Testing Algorithmic Welfare Improvability

Joseph Paul
Heriot-Watt University

December 2024

Abstract
The increasing use of algorithmic decision-making systems across society demands robust methods to evaluate their impacts. While the empirical welfare maximisation literature has made significant advances in deriving optimal treatment assignment policies, existing approaches often focus on asymptotic optimality within simplified policy classes. This paper develops a practical framework for testing whether existing algorithmic policies can be improved upon by alternatives within broad policy classes. The proposed testing procedure requires minimal constraints on the allowed policy class, accommodating state-of-the-art algorithms such as deep neural networks while providing interpretable results in terms of treatment effects. The framework extends naturally to testing for welfare improvements across demographic subgroups, offering an approach to assess and mitigate algorithmic discrimination. I establish the large-sample statistical properties of the testing procedure and demonstrate its finite-sample performance through Monte Carlo simulations. Finally, I illustrate its practical application by evaluating competing patient risk prediction algorithms using data from a healthcare intervention trial. The framework provides practitioners with a practical tool for evidence-based evaluation of algorithmic systems.

Keywords: Empirical Welfare Maximisation, Algorithmic Discrimination, Decision Theory

1 Introduction

Algorithms have emerged as powerful and valuable tools in complex decision problems, offering substantial benefits across various domains. Ludwig, Mullainathan, and Rambachan (2024) argues that algorithmic decision-making can provide a "free lunch in terms of public spending." For instance, in the criminal justice system, an algorithm applied to pretrial release decisions in New York City demonstrated the potential to reduce pretrial detentions by up to 40% without increasing failure-to-appear rates (Ludwig, Mullainathan, and Rambachan 2024). In healthcare, an algorithmic approach to diagnosing heart attacks could reduce unnecessary stress tests and catheterisations, leading to significant cost savings (Mullainathan and Obermeyer 2019).

The power of algorithms lies in their ability to extract signals from complex datasets, often outperforming human judgement in ranking and prediction tasks, allowing for a more efficient allocation of resources and more accurate decision-making. In education, Bergman, Rodriguez, and Kopko (2023) found that an algorithm for college course placement increased college-level class enrolment without compromising pass rates while reducing disparities across racial and ethnic groups. In workplace safety regulations, Johnson, Levine, and Toffel (2019) demonstrate that an algorithm better predicts which work sites will likely have future injuries, potentially preventing thousands of severe injuries and saving hundreds of millions of dollars in lost income.

However, these promising results should not lead to immediate large-scale implementation but encourage further research and development in algorithmic solutions to policy problems. Key challenges remain, such as understanding how human decision-makers will respond to algorithmic tools in practice (Albright 2024), addressing potential data drift over time, ensuring algorithms generalise sufficiently across different contexts, mitigating discrimination, and accurately assessing the impact of algorithmic policies. This paper contributes to the last two challenges.

Recent years have witnessed substantial progress in developing frameworks that aim to maximise the welfare impacts of algorithmic decisions, mainly through the lens of treatment effects. Most notable is the literature on empirical welfare maximisation, including the seminal work by Kitagawa and Tetenov (2018) and Athey and Wager (2020), who established the rigorous foundations for policy learning. They demonstrated the possibility of deriving optimal treatment assignment policies that maximise welfare from experimental and observational data settings. Empirical welfare maximisation aims to leverage treatment heterogeneity to maximise welfare using an algorithmic decision policy. However, the existing literature often focuses on asymptotic optimally within simplified policy classes, prioritising theoretical guarantees over their practical applicability in finite samples. While these contributions are important, there remains a need for more pragmatic approaches to creating and using algorithmic decision rules. Furthermore, it has been argued that using complex models is often preferable to simple models. Kleinberg and Mullainathan (2019) shows that a more complex function exists for every simple prediction function that is strictly more equitable and efficient.

This paper develops a framework for evaluating and improving algorithmic decision-making systems. I propose a testing procedure for *welfare improvability* that compares a status quo algorithm against alternative decision rules. An algorithm maps observed covariates (like a medical profile) to actions (such as treatment decisions). The framework tests whether welfare improvement is possible without compromising other pre-specified objectives such as accuracy, profit, or fairness. Given a pre-specified

class of permissible algorithms, for any given $\delta \in \mathbb{R}$, I define an algorithm to be δ-welfare improvable within the proposed class if some algorithm exists that strictly improves on its welfare by at least a factor of δ. Similarly, I define an algorithm as δ-accuracy improvable if some algorithm exists that has at least the same level of welfare while strictly improvable on accuracy by a δ factor.

In both cases, improvability is defined with respect to a fixed population distribution. However, in reality, we do not know this distribution. Instead, I assume the analyst has access to an i.i.d. sample of individual covariates and ground truth outcomes (i.e., the results from an experiment or the outcome from some observational setting). The sample is used to estimate and make inferences about the population's welfare and the accuracy of the status quo algorithm. The goal of the test procedure is to test whether there exists another algorithm that is a Pareto-improvement on at least one of the criteria.

An overview of the test is as follows. First, the data is split into two subsets: training and test sets. The training set is then used to identify a candidate algorithm that improves on the status quo. Once we identify such a candidate algorithm, we evaluate the impact on the test set to test whether improvements over the candidate are statistically significant in our dual objectives. To avoid computing standard errors case by case and having to specify a limiting distribution of the test statistic, I propose using a bootstrap to compute the critical values. This allows for greater flexibility in the choice of decision functions. Given sufficiently fast convergence in estimating nuisance parameters, I show that the proposed bootstrap procedure consistently estimates average welfare under different algorithmic policies.

While the bootstrap is relatively straightforward in its implementation, its justification is complicated by the fact that the allowed class of algorithms and the procedure for choosing an algorithm have very few restrictions placed on them. This means that the distribution of the test statistic may vary with the sample size and is not guaranteed to settle down in the limit. The bootstrap's validity in this setting follows from the work of Auerbach et al. (2024), who build on the triangular array results from Mammen (1992).

While these theoretical results hold for a single data split, relying on a single random split introduces additional uncertainty to the testing procedure. To reduce this uncertainty, I recommend using multiple data splits and combining their results. Specifically, I show that rejecting the null hypothesis when the median p-value across splits falls below $\alpha/2$ provides asymptotically valid coverage at significance level α. This multiple-split approach maintains the theoretical guarantees while reducing the variance any random split introduces. I show that this test is also consistent under suitable convergence assumptions on the selection rule. Intuitively, convergence assumption means the selection rule must asymptotically identify algorithms that improve upon the status quo when such improvements exist. This could be formalised through PAC learnability theory in future work.

This approach offers several key advantages over existing methods in the empirical welfare maximisation literature. First, it allows for exceptional flexibility in policy class selection, with minimal constraints placed on permitted policies a priori. The procedure readily accommodates complex state-of-the-art algorithms, including deep neural networks and large ensembles, which often outperform simpler, analytically tractable policy classes in practice. Second, it produces interpretable results expressed as treatment effects that are accessible to policymakers, legal experts, and other non-technical stakeholders. Third, it has valuable applications in legal and ethical contexts, providing a rigorous framework for demonstrating either the absence of discriminatory practices or the impossibility of Pareto improvements in welfare for specific subgroups.

Due to the flexibility in the allowable class of algorithms, this approach readily accommodates many practical challenges that arise when algorithms are deployed in real-world settings with exogenous constraints - whether legal, ethical, or logistical. These constraints can take various forms depending on the context, including capacity limitations on treatments, monotonicity requirements on decisions with respect to specific inputs or statistical restrictions needed to satisfy fairness criteria. Importantly, this procedure remains valid across different algorithm classes, constraints, and utility functions, subject only to regularity conditions on the selection rule.

Overall, this approach takes a more pragmatic view compared to traditional empirical welfare maximisation. Rather than solely identifying optimal policies within restricted classes, it provides a framework for evaluating algorithmic systems based on their real-world impacts. This fills a critical gap as algorithmic systems become increasingly influential in high-stakes decisions across society. The framework also serves as a practical tool for policymakers, legal professionals, researchers and ethicists to assess the efficacy and fairness of algorithmic decisions.

