2024

Algorithmic Bias, Statistical Notions of Fairness, and the Seldonian Framework

Dasha Asienga

Advised by Professor Katharine Correia

Table of Contents

{ }

1

Algorithmic Bias

The problem

{ }

Data-Driven Algorithms Can Be Unfair

SEPTEMBER 7, 2023 | 7 MIN READ

Algorithms Are Making Important Decisions. What Could Possibly Go Wrong?

Seemingly trivial differences in training data can skew the judgments of AI programs—and that's not the only problem with automated decision-making

BY ANANYA

Are Decision-Making Algorithms Always Right, Fair and Reliable or NOT?

Algorithmic decision-making (ADM) is swiftly changing our societies. But does it hold up its promise of objectivity, or in the end do more harm than good?

Should Algorithms Make Layoff Decisions?

Research shows more HR leaders are using AI to recommend workforce reductions.

May 30, 2023

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

AI can be sexist and racist — it's time to make it fair

Computer scientists must identify sources of bias, de-bias training data and develop artificial-intelligence algorithms that are robust to skews in the data, argue James Zou and Londa Schiebinger.

MACHINE BIAS

When Big Data Becomes Bad Data

Corporations are increasingly relying on algorithms to make business decisions and that raises new legal questions.

SCIENCE & TECH

Artificial Intelligence Is as Unfair as We Are.

The Problem With Biased AIs (and How To Make AI Better)

Algorithmic bias arises when an algorithm's decisions are skewed towards a particular group of people, either positively or negatively.

COMPAS Recidivism Risk Assessment

Two Drug Possession Arrests

BLACK	WHITE
Prior Offense: 1 resisting arrest without violence	Prior Offense: 1 attempted burglary
Subsequent Offenses: None	Subsequent Offenses: 3 drug possessions
HIGH RISK (10)	**LOW RISK (3)**

Credit: Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. ProPublica.

%

Recidivism Status	Predicted Risk	Black	White
Did Not Reoffend	Higher	37.7	16.3
Reoffended	Lower	39.0	62.3

A **race-blind model** results in such racially disparate outcomes.

2

Statistical Notions of Fairness

Defining fairness mathematically

Notation

X	**Predictors**	the set of all predictor variables
Y	**Response**	response variable: real-valued/ continuous (for regression) or discrete (for classification)
$\hat{Y}$	**Prediction**	predicted response variable
A	**Protected Attribute**	The demographic group along which fairness is required, ex. race, sex, religion

Group v Individual Notions of Fairness

continuum

Fix a few demographic groups and assess the **parity of some statistical measures across all the groups**.

*Does not guarantee fairness to individuals or structured subgroups.
*Focuses on "**average numbers**".

Ex: independence, sufficiency, separation.

There are **conflicts** in simultaneous enforcement

Similar individuals should be treated similarly along some defined similarity or inverse distance metrics.

*Can be **impractical**, relies on **strong assumptions** about the data, and approaches the realm of **causal inference**.

Images: https://www.linkedin.com/pulse/credit-scoring-puzzle-fairness-metrics-aidan-connolly/

Group v Individual Notions of Fairness

Group Fairness

continuum

Individual fairness

Fix a few demographic groups and assess the **parity of some statistical measures across all the groups**.

*Does not guarantee fairness to individuals or structured subgroups.
*Focuses on "**average numbers**".

Ex: independence, sufficiency, separation.

There are **conflicts** in simultaneous enforcement

Similar individuals should be treated similarly along some defined similarity or inverse distance metrics.

*Can be **impractical**, relies on **strong assumptions** about the data, and approaches the realm of **causal inference**.

Images: https://www.linkedin.com/pulse/credit-scoring-puzzle-fairness-metrics-aidan-connolly/

Group v Individual Notions of Fairness

Group Fairness

continuum

Individual fairness

Fix a few demographic groups and assess the **parity of some statistical measures across all the groups**.

*Does not guarantee fairness to individuals or structured subgroups.
*Focuses on “**average numbers**”.

Ex: independence, sufficiency, separation.

There are **conflicts** in simultaneous enforcement

Similar individuals should be treated similarly along some defined similarity or inverse distance metrics.

*Can be **impractical**, relies on **strong assumptions** about the data, and approaches the realm of **causal inference**.

Images: https://www.linkedin.com/pulse/credit-scoring-puzzle-fairness-metrics-aidan-connolly/

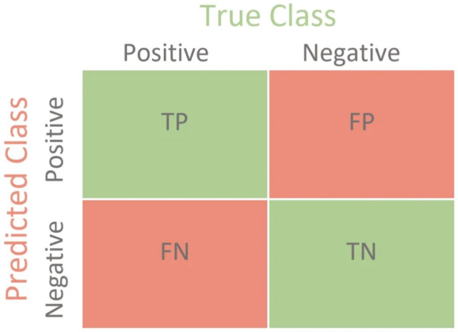
Separation

$$P(\hat{Y}=1|A=a,Y=y)=P(\hat{Y}=1|A=b,Y=y),\ \forall a,b\in A,\ \ y\in\{0,1\}$$

where *a, b* are two levels of the demographic group.

$$FPR=\frac{FP}{FP+TN}$$

$$FNR=\frac{FN}{TP+FN}$$



(Mohajon, 2021)

- Requires $\hat{Y} \perp\!\!\!\perp A|Y$.
- Also known as **equality of the odds** or **equality of the error rates**.
- The error rates should be the same across each level of the demographic group.

{ }

3

The Seldonian Framework

A theoretical overview

{ }

The Standard ML Approach

search for an optimal solution:

$$\theta^* \in \arg\max_{\theta \in \Theta} f(\theta).$$

optimal solution

objective/ cost function

$$\theta(X) = \beta_0 + \beta_1 X = \hat{Y}$$

$$f(\theta) := -E[(\theta(X) - Y)^2]$$

$$\cong$$

$$\hat{f(\theta)} = -\frac{1}{n}\sum_{i=1}^{n}(\theta(x_i) - y_i)^2$$

MSE (mean squared error)

{ }

Limitation of the Standard ML Approach

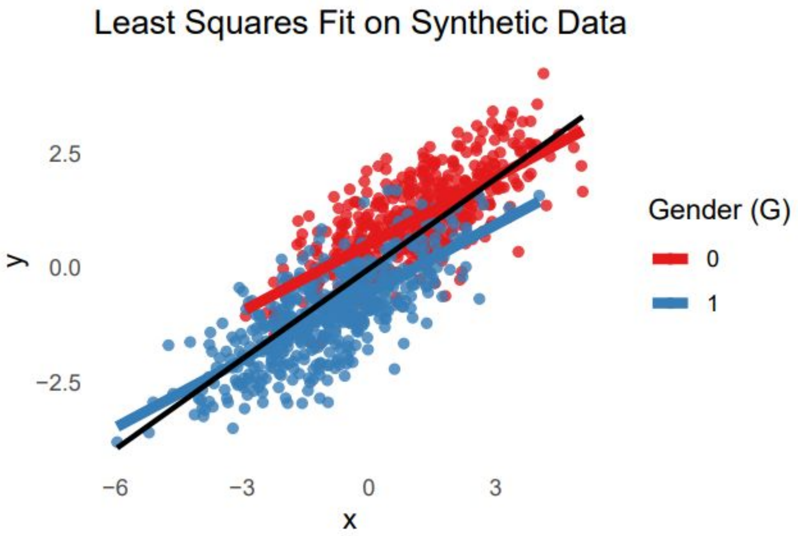
consider a linear regression example to **predict the qualifications of job applicants (Y)** based on the **job-relevant keywords on their resumes (X).**

Y ~ N(1, 1) if G = 0 (female)

Y ~ N(-1, 1) if G = 1 (male)

different distributions for different genders G

X ~ N(Y, 1) ← X strongly correlated with Y



{ }

The Seldonian framework is premised on the notion that if **‘unfair’** or **‘unsafe’** outcomes or behaviors can be **defined mathematically**, then it should be possible to create algorithms that can **learn from the data on how to avoid these unwanted results** with **high confidence**.

