CENTER FOR DEMOCRACY & TECHNOLOGY

Colorado AI Impact Task Force presentation

Transparency and Accountability in AI-Driven Decisions

How to ensure consumers and workers can vindicate their rights under Colorado's AI Act

Matt Scherer, Senior Policy Counsel for Workers' Rights and Technology

October 21, 2024

Agenda

- About CDT
- Risks of Automated Decision Systems (ADSs)
- Why transparency requirements are essential…
  - …backed by a clear, broad scope and effective enforcement without loopholes
  - The NYC Hiring ADS Law: A cautionary tale
- Why the Colorado AI law is a step in the right direction
- What changes are needed to ensure the law provides effective protection to consumers and workers

About CDT

- CDT is a nonprofit, nonpartisan organization founded in 1994
- Mission is to advance civil rights and civil liberties in the Digital Age
- Privacy & Data project focuses on centering the interests of consumers and workers in the face of a rapidly evolving environment for data-driven technologies through a mix of research and public policy advocacy

Risks when using Automated Decision Systems (ADSs)

Overview

- Violations of law
  - Discrimination
  - Disrupting right to organize
  - Privacy laws
- Threats to health and safety (automated management; health care)
- Manipulation of prices and wages
- Serious validity, reliability, and efficacy problems

Risks when using Automated Decision Systems (ADSs)

Overview

- Violations of law
  - Discrimination
  - Disrupting right to organize
  - Privacy laws
- Threats to health and safety (automated management; health care)
- Manipulation of prices and wages
- Serious validity, reliability, and efficacy problems

Risks when using AI to make consequential decisions

How ADSs can lead to discrimination

- Instead of identifying and measuring the things that matter, automated decision systems (ADSs) often measure:
  - Traits and characteristics that are *typical*, but not necessary or even important (resume screeners, financial services, housing)
  - Personality traits based on movements, vocal intonation, speech patterns, and other easily observable (but not necessarily relevant) characteristics (video interviews, facial recognition)
  - ??? (no one knows what goes into many of these systems)

- Results can be affected by gender and cultural norms, disability, and other irrelevant and/or unlawful attributes

Risks when using AI to make consequential decisions

How ADSs can lead to discrimination

- ADSs may not be provided in an accessible format, and consumers/workers may not be able to communicate with a person about format's inaccessibility
- Exacerbated by lack of transparency (more on this in a bit)
  - Often don't know what companies are using ADSs, what those ADSs are measuring, or how they are measuring it
  - Makes it hard to detect when ADS make discriminatory decisions or to anticipate when accommodations are needed
  - In short, hard for consumers, workers, and regulators to know whether a given system is in use, much less whether it poses a risk of violating the law

Risks when using AI to make consequential decisions

How ADSs can lead to discrimination

- Healthcare algorithm impacting millions found to be racially biased
- Researcher access like this is a rarity

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

There is growing concern that algorithms

researcher-created algorithms (*10*–*13*). With-

that rely on pas
future health ca
Our dataset d
rithm. It contair
tions as well as t
its inner workin
gredients used
objective functi
set of outcome
inputs, outputs
data allow us a
racial disparities
mechanisms by
emphasized that
Rather, it is emb
proach to risk p
tor, widely ado
non-profit med
agencies (*21*).
Our analysis l
we learn about t
the specific prob
has analogies in
dicted risk of s
case, health care

Risks when using AI to make consequential decisions

Sometimes they just don't work

MIT Technology Review

ARTIFICIAL INTELLIGENCE

We tested AI interview tools. Here's what we found.

One gave our candidate a high score for English proficiency when she spoke only in German.

By
Sheridan Wall
Hilke Schellmann

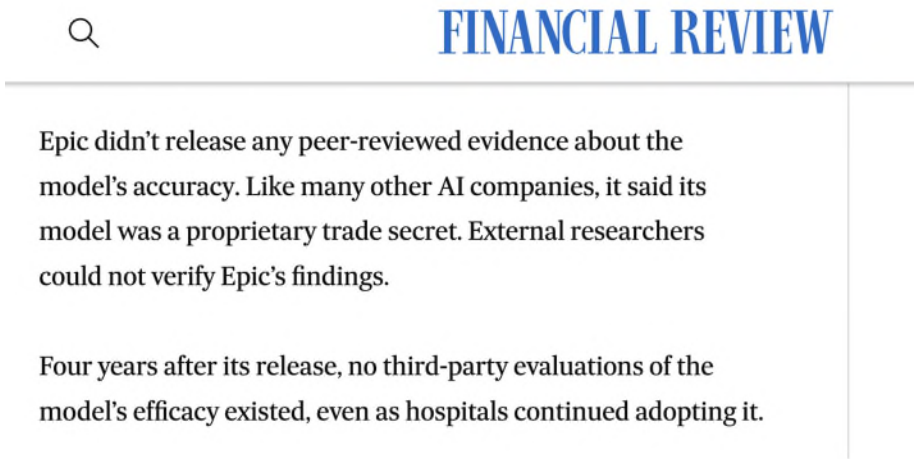
July 7, 2021

cdt

Risks when using AI to make consequential decisions

Sometimes they just don't work

- Widely adopted AI sepsis prediction tool not as accurate as claimed
- Slightly more accurate than coin flip
- Research access like this is a rarity



