W_y \prec W_x\}$$

Without loss of generality, we describe a minimization problem where $W_x \prec W_y$ if the following two conditions hold:

$$L_m(W_x) \le L_m(W_y) : \forall m \in \{1, \dots, J+F\}$$
$$L_m(W_x) < L_m(W_y) : \exists m \in \{1, \dots, J+F\}$$

Job Performance

Training a model to rank applicants for predictive job performance requires data that is unavailable in the applicant pool. Hiring algorithms are often trained using job incumbents for the initial validation (i.e., concurrent validity; EEOC Uniform Guidelines). The standard loss function for criterion validity is Pearson's correlation; however, correlation is ill-suited for top-k ordinal data (Li, Wang, and Xiao 2019). As we do not have a ground truth for who should be selected, top-ranking distances cannot be applied (Xia, Liu, and Li 2009). Instead, we propose using the sum of the true scores for the predicted top candidates as a distance metric. We then normalized the predicted ordinal sum by subtracting the ordinal sum of the known top-k candidates, which represents the maximum individual fairness in relation to the true criterion score. This metric was loosely based on the earthmovers distance outlined in Feldman et al (2015).

Adverse Impact

A commonly used heuristic for assessing adverse impact is the four-fifths rule, where all minorities must be selected at least four-fifths as often as the majority. A drawback of

using selection ratios as a fairness objective is their instability when there is extreme class imbalance. We addressed this by using the smoothed empirical differential fairness objective (Foulds et al. 2019; 2020). A secondary benefit of sEDF is its ability to incorporate intersectionality considerations; nevertheless, intersectional group evaluation is beyond the scope of this work.

Evolutionary Framework

EMOO methods offer three primary benefits for algorithmic hiring: (1) they generate an approximate pareto front, allowing for human-in-the-loop decision-making, (2) they can handle many-objectives, and (3) they can optimize on non-differentiable loss functions (e.g., selection). While encompassing many diverse approaches, most EMOO methods utilize four genetic operators for the creation of solutions: sampling, crossover, mutation, and repair (Algorithm 1).

Sampling

The sampling operator initializes the first generation of algorithms. A commonly used approach for problems with linear fixed sum constraints is to perform random uniform sampling or Latin hypercube sampling (LHS) and then to repair the generated solutions (Chiam, Tan, and Mamum 2008). These approaches result in non-uniform sampling on the constrained space and can degrade convergence. We propose sampling from a Dirichlet distribution using a uniform prior as an alternative. Figure 1 displays the sampling density (N=10000) on a three-dimensional fixed linear sum problem for two of the dimensions across the three approaches, demonstrating the improved uniformity of the proposed Dirichlet method.

Algorithm 1 Evolutionary Optimization

1: **Input:** data X, N_EMOO, m
2: $P \leftarrow$ initialize(X, N_EMOO)
3. **while** termination criterion not met **do**
4. $P_1^*, P_2^* \leftarrow$ matingSelection(P)
5. $P_O \leftarrow$ crossover(P_1^*, P_2^*)
6. $P_O \leftarrow$ mutation(P_O)
7. $P_O \leftarrow$ repair(P_O)
8. $P \leftarrow$ environmentalSelection($P \cup P_O$)
9: **Return**: P

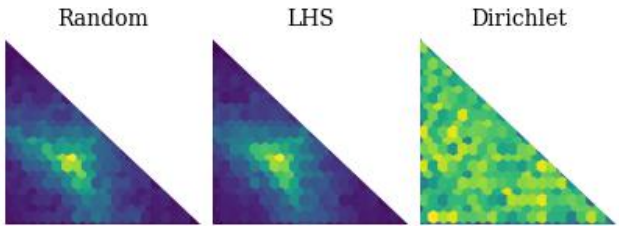

Figure 1: Density of sampling operators.

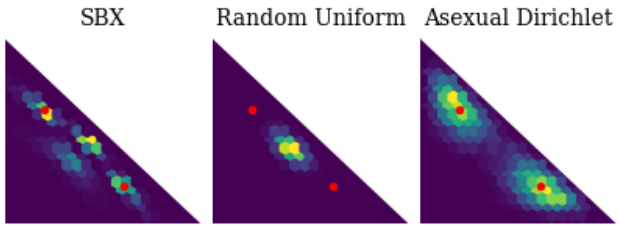

Figure 2: Density of crossover operators. Red points are parents.

