AI in Consequential Decisions: The Need for Transparency

Cris Moore, Santa Fe Institute

Interdisciplinary Working Group on Algorithmic Justice


Melanie Moses
CS, UNM / SFI


Cris Moore
SFI


Kathy Powers
Poli Sci, UNM/ SFI


Alfred Mathewson
Law, UNM


Sonia Rankin
Law, UNM


Mirta Galesic
SFI


Josh Garland
Arizona State


Matthew Fricke
CS, UNM


Gabe Sanchez
Poli Sci, UNM


Tina Eliassi-Rad
CS, NEU/SFI


Mahzarin Banaji
Psych, Harvard/SFI


Trilce Estrada
CS, UNM


Nadiyah Humber
Law, UConn

AI and Consequential Decisions

AI is being used in both the public and private sector to make decisions that have long-term effects on people's lives:

- Employment (automated hiring)
- Health care, education, social services, fraud detection
- Housing: credit, lending, tenant screening, public housing waiting lists
- Criminal justice: pretrial, sentencing, parole, predictive policing

Pros: evidence-based, objective, accurate, avoids stereotypes

Cons: based on historical data, treats people as statistics, black boxes

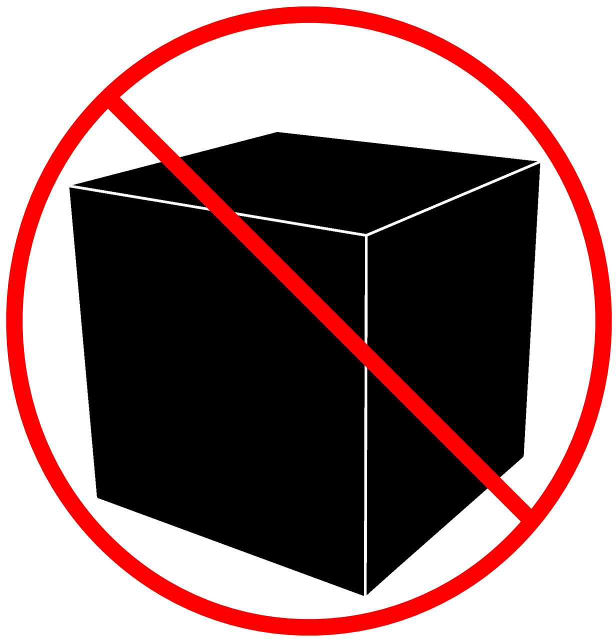
What do citizens and governments need to know about these systems?

Transparency vs. Black Boxes

What data does an AI use about a defendant or applicant?

Where does this data come from?

What does the AI do with this data to make a decision, a score, or a recommendation?



Do the people affected by an AI, and the decision makers advised by it, understand the logic behind its decisions?

Do they know what its limitations are, and what kinds of errors it can make?

Can we independently assess AIs for accuracy and fairness, or do we just have to take the vendor’s word for it?

Example #1:
Pretrial Supervision

Public Safety Assessment:
Simple point system,
publicly known weights

Based on criminal record:
Past convictions,
past failures to appear

Uses age, but not race,
gender, employment,
education, or environment

PUBLIC SAFETY ASSESSMENT RISK FACTORS

RISK FACTOR	WEIGHTS
FAILURE TO APPEAR maximum total weight = **7 points**	
Pending charge at the time of the offense	No = 0 Yes = 1
Prior conviction	No = 0 Yes = 1
Prior failure to appear pretrial in past 2 years	0 = 0 1 = 2 2 or more = 4
Prior failure to appear pretrial older than 2 years	No = 0 Yes = 1
NEW CRIMINAL ACTIVITY maximum total weight = **13 points**	
Age at current arrest	23 or older = 0 22 or younger = 2
Pending charge at the time of the offense	No = 0 Yes = 3
Prior misdemeanor conviction	No = 0 Yes = 1
Prior felony conviction	No = 0 Yes = 1
Prior violent conviction	0 = 0 1 or 2 = 1 3 or more = 2
Prior failure to appear pretrial in past 2 years	0 = 0 1 = 1 2 or more = 2
Prior sentence to incarceration	No = 0 Yes = 2
NEW VIOLENT CRIMINAL ACTIVITY maximum total weight = **7 points**	
Current violent offense	No = 0 Yes = 2
Current violent offense & 20 years old or younger	No = 0 Yes = 1
Pending charge at the time of the offense	No = 0 Yes = 1
Prior conviction	No = 0 Yes = 1
Prior violent conviction	0 = 0 1 or 2 = 1 3 or more = 2