In the following sections, I first review the relevant literature, notably empirical welfare maximisation and algorithmic fairness. I then formally define welfare improvability along with the accompanying notions of accuracy, where "accuracy" is used as an umbrella term for any objective not directly related to welfare. I also consider how to measure and test inequalities across groups for settings where we might be concerned with disparate impact. The framework's applications to legal contexts, particularly in addressing disparate impact cases, are discussed in detail in Section 1.1.

1.1 Legal Justification

This framework has direct applications to policy evaluation and legal compliance in assessing whether algorithmic policies can be improved while maintaining other critical objectives.

The framework is especially relevant for analysing disparate impact across demographic groups. Legal frameworks in Europe, the UK, and the US prohibit discrimination based on protected characteristics, with US law specifically addressing two forms of discrimination. Disparate treatment concerns intentional discrimination based on protected class membership, while disparate impact focuses on policies that disproportionately benefit or harm certain groups, even without discriminatory intent. This procedure can help in establishing and evaluating disparate impact cases.

Disparate impact cases are usually assessed in three steps. The first step is to establish whether the harms or benefits of some policy disproportionality affect some protected class members. If it is established that there is some disparate impact then the next step is determining whether there exists business necessity, i.e. the observed difference is necessary to achieve some legitimate, non-discriminatory aim. If the business necessity case determines that there is business necessity, then the final step is to determine whether there are some less discriminatory alternatives that could achieve the same aim. This framework provides statistical tools for evaluating the second and third steps - quantifying both the necessity of observed disparities and the existence of less discriminatory alternatives.

In the United Kingdom, the UK Government's Data Ethics Framework encourages practitioners to evaluate their transparency, fairness, and accountability (**REF**). In contrast, the computer science literature, which primarily influences these ideas, often overlooks the discussion of "real-world" impacts or methods for measuring them. While the focus in computer science has largely been on the technical

aspects of fairness, the framework introduces an approach to assessing real-world effects, fulfilling the UK Government's Data Ethics Framework's emphasis on empirical evaluation of algorithmic systems.

1.2 Related Literature

This paper connects several distinct literatures across economics, statistics, and computer science that study algorithmic policies. The evaluation of treatment assignment rules has a long history in economics. Classic examples include enrolment in government welfare programs (Dehejia 2005), job training programs programs (Black et al. 2003; Frölich 2008) and judge sentencing decisions (Bushway and Smith 2007). This literature has often used a minimax regret framework, an approach pioneered by Manski (2004).

The more recent literature on empirical welfare maximisation is directly related to this paper, which develops statistical methods to identify optimal policies from predefined policy classes and can be broadly categorised into model-based and direct-search approaches. Model-based approaches from computer science include Q-learning (Nahum-Shani et al. 2012) and A-learning (Shi et al. 2018), which estimate conditional expectation functions to then determine optimal policies. However, these methods can be sensitive to misspecification of the underlying functional forms.

Direct-search methods, in contrast, optimise policy parameters directly without intermediate modelling steps. Outcome weighting (Zhao et al. 2012) attempts to learn the optimal policy non-parametrically using an inverse probability weighting estimator (IPWE). However, it is well known that the potential for instability is caused by extreme propensity scores and model specifications. To increase, stability some methods use the augmented IPWE (Zhang et al. 2012; Zhao et al. 2019; Athey and Wager 2020; Pan and Zhao 2021). Athey and Wager (2020) 's seminal work defines a minimax optimal policy learning algorithm with optimal regret guarantees under suitable regularity conditions on the estimators. These methods generally come with stronger theoretical guarantees than IPWE but can still suffer from extreme weights.

Similar to Athey and Wager (2020), I propose using debiased double machine learning to estimate the causal effects of an algorithmic decision function, using cross-fitting and classical estimators or flexible machine learning ones for the estimation of nuisance functions. This paper differs in that it allows a much larger and broader class of algorithms and is a lot more flexible, but it comes with no guarantees of asymptotic optimality in terms of regret.

The second strand of literature is that of algorithmic fairness, principally from computer science. This literature often conceptualises fairness through statistical constraints on prediction functions. Key works include Zemel et al. (2013), which proposes "statistical parity" as a fairness measure, requiring equal probability of classifications across groups. Feldman et al. (2015) define fairness through equal classification accuracy across groups, while Hardt, Price, and Srebro (2016) introduce "equalised odds" and "equal opportunity" criteria focusing on equal error rates. See Mitchell et al. (2021) for a comprehensive survey.

These fairness definitions are typically imposed as constraints during algorithm training. However, Kleinberg, Mullainathan, and Raghavan (2016) shows that most fairness definitions are mathematically incompatible except in trivial cases. Furthermore, Liu et al. (2018) demonstrates that while these methods may achieve static optimality, they can break down and potentially harm disadvantaged groups in dynamic settings.

More recent work bridges predictive fairness with welfare economics concepts. Hu and Chen (2018) evaluate how fairness criteria affect group and individual welfare. Balashankar et al. (2019) introduce "Pareto-Efficient Fairness" to identify optimal rules maximising subgroup accuracy without sacrificing overall performance. Viviano and Bradic (2023) advocate for Pareto optimal treatment rules balancing welfare across sensitive groups.

Most closely related to this paper is that of Auerbach et al. (2024) who defined a testing procedure based on the error rate of a given algorithm and focused on optimising these statistical fairness objectives. This paper differs in its focus on measured welfare.

Several economics papers have theoretically integrated welfare analysis into algorithmic decision-making. Rambachan et al. (2020) examines algorithmic bias through a welfare economics lens, showing how social planners' equity preferences affect prediction use rather than construction. Cowgill and Stevenson (2020) explores designing algorithms to influence human decision-makers while maintaining incentive compatibility.

This paper builds on this literature by developing a practical framework for evaluating algorithmic policies that incorporate both welfare and fairness considerations. Rather than imposing fairness constraints during training, I assess algorithms based on their realised welfare impacts across groups.

2 Algorithm Evaluation

2.1 Set-up

I assume the analyst has access to independent and identically distributed samples $\{(X_i, Y_i, D_i, Z_i)\}_{i=1}^n$ which we collect in W_i, where $X_i \in \mathcal{X}$ describes the characteristics of individual i, $D_i \in \mathcal{D}$ is the observed treatment assignment, $Z_i \in \mathcal{Z}$ is in an (optional) instrument, and $Y_i \in \mathbb{R}$ is as the outcome we care about and which we wish to intervene on. I use $\mathbb{P}$ to denote the joint distribution of (X, Y, D, Z) across individuals. A decision-making algorithm $a \in \mathcal{A}$ is a mapping from a subject's features to a decision $a : \mathcal{X} \to \mathcal{D}$. I evaluate algorithms with respect to a welfare utility function $u_w : \mathcal{X} \times \mathcal{Y} \times \mathcal{D} \to \mathbb{R}$ and an accuracy utility function $u_a : \mathcal{X} \times \mathcal{Y} \times \mathcal{D} \to \mathbb{R}$, discussed in Section 3.2.

2.2 Welfare

I follow the existing literature (Manski 2004; Hirano and Porter 2009; Kitagawa and Tetenov 2018) and define the welfare induced by an algorithm $a(\cdot)$ as the expected utility under the algorithmic policy relative to the expected utility under no treatment. Formally, in the case where a decision space is discrete, the welfare induced by an algorithm a is given by

$$V(a) = \mathbb{E}[Y_i(a(X_i))] - \mathbb{E}[Y_i(0)] \tag{1}$$

where $Y_i(a(X_i))$ is the potential outcome for individual i under.

When the decision space is continuous, I define *welfare* as infinitesimal interventions (Athey and Wager 2020):

$$V(a) = \left[\frac{d}{d\nu}\mathbb{E}[Y_i(D_i + \nu a(X_i))]\right]_{\nu=0}.$$

In both of these cases, we can equivalently define welfare in terms of conditional average treatment effects:

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] \quad \text{or} \quad \tau(x) = \left[\frac{d}{d\nu}\mathbb{E}[Y_i(D_i + \nu) \mid X_i = x]\right]_{\nu=0},$$

with welfare then being defined as

$$V(a) = \mathbb{E}[a(X_i)\tau(X_i)]. \tag{2}$$

2.3 Policy Learning

The approach I take in this paper builds on the work of Athey and Wager (2020) and Chernozhukov et al. (2022). Chernozhukov et al. (2022) show that we can construct semi-parametric efficient estimates of the average treatment effect $\theta = \mathbb{E}[\tau(X_i)]$ using

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i$$

where $\hat{\psi}_i$ is an the appropriate doubly robust score for the target estimated. Given that we can estimate an average treatment effect, I show that we use these estimates to get unbiased estimates of the expected welfare under a fixed algorithmic policy.