Crossover

Crossover functions serve to create new algorithms (children) through the combination of two existing algorithms (parents). Offspring should remain near parents in many-objective problems due to the large search space (Sat, Aguirre, and Tanaka 2011). Due to our summation constraint, traditional crossover methods do not apply directly because they require repair that alters the algorithms in an undesirable way. Figure 2 demonstrates the complications with repaired SBX and random uniform crossover, where children tend toward the center of the distribution. We propose the use of a constrained crossover operator based on a Dirichlet distribution centered around a single parent ("asexual"). Figure 2 demonstrates that asexual crossover generates offspring that are more similar to their parents than traditional methods.

Mutations

Mutation operators serve to create children through creating a random alteration in an existing parent algorithm. Due to the summation constraints, our proposed mutation operator was based on a Dirichlet distribution centered around the parent algorithm and applies to the entire algorithm, rather than effecting a single weight.

Repair

Algorithms that failed to meet our rounding and summation constraints underwent one of two repair operations. First, the summation repair activated when $\sum_i^K \boldsymbol{w_i} \neq \mathbf{1}$ and divided each weight by their sum. Next, the rounding repair operation rounded weights to the nearest $\boldsymbol{\psi}$ if $\exists \boldsymbol{w_i}: \boldsymbol{\psi} \nmid \boldsymbol{w_i}$. When the rounding repair operation caused $\sum_i^K \boldsymbol{w_i} \neq \mathbf{1}$, we randomly distributed $(\sum_i^K \boldsymbol{w_i} - \mathbf{1})/\boldsymbol{\psi}$ increments of $\boldsymbol{\psi}$ across the available weights, similar to the round-lot constraints used in financial problems (Chiam, Tan, and Mamum 2008).

Datasets

Evaluation of selection systems requires data on a candidate's protected class (e.g., race, gender), responses to the hiring assessment, and job-related performance criteria. As performance data cannot be acquired from applicants, selection systems are typically trained on data collected from current employees (SIOP 2018). We collected manager and supervisor ratings on a 1–5 Likert scale along multiple

attributes of performance criteria (e.g., overall performance, technical skills). Raters provided estimates of their rating confidence and knowledge of the subordinate on 7-point Likert scales, with values less than 5 removed from analysis. We evaluated adverse impact for self-reported race (i.e., Black and Latinx) and binary gender (i.e., female), as they had sufficient representation in our datasets (N > 30). Certain activities, such as personality, are multifaceted constructs composed of multiple correlated subcomponents (e.g., optimism, self-efficacy, competitiveness). Due to the limited capacity of rational weighting, multifaceted constructs are often aggregated into a single score to reduce dimensionality.

We conducted experiments on three datasets across diverse positions at different Fortune 500 companies. Datasets were limited in size due to the cost of collecting criterion data and the size of incumbent populations; however, these sample sizes are common in industry applications and can be sufficient for generalizable results (Song, Wee, and Newman 2017). This is in part due to the simplicity of constrained, linear, hiring algorithms and the regularizing impact of multiple objectives (Sener and Koltun 2018).

Leadership Dataset

Data were collected for an entry level management position at an international service organization. The performance criteria were supervisor ratings on employees being a best performer over time, whether the employee is a currently a top performer, and the likelihood that the manger would rehire the employee, resulting in a final set of six objectives (3 criteria, 3 diversity). The total sample size was 377 after removing 18 observations with insufficient rater confidence. The feature set was composed of five activity scores across multiple questions for situational judgement, priority setting, problem solving, personality, and work history.

Sales Dataset

Data were collected for commercial sales positions at a large international retail sales organization. The performance criteria were manager ratings on employees being a current top performer, their likelihood of promotion, and the likelihood the manager would rehire the employee, resulting in a final set of six objectives (3 criteria, 3 diversity). The total sample size was 922 after removing 345 observations with insufficient rater confidence. The feature set was composed of five activity scores across multiple questions for situational judgement, data input verification, mathematical ability, personality, and work history

Banker Dataset

Data were collected for a financial banking position at a large U.S. financial institution. The performance criteria were supervisor ratings on individuals being a best employee over time, being a current top performer, having promotion potential, being a team player, and having the ability to ramp up, resulting in a set of 8 objectives (5 criteria, 3 diversity). The total sample size was 416 after removing 137 observations due to insufficient rater confidence. The feature set was composed of six activity scores across multiple questions for situational judgement, mathematical ability, problem solving, culture fit, personality, and work history.

