Algorithms as discrimination detectors

Abraham Guerrero, Carlos Orozco, Diego Sanchez, Braian Moreno.

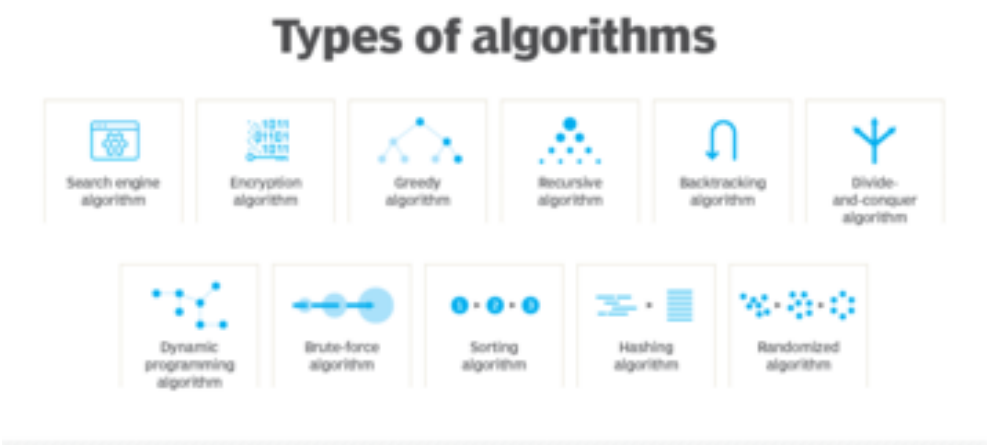
June 6, 2023

Outline

Outline

Motivation

- Today, algorithms can increase the risk of discrimination.
- But, algorithms require a far greater level of specificity than is usually possible with human decision making.
- This specificity makes it possible to prove aspects of the decision in additional ways.



Motivation

With the right changes to legal and regulatory systems, algorithms can thus potentially make it easier to detect (and hence to help prevent) discrimination.

Problems

Problems of Discrimination Detectors

1. First, the existing legal, regulatory, and related systems for detecting discrimination were originally built for a world of human decision makers, unaided by algorithms.
2. Algorithms by their nature require a far greater level of specificity than is usually involved with human decision making, which in some sense is the ultimate "black box."

Problems

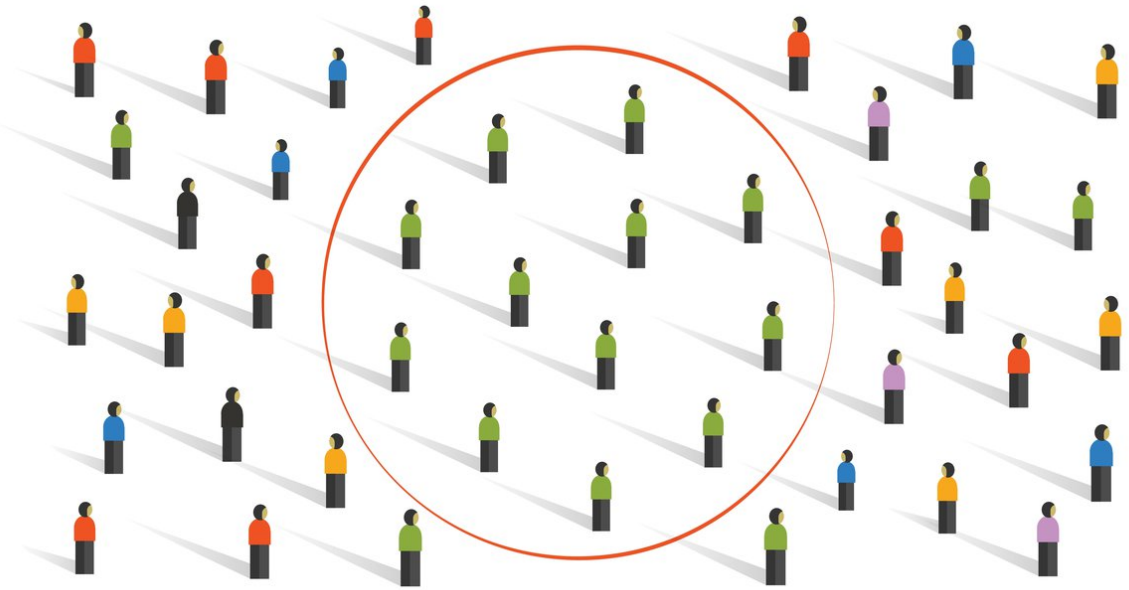
With the right legal and regulatory systems in place, algorithms can serve as something akin to a Geiger counter that makes it easier to detect—and hence prevent—discrimination [1].

[1] Geiger counter is an electronic instrument used for detecting and measuring ionizing radiation

Problems

However...

Aspirationally, such systems can help not only un detecting discrimination as it is now understood in law, but also in clarifying the normative questions raised by debates over that contested concept.



Some questions are:

- What can be done to test whether practices of that kind can be justified?

- What if a private or public institution uses a factor (such as credit ratings or criminal records) that may be bound up with past discrimination?