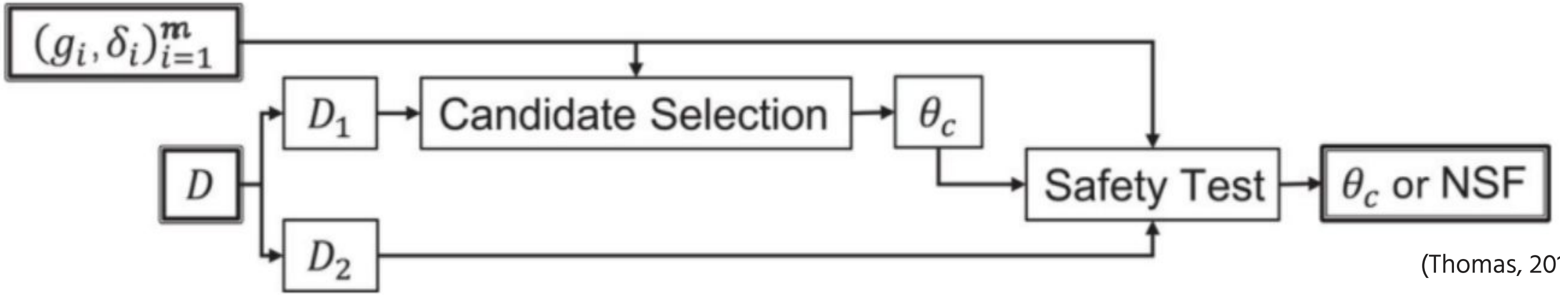
{ }

{ }

Notation

D	Data	The set of all possible inputs (data sets) *D*.
Θ	Solutions	The set of all possible outputs (solutions) Θ.
$\mathcal{A}$	Algorithms	The set of all possible machine learning algorithms a: $\mathbf{D}\rightarrow\mathbf{\Theta}$.
f(a)	Objective function	$f{:}\mathcal{A}\rightarrow\mathbb{R}$ is a real-valued measure of the solution θ returned by an algorithm *a*.

The Seldonian Framework



(Thomas, 2019).

- $g_i(\theta) > 0$ signifies **undesirable behavior**.
- $1 - \delta_i$ is the **minimum probability** that **desirable behavior** $(g_i(\theta) \leq 0)$ is met.

The Seldonian Optimization Problem (SOP)

putting it all together, ...

$$\arg\max_{a \in \mathcal{A}} f(a)$$

$$\text{s.t. } \forall\, i \in \{1, 2, \ldots, m\},\ P(g_i(a(D)) \leq 0) \geq 1 - \delta_i.$$

* the **fairness constraints** $g_i(\theta)$ and **desired confidence levels** δ_i need to be chosen appropriately,

BUT finding exact confidence bounds can be impractical and require large amounts of data. We need to rely on <u>standard statistical tools</u> to estimate them.

***quasi*-Seldonian algorithms** are Seldonian algorithms that make **assumptions** about the data (such as <u>normality</u>) and employ **statistical estimation techniques** to estimate the <u>confidence bounds</u> in the SOP.

{ }

{ }

The Student's *t*

Let $X = (X_1, \ldots, X_n)$ be *n* i.id. random variables.

We know, under the assumption that:

$\frac{1}{n}\sum_{i=1}^{n} X_i$ is *normally distributed*

or

the sample size *n is sufficiently large* (CLT), then:

$$P\Big(E[X] \leq \underbrace{\hat{\mu}(X) + \frac{\hat{\sigma}(X)}{\sqrt{n}} t_{1-\delta, n-1}}_{\text{upper bound of } E[X] \text{ with } 1-\delta \text{ confidence}}\Big) \geq 1 - \delta$$

upper bound of E[X] with 1 - δ confidence

The Safety Test Mechanism

... extending this idea, under the assumption that the $\hat{\mu}(\hat{g}_i(\theta_c, D_2))$ is *normally distributed* or that the size of the safety data set (D_2) is *sufficiently large* (CLT), for each $g_i(\theta)$,

$$P(g_i(\theta_c) \leq \hat{\mu}(\hat{g}_i(\theta_c, D_2)) + \frac{\hat{\sigma}(\hat{g}_i(\theta_c, D_2))}{\sqrt{|D_2|}} t_{1-\delta_i, |D_2|-1}) \geq 1 - \delta_i.$$

safety data

candidate solution

if **less than, or equal to, 0**, then the i^{th} behavioral constraint is **satisfied** with at least probability $1 - \delta_i$, and Θ_c will be returned!

- $\hat{g}_i(\theta_c, D_2) = (\hat{g_{i,1}}(\theta_c, D_2), \ldots, \hat{g_{i,n}}(\theta_c, D_2))$
- $E[\hat{g}_i(\theta_c, D_2)] = g_i(\theta_c)$ for each behavioral constraint $i \in \{1, 2, \ldots, m\}$

The Candidate Selection Mechanism

with the safety test in place, *any* algorithm will be Seldonian!

BUT, if θ_c is computed using the standard ML process, it will likely fail the safety test (**NSF!**).
Instead, θ_c will be computed as follows:

$$\theta_c \in \arg\max_{\theta \in \Theta} \hat{f}(\theta, D_1)$$

s.t. θ_c is predicted to pass the safety test.

candidate data

$$\theta_c \in \arg\max_{\theta \in \Theta} \hat{f}(\theta, D_1)$$

$$\text{s.t. } \forall\, i \in \{1, 2, \ldots, m\},\ \hat{\mu}(\hat{g}_i(\theta_c, D_1)) + \frac{\hat{\sigma}(\hat{g}_i(\theta_c, D_1))}{\sqrt{|D_2|}} t_{1-\delta_i, |D_2|-1} \leq 0.$$

ensure sol'n is properly predicted to pass the safety test

A *quasi*-Seldonian Linear Regression

consider the linear regression set-up to **predict Y ~ N(X,1) dependent on X** based on **X ~ N(0,1)**

Goals:

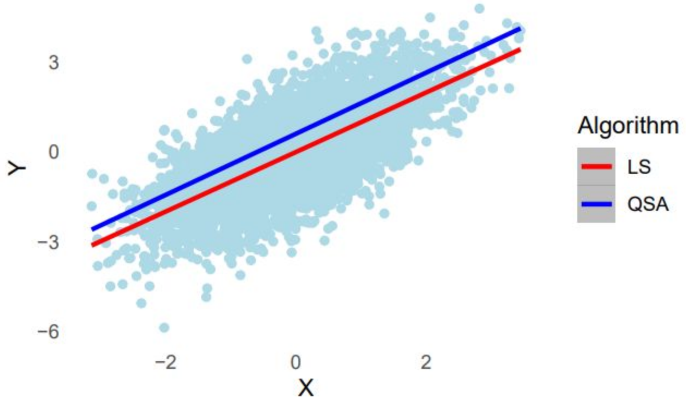
1. minimize MSE (maximize -MSE)

2. ensure, with probability at least 0.9, 1.25 < MSE < 2 *

in conflict

- $g_1(\theta) = MSE(\theta) - 2.0;\ \delta_1 = 0.1.$
- $g_2(\theta) = 1.25 - MSE(\theta);\ \delta_2 = 0.1.$

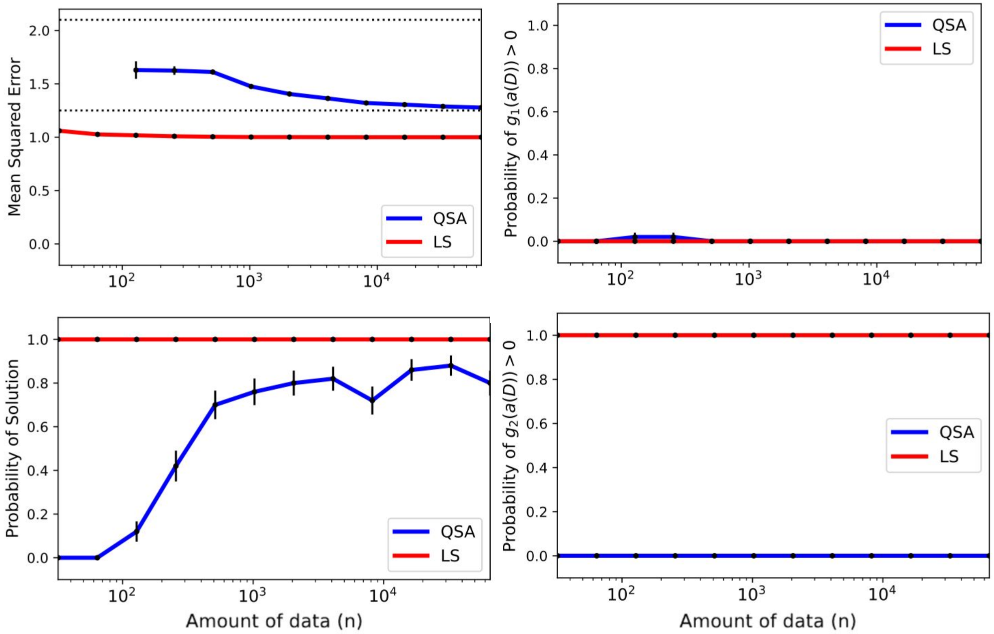
a solution was found!



* **impractical**, but allows us to test behavior when the goals are in conflict.

Experimentation with the QSLR

… scaling this process for different sample sizes *n* and 50 trials for each *n*,



- better performance with **more data**.
- probability of a Seldonian solution stabilizes at **~0.8**.
- there is a **performance tradeoff**.
- the solution returned almost always satisfied the constraint with **100% confidence**.

{ }

4

Application

Applying the Seldonian framework to the COMPAS data set

{ }

Formulating the Seldonian Classification Problem

1. Define the discrimination statistic [measure of (un)fairness]:

$$d(\theta) = abs[(FPR|\text{Black} - FPR|\text{White}) + (FNR|\text{White} - FNR|\text{Black})] \longleftarrow \begin{cases} d(\theta_{COMPAS}) = 0.45 \text{ (or } 44.7\%) \\ d(\theta_{LR}) = 0.34 \text{ (or } 34.18\%) \end{cases}$$

2. Define the behavioral/ fairness constraint as desired:

$$g(\theta) = abs[(FPR|\text{Black} - FPR|\text{White}) + (FNR|\text{White} - FNR|\text{Black})] - \epsilon.$$

define margin of difference

3. Formulate the Seldonian objective:

minimize logistic loss

such that

$$P\{abs[(FPR|\text{Black} - FPR|\text{White}) + (FNR|\text{White} - FNR|\text{Black})] - \epsilon \leq 0\} \geq 1 - \delta;\ \delta = 0.05$$

{ }

Evaluating Performance and Fairness

$(\epsilon = 0.2)$

Recidivism Status	Predicted Risk	Black	White
Did Not Reoffend	High	8.88	2.24
Reoffended	Low	73.82	88.82

accuracy: 68.2%, d(θ) = 0.22

$(\epsilon = 0.1)$

Recidivism Status	Predicted Risk	Black	White
Did Not Reoffend	High	1.87	0.3
Reoffended	Low	93.13	97.4

accuracy: 65.59%, d(θ) = 0.06

$(\epsilon = 0.05)$

Recidivism Status	Predicted Risk	Black	White
Did Not Reoffend	High	0.29	0.39
Reoffended	Low	98.73	98.17

accuracy: 64.7%, d(θ) = 0.007

$(\epsilon = 0.01)$

Recidivism Status	Predicted Risk	Black	White
Reoffended	Low	100	100

accuracy: 64.4%, d(θ) = 0

*** logistic regression: accuracy = 70.2%, d(θ) = 0.34**

{ }

5

Simulation Study

Evaluating efficacy and applicability in practical classification settings

The **aim** of the simulation study was to assess the **efficacy** and **applicability** of Seldonian algorithms in **practical classification settings** along three key performance measures:

- convergence
- fairer (less discriminatory) outcomes
- predictive accuracy

Data Generation Mechanism

for simplicity

to emulate complex, social relationships

retained <u>two</u> of the most informative <u>COMPAS variables</u> and searched through possible values to choose a linear combination with <u>improved predictive performance</u>:

$$logit(p_i) = 5 - 0.2\ Age_i + 0.5\ PriorOffense_i \mid i \in \{1, 2, \ldots, 9387\}$$

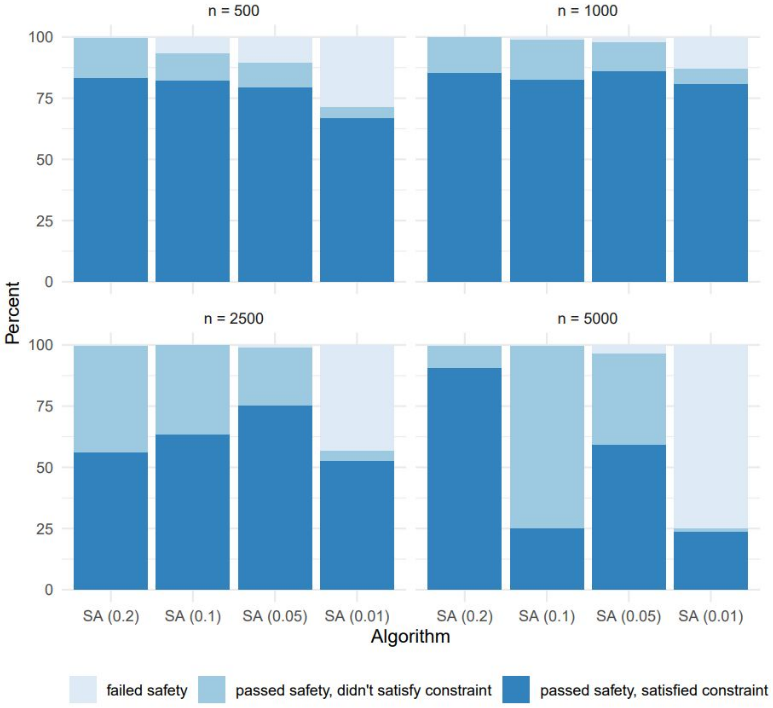
probability of defendant ***i*** recommitting a crime within 2 years

# of obs in the data

The response variable, *Y*, was drawn from a **Bernoulli distribution** and **class balance** was induced by randomly drawing, with replacement, the same number of observations in each class.