Experiments

Models

We included two baseline methods commonly used in industry settings.

- *Unit weighting*: This is a standard method used as a baseline for comparisons (De Corte, Lievens, and Sackett 2007).
- *Rational weighting (RW)*: I/O Psychologists generated a pool of algorithms based on the full dataset (no cross-validation), from which they select a top k subset. The experts created these algorithms, unaware of the current study, as part of the normal validation procedure for live hiring systems.

We investigated several well-established evolutionary models and one proposed model that modifies an existing aggregation method.

- *NSGA-II* (Deb et al. 2002): This model is a classic baseline for multi-objective optimization. It uses non-dominated sorting and crowding distance for environmental selection.
- *MOEA/D* (Zhang and Li 2007): This model employs reference direction decomposition and applies pressure along a neighborhood for each subproblem using Euclidean distances.
- *SPEA2-SDE* (Li, Yang, Liu 2014): This is a modified form of SPEA-2 designed for many-objective optimization. It uses shift-based density estimation and the number of dominated solutions for environmental selection.
- *NSGA-III* (Deb and Jain 2013): This method modifies the NSGA-II framework with the addition of reference directions, which effectively replaces crowding-distance, to improve the search process for many-objective problems.
- *BiGE* (Li, Yang, Liu 2015): Bi-goal evolution uses an aggregation approach to many-objective problems by converting all objectives into a bi-goal problem regarding proximity and diversity.
- *TriGE*: We propose a simple modification of BiGE to create three-objectives: criterion performance, fairness, and solutions diversity. This modification allows for a direct tradeoff between fairness and utility.

Experimental Settings

We trained and evaluated models using 10-repetitions of a 70-30 train-test split (Bischl et al. 2012). For each repetition, EMOO models trained on standardized training data and generated an approximate pareto set of hiring algorithms from the final generation's non-dominated solutions. Hiring algorithms that failed to meet the four-fifths ratio for any protected class were pruned and the remaining algorithms were evaluated on the test set. Data splits varied across repetitions but were held constant between models to calculate performance indicators. Models were constructed in python using the pymoo module (Blank and Deb 2020) and run on 32gb of RAM and a 2.90GHz processor. Repetitions took between 0–5 hours depending on the model and dataset.

Evolutionary hyperparameters were selected based on the original papers and were not tuned due to the limited sample size of the datasets. Reference direction for NSGA-III and MOEA/D were generated using Riesz s-Energy (Blank et al. in press), with NSGA-III using a multi-layer morphology (Deb and Jain 2013). The number of reference directions was a function of the number of objectives in each of the datasets. Training terminated either after 30 generations of no improvement or when 800 generations finished. The mutation rate was 0.75, which was substantially higher than the standard 1/K because our mutation operator acts on the entire chromosome rather than each independent weight. The crossover rate was 1. The crossover and mutation Dirichlet intensity parameters were 0.25. All models used differential fairness and ordinal sum for the loss functions.

Results

We selected a diverse set of unary performance indicators to evaluate the approximate pareto sets from each model along their objective diversity, proximity, and cardinality (Cheng, Shi, and Qin 2015). Objective diversity was assessed along the *spread* (S) of the objective space (average range of objectives in pareto front). Proximity was assessed with *inverted generational distance* plus (IGD+; Ishibuchi et al. 2015) in reference to the concatenation of all model's non-dominated fronts for each repetition, forming an approximate true pareto front. Cardinality was calculated as the number of non-discriminatory non-dominated solutions generated by an algorithm (C). The number of non-dominated solutions (N_{NDS}) is the number of solutions generated by an algorithm that were present in the approximated true pareto front. *Hypervolume* (HV), measuring both diversity and proximity, was calculated using an approximation with an epsilon and delta of .01 (Bringmann and Friedrich 2010). The hypervolume reference point was 1.5 for differential fairness and 1 for normalized ordinal sum.