The use of algorithms cannot answer such questions, but it can help produce another level of clarity about the stakes and about potential tradeoffs.

Outline

Hypothesis

Hypothesis

Is difficult to detect bias in an entirely human-driven decision system.

For example, the problem of discrimination in a tech firm.

Hypothesis

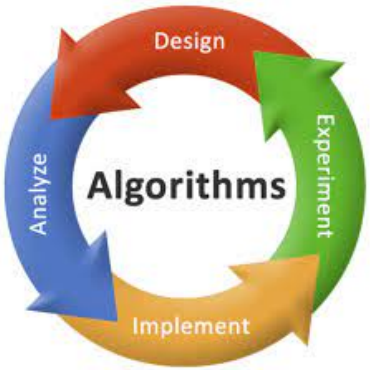
Challenges in using statistical evidence to show discrimination,combine to create a fog of ambiguity, which prevents us from stopping a behavior that we know to be widespread yet for which in any one instance there may well be plausible alternative explanations.

Outline

Their Scope

How algorithms might affect the current state of affairs?

The word "algorithms" encompasses a wide range of tools from optimization to search, which in turn influence a variety of decisions.



The screening decisions do not capture all of the potential uses of algorithms for decision making, but are among the most dominant uses of algorithms.

Their Scope

In this context, they focus on algorithms that build prediction functions from training data; the prediction function produced, in turn, takes "inputs" (like the characteristics of a college applicant) and predicts some outcome (like college grade point average).

But...

Regardless of the complexity of the function class that the learning procedure allows, all standard techniques share the property that the algorithm builder will need to specify an outcome to be predicted, what candidate predictors will be made available to the learning procedure, and a dataset.

Their Scope

Notwithstanding this point, they are keenly aware that the idea of "discrimination" is contested and ambiguous. Authors have focused thus far on intentional discrimination of this kind, which is the canonical form known as "disparate treatment," and it is their emphasis here.

Outline

Sources of Algorithmic Discrimination

Algorithms can be powerful tools that help people use data to accomplish their objectives more effectively. Sometimes the objective humans have in mind is to discriminate on some forbidden ground.

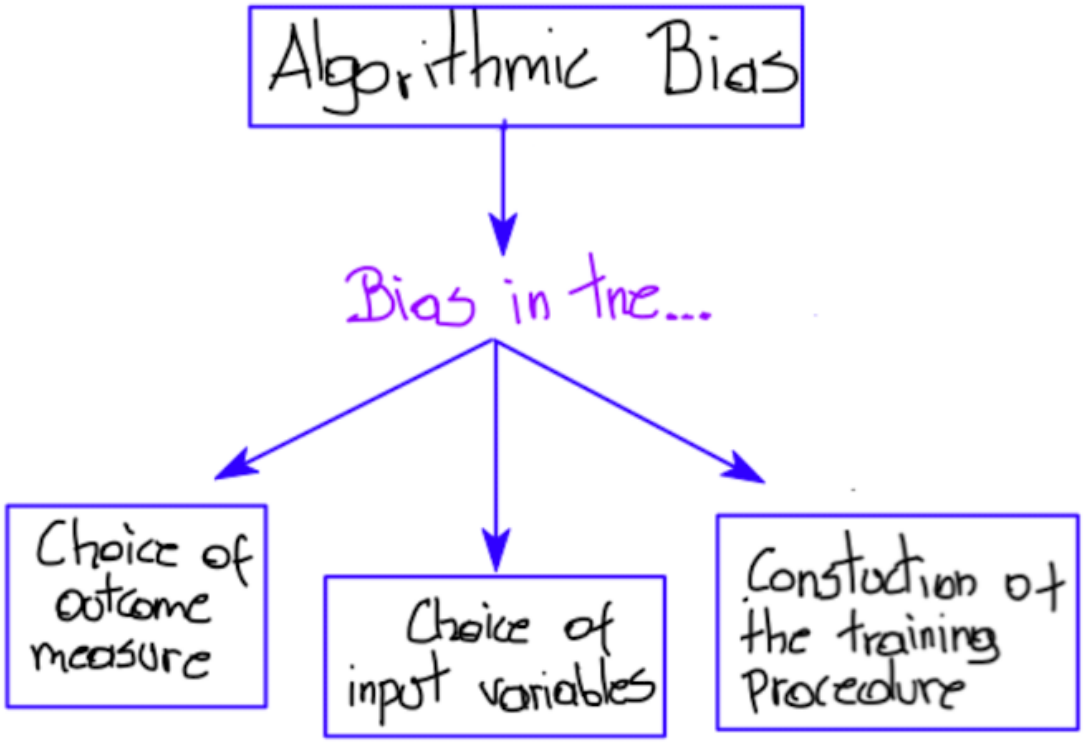
Sources of Algorithmic Discrimination

Type of Separate Algorithms at Work in Screening

1. The screening algorithm (or screener): Simply takes the characteristics of an individual and returns a prediction of this individual's outcome. This prediction then informs a decision. (Is just the mechanical result of running the training algorithm on a set of training data).
2. The training algorithm (or trainer): Is what produces the screening algorithm. Constructing the training algorithm involves assembling past instances to use as training data, defining the outcome to predict, and choosing candidate predictors to consider.