FINANCIAL REVIEW

Epic didn't release any peer-reviewed evidence about the model's accuracy. Like many other AI companies, it said its model was a proprietary trade secret. External researchers could not verify Epic's findings.

Four years after its release, no third-party evaluations of the model's efficacy existed, even as hospitals continued adopting it.

Risks when using AI to make consequential decisions

Sometimes they just don't work

FINANCIAL REVIEW

Epic had claimed that its model had a relative accuracy between 76 per cent and 83 per cent. (Relative accuracy refers to the probability that a patient who would go on to develop sepsis would be rated as higher risk than a patient who wouldn't.)

But the study found that the relative accuracy was actually 63 per cent – far worse than originally claimed. A relative accuracy of 50 per cent means it is as good as flipping a coin. So a 63 per cent relative accuracy means that the model is only slightly better than random.

AI Snake Oil

Why consequential decisions about the future are hard to automate

- Arvind Narayanan and Sayash Kapoor: *AI Snake Oil*
  - AI is rapidly improving at perception (recognizing faces, voices) and generating text/images, but not at predicting human behavior or social outcomes or at performing tasks where human judgments vary widely
  - Example: Determining the best candidate for a job involves both predicting social outcomes ("fit") and widely varying human judgments (recruiters *frequently* disagree on best candidates)

- This is because ADSs look for correlation rather than causation
  - It's often easier for an AI system to pick up on patterns that relate to our society's biases as it is for it to pick up on the things that actually matter in predicting who will be a good employee, tenant, student, etc

- Fortunately, these predictive ADS are precisely what Colorado's AI Law targets

The Need for Transparency

…with broad definitions and without loopholes

(1) Consumers and regulators often do not know which companies are using automated decision systems (ADSs), much less how those companies are using them.

(1) Companies have strong incentives to keep ADS use hidden--both to maintain their information advantage and to avoid regulatory scrutiny under civil rights and consumer protection laws *(and, if it's ineffective, to keep the outside world from finding out)*

(1) Consequently, companies are likely to take advantage of any narrow definitions or other loopholes in ADS laws that give them discretion to wiggle out of ADS disclosure requirements which would allow companies to ignore such laws.

(1) The only way to avoid this policy failure is to include ironclad disclosure obligations backed by strong, loophole-free enforcement provisions that prevent ADS developers and deployers from using their information advantage to avoid accountability.

A cautionary tale: NYC's AI Hiring Ordinance

- NYC passed a hiring ADS ordinance that went into effect last year--but a detailed study by academic and public interest researchers showed that companies have almost totally ignored it
- Problem is that the ordinance applies to only to ADSs that effectively replace human decision-making or otherwise dominate the decision process (combined with weak enforcement)
- The law's standard basically allows companies to decide for themselves whether their ADS use triggers the law's disclosure requirements
  - Employers might say that ADS output is one factor among many and that humans have final say--even if, in reality, the hiring managers are actually just rubber stamping or deferring to ADS "recommendations."
  - Remember: outsiders usually have no way of knowing whether a company is using an ADS unless the company tells the outside world
- Note that trade secret exemptions create a similar loophole

Example
Cigna's secret automated rejection of insurance claims

Health Care

How Cigna Saves Millions by Having Its Doctors Reject Claims Without Reading Them

by Patrick Rucker, The Capitol Forum, and Maya Miller and David Armstrong, ProPublica

March 25, 2023, 5 a.m. EDT

A Doctor at Cigna Said Her Bosses Pressured Her to Review Patients' Cases Too Quickly. Cigna Threatened to Fire Her.

Internal documents and former company executives reveal how Cigna doctors reject patients' claims without opening their files. "We literally click and submit," one former company doctor said.

The Bottom Line

- Failing to pair strong transparency with strong definitions creates a catch-22: Once a company decides that it needn't disclose an ADS's existence, the outside world may not even be aware of it. Thus, no one will be able to challenge the company's behind-closed-doors determination that the ADS is exempt from disclosure.