In Experiment 1, we compared models under the constraints required by the rational weighting baseline ($\psi = 0.05, L \approx 5$). The population and offspring size were set to 300. EMOO models outperformed baseline methods across all datasets (Table 1). SPEA2-SDE and NSGA-II consistently demonstrated the best overall fit, yielding proximal and diverse approximate pareto fronts that substantially outperformed rational weighting. BiGE, TriGE, and NSGA-III generated null sets for some repetitions on the Leadership dataset, resulting in a lower cardinality. A subset of the best performing SPEA2-SDE algorithms, equal to the number of rational algorithms (k), were selected to mimic the process of selecting hiring algorithms. Algorithms were selected by taking the top-k average across the ranking for each objective. These top-k SPEA2-SDE and rational weighting hiring algorithms were compared along their average criterion (C) and fairness (F) performance (Figure 3), with SPEA2-SDE providing significant gains on fairness objectives.

In Experiment 2, we evaluated EMO models under relaxed constraints that are infeasible for rational weighting methods using the Sales dataset ($\boldsymbol{\psi = 0.01, L = 16}$). Personality and Situational Judgement were decomposed into their respective subscales. Population and offspring size were raised to 600 to support the increased size of the feature space. The partially constrained EMOO results demonstrated improved fit over the human constrained results, with NSGA-II yielding the strongest performance (Table 2). The relaxed constraints degraded performance on the Leadership and Banker datasets, where the personality subscales were generally forced to the minimum weight ($\boldsymbol{\psi = 0.01}$). The increased volume of low utility subscales, even when suppressed, consumed a large proportion of overall weight. Detailed results are omitted due to space constraints.

	Sales					Leadership					Banker				
Model	HV	IGD+	S	C	N_{NDS}	HV	IGD+	S	C	N_{NDS}	HV	IGD+	S	C	N_{NDS}
Unit weights	5.37	0.164	0.00	1.0	0.2	5.54	0.382	0.00	1.0	0.4	1.68	0.19	0.00	1.0	0.2
RW	6.93	0.049	0.09	20.0	10.6	7.27	0.131	0.13	5.0	1.7	2.12	0.16	0.09	5.0	1.5
NSGA-II	**7.68**	**0.003**	**0.22**	**152.5**	**28.6**	**8.69**	**0.006**	**0.33**	**65.7**	**16.1**	3.27	0.04	0.32	141.4	32.1
MOEA/D	7.36	0.019	0.19	62.8	12.6	8.07	0.044	0.29	32.2	6.4	3.05	0.04	0.26	79.9	20.4
SPEA2-SDE	**7.68**	**0.003**	**0.22**	**153.1**	**28.7**	**8.69**	**0.006**	**0.33**	**65.8**	**16.1**	**3.37**	**0.01**	**0.33**	**249.7**	**65.9**
NSGA-III	7.14	0.033	0.17	33.8	5.9	8.05	0.037	0.28	18.2	5.3	2.86	0.08	0.19	17.8	4.6
BiGE	6.04	0.095	0.05	2.3	0.7	6.67	0.146	0.12	3.3	0.8	2.06	0.23	0.09	4.4	0.6
TriGE	6.69	0.055	0.13	13.5	2.4	7.72	0.050	0.25	13.4	3.0	2.77	0.09	0.22	19.3	4.2

Table 1: Rational Weighting Constrained Performance Indicators

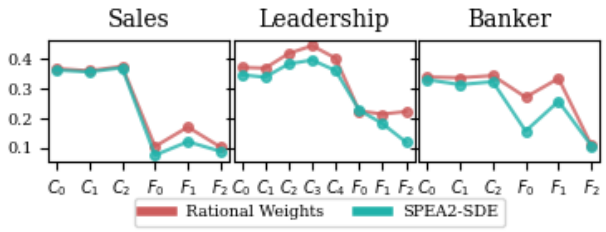

Figure 3: Average Pareto Front for Top-k Algorithms

Model	HV	IGD+	Spread	C	N_{NDS}
Unit Weights	5.95	0.167	0	1	0.4
NSGA-II	**8.47**	**0.006**	**0.28**	**396.7**	**39.7**
MOEA/D	7.41	0.044	0.14	41.7	5.6
SPEA2-SDE	8.31	0.047	0.27	370.3	18.4
NSGA-III	7.80	0.029	0.19	85.8	6.9
BiGE	6.93	0.087	0.13	5.4	0.7
TriGE	7.45	0.068	0.19	24.2	3.6

Figure 4: Partially Constrained Performance Indicators on Sales Dataset

Discussion

In this study, we present the first algorithmic hiring framework that adheres to legal fairness standards. The proposed EMOO framework generates an approximate pareto front of interpretable, predictive, and legally compliant hiring algorithms. Evaluation results across three diverse, real-world datasets demonstrated that EMOO models substantially outperformed competitive baselines at fairly selecting high-performing candidates.