Assumption 1: Let $m(x, d) = \mathbb{E}[Y_i(d) \mid X_i = x] \in \mathcal{M}$ represent the counterfactual response surface. I assume this function induces a treatment effect function $\tau_m(x, d)$ with the following two properties.

1. The mapping $m \mapsto \tau_m(\cdot)$ is linear in m, and there exists a weighting function $g(x, z)$ such that
$$\mathbb{E}[\tilde{\tau}_m(X_i, D_i) - g(X_i, Z_i)\tilde{m}(X_i, D_i) \mid X_i] = 0 \tag{3}$$
for any response surface $\tilde{m}$.
2. Algorithm welfare can be defined in terms of moments of $\tau_m(X_i, D_i)$, such that $V(a) = \mathbb{E}[\tau_m(X_i, a(X_i))]$, for any $a : \mathcal{X} \to \{0, 1\}$.

We can construct the estimate of $\hat{\theta}$ using the well-known augmented inverse-propensity weighted scores of Robins, Rotnitzky, and Zhao (1994)

$$\psi_i = \psi(W_i) \coloneqq \tau_m(X_i, D_i) + \hat{H}(D_i, X_i, Z_i)(Y_i - \hat{g}(D_i, X_i))$$

where

$$\hat{H}(D, X, Z) \coloneqq \frac{D}{\hat{m}(X, Z)} - \frac{1 - D}{1 - \hat{m}(X, Z)}.$$

Chernozhukov et al. (2022) shows that $\frac{1}{n}\sum_i \hat{\psi}$ is a $\sqrt{n}$-consistent estimate of θ and is asymptotically unbiased Gaussian, given that we estimate $\hat{g}$ and $\hat{m}$ using cross-fitting and that they converge sufficiently fast in terms of root mean squared error. This is known as a double robust estimator due to

having the property that it is consist if either the propensity score or the counterfactual response surface needs to be adequately modelled, but not both necessarily.

Flexibility in estimating the nuisance functions lies at the heart of this approach. Following Chernozhukov et al. (2022), I remain agnostic about the specific method used to estimate the nuisance components. It is only required that these estimates meet certain convergence rates, given in Assumption 2, and are estimated using sample splitting. Sample splitting avoids the need for Donsker conditions on the nuisance functions (Chernozhukov et al. 2022), which complicated much of the early work in this area. *(NB: Discussion on choices of estimators (classical and ML) and ensembling needed.)*

I follow Athey and Wager (2020) and impose the following high-level conditions on the rate of convergence of the nuisance functions.

Assumption 2: I assume that the second moments of m_n, g_n, and H_n are uniformly bounded for all $n = 1, ..., \infty$. Furthermore, I assume that our estimators are uniformly consistent estimates of the nuisance components, such that

$$\sup_{x,d}\{|\hat{m}_n(x,d) - m_n(x,d)|\} = o_p(1), \quad \sup_{x,z}\{|\hat{g}_n(z,x) - g_n(z,x)|\} = o_p(1)$$

$$\sup_{x,d}\Big\{|\tau_{\widehat{m}_n}(x,d) - \tau_{m_n}(x,d)|\Big\} = o_p(1),$$

where the L_2 erros decay, for some $0 < \zeta_m, \zeta_g < 1$ and $\zeta_m + \zeta_g \geq 1$[1] and some sequence $a(n) \to 0$:

$$\mathbb{E}\Big[\Big(\hat{m}_{n(X,D)} - m_{n(X,D)}\Big)^2\Big] \leq \frac{a(n)}{n^{\zeta_m}}, \quad \mathbb{E}\Big[\Big(\hat{g}_{n(Z,X)} - g_{n(Z,X)}\Big)^2\Big] \leq \frac{a(n)}{n^{\zeta_g}},$$

$$\mathbb{E}\Big[\Big(\tau_{\widehat{m}_n}(X,D) - \tau_{m_n}(X,D)\Big)^2\Big] \leq \frac{a(n)}{n^{\zeta_m}}.$$

I next show how we can use this setup to estimate the welfare in several settings of interest.

2.3.1 Binary treatment with selection on observables

In the case of binary treatment and uncounfoundedness $(Y_i(0), Y_i(1) \perp\!\!\!\perp D_i \mid X_i)$, then weighing by the inverse propensity scores will estimate the average treatment effect. To do so, define $g(x,d) = (d - e(x))/e(x)(1 - e(x))$ where $e(x) = \mathbb{P}[D_i \mid X_i = x]$ is the propensity score. Then

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}(X_i, 1) - \hat{m}(X_i, 0) + \frac{D_i - \hat{e}(X_i)}{\hat{e}(X_i)(1 - \hat{e}(X_i))}(Y_i - \hat{m}(X_i, D_i))\right),$$

is the augmented inverse propensity weighting estimator (Robins, Rotnitzky, and Zhao 1994).

[1] If $\zeta_m = \zeta_g = 1/2$, then this is equivalent to the standard assumption from the efficient semi-parametric estimation literature in that all nuisance components are $o\big(n^{-1/4}\big)$-consistent in terms of L_2 error. This is a weaker assumption that comes from the robustness properties of doubly robust estimators and allows us to trade-off the error rates of the two nuisance functions, provided that the product of the two are controlled (Farrell 2015).

2.3.2 Continuous treatment with selection on observables

When the treatment is continuous and unconfounded ($\{Y_i(d)\}_{d\in\mathcal{D}} \perp\!\!\!\perp D_i \mid X_i$), then the conditional treatment effect $\tau_{m(x,d)} = \left[\frac{d}{d\nu} m(x, d+\nu)\right]_{\nu=0}$ can be identified via

$$\int\int g(X_i, D_i) m(X_i, D_i) dF_{D_i \mid X_i} dF_{X_i},$$

where $g(X_i, D_i) = -\frac{\partial}{\partial d}[\log(f(w \mid X_i))]_{d=D_i}$, and $f(\cdot \mid x)$ is the conditional density of D_i given $X_i = x$ (Athey and Wager 2020).

2.3.3 Binary treatment with binary instrument

When the treatment is binary and there is an instrument Z_i that satisfies the exclusion restriction ($Y_i(d) \perp\!\!\!\perp Z_i \mid X_i$), then we can use a weighting function $g(\cdot)$ (Abadie 2003) that are estimates of scores, and identify τ_m from Eq. (3) using

$$g(X_i, Z_i) = \frac{1}{\Delta(X_i)} \frac{Z_i - z(X_i)}{z(X_i)(1 - z(X_i))}, \quad z(x) = \mathbb{P}[Z_i = 1 \mid X_i = x]$$
$$\Delta(x) = \mathbb{E}[D_i \mid Z_i = 1, X_i = x] - \mathbb{E}[D_i \mid Z_i = 0, X_i = x].$$

2.4 From average to conditional effects

Having shown that we can use the above methods to estimate the average treatment effect, we can use these estimates to get unbiased estimates of the expected welfare under a fixed algorithmic policy.

Theorem 1 (Algorithmic Welfare Estimation): Under Assumption 1 and Assumption 2, for any measurable fixed policy a:

$$\sqrt{n}\big(\hat{V}_n(a) - V_n(a)\big) \underset{p}{\to} 0$$

2.5 Scores Estimation

As in Chetverikov et al. (2016), Chernozhukov et al. (2022), I assume that scores are obtained via cross-fitting, where the sample is split into K folds. The k-th fold is used to estimate the nuisance functions, and the remaining $K-1$ folds are used to estimate the treatment effect. This is used to generate asymptotic normality, given only high-level assumptions on predictive accuracy and convergence rates of the nuisance functions. This allows us to use machine learning methods or classical estimators as long as they are sufficiently accurate. Then, we estimate the scores for individual i as

$$\hat{\psi}_i = \hat{g}_n^{-k(i)}(1, X_i) - \hat{g}_n^{-k(i)}(0, X_i) + \hat{H}_n^{-k(i)}(D_i, X_i, Z_i)\big(Y_i - \hat{g}_n^{-k(i)}(D_i, X_i)\big)$$

where $k(i)$ denotes the fold that observation i belongs to.