Sources of Algorithmic Discrimination

They showed formally that algorithmic bias can be decomposed completely into three components:



Sources of Algorithmic Discrimination

Any remaining disparity corresponds to the structural disadvantage of one group relative to another.

A Screening problem (Hiring) With Applicants From two Distinct Groups

- Each applicant is described by a feature vector x.
- The applicant's true productivity for the task hand is a function $f(x)$ of this vector.
- Fraction of applications with feature vector x in the advantage group $p(x)$.
- Fraction of applications with feature vector x in the disadvantage group $q(x)$.

A Screening problem (Hiring) With Applicants From two Distinct Groups

- Average productivity of applicants in the advantage group

$$\sum_x p(x) f(x)$$

- Average productivity of applicants in the disadvantage group

$$\sum_x q(x) f(x)$$

A Screening problem (Hiring) With Applicants From two Distinct Groups

For an arbitrary function v: Difference in the average value of v between the advantaged and the disadvantaged groups $D(v)$

$$\begin{aligned} D(v) &= \sum_x p(x)v(x) - \sum_x q(x)v(x) \\ &= \sum_x \left[p(x) - q(x)\right] v(x) \end{aligned}$$

Applying this notation to the function f,we see that $D(f)$ is the difference in average productivity between the two groups.

A Screening problem (Hiring) With Applicants From two Distinct Groups

Then:

This constitutes the structural disadvantage of one group relative to the other.

Process of building an algorithm

The designers of the algorithm know neither the true function f nor the applicant's full feature vector x.

1. The designer specifies applicant performance as $g(x)$, which is usually not exactly $f(x)$.
2. The designer constructs the algorithm based on a reduced feature vector $r(x)$.
3. since the function g is designed to be applied to vectors whose dimension is the same as x, a different function h must be used on the reduced vector $r(x) = h(r(x))$. the value for an applicant with feature vector x is $(h \circ r)(x)$.

Process of building an algorithm

4. The function h must be estimate from the available training data (t).
5. t is applied to the reduced feature vector $r(x)$, resulting in the value $(t \circ r)(x)$.

Result

the design process results in an algorithm that takes an applicant with feature vector x and produces a score $(t \circ r)(x)$. The screening process ranks applicants by this score and selects the highest-ranked ones.

In the Example:

A central question in evaluating the possible biases introduced by the algorithm design process is to compare this disparity $D(t \circ r)(x)$ with the underlying structural disadvantage $D(f)$

$$\begin{aligned} D(t \circ r) &= D(f) + (D(g) - D(f)) + \\ &+ (D(h \circ r) - D(g)) + (D(t \circ r) - D(h \circ r)) \end{aligned} \quad (1)$$

In the Example:

- $D(f)$: Underlying structural disadvantage.
- $(D(g) - D(r))$: Bias added by using g as an outcome measure instead of f.
- $(D(h \circ r) - D(g))$: Bias added by using $r(x)$ as the feature vector instead of x.
- $(D(t \circ r) - D(h \circ r))$: Bias added by the fact that we are using an estimated function t rather than the function h.

Process of building an algorithm

To take into account

Human decisions that go into building an algorithm can, either intentionally or inadvertently, produce discrimination.

Moreover

On the other hand, the right terms in the equation (1) can be quantified in ways that bias in purely humandriven decision systems can never be. There is no way to know the true functions f , p, or q.

Process of building an algorithm

The use of an algorithm will now make detection easier by those with the authority to prevent them from discriminating.

Process of building an algorithm

Strong Affirmation

In principle (and we emphasize those cautionary words), algorithms therefore have the potential to become a force for social justice by serving as powerful detectors of human discrimination.

Importance of Algorithms

"The transparency and specificity of algorithms now create radically different opportunities to uncover discrimination. Rather than asking humans unanswerable questions, we have a clearer set of targets for inquiry that we can answer more precisely."

Importance of Algorithms

We have, in sum, gone from a scenario with purely humandriven decision making where getting any useful insights is difficult to one where, if the right laws and regulations are in place, we have opportunities to carry out concrete tests for each of the ways in which humans might introduce discrimination.

Outline

Conclusions

- It is tempting to think that human decision making is transparent and that algorithms are opaque. But, with respect to discrimination the opposite is true—or could be true, if we put the right laws and regulations in place to capitalize on the far greater specificity and transparency that algorithms make possible. That is an essential task.
- It is impossible to design an algorithm that leaves human bias isolated, because, if this were possible, it is most likely that new biases will come along that will make this isolation impossible.

Conclusions

- The risk that algorithms introduce is not from their use per se, but rather the risk that our regulatory and legal systems will not keep pace with the changing technology. But if we make the necessary adjustments to account for the different world we are in, algorithms have enormous potential to be not just a risk to be managed but actually a force for social good.

Thanks

Thank you for your attention.

Contact:
aguerreroj@unal.edu.co
corozcom@unal.edu.co
disanchezg@unal.edu.co
bhmorenoc@unal.edu.co