When operating within the constraints of rational weighting, SPEA2-SDE and NSGA-II yielded the best performance among EMOO models across all performance indicators. NSGA-II and SPEA2-SDE demonstrated nearly identical performance in the constrained condition for two of the datasets. A potential explanation is that both algorithms rely on similar domination-based proximity mechanisms, resulting in identical non-dominated solutions. Therefore, their primary distinction lies in their selection of dominated solutions based on their diversity preservation mechanisms. The rational weighting constraints used in Experiment I severely limit the size of the feature space, potentially causing the differences in diversity mechanisms to become negligible. One explanation for the surprisingly strong performance of NSGA-II compared to some of the many-objective optimization models could be that there was a relatively small number of correlated objectives (6–8). This prevented one of NSGA-II's primary drawbacks, which is the degradation of non-dominated sorting that occurs when the entire population belongs to the non-dominated front. The proposed TriGE model outperformed the original BiGE; however, it did not produce competitive results.

In Experiment 2, we relaxed the rational weighting constraints on the number of features and the weight increments using the Sales dataset. Similar to Experiment 1, NSGA-II was the best performing models, and all EMOO models except for BiGE outcompeted both baselines. Additionally, relaxing these constraints resulted in overall better performance compared to Experiment 1.

This study substantiates the claims that algorithmic hiring can transform employment decision-making to be a more fair, valid, and interpretable process (Houser et al. 2019). The hiring algorithms generated from the EMOO framework are glass-box models that are well-suited for high-stakes decision making (Rudin 2019). NSGA-II was able to identify hiring algorithms with equal or better prediction while improving fairness for female, Black, and Latinx applicants compared to rational weighting (Figure 3). The advent of algorithmic hiring opens the possibility of tackling critical challenges in the employment selection industry, such as intersectionality, which have remained out of grasp for traditional methods.

There were some limitations with this study. Firstly, not all demographic classes were sufficiently represented in the data for conducting fairness analysis, which limited the number of objectives in the current study. Secondly, the current study trained models using concurrent validation data; however, the distribution of job incumbents may differ from the applicant pool, which could potentially create issues from covariate shift and distributional differences.

Conclusion and Future Work

Hiring is an area of research and practice that can greatly benefit from meaningful applications of AI. Nonetheless, there currently exists no published solution for models that can operate within the legal framework of hiring. The contribution of this study is in the unique application of EMOO models that adhere to legal constraints and produce human interpretable hiring algorithms. Results demonstrated across three industry data sets show that EMOO models outperformed industry baselines. Most promisingly, results indicate that EMOO models could have substantial impact on hiring fairness, creating more equitable outcomes for candidates across protected classes.

Future work on this subject should continue to investigate important fairness topics, such as intersectionality. In reality, candidates often belong to multiple protected groups, such as Black and female; however, the intersection of an individual's multi-group membership is often not assessed in typical fairness analysis. EMOO algorithms that can assess and balance intersectional fairness effectively would be even more impactful and beneficial to society and organizations. Similarly, future research should investigate predictive parity, or equality of prediction across protected groups.

Ethics Statement

The present study was explicitly focused on the utility of AI for increasing predictive accuracy and human well-being (fairness) in a major area of society, the workplace. Results from this study indicate evolutionary algorithmic hiring can substantially improve fairness outcomes while maintaining or enhancing selection of top ranked candidates. Congruent with both AAAI and SIOP ethics statements, we strove to avoid harm to minority groups, provide trustworthy results, and to honor the privacy of our candidates and clients. Importantly, our focus was on creating models that increase fairness across all protected groups, which discourages the models from harming one group to benefit another. Furthermore, our presented models were designed for human interpretability, which enhances accountability and transparency as well as aids in communication to society. Candidate information for the three hiring systems was collected in a secure and ethical manner consistent with the Society for Human Resource Management (SHRM).

References

Bader, J.; and Zitzler, E. 2011. HypE: An Algorithm for Fast Hyper Volume-Based Many-Objective Optimization. *Evolutionary Computation*, 19(1): 45–76. doi.org/10.1162/EVCO_a_00009.