Cross-fitting is closely related to cross-validation and is a common technique in machine learning for choosing parameters. Instead, we use it to estimate the nuisance functions in a way that preserves the independence of the scores and is essential for valid inference.

2.5.1 Bootstrapping Scores

For each i, we need to generate a set of bootstrapped scores $\left\{\hat{\psi}_i^b\right\}_{b=1}^{B_1}$. This forms the inner loop of our double bootstrap procedure. These bootstrapped scores will be used to estimate the sampling distribution of our test statistics under the null hypothesis.

The procedure is as follows:

1. For each bootstrap iteration $b = 1, ..., B_1$:
   - Generate a bootstrap sample by resampling with replacement from the original data
   - Re-estimate the nuisance parameters $\left(\hat{m}_{n,b}, \hat{g}_{n,b}\right)$ using the bootstrap sample
   - Calculate the bootstrapped score for each i:

$$\hat{\psi}_i^b = \hat{g}_{n,b}^{-k(i)}(1, X_i) - \hat{g}_{n,b}^{-k(i)} + \hat{H}_{n,b}^{-k(i)}(D_i, X_i, Z_i)\left(Y_i - \hat{g}_{n,b}^{-k(i)}(D_i, X_i)\right)$$

2. Store the collection of bootstrapped scores $\left\{\hat{\psi}_i^b\right\}_{b=1}^{B_1}$ for each individual

These bootstrapped scores will then be used in the outer bootstrap loop to estimate the critical values of our test statistics, as discussed in the following sections.

Note that this procedure maintains the cross-fitting structure discussed earlier - the nuisance parameters for each bootstrap iteration are estimated using data excluding observation i to preserve the independence needed for the theoretical results.

3 Model

Having shown how we can estimate the expected welfare of an algorithm, I will now outline the testing procedure followed by the main results.

I propose that we evaluate algorithms in terms of welfare and accuracy. In many cases, we might not have any reason to believe these two objectives are at odds. However, there might be cases where this is so, such as considering a population's welfare and a supplier's profit. This will become more salient when discussing group differences, where a higher cost might be associated with reducing discrepancies in observed welfare differences.

3.1 Definitions

As above, I will denote the expected welfare induced by the algorithm a as $V(a)$. Other objectives will enter into a utility function $u(x, y, d)$ that could include the algorithm's accuracy, the expected profit/cost, or some other preferences of the social planner. The expectation of this utility will be given by

$$U(a) = \mathbb{E}[u(X, Y, a(X))]$$

Definition 1: An algorithm a_1 has **higher welfare** than a_0 if

$$V(a_1) > V(a_0)$$

When we are considering groups, an algorithm a_1 has **superior joint group welfare** compared to a_0 if, for a set of groups $\mathcal{G} = \{g_1, g_2, ..., g_k\}$, we have:

$$\forall g \in \mathcal{G} : V^g(a_1) > V^g(a_0)$$

Similarly, algorithm a_1 is more accurate if

$$U(a_1) > U(a_0)$$

and when considering groups,

$$\forall g \in \mathcal{G} : U^g(a_1) > U^g(a_0)$$

As mentioned above, *accuracy* is an umbrella term for any objective not directly related to welfare. Below, I list some examples.

Example (Profit): Consider a credit scoring algorithm where the utility function is defined as profit:

$$U(a) = \mathbb{E}[a(X)(\text{Loan Interest} - \text{Expected Default Loss})]$$

Here, $D = 1$ represents issuing a loan, and $D = 0$ represents rejecting the loan application.

Example (False Positive Rate): Both Y and the decision are binary. The per-group false positive rate is

$$U^g(a) = P(a(X) = 1 \mid Y = 0, G = g)$$

Example (Revenue): An algorithm offers a good to an individual at a price d. The individual has an unobserved willingness to pay Y. The revenue for the firm is given by

$$u(x, y, d) = \begin{cases} d \text{ if } \ y \geq d \\ 0 \text{ otherwise} \end{cases}$$

$U(a)$ is thus the expected revenue under algorithm a.

3.2 Welfare/accuracy improvability

I assume there is a *status quo algorithm* $a_0 : \mathcal{X} \to \mathcal{D}$ that is an algorithm that has been proposed for use or is already in use.

Definition 2 (Welfare Improvement): Fix any $\delta \in \mathbb{R}_+$. An algorithm $a_1 \in \mathcal{A}$ constitutes a δ-weflare improvement if and only if

$$\frac{V(a_1)}{V(a_0)} \geq 1 + \delta.$$

An algorithm a_1 leads to a δ-improvement on the status-quo algorithm a_0 if it increases expected welfare by a factor of δ. I define accuracy similarly.

Definition 3 (Accuracy Improvement): Fix any $\delta \in \mathbb{R}_+$. An algorithm $a_1 \in \mathcal{A}$ constitutes a δ-welfare improvement if and only if

$$\frac{U(a_1)}{U(a_0)} \geq 1 + \delta.$$

Our main objective is to test whether we can improve on our status-quo algorithm without harming some other objective.

The following definition of welfare-accuracy-improvement formally defines the joint improvement across both welfare and accuracy.

Definition 4 (Welfare-Accuracy Improvement): Fix any $\delta_w, \delta_a \in \mathbb{R}_+$. An algorithm $a_1 \in \mathcal{A}$ constitutes a δ_w, δ_a-welfare accuracy improvement if and only if[2]

$$\frac{V(a_1)}{V(a_0)} \geq 1 + \delta_w \text{ and } \frac{U(a_1)}{U(a_0)} \geq 1 + \delta_a. \tag{4}$$

An additional test we might be interested in if fairness is a concern is the absolute difference between the two welfare estimates $\mid V^g(a) - V^{g'}(a) \mid$ for $a \in \{a_0, a_1\}$. In this case, we will assume the following:

Assumption 3: The status quo algorithm is not perfectly fair. More specifically, the absolute difference in welfare between different groups is bounded away from 0:

$$\mid V^{g'}(a_0) - V^{g'}(a_0) \mid \geq \eta,$$

for some $\eta > 0$.

4 Main Result

4.1 Proposed Approach

I consider a setting where an analyst does not know the joint distribution of (X, Y, D, Z), but instead observes an independently and identically distributed sample $\{(X_i, Y_i, D_i, Z_i)\}_{i=1}^n$ of size n. The analyst wishes to test the welfare improvability and accuracy improvability of a status quo algorithm a_0 that is is already in use or has been proposed for use. Formally, the analyst proposes a class of algorithms $\mathcal{A}$, as well as specifies $\Delta = (\delta_w, \delta_a)$. They then test the following null hypothesis:

$$H_0: \text{ The algorithm } a_0 \text{ is not } \Delta\text{-improvable within class } \mathcal{A},$$

[2]NB: I still need to formally define the group setting.

against

$$H_1 : \text{There exists an algorithm in } \mathcal{A} \text{ that is a } \Delta\text{-improvement over } a_0.$$

4.2 Procedure

The analyst first estimates a series of bootstrapped estimates of ψ_i ($\left\{\psi_i^b\right\}_{b=1}^{B_1}$)[3], by following the steps outlined in section REF.

The analyst then chooses a *selection rule* ρ that maps a sample into an algorithm from $\mathcal{A}$:

$$\rho : \mathcal{S} \to \mathcal{A},$$

where $\mathcal{S} = \bigcup_{m \geq 1} \mathcal{S}_m \equiv \bigcup (\mathcal{X} \times \mathcal{Y})^m$ is the set of all finite samples of observations.