Example #2: Prison Classification

Validation study of 2003 system at NMCD's request

Reduce medical and mental health overrides

Recent misconduct is more predictive

LFC 2020 recommendation: 10-year history is too long, one year too short

New policy is 3-5 years

Form CD-081200.1
Reviewed/Revised 05/28/24

NEW MEXICO CORRECTIONS DEPARTMENT
INITIAL CUSTODY SCORING FORM

Inmate's Name:______________________________NMCD#__________
Last *First* *MI*

Classification Officer:__________________________Classification Date:____________

1. **HISTORY OF INSTITUTIONAL ADJUSTMENT/VIOLENCE.** (Review individual's entire background for 5 years prior to classification date to include juvenile incidents) (Include date of incident; rate most severe)
   None__________ 0
   Ten or more non-violent disciplinary reports __________ 2
   Non-Violent /Serious Class A level incidents__________ 2
   Violent incident with no weapon, serious injury or death__________ 6
   Violent incident involving a weapon, serious injury or death__________ 8
2. **CURRENT CONVICTION SEVERITY** (score the most serious conviction, list offense and date)
   Low __________ 0
   Moderate __________ 1
   High __________ 2
   Highest__________ 3
3. **ESCAPE HISTORY** (Last 3 years from this rating date. List date of escape)
   None__________ 0
   Escape/Attempted escape from level I or II, county jail, juvenile facility, or peace officer (no violence) __________ 3
   Escape/Attempted escape from level III facility or above (no violence) __________ 5
   Escape/Attempted escape (with violence) __________ 10
4. **PRIOR # OF FELONY CONVICTIONS** (Do not include current conviction; list offenses and dates.) None _ 0 One or more _____ 1
5. **PRIOR CONVICTION SEVERITY** (Score the most serious offence; list offense and dates)
   None/Low ___ 0 Moderate ________ 2 High ____ 4 Highest _ 6
6. **CURRENT AGE**
   21 and under___8 22 to 25 ___5 26 to 34 ___4 35 to 44____2 45 and above ____ 0
7. **GANG MEMBERSHIP or ACTIVITIES IN THE PAST 3 YEARS**
   Yes ________ 3 No _____ 0

TOTAL SCORE (Add 1 through 7)

Example #3: Predictive Policing

1) Finding “hot spots” — places and times where crime is more likely
2) Finding people likely to commit crimes or be victims

PREDPOL®

CITY HALL NEWS CHICAGO

CPD decommissions ‘Strategic Subject List’

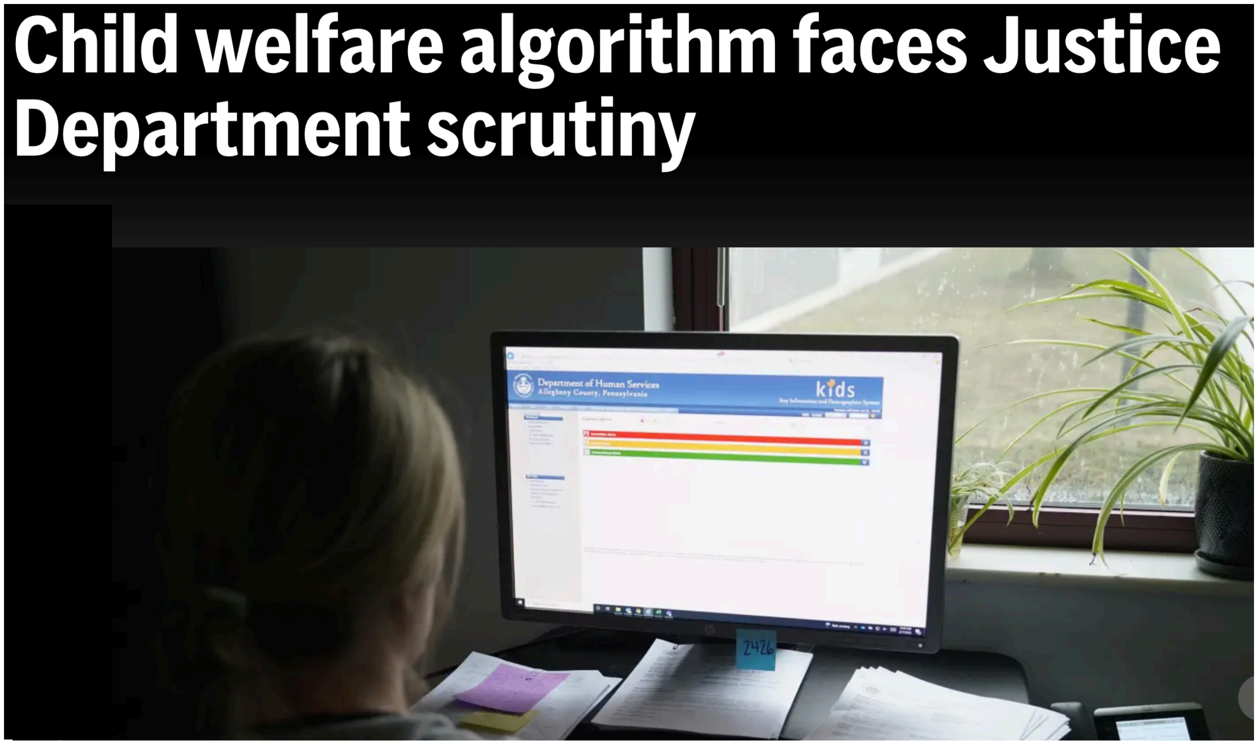
The Chicago Police Department had used analytics to identify which prior arrestees would be most likely to carry out — or be victims of — shootings.