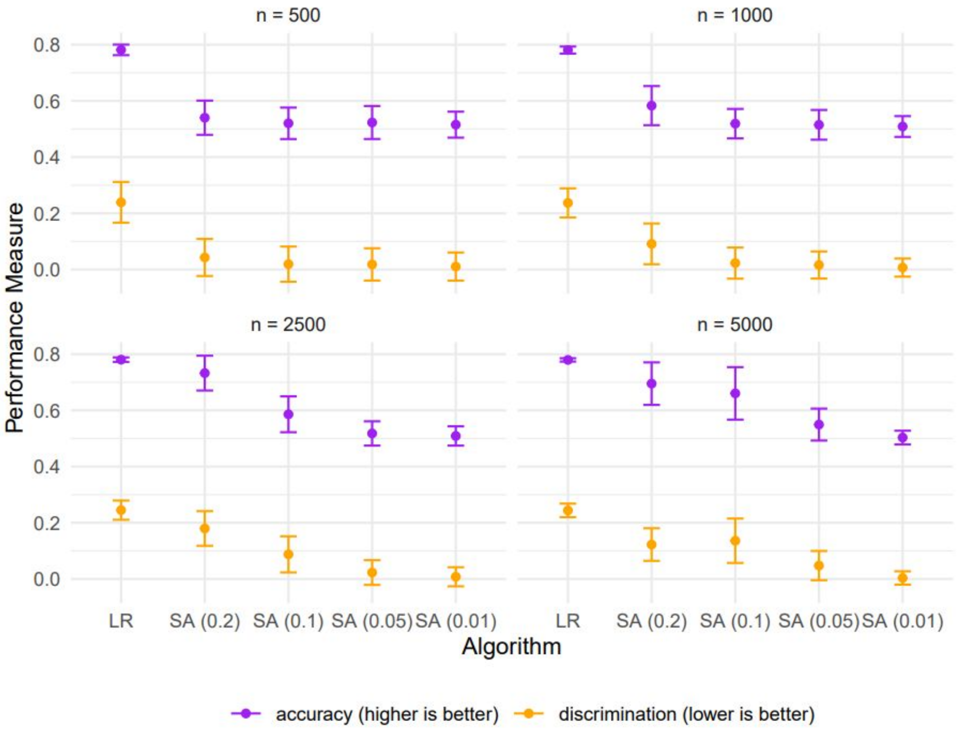
We generated **1000 total data sets** (250 each of size ***n = 500, 1000, 2500, 5000***).

Convergence



- **100% convergence** rate for the **logistic regression** models.
- For ϵ **= 0.2, 0.1, 0.05, the probability of a Seldonian solution** (passed safety test, **light blue**) was **>90%** across all sample sizes & **>96% for n >= 1000**.
- But for ϵ **= 0.01, the probability of passing the safety test drops** drastically.
- However, looking at how many of these solutions actually satisfied the constraint (**dark blue**), we notice that **larger sample sizes have a harder time constraining the discrimination statistic**, likely because of increased variability.
- For ϵ = 0.01, most returned solutions satisfy the fairness constraint. For ϵ **= 0.2, 0.1, 0.05, the defined δ = 0.05 was violated**.

Accuracy - Fairness TradeOff



- The value of the **discrimination statistic decreases** as the constraint is tightened across all sample sizes.
- The **accuracy decreases** as the constraint is tightened across all sample sizes, stabilizing at 50% accuracy (random, coin-flip model).
- The **drop in accuracy is more drastic for smaller sample sizes** (n = 500, 1000).
- Seldonian algorithms with **looser fairness constraints** (ϵ = 02, 0.1) and **larger sample sizes** (n = 2500, 5000) **perform comparably to logistic regression** while offering **some improvement in model fairness**.

{ }

6

Conclusions

Main takeaways and future directions

{ }

Key Takeaways

{ }

- The **time** it takes to understand and implement the Seldonian algorithm Python code poses a huge **barrier to implementation** and experimentation.
- There are **a lot of nuances to consider** when creating solutions to fair ML and AI, such as balancing tradeoffs with predictive performance, defining what fairness means, and incorporating them into current technologies.
- The Seldonian framework is **a step in the right direction** and allows us to have fairer outcomes (though not the fairest!), but it is far from the perfect solution.

{ }

Suggestions for Future Work

- Investigating ways to **balance fairness and predictive performance** within the Seldonian framework.
- Assessing performance of Seldonian algorithms in **practical continuous settings** as well as with **different group, subgroup, and individual notions** of fairness.
- Performing a **holistic comparison** of Seldonian outcomes with other **state-of-the-art fair ML tools**, such as Microsoft's Fairlearn and IBM's Fairness 360 AI.