The analyst then chooses the number of times the split and test is repeated K, the training sample size β where $\ell_n = \lfloor n\beta \rfloor$ is the testing sample size. Then, for each round in $k = 1, ..., K$[4]: 1. Split the sample into a train and test set 2. Find a candidate algorithm using the training sample 3. Test whether $\hat{a}_{1n}^{\rho}$ constitutes a Δ-improvement on a_0 using the test outlined in Section REF. Let p_k be the p-value associated with this test 4. Repeat steps 1-3 K times, returning us a vector of p-values $\vec{p} = [p_1, ..., p_k$ 5. Reject the test if med $(\vec{p}) < \frac{\alpha}{2}$.

4.2.1 Algorithm: Split and Test Procedure

Inputs: - Number of repetitions: K - Training sample size fraction: β - Significance level: α - Total sample size: n

Outputs: - Decision to reject the null hypothesis

Procedure:

1. **Determine Sample Sizes**
   - Compute the training sample size:

     $$\ell_n = \lfloor n\beta \rfloor$$

   - The testing sample size is then $n - \ell_n$.
2. **Repeat for Each Round $k = 1$ to K**

   1. **Split the Data**

   - Divide the dataset into:
   - Training set of size ℓ_n
   - Testing set of size $n - \ell_n$

   1. **Train the Model**

   - Identify a candidate algorithm with the selection rule ρ using

   the training sample S_{train}.

   1. **Perform the Test**

   - Evaluate whether $\hat{a}_{1n}^{\rho}$ constitutes a Δ-improvement over a_0 using the test outlined in Section

[3]NB: When evaluating the algorithms, should they use different bootstrapped samples to decorrelate the welfare estimates?

[4]Test follows the procedure proposed by REF

REF.

- Let p_k be the p-value obtained from this test.

3. **Aggregate P-Values**
   - Collect all p-values into a vector:

$$\vec{p} = [p_1, p_2, ..., p_K]$$

4. **Make a Decision**
   - Calculate the median of the p-values:

$$\text{med } (\vec{p})$$

   - **Reject the null hypothesis** if:

$$\text{med } (\vec{p}) < \frac{\alpha}{2}$$

This procedure is visually represented in Figure 1.

SPLIT AND TEST PROCEDURE(K, β, α):
let $\ell_n = \lfloor n\beta \rfloor$ be the training sample size
for $k = 1, ..., K$ **do**
    Split the data into S_{train} **and** S_{test}
    Let $\alpha^{\rho}_{1,n,k} = \rho(S_{\text{train}})$
    perform the test using S_{test}
    collect the p-value
let $\vec{p} = [p_1, p_2, ..., p_K]$ be the vector of p-values
if med $(\vec{p}) < \frac{\alpha}{2}$
    reject the null hypothesis
end

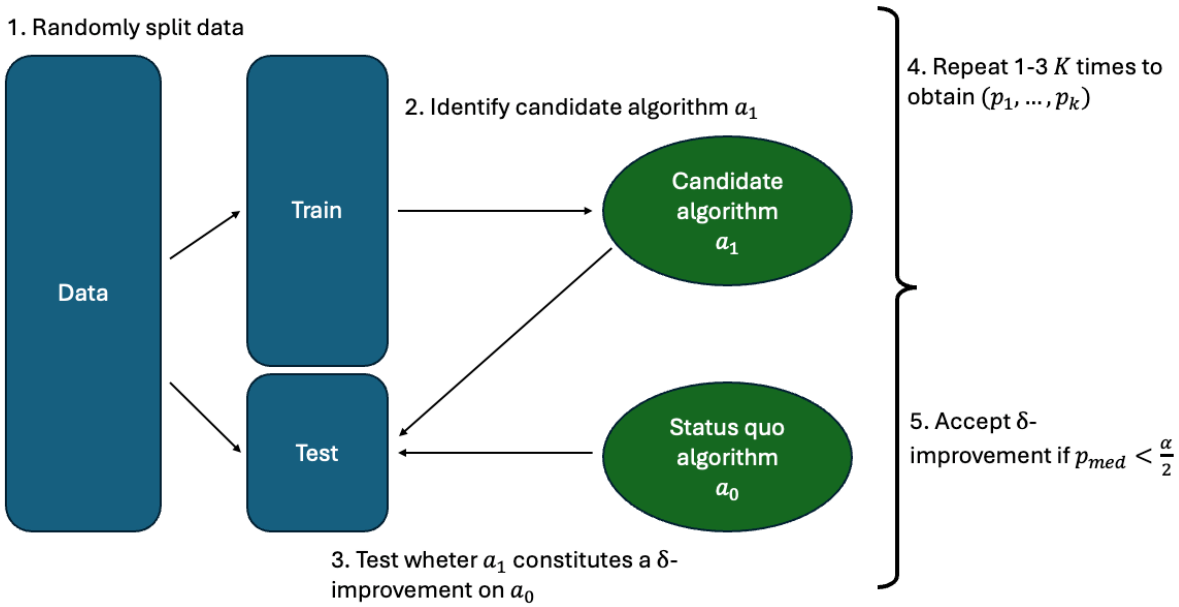

Figure 1: The Split and Test Procedure

4.3 The Double Bootstrap Test for Δ-Improvability

The below procedure outlines steps for the double-bootstrap for testing if there exists a Δ-improvable function within the class $\mathcal{A}$ over a_0. The procedure has an inner and outer bootstrap loop. The inner bootstrap accounts for uncertainty in the estimate of $\hat{V}(a)$, while the outer bootstrap is used to estimate the distribution of the test statistic under the null.

In the remainder of the Section, I provide a detailed treatment of the testing procedure from step 3 and the outer loop of the bootstrap. Since $\hat{a}_{1n}^{\rho}$ was selected using the training sample S_{train}, we can treat $\hat{a}_{1n}^{\rho}$ as a fixed function. Our null hypothesis is then given by

$$H_0 : V(\hat{a}_{1n}^{\rho}) \leq V(a_0)(1+\delta_w) \quad \text{OR} \quad U(\hat{a}_{1n}^{\rho}) \leq U(a_0)(1+\delta_u), \tag{5}$$

with the alternative being defined as

$$H_1 : V(\hat{a}_{1n}^{\rho}) > V(a_0)(1+\delta_w) \quad \text{AND} \quad U(\hat{a}_{1n}^{\rho}) > U(a_0)(1+\delta_u). \tag{6}$$

Note that the null hypothesis is a combination of two tests (one for welfare, one for accuracy). Therefore, we can test these conditions separately and use the intersection-union method to combine the tests. Intuitively, we can test each condition separately and reject the null only if both tests reject the null.

Define our test statistics as

$$T_{w,n} = \sqrt{\ell_n}(V_n(a_1) - (1+\delta_w)V_n(a_0))$$
$$T_{u,n} = \sqrt{\ell_n}(U_n(a_1) - (1+\delta_u)U_n(a_0))$$

The test statistic of welfare can be computed from the asymptotic properties of the DDML estimator (Chernozhukov et al. 2022). However, the asymptotic distribution of the test statistic for the accuracy would vary depending on the specific choice of the utility function. I, therefore, propose using the nonparametric bootstrap to generate critical values of the test and makes up the outer loop of the double bootstrap.

Denote $\phi_n^{(t)}$ the rejection rule for each test t. For a nominal significance level α, the reject rule is given by $\phi_n^{(t)} := \mathbb{1}\left\{\hat{T}_{w,n} > \Psi_{t,n}^{-1}(1-\alpha)\right\}$ where $\Psi_{t,n}^{-1}(1-\alpha)$ is the $(1-\alpha)$ quantile of the bootstrap distribution of the test statistic t under the null, which I define below. The p-value can associated with the test $\phi_n^{(t)}$ is given by $1 - \Psi_{t,n}\left(\hat{T}_{w,n}\right)$.

Our full test of the joint null hypothesis is given by

$$\phi_n(a_0, a_1) = \phi_n^{(w)} \cdot \phi_n^{(u)}.$$

We only reject if both tests reject the null. Furthermore, the p-value associated with the test ϕ_n is the maximum of the p-values of the two individual tests.

4.4 Main Results

Due to the flexibility of the selection rule ρ, there are a number of challenges with the asymptotic theory of the testing procedure. Without additional regularity assumptions on ρ, such asymptotic theory would not exist for the general case.

As well as the assumptions placed on the data to get consistent estimates of welfare, we will also make the following assumption.