By Sam Charles | Jan 27, 2020, 1:11pm MST

“The police say the risk scores were based on eight factors, including arrests for gun crimes, violent crimes or drugs, the number of times the person had been assaulted or shot, age at the time of the last arrest, gang membership and a formula that rated whether the person was becoming more actively involved in crime.

But the database doesn’t indicate — and the police won’t say — how much weight is given to each factor in computing the scores, which are produced using an algorithm developed at the Illinois Institute of Technology.”

Example #4: Child Welfare and Protective Services



Allegheny County, PA (Pittsburgh)

Uses prior allegations, publicly funded mental health and drug/alcohol services, jail bookings

Predicts removal from home within 2 years, re-referral after initially being screened out, or injury

Oregon Department of Human Services to End Its Use of Child Abuse Risk Algorithm

Example #5: Fraud Detection

Government's Use of Algorithm Serves Up False Fraud Charges

Using a flawed automated system, Michigan falsely charged thousands with unemployment fraud and took millions from them.

"Over a two-year period, the agency charged more than 40,000 people, billing them about five times the original benefits, which included repayment and fines of 400 percent plus interest. Amid later outcry, the agency later ran a partial audit and admitted that **93 percent of the changes had been erroneous** — yet the agency had already taken millions from people and failed to repay them for years. So far, the agency has made no public statements explaining what, exactly, went wrong."

Example #6: Tenant Screening

Why did the system say “no”?

Eviction records, data brokers

- Was this you? Name mismatches
- Were you at fault?
- Building sale, condos
- Maintenance, disputes
- Were charges dropped?
- Were records expunged?

The Markup

Big Tech Is Watching You. We’re Watching Big Tech.

Locked Out

The Obscure Yet Powerful Tenant-Screening Industry Is Finally Getting Some Scrutiny

Reforms have been in the works for years, but a looming eviction crisis has made them urgent

By Lauren Kirchner

AI can help inform consequential decisions *if*…

People affected by them understand what data about them is used and what the AI does with this data

Decision makers advised by them understand what they mean and what mistakes they can make

Policymakers understand their strengths and weaknesses

They are regularly and independently assessed for accuracy and fairness, rather than relying on vendor's claims

All this requires transparency!

Types of Transparency

“Where constitutional rights are involved, transparency is paramount.”
— Computing Community Consortium

Simple notice: Alert consumers or applicants that an AI is being used

Applicant Challenges: Allow applicants to see their data and correct it (e.g. FCRA)

Self-assessment: Require AI developers to assess their own product for bias, and perform due diligence to avoid it (like an impact statement)

Local studies: Require AI deployers to periodically test the AI for accuracy and bias on local data to make sure it works well for local populations

Independent assessments: Independent third parties (e.g. ISR at UNM)

Full transparency: Public disclosure of design and methods, sources of data, and how the AI uses that data to produce its output