Bigu, D.; and Cernea, M. 2019. Algorithmic Bias in Current Hiring Practices: An Ethical Examination. *In Proceedings of the 13th International Management Conference*, 13(1), 1068–1073.

Bischl, B.; Mersmann, O.; Trautmann, H.; and Weihs, C. 2012. Resampling Methods for Meta-Model Validation with Recommendations for Evolutionary Computation. *Evolutionary Computation*, 20(2): 249–275. doi.org/10.1162/EVCO_a_00069.

Blank, J.; and Deb, K. 2020. Pymoo: Multi-Objective Optimization in Python. *IEEE Access*, 1–18. doi.org/10.1109/ACCESS.2020.2990567.

Blank, J.; Deb, K.; Dhebar, S.; Bandaru, S.; and Seada, H. In Press. Generating Well-Spaced Points on a Unit Simplex for a Evolutionary Many-Objective Optimization. *IEEE Transactions on Evolutionary Computation*. doi.org/0.1109/TEVC.2020.2992387.

Bornstein, R. 1996. Face Validity in Psychological Assessment: Implications for a Unified Model of Validity. *American Psychologist*, 51(9): 983–984. doi.org/10.1037/0003-066X.51.9.983.

Boyd, S.; and Bandenberghe, L. 2004. Convex Optimization Problems. *In Concave Optimization*, 127–189. New York: Cambridge University Press.

Bringmann, K.; and Friedrich, T. 2010. Approximating the Volume of Unions and Intersections of High-Dimensional Geometric Objects. *Computational Geometry*, 43(6): 601–610. doi.org/10.1016/j.comgeo.2010.03.004.

Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K.; and Varshney, K. 2017. Optimized Pre-Processing for Discrimination Prevention. In *31st Proceedings of NeurIPS*, 1–10.

Cheng, S.; Shi, Y.; and Qin, Q. 2015. On the Performance Metrics of Multi-Objective Optimization. *In Proceedings of the International Conference in Swarm Intelligence*, 504–512. doi.org/10.1007/978-3-642-30976-2_61.

Chiam, C.; Tan, C.; and Al Mamum, A. 2008. Evolutionary Multi-Objective Portfolio Optimization in Practical Context. *International Journal of Automation and Computing*, 5(1): 67–80. doi.org/10.1007/s11633-008-0067-2.

Deb, K.; Pratap, A.; Agarwal, S.; and Meyarivan, T. 2002. A Fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2): 182–197. doi.org/10.1109/4235.996017.

Deb, K.; and Jain, H. 2013. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems with Box Constraints. *IEEE Transactions on Evolutionary Computation,* 18(4): 577–601. doi.org/10.1109/TEVC.2013.2281535.

De Corte, W.; Sackett, P.; and Lievens, F. 2011. Designing Pareto-Optimal Selection Systems: Formalizing the Decisions Required for Selection System Development. *Journal of Applied Psychology*, 96(5): 907–923. doi.org/10.1037/a0023298.

De Corte, W.; Sackett, P.; and Lievens, F. 2010. Selecting Predictor Subsets: Considering Validity and Adverse Impact. *International Journal of Selection and Assessment*, 18(3): 260–270. doi.org/10.1111/j.1468-2389.2010.00509.x.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. and Zemel, R. 2012. Fairness Through Awareness. *In Proceedings of the 3rd ITCS*, 214–226. doi.org/10.1145/2090236.2090255.

Einhorn, H.; and Hogarth, R. 1975. Unit Weighting Schemes for Decision Making. *Organizational Behavior and Human Performance*, 13(2): 171–192. doi.org/10.1016/0030-5073(75)90044-6.

Feldman, M.; Friedler, S.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. *In Proceedings of KDD*, 259–268. doi.org/10.1145/2783258.2783311.

Foulds, J.; Islam, R.; Keya, K.; and Pan, S. 2019. Differential Fairness. In *Proceedings of NeurIPS Workshop on Machine Learning with Guarantees*, 1–16.

Foulds, J.; Islam, R.; Keya, K.; and Pan, S. 2020. An Intersectional Definition of Fairness. *In Proceedings of ICDE*, 1918–1921. doi.org/10.1109/ICDE48307.2020.00203.