Assumption 4: Let $\mathcal{M}$ be the set of covariance matrices of the nuisance parameters and objective functions for every candidate algorithm $a_1 \in \mathcal{A}$ that can be realised by the selection rule ρ. Then

$$\inf_{\Sigma \in \mathcal{M}} \lambda_{\min}(\Sigma) \geq v,$$

for some $v > 0$, where $\lambda_{\min}(\Sigma)$ denotes the smallest eigenvalue of Σ.

This assumption states that the limiting variances of the test statistics are non-degenerate in the set of realisable candidate algorithms in $\mathcal{A}$. This uniform non-degeneracy of limiting variance is important to ensure the test statistic remains well-behaved even as we allow the candidate algorithms to vary arbitrarily through any selection rule. Without this assumption, certain selection rules could choose algorithms that are so similar to the status quo that their difference in utilities becomes arbitrarily close to zero, leading to degenerate test statistics and unreliable inference. Intuitively, this assumption requires that the variance of utilities between any candidate algorithm and the status quo remains bounded away from zero, ensuring that we can meaningfully distinguish between them in our statistical tests.

Theorem 2: Suppose P and $\mathcal{A}$ satisfy the above assumptions and suppose the null hypothesis holds. Then, for a nominal significance level $\alpha \in (0, 1)$

$$\lim_{n \to \infty} \sup \mathbb{E}_P \left[\phi_n \left(a_0, \hat{a}^{\rho}_{1,n} \right) \right] \leq \alpha$$

While this result holds for a single data split, relying on one random split in practice introduces additional uncertainty. To reduce this, we can extend the framework to multiple data splits. Let $\hat{a}^{\rho}_{1,n,k} = \rho\left(S^{n,k}_{\text{train}}\right)$ denote the algorithm chosen on split k for $k = 1, ..., K$. We can then define an aggregate rejection rule $\varphi_n = \mathbb{1}\left\{ \frac{1}{K} \sum_{k=1}^{K} \phi_{n,k}(a_0, \hat{a}_{1,n,k}) \geq \frac{\alpha}{2} \right\}$. This rule rejects the null hypothesis when the median test statistic across splits is less than $\frac{\alpha}{2}$, which I show in the following corollary to Theorem 2.

Corollary 1: Under the assumptions of Theorem 2, for a nominal significance level $\alpha \in (0, 1)$, we have

$$\lim \sup_{n \to \infty} \mathbb{E}_P[\varphi_n] \leq 2\alpha$$

Corollary 1 implies that to achieve a significance level α, each individual test must be of level $\alpha/2$. We, therefore, reject the null if $\text{med}(\vec{p}) < \alpha/2$. This does imply a loss of power in the test. However, the local power analysis of DiCiccio, DiCiccio, and Romano (2020) reveals this may not be as large an issue as it first appears. I will explore this further in later revisions of the paper.

Finally, I wish to show that the test is consistent. However, this requires additional assumptions on the selection rule ρ. Intuitively, we have yet to place any assumptions thus far that the selection rule ρ will asymptotically identify algorithms that improve upon the status quo when such improvements exist. I will, therefore, have to make the following assumption.[5]

[5] I follow the definition of Auerbach et al. (2024) in defining the consistency of the test.

Assumption 5: Suppose the null hypothesis does not hold, i.e. there exists an algorithm $a \in \mathcal{A}$ that is a Δ-improvement over a_0. A selection rule ρ is said to be **improvement consistent** if $V\left(\hat{a}^{\rho}_{1,n}\right) \underset{p}{\to} V(\gamma)$ and $U\left(\hat{a}^{\rho}_{1,n}\right) \underset{p}{\to} U(\gamma)$ for some algorithm γ that is a Δ-improvement on a_0.

Assumption 5 intuitively states that if there exists an algorithm in $\mathcal{A}$ that leads to a Δ improvement over a_0, then the selection rule $\rho(\cdot)$ will find it in the limit. Under this assumption,

Theorem 3: Suppose P, $\mathcal{A}$ and γ satisfy Assumption 4-Assumption 5, and that the null distribution in Eq. (5) doesn't hold. Then

$$\lim_{n\to\infty} \mathbb{E}_{\mathbb{P}}\left[\phi_n\left(a_0, a^{\rho}_{1,n}\right)\right] = 1$$

It is worth noting that while Assumption 5 is sufficient for consistency, it may not be necessary in practice. I believe many selection rules that do not strictly satisfy this assumption may still lead to rejection of the null hypothesis when improvements exist. Indeed, the previous theoretical results on size control (Theorem 2 and Corollary 1) remain valid regardless of whether the selection rule satisfies Assumption 5. One may be able to rationalise assumption Assumption 5 through PAC (Probably Approximately Correct) learnability theory - if the class of algorithms $\mathcal{A}$ is PAC learnable, then a selection rule that minimises empirical risk over the training data should converge to identifying algorithms that approach optimal welfare with high probability. This connection between PAC learnability and improvement consistency could be a promising direction for future research.

5 Monte-Carlo Simulation

For my first simulation, I abstract away from the details of the procedure and directly simulate values for the average utility. Let μ_0 be the true average welfare under algorithm a_0. This is a fixed value. Let λ be the true difference between the average welfare between a_1 and a_0. Varying this parameter will vary the degree of difference between the algorithms. The true welfare under algorithm a_1 is then $\mu_1 = \mu_0 + \lambda$. As our DDML estimates are noisy, we will control the degree of error with the following parameters.

- Let σ_0 be the standard deviation of the DDML estimate for welfare under a_0.
- Let σ_1 be the standard deviation of the DDML estimate of welfare under a_1.
- Let ϱ be the correlation between the DDML estimation errors for a_0 and a_1.

I generate simulated values in the following manner. For each simulation run i:

1. Generate a random error vector $\varepsilon_i = [\varepsilon_{0i}, \varepsilon_{1i}]'$ from a bivariate normal distribution with mean zero and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_0^2 & \varrho \\ \varrho & \sigma_1^2 \end{bmatrix}$$

2. Generate estimated welfare

$$\hat{\Gamma}_{0i} = \mu_0 + \varepsilon_{0i}$$
$$\hat{\Gamma}_{1i} = \mu_0 + \lambda + \varepsilon_{1i}$$

3. Apply the testing procedure to the pair $\left(\hat{\Gamma}_{0i}, \hat{\Gamma}_{1i}\right)$ for the hypothesis $\mathbb{E}[\Gamma_1] \leq \mathbb{E}[\Gamma_0]$.

5.1 Results

I parameterise the simulation as follows: I set $\mu_0 = 1.25$, $\sigma_0 = \sqrt{10}$, $\sigma_1 = \sqrt{11}$. These values were chosen to be similar to the moments in the empirical example. I perform 1000 MC simulation for each parameter combination using $10,000$ bootstrapped samples.