- Provisions that give companies discretion in deciding whether and what to disclose is thus tantamount to giving companies the unilateral ability to opt out of complying with ADS laws (same if enforcement provisions aren't strong enough to compel compliance)

Colorado AI Act (SB 24-205)

The things we like

- **Broad definition of covered systems**, making it harder for companies to evade the law;
- **Direct, proactive notice** to consumers subjected to AI-driven decisions about the purpose of the system and the role it plays in the decision process;
- **Impact assessments** that test AI decision systems for discrimination risks and document the AI decision system's purpose, intended uses, data used and produced, performance, and post-deployment monitoring;
- A **right to an explanation** of the principal reasons behind decisions and **a right to appeal** such decisions to a human decision-maker; and
- Giving the **Attorney General authority to issue rules** interpreting and clarifying the law.

Colorado AI Act (SB 24-205)

The things we like

- **Broad definition of covered systems**, making it harder for companies to evade the law;
  - If the use of the system could alter a decision, companies must disclose it. Helps avoid the NYC issue.
- **Direct, proactive notice** to consumers subjected to AI-driven decisions about the purpose of the system and the role it plays in the decision process;
  - But--more is needed.
- **Impact assessments** that test AI decision systems for discrimination risks and document the AI decision system's purpose, intended uses, data used and produced, performance, and post-deployment monitoring;
- A **right to an explanation** of the principal reasons behind decisions and **a right to appeal** such decisions to a human decision-maker; and
- Giving the **Attorney General authority to issue rules** interpreting and clarifying the law.

Colorado AI Act (SB 24-205)

The things that need to get better

- **Building on existing civil rights protections** by prohibiting the sale or use of discriminatory AI decision systems;
- **Expanding the law's transparency provisions** so that consumers understand why companies are using AI decision systems and what these tools measure and how, including requiring explanations to be actionable;
- **Strengthening impact assessment provisions** to require companies to test AI decision systems for validity and the risk that they violate consumer protection, labor, civil rights, and other laws;
- **Eliminating the many loopholes** that exclude numerous consumers, workers, and companies from the law's protections and obligations; and
- **Strengthening enforcement** by giving consumers and local district attorneys the right to seek redress in court when companies fail to comply with the law.

Colorado AI Act (SB 24-205)

The things that need to get better

- **Building on existing civil rights protections** by prohibiting the sale or use of discriminatory AI decision systems;
- **Expanding the law's transparency provisions** so that consumers understand why companies are using AI decision systems and what these tools measure and how, including requiring explanations to be actionable;
- **Strengthening impact assessment provisions** to require companies to test AI decision systems for validity and the risk that they violate consumer protection, labor, civil rights, and other laws;
- **Eliminating the many loopholes** that exclude numerous consumers, workers, and companies from the law's protections and obligations; and
- **Strengthening enforcement** by giving consumers and local district attorneys the right to seek redress in court when companies fail to comply with the law.

Colorado AI Act (SB 24-205)

The things that need to get better

- ***Building on existing civil rights protections*** *by prohibiting the sale or use of discriminatory AI decision systems;*

- The AI Law currently creates only a duty of care

- But the country's – and Colorado's – approach to discrimination is to prohibit it, regardless of whether "care" was taken to prevent it.
  - The Supreme Court said 50 years ago that there is no intent requirement for discrimination

- Having a mere "duty of care" sends a signal that ADSs are somehow going to be held to a laxer standard for discrimination--precisely the wrong message since a single biased AI system could violate the rights of thousands (or millions) of consumers and workers

Colorado AI Act (SB 24-205)

The things that need to get better

- ***Expanding the law's transparency provisions*** *so that consumers understand why companies are using AI decision systems and what these tools measure and how;*

- Right now, the notice that consumers and workers receive is very barebones, and the language leaves lots of rooms for companies to provide notice that fails to tell consumers how AI systems will make their decisions/recommendations

- Upfront description should include:
  - The things the ADS will measure or assess;
  - How it will measure/assess them;
  - How those things are relevant to the consequential decision; and
  - The respective roles of the ADS and humans in the decision-making process.

- This improved upfront notice is especially essential for individuals with disabilities

- **This is all information that ADS vendors routinely provide to prospective clients (deployers)**

Colorado AI Act (SB 24-205)

The things that need to get better

- ***…including requiring explanations to be actionable;***

- Explanations should be required to be actionable when possible, similar to explanations in credit denials
  - Those tell consumers what specifically led to the credit denial, which effectively tells consumers what steps they could have taken--or might take in the future--that might have secured a different decision (credit card balances, missed payments, etc)

- If the decision *cannot* be explained in this manner, then companies should not be allowed to use the system.

Bottom line

- ★ Prohibit discrimination
- ★ Transparency provisions must be specific, accurate, and user friendly
- ★ Robust enforcement key for compliance