Modeling Illegal Discrimination Using the IEEE P3591 Standard

Chris Lam

Epistamai

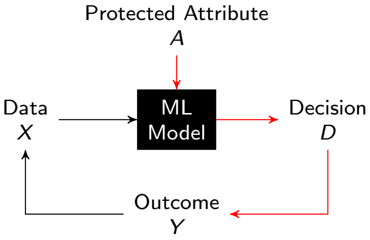
August 17, 2026

Legal Disclaimer

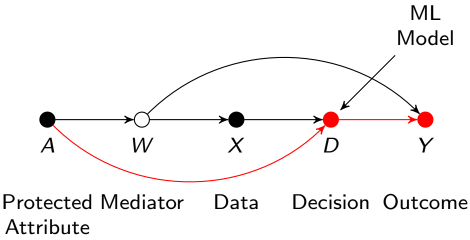
The views represented are my own and do not necessarily represent a position of the IEEE P3591 Working Group, the IEEE AI Standards Committee, the IEEE Standards Association, or IEEE.

Modeling Illegal Discrimination in ML Using Causal AI Models

Supervised Machine Learning
Block Diagram



Causal Inference
Causal Bayesian Network



Application	Protected Attribute A	Mediator W	Data X	Decision D	Outcome Y
Credit scoring	Race, gender	Creditworthiness	Income, credit history	Deny loan?	Loan default?
Resume screening	Race, gender	Qualifications	Experience, education	Screen resume?	Employee turnover?
College admissions	Race, gender	Merit	Grades, test scores	Reject applicant?	Student failure?

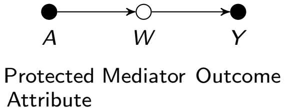
Table of Contents

What is the Basis for Using New Data for Credit Underwriting?

Keeping Fintech Fair: Thinking About Fair Lending and UDAP Risks (Federal Reserve System)

The first question to ask before using new data is the basis for considering the data. If the data are used in the credit decision-making process, what is the nexus with creditworthiness? Some data have an obvious link to creditworthiness and are logical extensions of current underwriting practices, while others are less obvious. – Carol A. Evans, Consumer Compliance Outlook: Second Issue 2017

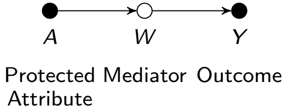
We can describe this concept using a causal graph with three nodes: the protected attribute A (e.g. race), the mediator W (e.g. creditworthiness), and the outcome Y (e.g. loan default).



The protected attribute A has a **direct effect** on the mediator W (e.g. due to historical or structural discrimination), while the mediator W has a **direct effect** on the outcome Y.

Nexus with Creditworthiness

This causal graph shows that the mediator W (creditworthiness) “explains away” the causal relationship between the protected attribute A and the outcome Y.



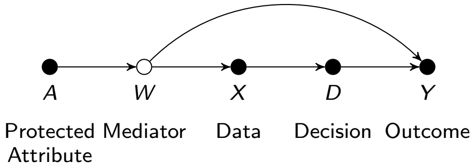
The mediator W has a **direct effect** on the outcome Y. On the other hand, the protected attribute A has an **indirect effect** on the outcome Y through the mediator W.

In causal inference, determining nexus with creditworthiness is **mediation analysis**. The outcome Y is **conditionally independent** of the protected attribute A given the mediator W.

Note that we can never fully measure the mediator W, which is why the node has been filled white (i.e. it is a latent variable).

Traditional Credit Scoring (e.g. FICO)

In traditional credit scoring, proxies for a borrower's creditworthiness W are measured as data X (e.g. credit history), which is then used to make a decision D (e.g. approve loan).



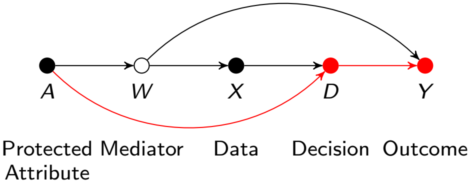
Note that the protected attribute A has **no direct effect** on the decision D (i.e. there is no direct path $A \to D$). Therefore a loan applicant cannot be treated differently based on their membership within a protected group.

Disparate Treatment

Interagency Fair Lending Examination Procedures, page iii

Overt evidence of disparate treatment. *A lender offered a credit card with a limit of up to \$750 for applicants aged 21-30 and \$1500 for applicants over 30. This policy violated the ECOA's prohibition on discrimination based on age.*

The protected attribute A (age) has a **direct negative effect** on the decision D (limit). We show that the decision D is negatively biased by coloring the nodes and edges red.

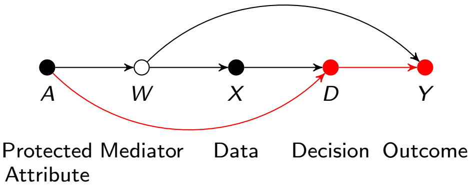


Disparate Treatment

Interagency Fair Lending Examination Procedures, page iii

Comparative evidence of disparate treatment. *A non-minority couple applied for an automobile loan. The lender found adverse information in the couple's credit report. The lender discussed the credit report with them and determined that the adverse information, a judgment against the couple, was incorrect because the judgment had been vacated. The non-minority couple was granted their loan. A minority couple applied for a similar loan with the same lender. Upon discovering adverse information in the minority couple's credit report, the lender denied the loan application on the basis of the adverse information without giving the couple an opportunity to discuss the report.*

The protected attribute A (race) has a **direct negative effect** on decision D (approve loan).