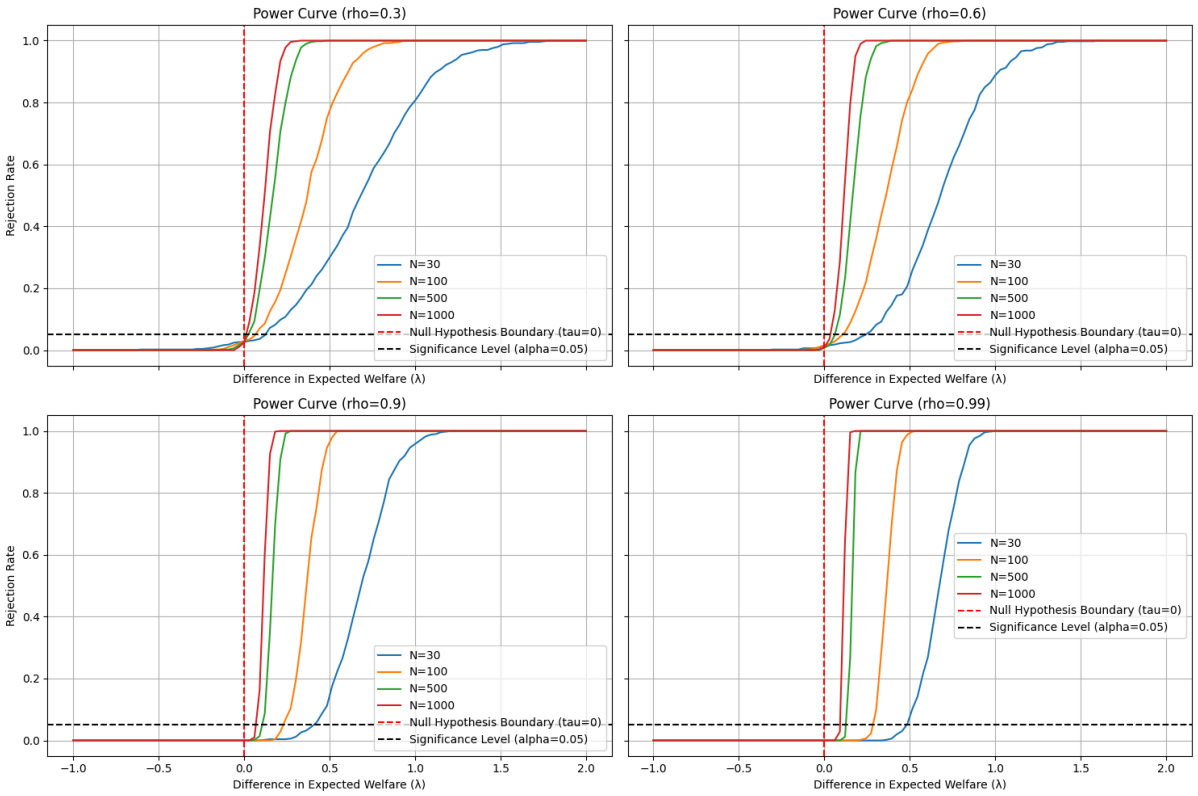

Figure 2: Monte-Carlo Simulation

The power curves in the diagram above illustrate the probability of rejecting the null hypothesis, plotted against the expected difference in welfare between the status quo algorithm and a candidate algorithm λ, under different sample sizes N and correlation in errors ϱ. The plots indicate that for data sets of reasonable size the test is suitably powered. Unsurprisingly, for all values of ϱ, larger sample sizes consistently yield higher power. In contrast, smaller sample sizes result in lower power, making it harder to detect an improvement when one exists. We can see that the test is sufficiently powered for medium and large data sets.

When the correlation (ϱ) is higher, the errors between the two algorithms' welfare estimates are more similar, leading to shared variance that complicates detecting meaningful differences. This shared error structure effectively reduces the signal-to-noise ratio, making it challenging to discern whether observed differences are due to actual improvements or just noise. High correlation diminishes the distinctiveness of the candidate algorithm's performance, making it harder to demonstrate clear welfare improvements.

These simulations are a very challenging case due to the size of the error variance and the degree of correlation. In practice, we might expect better improvement, which will be reduced by using separate bootstrap samples for each algorithm. This will be explored further in a more extensive Monte Carlo study.

6 Empirical Example

In my first empirical example, I use data from a randomised control trial conducted by Finkelstein et al. (2020) on the Camden Coalition of Healthcare Providers' "hot spotting" program. The program targeted "super-utiliser" - patients with exceptionally high healthcare usage - and aimed to reduce hospital readmission rates by employing a care-transition model involving multidisciplinary teams of nurses, social workers, and community health workers. The intervention group consisted of 800 hospitalised patients with medially and social complex conditions, all of whom had experienced at least one additional hospitalisation in the preceding six months.

The main variables of interest in the study include hospital readmission within 180 days and the costs associated with the readmission and other health-related expenditures. The data collected included patient demographics, hospitalisation history, and post-discharge interactions with care teams. I report a complete list of variables used in the appendix. Despite the widespread optimism surrounding hot-spotting, the RCT results indicated a significant difference in the average treatment effect of readmission rates between the intervention and control groups, challenging prior observational claims of the program's efficacy.

I use this data to evaluate the impact of algorithmic decision-making in healthcare settings. Specifically, I investigate heterogeneity in treatment effects and compare two automated algorithms in their ability to identify high-risk patients who would benefit most from the target interventions.

In the original paper, they get an estimate of ATE, which is reported in Table 1.

Variable	Coefficient	Std. Error	z-value	p-value	95% Confidence Interval
Treatment	0.8215	3.465	0.237	0.813	[−5.969, 7.612]

Table 1: OLS Regression Results for Average Treatment Effect

Using the DDML procedure, I estimate an ATE effect of 0.44 but find substantial heterogeniety in treatment effects, plotted in Figure 3.

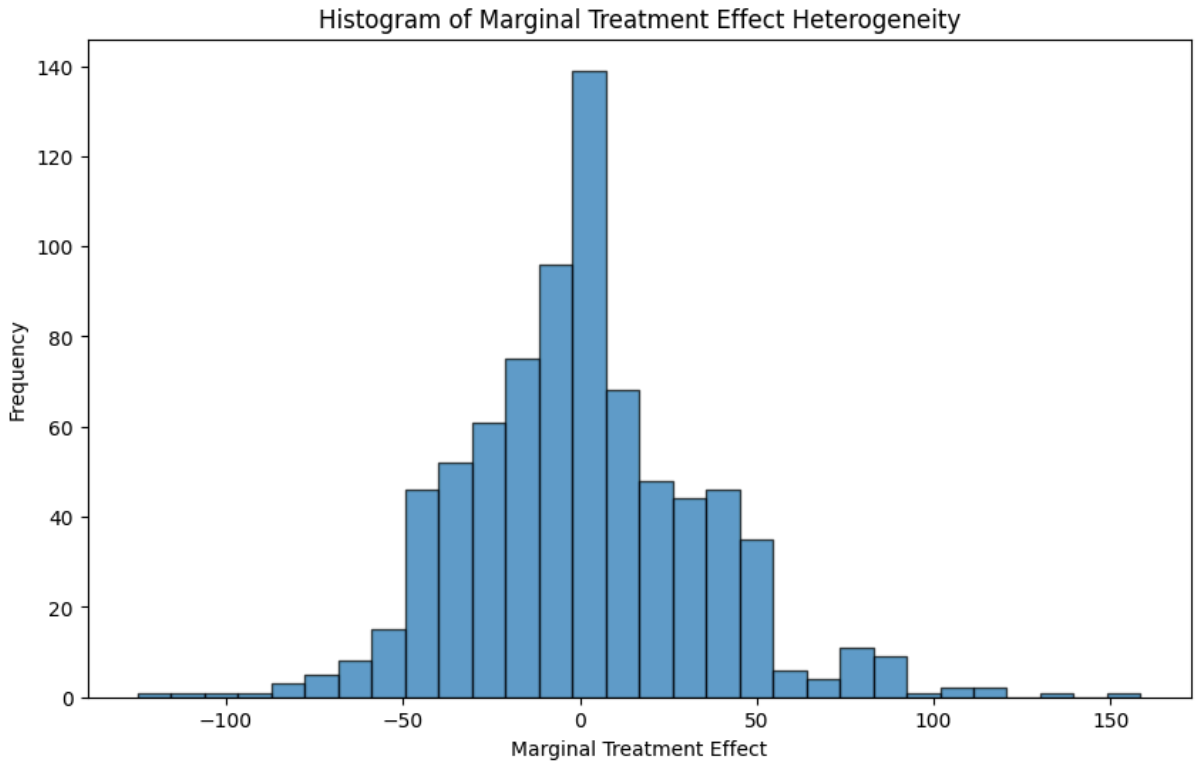

Figure 3: Treatment Heterogeneity

G Appendix 1: Proofs

Proof of Theorem 1: We define the welfare of the algorithm a as

$$\begin{aligned} V(a) &= \mathbb{E}[a(X_i)\tau(X_i)] \\ &= \mathbb{E}[a(X_i)\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i]] \\ &= \mathbb{E}[a(X_i)\psi_i] \end{aligned}$$

where the last equality is due to the law of iterated expectations. Recall the influence score representation:

$$V_n(a) = \frac{1}{n}\sum_{i=1}^{n}(2a(X_i) - 1)\psi_i$$

where $\psi_i = \tau_{m_n}(X_i, W_i) + g_n(X_i, Z_i)(Y_i - m_n(X_i, W_i))$

By Lemma 15 from Chernozhukov et al. (2022) and our Assumptions 1-3:

$$\sqrt{n}\big(\hat{V}(a) - V(a)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(2a(X_i) - 1)\psi_i^0 + o_p(1)$$

where ψ_i^0 is the influence score evaluated at the true functions (m_0, g_0)