Bibliographical References

- Agarwal, A., Dudík, M., & Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. In International conference on machine learning (pp. 120–129). PMLR.
- Andrews, E. L. (2021, August). How flawed data aggravates inequality in credit. Retrieved October 29, 2023, from https://hai.stanford.edu/news/how-flawed-dataaggravates-inequality-credit
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: Risk assessments in criminal sentencing. Retrieved November 1, 2023, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Asimov, I. (1994). Forward the foundation (Vol. 7). Spectra.
- Basu, J., Hanchate, A., & Bierman, A. (2018). Racial/ethnic disparities in readmissions in US hospitals: The role of insurance coverage. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 55, 0046958018774180.
- Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., . . . Mojsilović, A. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63 (4/5), 4–1.
- Bird, S., Dudik, M., Edgar, R., Horn, B., Lutz, R., & Milan, V. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft.
- Boyd, S. P., & Vandenberghe, L. (2004). Convex optimization. Cambridge university press.
- Broward County Clerk’s Office, Broward County Sheriff's Office, Florida Department of Corrections, & ProPublica. (2024). COMPAS Recidivism Risk Score Data [Data set]. ProPublica Data Store. Retrieved from https://www.propublica.org/ datastore/dataset/compas-recidivism-risk-score-data-and-analysis
- Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. Scientific Reports, 12 (1), 4209.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5 (2), 153–163.
- Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. Retrieved October 1, 2023, from https://arxiv.org/abs/1810.08810
- Durahly, L. (2023). A gentle introduction to ML fairness metrics. Retrieved October 14, 2023, from https://superwise.ai/blog/gentle-introduction-ml-fairness-metrics/
- Frederick, S. (2023, June). Supreme court rejects use of race in college admissions process. Retrieved April 13, 2024, from https://www.ncsl.org/state-legislaturesnews/details/supreme-court-rejects-use-of-race-in-college-admissions-process
- Kubiak, E., Efremova, M. I., Baron, S., & Frasca, K. J. (2023). Gender equity in hiring: Examining the effectiveness of a personality-based algorithm. Frontiers in Psychology, 14:1219865.
- Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. Retrieved December 2, 2023, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54 (6), 1–35.
- Mohajon, J. (2021). Confusion matrix for your multi-class machine learning model. Retrieved November 28, 2023, from https://towardsdatascience.com/confusionmatrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826
- Noble, S. U. (2018). Algorithms of oppression. New York University Press.
- O’Donnell, R. M. (2019). Challenging racist predictive policing algorithms under the equal protection clause. NYUL Rev., 94, 544.
- O’Neil, C. (2016). Weapons of math destruction. Crown Publishing Group.
- Saeed, S., Alireza, B., Mohamed, E., & Ahmed, N. (2015). Evidence based emergency medicine part 2: Positive and negative predictive values of diagnostic tests. Emergency (Tehran, Iran), 3 (3), 87–88.
- Silva, B. C. da. (2019). UFRGS Entrance Exam and GPA Data (Version V2) [Data set]. Harvard Dataverse. http://doi.org/10.7910/DVN/O35FW8
- The Sentencing Project. (2018). Report to the united nations on racial disparities in the US criminal justice system. Retrieved March 18, 2024, from https://www.sentencingproject.org/reports/report-to-the-united-nations-onracial-disparities-in-the-u-s-criminal-justice-system/
- Thomas, P. (2020). Testimony to the house committee on financial services task force on artificial intelligence hearing: “Equitable algorithms: Examining ways to reduce AI bias in financial services.” Retrieved September 30, 2023, from https://www.congress.gov/116/meeting/house/110499/witnesses/HHRG-116- BA00-Wstate-ThomasP-20200212.pdf
- Thomas, P. S. (2020). AI safety. Retrieved April 12, 2024, from /url%7Bhttps: //aisafety.cs.umass.edu/index.html%7D
- Thomas, P. S., Castro da Silva, B., Barto, A. G., Giguere, S., Brun, Y., & Brunskill, E. (2019a). Preventing undesirable behavior of intelligent machines. Science, 366 (6468), 999–1004.
- Thomas, P. S., Castro da Silva, B., Barto, A. G., Giguere, S., Brun, Y., & Brunskill, E. (2019b). Supplementary materials for preventing undesirable behavior of intelligent machines. Science, 366 (6468), 999–1004.

Acknowledgements

Special thanks to Professor Katharine Correia, as well as the team from FrostByte (Amherst College's High Performance Computing Cluster) and all Statistics faculty.

This work was performed in part using high-performance computing equipment at Amherst College obtained under National Science Foundation Grant Number 2117377. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Science Foundation.

Thanks!

Any questions?

dasienga24@amherst.edu

Appendix

Additional Slides

More Examples of Algorithmic Bias

Facial Recognition:

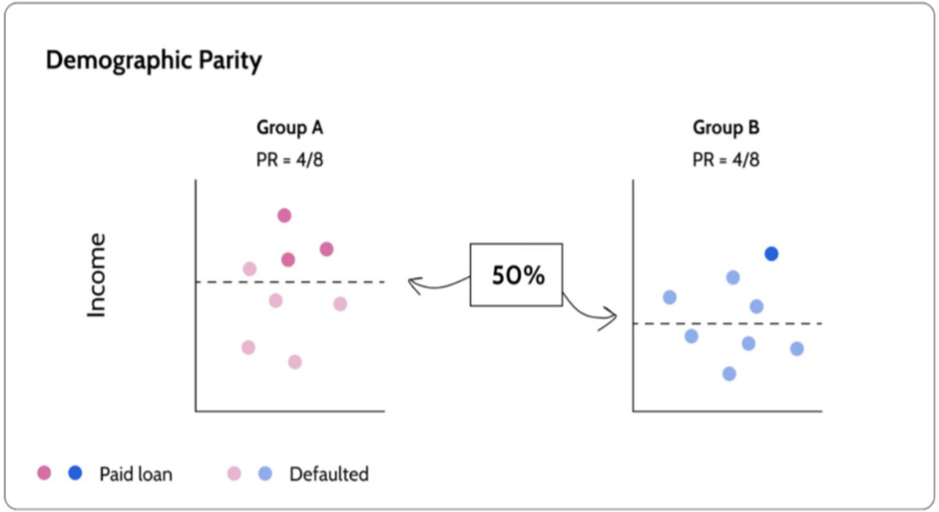
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

{ } { }

Independence

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = b), \ \forall a, b \in A,$$

where a, b are the two demographic groups in question.



- Requires $\hat{Y} \perp\!\!\!\perp A$
- Also known as demographic parity or statistical parity.
- The likelihood of a positive outcome should be the same across each demographic group.

Conditional Demographic Parity

$$P(\hat{Y} = 1 | A = a, R = r) = P(\hat{Y} = 1 | A = b, R = r), \ \forall a, b \in A, \forall r.$$

with R as the set of possible ratings.

$$\downarrow$$

Individual Fairness

$$\boxed{P(\hat{Y} = 1 | A = a, X = x) = P(\hat{Y} = 1 | A = b, X = x), \ \forall a, b \in A, \forall x \in X.}$$

{ } { }

Sufficiency

$$P(Y = 1 | A = a, \hat{Y} = 1) = P(Y = 1 | A = b, \hat{Y} = 1), \ \forall a, b \in A.$$

- Requires $Y \perp\!\!\!\perp A | \hat{Y}$.
- Also known as predictive parity.
- The precision of the model should be equal across all demographic groups.

{ } { }

Fairness Conflict – Classification

sufficiency: [equal positive predictive values (PPV)]

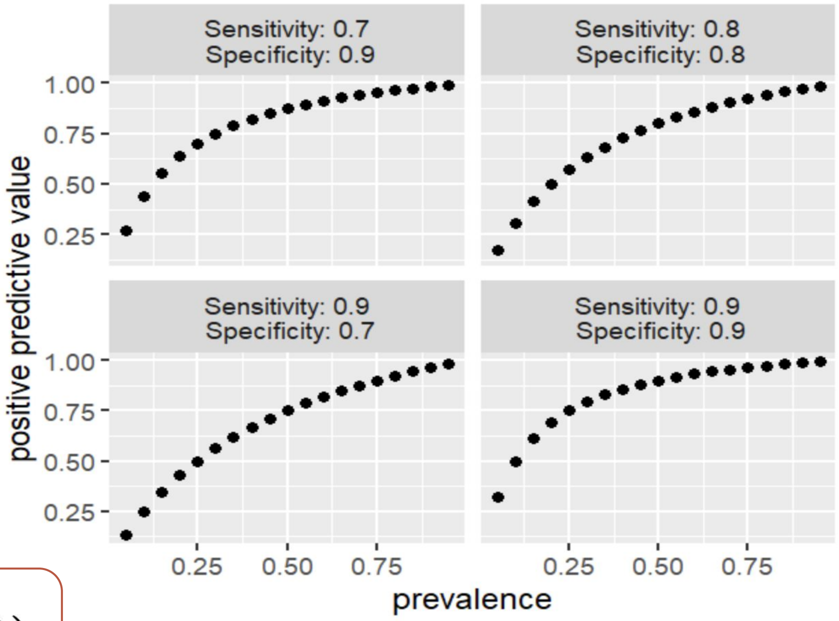
separation: [equal FPR and FNR]

<u>Define</u>:

- ➔ FNR = 1 - *sensitivity*
- ➔ FPR = 1 - *specificity*

Then, given values of PPV $\in (0, 1)$ and prevalence $p \in (0, 1)$, we can show that:

$$\text{FPR} = \frac{p}{1-p}\frac{1-\text{PPV}}{\text{PPV}}(1-\text{FNR}).$$



Fairness Conflict – Regression

Consider a case with **gender G** as the protected attribute.