Friedler, S.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E.; and Roth, D. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. *In Proceedings of FAccT*, 329–338. doi.org/10.1145/3287560.3287589.

Goodman, B.; and Flaxman, S. 2016. EU Regulations on Algorithmic Decision-Making and a "Right to Explanation". *In Proceedings of ICML Workshop on Human Interpretability in Machine Laerning*, 26–30.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. *In Proceedings of NeurIPS*, 3315–3323.

He, Y.; Burghardt, K.; and Lerman, K. 2020. A Geometric Solution to Fair Representations. *In Proceedings of AIES*, 279–285. doi.org/10.1145/3375627.3375864.

Hemamou, L; Felhi, G.; Vanderbusshe, V.; Martin, J.; and Clavel, C. 2019. HireNet: A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. *In Proceedings of AAAI*, 33, 573–581. doi.org/10.1609/aaai.v33i01.3301573.

Houser, K. 2019. Can AI Solve the Diversity Problem in the Tech Industry: Mitigating Noise and Bias in Employment Decision-Making. *Stanford Tech. Law Review*, 22: 290–353.

Ishibuchi, H.; Masuda, H.; Tanigaki, Y.; and Nokima, Y. 2015. Modified Distance Calculation in Generational Distance and

Inverted Generational Distance. In *Evolutionary Multi-Criterion Optimization* by Obayashi, S; Deb, K.; Poloni, C.; and Murata, T., 110–125, Cham, Springer International Publishing.

Ishibuchi, H.; Tanigaki, Y.; Masuda, H.; and Nojima, Y. 2014. Distance-Based Analysis of Crossover Operators for Many-Objective Knapsack Problems. *In Proceedings of the International Conference on Parallel Problem Solving from Nature*, 600–610. doi.org/10.1007/978-3-319-10762-2_59.

Ishibuchi, H.; Tsukamoto, N.; and Nojima, Y. 2008. Evolutionary Many-Objective Optimization: A Short Review. *In Proceedings of the IEEE Congress on Evolutionary Computing*, 2419–2426. doi.org/ 10.1109/CEC.2008.4631121.

Kukkonon, S.; and Lampinen J. 2007. Ranking-Dominance and Many-Objective Optimization. *In Proceedings for the IEEE Congress on Evolutionary Computation*, 3983–3990. doi.org/10.1109/CEC.2007.4424990.

Li, B.; Li, J.; Tang, K.; and Yao, X. 2015. Many-Objective Evolutionary Algorithms: A Survey. *ACM CSUR*, 48(1), 1–35. doi.org/10.1145/2792984.

Li, X.; Wang, X.; and Xiao, G. 2019. A Comparative Study of Rank Aggregation Methods for Partial and Top Ranked Lists in Genomic Applications. *Briefings in Bioinformatics*, 20(1): 178–189. doi.org/10.1093/bib/bbx101.

Li, M.; Yang, S.; and Liu, X. 2015. Bi-Goal Evolution for Many-Objective Optimization Problems. *Artificial Intelligence*, 228: 45–65. doi.org/10.1016/j.artint.2015.06.007.

Li, M.; Yang, S.; and Liu, X. 2014. Shift-Based Density Estimation for Pareto-Based Algorithms in Many-Objective Optimization. *IEEE Transactions on Evolutionary Computation*, 18(3): 348–365. doi.org/10.1109/TEVC.2013.2262178.

Meijer, R.; Neumann, M., Hemker, B.; and Niessen, S. 2020. A Tutorial on Mechanical Decision-Making for Personnel and Educational Selection. *Frontiers in Psychology*, 10: 1–8. doi.org/10.3389/fpsyg.2019.03002

O'Neill, T. and Steel, P. 2018. Weighted Composites of Personality Facets: An Examination of Unit, Rational, and Mechanical weights. *Journal of Research in Personality*, 73: 1–11. doi.org/10.1016/j.jrp.2017.10.003.

Ployhart, R. and Holtz, B. 2008. The Diversity-Validity Dilemma: Strategies for Reducing Racioethnic and Sex Subgroup Differences and Adverse Impact in Selection. *Personnel Psychology*, 61(1): 153–172. doi.org/10.1111/j.1744-6570.2008.00109.x.

Pyburn, K.; Ployhart, R.; and Karvitz, D. 2008. The Diversity-Validity Dilemma: Overview and Legal Context. *Personnel Psychology*, 61(1): 143–151. doi.org/10.1111/j.1744-6570.2008.00108.x.