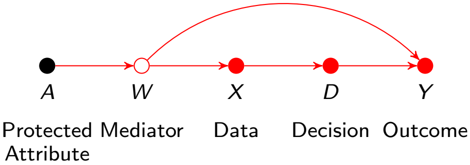


Disparate Impact

Interagency Fair Lending Examination Procedures, page iv

A lender's policy is not to extend loans for single family residences for less than \$60,000.00. This policy has been in effect for ten years. This minimum loan amount policy is shown to disproportionately exclude potential minority applicants from consideration because of their income levels or the value of the houses in the areas in which they live.

Even though the protected attribute A (race) has no direct effect on the decision D (approve loan), it still has an **indirect negative effect** through the mediator W.

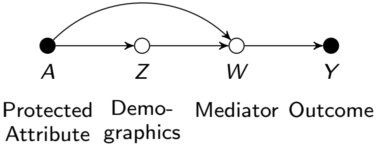


Redlining

Interagency Fair Lending Examination Procedures, page iv

Redlining is a form of illegal disparate treatment in which a lender provides unequal access to credit, or unequal terms of credit, because of the race, color, national origin, or other prohibited characteristic(s) of the residents of the area in which the credit seeker resides or will reside or in which the residential property to be mortgaged is located.

To understand this problem, we first create a new demographics variable Z, which can be used to identify the area where the credit seeker resides.

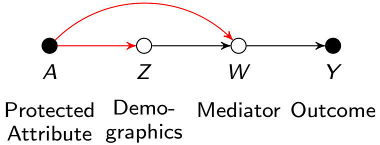


The protected attribute A (e.g. race) has a **direct effect** on the demographics variable Z (e.g. zip code). For example, historical redlining has caused minorities to live in certain areas today.

Redlining

Demographics variable Z is associated with the mediator W (creditworthiness) through both:

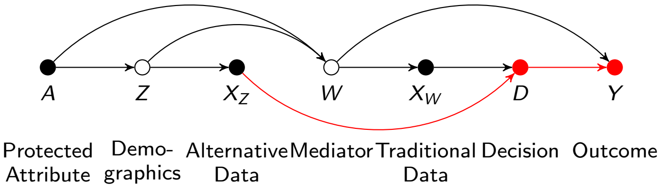
1. A **direct effect** on the mediator W ($Z \rightarrow W$). This is a direct path that would be considered its nexus with creditworthiness.
2. A **spurious effect** on mediator W through the protected attribute A ($Z \leftarrow A \rightarrow W$). This is a "backdoor" path and has been colored red. It is what causes the area where a credit seeker resides to act as a proxy for a protected attribute A (race).



Directly using data from demographics variable Z would lead to a biased decision because it would use the variable's **total effect**, which includes both its direct effect and spurious effect.

Redlining

We measure the mediator W to get traditional data X_W. We also measure demographics Z to get alternative data X_Z, which may act as a proxy for the protected attribute A.



Alternative data X_Z may have a **direct negative effect** on the decision D, which is triggered from the proxy effect of demographic variable Z through the protected attribute A. This can negatively bias the decision D, so we again color the nodes and edges red.

This is caused by the exclusion of the protected attribute A from a machine learning model ("fairness through unawareness"). In causal inference, this is called **confounding bias**. In statistics and economics, this is called **omitted variable bias**.

Educational Redlining

White House Blueprint for an AI Bill of Rights: Algorithmic Discrimination Protections

An automated system using nontraditional factors such as educational attainment and employment history as part of its loan underwriting and pricing model was found to be much more likely to charge an applicant who attended a Historically Black College or University (HBCU) higher loan prices for refinancing a student loan than an applicant who did not attend an HBCU. This was found to be true even when controlling for other credit-related factors.

A FinTech used alternative data X_Z (attendance at an HBCU) as a proxy for a prohibited basis (race), which led to a biased decision D (pricing).

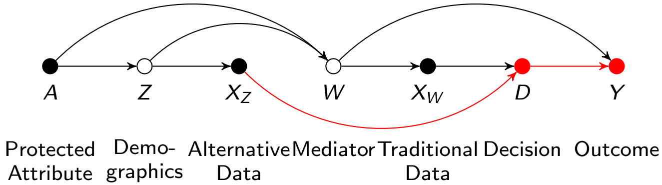
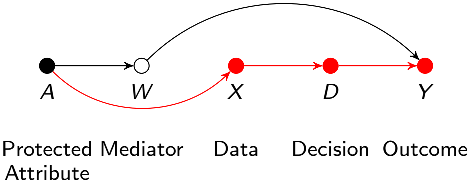


Table of Contents

Disparate Impact: Inclusion of Invalid Data

Griggs v. Duke Power (1971)

Established that neutral employment practices violate Title VII if they disproportionately exclude minority groups, regardless of intent. If a test or requirement creates a racial disparity, the employer must prove it is strictly job-related and a business necessity.

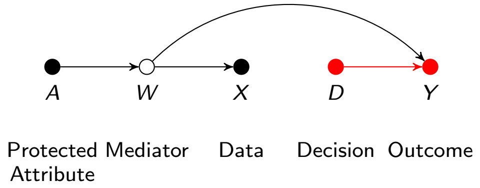


Duke Power failed to establish that the data X (test) was a valid measure of mediator W (job qualifications) or predictive of outcome Y (job performance). Therefore, the protected attribute A (race) had an **indirect negative effect** on the employment decision D through biased test data X.

Disparate Impact: Exclusion of Valid Data

Ricci v. DeStefano (2009)

Ruled that discarding exam results solely because of the racial makeup of the high scorers constitutes intentional discrimination. Employers cannot throw out tests to avoid a disparate-impact lawsuit unless they have a "strong basis in evidence" that the test itself was legally invalid.



New Haven failed to establish that the data X (test) was invalid or unlawfully discriminatory. Therefore, the exclusion of the test X could lead to a **biased** promotion decision D.