These two cannot hold simultaneously
if the average distribution of Y is different for both groups.

A model that satisfies independence:

$$E[\hat{Y}|G = Male] = E[\hat{Y}|G = Female]$$

cannot simultaneously satisfy
equal error rates:

$$E[\hat{Y} - Y|G = Male] = E[\hat{Y} - Y|G = Female]$$

{ } { }

COMPAS Tool Performance (by race)

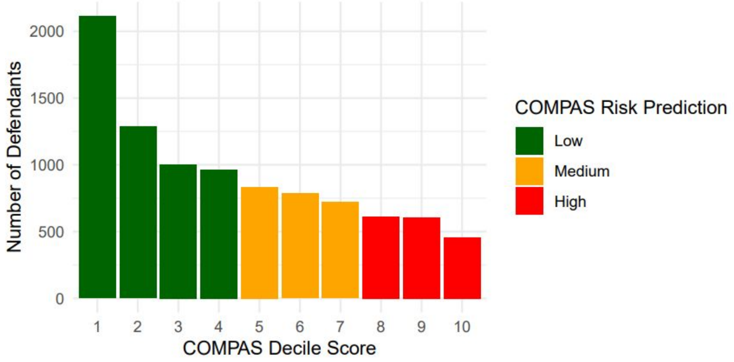


Figure 3.4: Distribution of COMPAS Tool Decile Scores among 9387 Defandants in Broward County Florida, 2013-2014

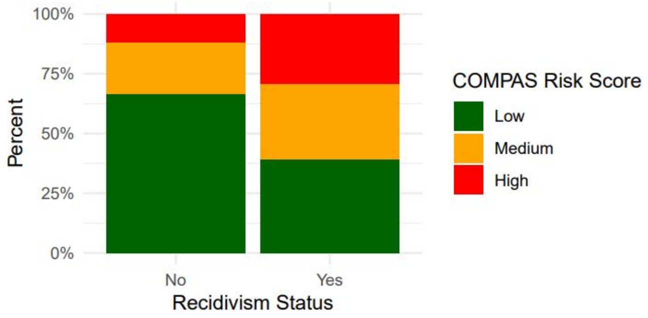


Figure 3.5: Distribution of COMPAS Tool Risk Scores among 9387 Defandants in Broward County Florida, 2013-2014

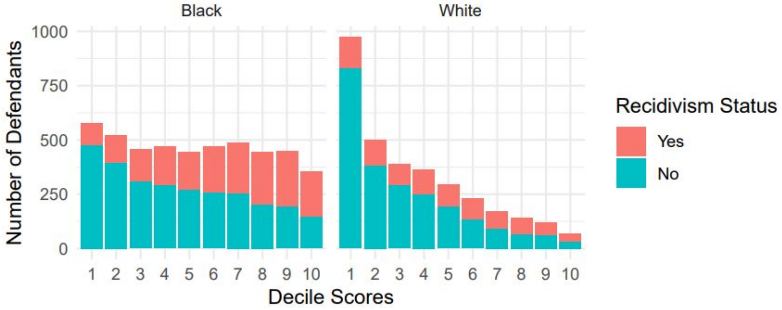
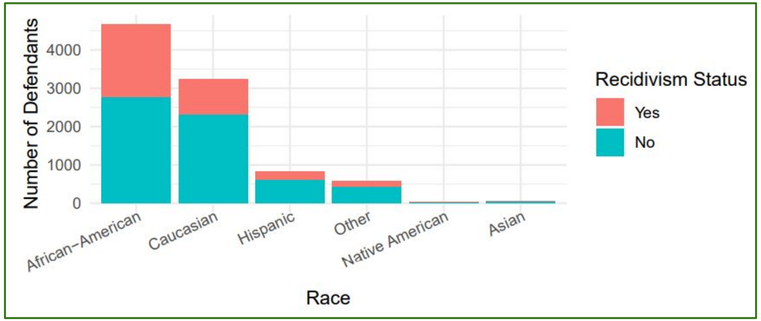


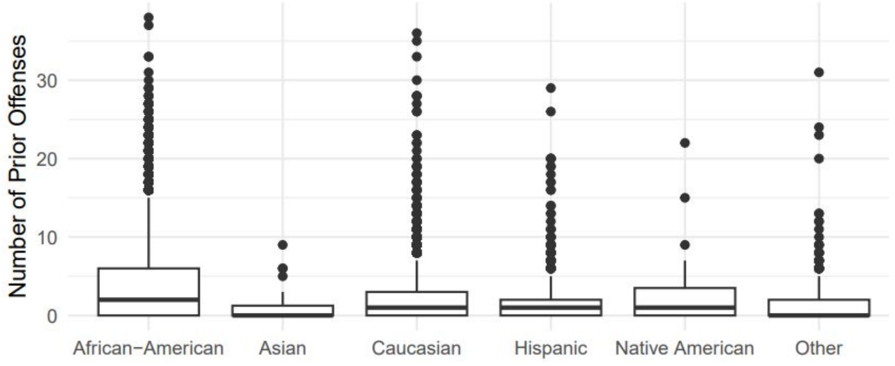
Figure 3.6: Distribution of COMPAS Tool Decile Scores Stratified by Race among 9387 Defandants in Broward County Florida, 2013-2014



The COMPAS Data Set

Recidivism Status	Predicted Risk	Black	White
Did Not Reoffend	High	13.9	5.09
Reoffended	Low	60.8	86.17

Fitting a **logistic regression model** on this data set results in the same racially disparate outcomes, although slightly less extreme.



Highlights the role that **proxy variables** (correlated with race & likely contain societal racial biases) play in incorporating information about race into race-blind models.