Similarly:

$$\sqrt{n}(V_n(a) - V(a)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(2a(X_i) - 1)\psi_i^0 + o_p(1)$$

Therefore:

$$\sqrt{n}\big|\hat{V}(a) - V_n(a)\big| \leq \big|\sqrt{n}\big(\hat{V}(a) - V(a)\big) - \sqrt{n}(V_n(a) - V(a))\big| \underset{p}{\to} 0$$

□

References

Abadie, Alberto. 2003. “Semiparametric Instrumental Variable Estimation of Treatment Response Models”. *Journal of Econometrics* 113 (2): 231–63. https://doi.org/10.1016/S0304-4076(02)00201-4

Albright, Alex. 2024. “The Hidden Effects of Algorithmic Recommendations”

Athey, Susan, and Stefan Wager. 2020. “Policy Learning with Observational Data”. arXiv

Auerbach, Eric, Annie Liang, Max Tabord-Meehan, and Kyohei Okumura. 2024. “TESTING THE FAIRNESS-IMPROVABILITY OF ALGORITHMS”

Balashankar, Ananth, Alyssa Lees, Chris Welty, and Lakshminarayanan Subramanian. 2019. “What Is Fair? Exploring Pareto-Efficiency for Fairness Constrained Classifiers”. arXiv

Bergman, Peter, Julio E. Rodriguez, and Elizabeth Kopko. 2023. “A Seven-College Experiment Using Algorithms to Track Students: Impacts and Implications for Equity and Fairness | NBER”

Black, Dan A., Jeffrey A. Smith, Mark C. Berger, and Brett J. Noel. 2003. “Is the Threat of Reemployment Services More Effective Than the Services Themselves? Evidence from Random Assignment in the UI System”. *American Economic Review* 93 (4): 1313–27. https://doi.org/10.1257/000282803769206313

Bushway, Shawn, and Jeffrey Smith. 2007. “Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt Us”. *Journal of Quantitative Criminology* 23 (4): 377–87. https://doi.org/10.1007/s10940-007-9035-1

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. 2022. “Locally Robust Semiparametric Estimation”. *Econometrica* 90 (4): 1501–35. https://doi.org/10.3982/ECTA16294

Chetverikov, Denis, Mert Demirer, Esther Duflo, Christian Hansen, Whitney K. Newey, and Victor Chernozhukov. 2016. “Double Machine Learning for Treatment and Causal Parameters”. https://doi.org/10.1920/wp.cem.2016.4916

Cowgill, Bo, and Megan T. Stevenson. 2020. “Algorithmic Social Engineering”. *AEA Papers and Proceedings* 110:96–100

Dehejia, Rajeev H. 2005. “Program Evaluation as a Decision Problem”. *Journal of Econometrics* 125 (1–2): 141–73

DiCiccio, Cyrus J., Thomas J. DiCiccio, and Joseph P. Romano. 2020. “Exact Tests via Multiple Data Splitting”. *Statistics & Probability Letters* 166 (November):108865. https://doi.org/10.1016/j.spl.2020.108865

Farrell, Max H. 2015. “Robust Inference on Average Treatment Effects with Possibly More Covariates Than Observations”. *Journal of Econometrics* 189 (1): 1–23. https://doi.org/10.1016/j.jeconom.2015.06.017

Feldman, Michael, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. “Certifying and Removing Disparate Impact”. arXiv

Finkelstein, Amy, Annetta Zhou, Sarah Taubman, and Joseph Doyle. 2020. “Health Care Hotspotting — A Randomized, Controlled Trial”. *New England Journal of Medicine* 382 (2): 152–62. https://doi.org/10.1056/NEJMsa1906848

Frölich, Markus. 2008. “Statistical Treatment Choice: An Application to Active Labor Market Programs”. *Journal of the American Statistical Association* 103 (482): 547–58. https://doi.org/10.1198/016214507000000572

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. “Equality of Opportunity in Supervised Learning”. arXiv

Hirano, Keisuke, and Jack R. Porter. 2009. “Asymptotics for Statistical Treatment Rules”. *Econometrica* 77 (5): 1683–1701

Hu, Lily, and Yiling Chen. 2018. “Welfare and Distributional Impacts of Fair Classification”. arXiv

Johnson, Matthew, David Ian Levine, and Michael W. Toffel. 2019. “Improving Regulatory Effectiveness Through Better Targeting: Evidence from OSHA”. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3443052

Kitagawa, Toru, and Aleksey Tetenov. 2018. “Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice”. *Econometrica* 86 (2): 591–616. https://doi.org/10.3982/ECTA13288

Kleinberg, Jon, and Sendhil Mullainathan. 2019. “Simplicity Creates Inequity: Implications for Fairness, Stereotypes, And Interpretability”

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. “Inherent Trade-Offs in the Fair Determination of Risk Scores”. *Arxiv:1609.05807 [Cs, Stat]*, November

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. “Delayed Impact of Fair Machine Learning”. In *Proceedings of the 35th International Conference on Machine Learning*, 3150–58. PMLR

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2024. “The Unreasonable Effectiveness of Algorithms”

Mammen, E. 1992. *When Does Bootstrap Work? Asymptotic Results and Simulations*. Lecture Notes in Statistics. New York: Springer-Verlag

Manski, Charles F. 2004. “Statistical Treatment Rules for Heterogeneous Populations”. *Econometrica* 72 (4): 1221–46

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. “Algorithmic Fairness: Choices, Assumptions, And Definitions”. *Annual Review of Statistics and Its Application* 8 (1): 141–63. https://doi.org/10.1146/annurev-statistics-042720-125902

Mullainathan, Sendhil, and Ziad Obermeyer. 2019. “A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-Predictions”

Nahum-Shani, Inbal, Min Qian, Daniel Almirall, William E. Pelham, Beth Gnagy, Gregory A. Fabiano, James G. Waxmonsky, Jihnhee Yu, and Susan A. Murphy. 2012. “Q-Learning: A Data Analysis Method for Constructing Adaptive Interventions.”. *Psychological Methods* 17 (4): 478–94. https://doi.org/10.1037/a0029373

Pan, Yinghao, and Ying-Qi Zhao. 2021. “Improved Doubly Robust Estimation in Learning Optimal Individualized Treatment Rules”. *Journal of the American Statistical Association* 116 (533): 283–94. https://doi.org/10.1080/01621459.2020.1725522

Rambachan, Ashesh, Jon Kleinberg, Sendhil Mullainathan, and Jens Ludwig. 2020. “An Economic Approach to Regulating Algorithms”

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed". *Journal of the American Statistical Association* 89 (427): 846–66. https://doi.org/10.2307/2290910

Shi, Chengchun, Ailin Fan, Rui Song, and Wenbin Lu. 2018. "High-Dimensional A-Learning for Optimal Dynamic Treatment Regimes". *The Annals of Statistics* 46 (3). https://doi.org/10.1214/17-AOS1570

Viviano, Davide, and Jelena Bradic. 2023. "Fair Policy Targeting". *Journal of the American Statistical Association*, January, 1–14. https://doi.org/10.1080/01621459.2022.2142591

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. "Learning Fair Representations". In *Proceedings of the 30th International Conference on Machine Learning*, 325–33. PMLR

Zhang, Baqun, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. 2012. "A Robust Method for Estimating Optimal Treatment Regimes". *Biometrics* 68 (4): 1010–18. https://doi.org/10.1111/j.1541-0420.2012.01763.x

Zhao, Ying-Qi, Eric B Laber, Yang Ning, and Sumona Saha. 2019. "Efficient Augmentation and Relaxation Learning for Individualized Treatment Rules Using Observational Data"

Zhao, Yingqi, Donglin Zeng, A. John Rush, and Michael R. Kosorok. 2012. "Estimating Individualized Treatment Rules Using Outcome Weighted Learning". *Journal of the American Statistical Association* 107 (499): 1106–18