Raub, M. 2018. Bots, Bias, and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices. *Arkansas Law Review*, 71(2): 529–570.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. *In Proceedings of the FAccT*, 469–481. doi.org/10.1145/3351095.3372828.

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5): 206–215. doi.org/10.1038/s42256-019-0048-x.

Sánchez-Monedero, J.; Dencik, L.; and Edwards, L. 2020. What Does it Mean to 'Solve' the Problem of Discrimination in Hiring? Social, Technical and Legal Perspectives from the UK on Automated Hiring Systems. *In Proceedings of the FAccT*: 458–468. doi.org/10.1145/3351095.3372849.

Sato, H.; Aguirre, H.; and Tanaka, K. 2011. Genetic Diversity and Effective Crossover in Evolutionary Many-Objective Optimization. *In Proceedings of the International Conference on Learning and Intelligent Optimization*, 91–105. doi.org/10.1007/978-3-642-25566-3_7.

Saxena, D.; Duro, J. Tiwari, A.; Deb, K.; and Zhang, Q. 2012. Objective Reduction in Many-Objective Optimization: Linear and Nonlinear Algorithms. *IEEE Transactions on Evolutionary Computation, 17(1)*: 77–99. doi.org/10.1109/TEVC.2012.2185847.

Schumann, C.; Foster, K., Mattei, N.; and Dickerson, J. 2020. We need Fairness and Explainability in Algorithmic Hiring. *In Proceedings of the 19th International Conference on AAMAS*, 1716–1720.

Sener, O.; and Koltun, V. 2018. Multi-Task Learning as Multi-Objective Optimization. *In Proceedings of NeurIPS*, 527–538.

Shute, V.; Leighton, J.; Jang, E.; and Chu, M. 2016. Advances in the Science of Assessment. *Educational Assessment*, 21(1): 34–59. doi.org/10.1080/10627197.2015.1127752

Society for Industrial, Organization Psychology and American Psychological Association 2018. Principles for the Validation and use of Personnel Selection Procedures, 11(1): 2–97. doi.org/10.1017/iop.2018.195.

Song, Q.; Wee, S.; and Newman, D. 2017. Diversity Shrinkage: Cross-validating Pareto-Optimal Weights to Enhance Diversity Via Hiring Practices. *Journal of Applied Psychology*, 102(12): 1636–1657. doi.org/10.1037/apl0000240.

Sun, Y.; Yen, G.; and Yi, Z. 2019. IGD Indicator-Based Evolutionary Algorithm for Many-Objective Optimization Problems. *IEEE Transactions on Evolutionary Computation*, 23(2): 173–187. doi.org/10.1109/TEVC.2018.2791283.

Wagner, T.; Beume, N.; and Naujoks, B. 2007. Pareto-, Aggregation-, and Indicator-Based Methods in Many-Objective Optimization. *In Proceedings of EMO*, 742–756. doi.org/10.1007/978-3-540-70928-2_56.

Wang, R.; Zhou, Z.; Ishibuchi, H., Liao, T.; and Zhang, T. 2018. Localized Weighted Sum Method for Many-Objective Optimization. *IEEE Transactions on Evolutionary Computation*, 22(1): 3–18. doi.org/10.1109/TEVC.2016.2611642.

Xia, F.; Liu, T.; and Li, H. 2009. Statistical Consistency of Top-k Ranking. *In Proceedings of NIPS*, 2098–2106.

Yuan, Y.; Xu, H.; Wang, B.; Zhang, B.; and Yao, X. 2016. Balancing Convergence and Diversity in Decomposition-Based Many-Objective Optimizers. *Transactions on Evolutionary Computation*, 20(2): 180–198. Doi.org/10.1109/TEVC.2015.2443001.

Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; and Baeza-Yates, R. 2017. Fa*ir: A Fair Top-k Ranking Algorithm. *In Proceedings of CIKM*, 1569–1578.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork C. 2013. Learning Fair Representations. *In Proceedings of ICML*, 325–333.

Zhang, Q.; and Li, H. 2007. MOEA/D: A Multi-Objective Evolutionary Algorithm Based on Decomposition. *IEEE Transactions on Evolutionary Computation*, 11(6): 712–731. doi.org/ 10.1109/TEVC.2007.892759.