Proxy Variables

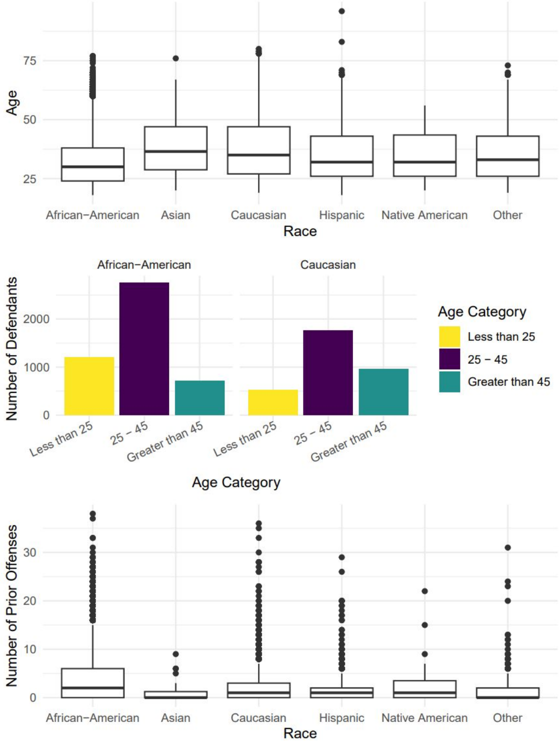


Figure 3.7: Distribution of COMPAS Tool Proxy Variables Stratified by Race among 9387 Defendants in Broward County Florida, 2013-2014

Probability of a Seldonian Solution

Table 4.1: Probability of Obtaining a Seldonian Solution

Sample Size	LR	SA (0.2)	SA (0.1)	SA (0.05)	SA (0.01)
500	100	99.6	93.2	89.6	71.2
1000	100	100.0	98.8	97.6	86.8
2500	100	99.6	100.0	98.8	56.8
5000	100	99.6	99.6	96.4	24.8

Table 4.2: Satisfaction of the Behavioral Constraint by Seldonian Solutions that Passed the Safety Test

Sample Size	SA (0.2)	SA (0.1)	SA (0.05)	SA (0.01)
500	83.5	88.0	88.4	93.8
1000	85.2	83.4	88.1	93.1
2500	56.2	63.2	76.1	92.2
5000	90.8	24.9	61.4	95.2

{ } { }

Discrimination

Table 4.3: Mean Discrimination Statistic of Convergent Seldonian Solutions

Sample Size	LR	sd	SA (0.2)	sd	SA (0.1)	sd	SA (0.05)	sd	SA (0.01)	sd
500	0.24	0.07	0.04	0.07	0.02	0.06	0.02	0.06	0.01	0.05
1000	0.24	0.05	0.09	0.07	0.02	0.06	0.02	0.05	0.01	0.03
2500	0.24	0.03	0.18	0.06	0.09	0.06	0.02	0.04	0.01	0.03
5000	0.24	0.02	0.12	0.06	0.14	0.08	0.05	0.05	0.00	0.02

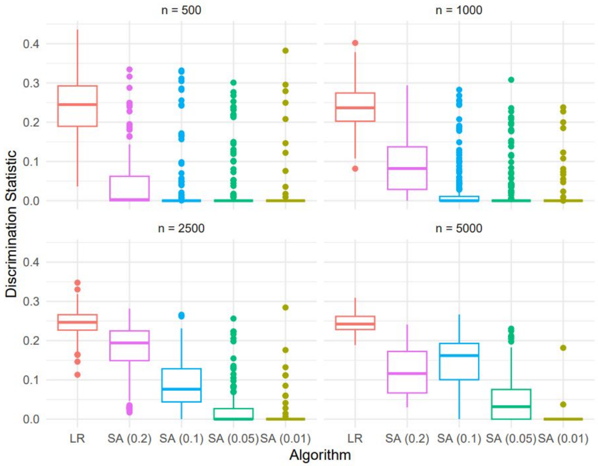


Figure 4.2: The Distribution of the Discrimination Statistic of Convergent Seldonian Solutions by Sample Size

Accuracy

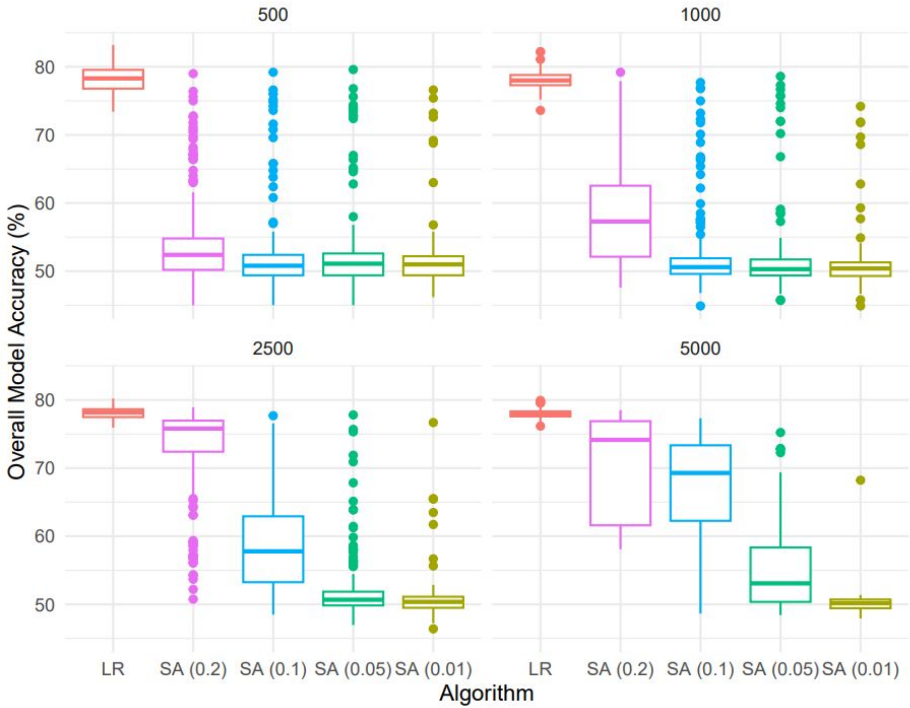


Figure 4.3: The Distribution of the Overall Accuracy of Convergent Seldonian Solutions by Sample Size

Non-Convergent Seldonian Models

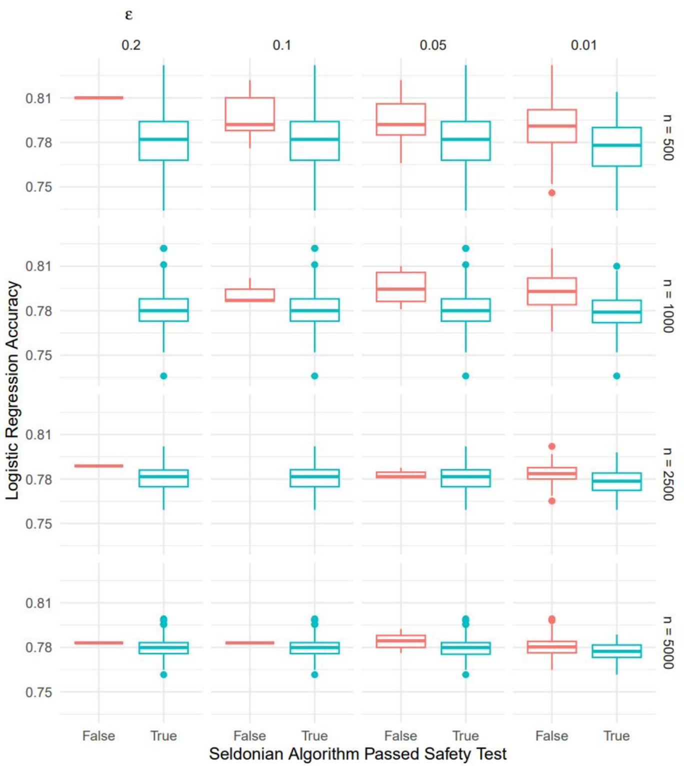


Figure 4.5: Evaluating Logistic Regression Accuracy on the Data Sets by Seldonian Convergence

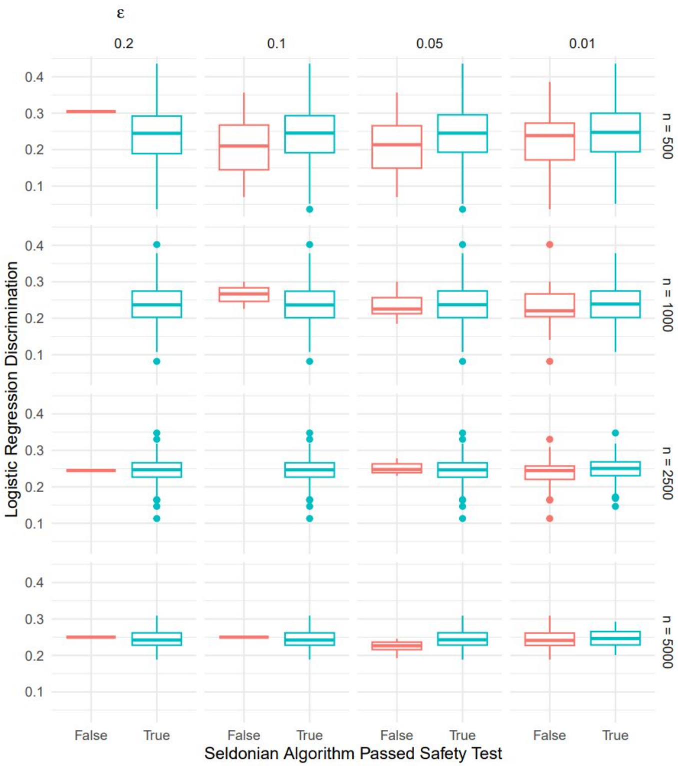
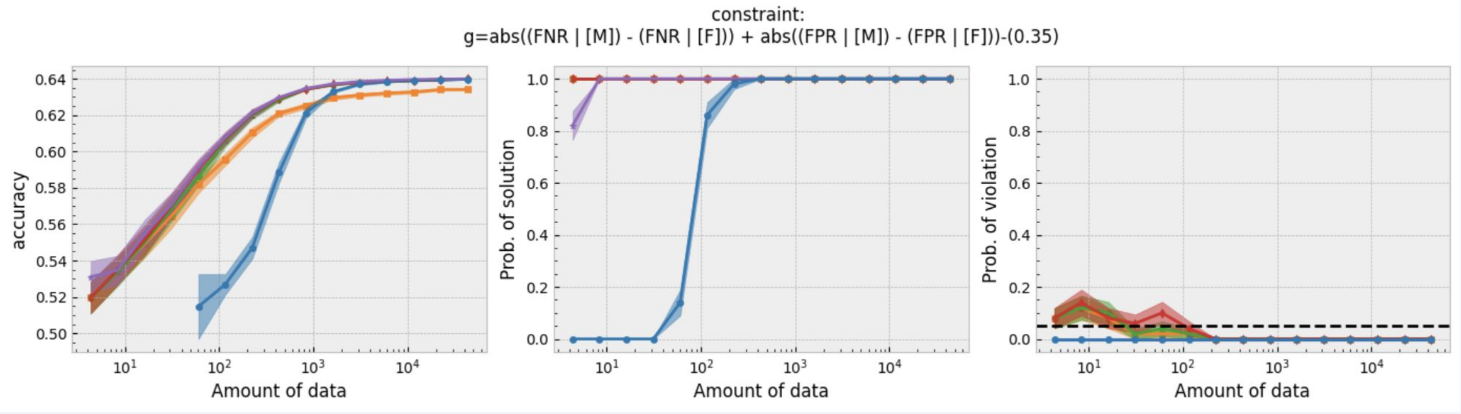
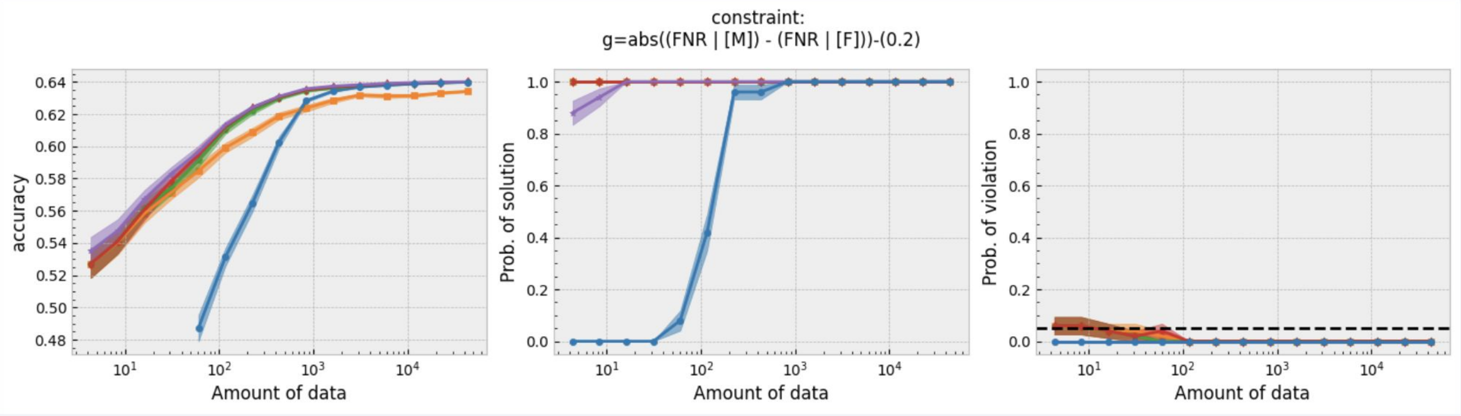


Figure 4.6: Evaluating Logistic Regression Discrimination on the Data Sets by Seldonian Convergence

Seldonian Classification Testing





Headlines on Slide 4

https://www.scientificamerican.com/article/algorithms-are-making-important-decisions-what-could-possibly-go-wrong/#:~:text=Despite%20their%20known%20shortcomings%2C%20algorithms,what%20task%2C%20among%20other%20significant

https://www.shrm.org/hr-today/news/hr-magazine/summer-2023/pages/should-algorithms-make-layoff-decisions-.aspx

https://www.liberties.eu/en/stories/decision-making-algorithm/44109

https://www.theregreview.org/2021/11/11/adams-algorithmic-decisions-human-consequences/

https://www.propublica.org/article/when-big-data-becomes-bad-data

https://www.nature.com/articles/d41586-018-05707-8

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

https://www.brownalumnimagazine.com/articles/2024-01-29/artificial-intelligence-biases-suresh-venkatasubramanian#:~:text=A%20new%20course%20asks%20how,our%20biases%20and%20automating%20oppression.&text=Can%20an%20algorithm%20be,program%20is%20being%20trained%20on.

https://www.forbes.com/sites/bernardmarr/2022/09/30/the-problem-with-biased-ais-and-how-to-make-ai-better/?sh=3f620ea94770

{ }

{ }