AI and Algorithms in High-Stakes Decisions: The Need for Transparency

Cris Moore, Santa Fe Institute

High-Stakes Decisions in the Algorithmic Age

AI and Algorithms are 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
- Housing: credit, lending, tenant screening
- Criminal justice: pretrial, sentencing, parole

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

Cons: arbitrary, unable to deal with exceptions, possibly biased, and opaque

Algorithms in Government

In government, algorithms are used in ways that affect both public safety and constitutional rights:

- Pretrial detention, sentencing, prison classification, parole: estimate risk of recidivism or failure to appear, recommending detention or release
- Health care, social services, child protection: fraud detection, recommendations to case workers
- Predictive policing: place-based and network-based, “strategic subjects”
- Housing: public housing waiting lists

Should we use these algorithms (and spend taxpayer $$$ on them) if we don’t know how they work, or if they haven’t been independently tested for accuracy and fairness?

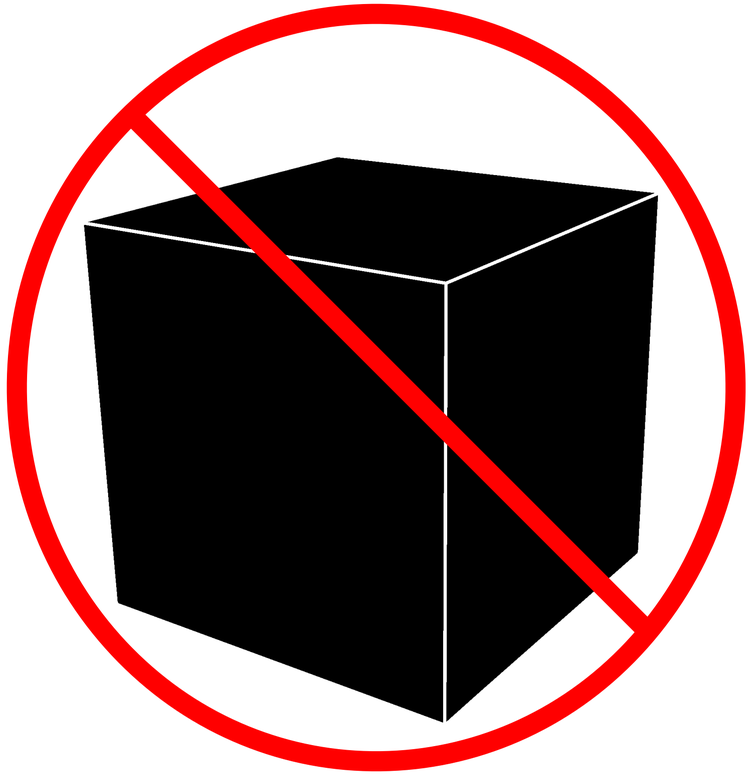
Transparency vs. Black Boxes

What data does the algorithm use about a defendant or applicant?

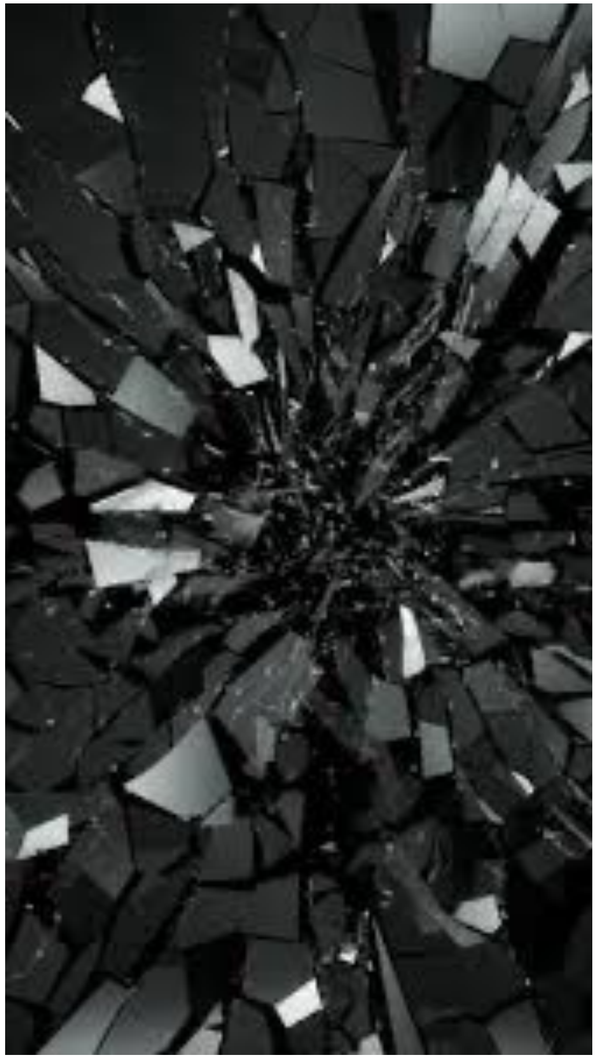
How does it weight and combine these factors?

Where does this data come from? How was it collected and curated?

How was the algorithm designed or trained?



Transparency vs. Black Boxes



Do people affected by algorithms (defendants, applicants) and people advised by them (judges, caseworkers) understand what an AI's outputs mean, and what kinds of errors it can make?

Do policymakers understand AI's strengths and weaknesses, so we can decide whether to use it?

Can we audit the AI for accuracy and fairness in New Mexico, or do we just take the vendor's word for it?

Pretrial Detention and Supervision: Public Safety Assessment (PSA)

Simple point system, publicly known weights

Based on criminal record: Past convictions, not arrests

Doesn't use juvenile record

Uses age, but not 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

Predictive Policing 1: Places and Times

Table 2. Successfully predicted crimes under deployed conditions

	Algorithm				Human Analyst					
	Success	Total	Rate	PAI	Success	Total	Rate	PAI	Boost	*P*-value
Foothill	22	346	6.4%	16.9	11	347	3.2%	8.4	2.0	0.0244
N. Hollywood	21	611	3.4%	4.9	12	732	1.6%	2.4	2.1	0.0170
Southwest	38	981	3.9%	2.9	21	936	2.2%	1.7	1.7	0.0194
Total	81	1938	4.2%	6.8	44	2015	2.2%	3.5	1.9	0.0002

Mohler et al., Randomized Controlled Field Trials of Predictive Policing
Journal of the American Statistical Association (2015)

a 6 month randomized controlled trial found that crime analysts using PredPol technology in addition to their existing tools are **twice as effective** as experienced crime analysts using hotspot mapping alone.

PREDPOL®

Predictive Policing 2: People and Networks

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

CHICAGO DATA PORTAL Chicago Data Portal

Strategic Subject List - Dashboard

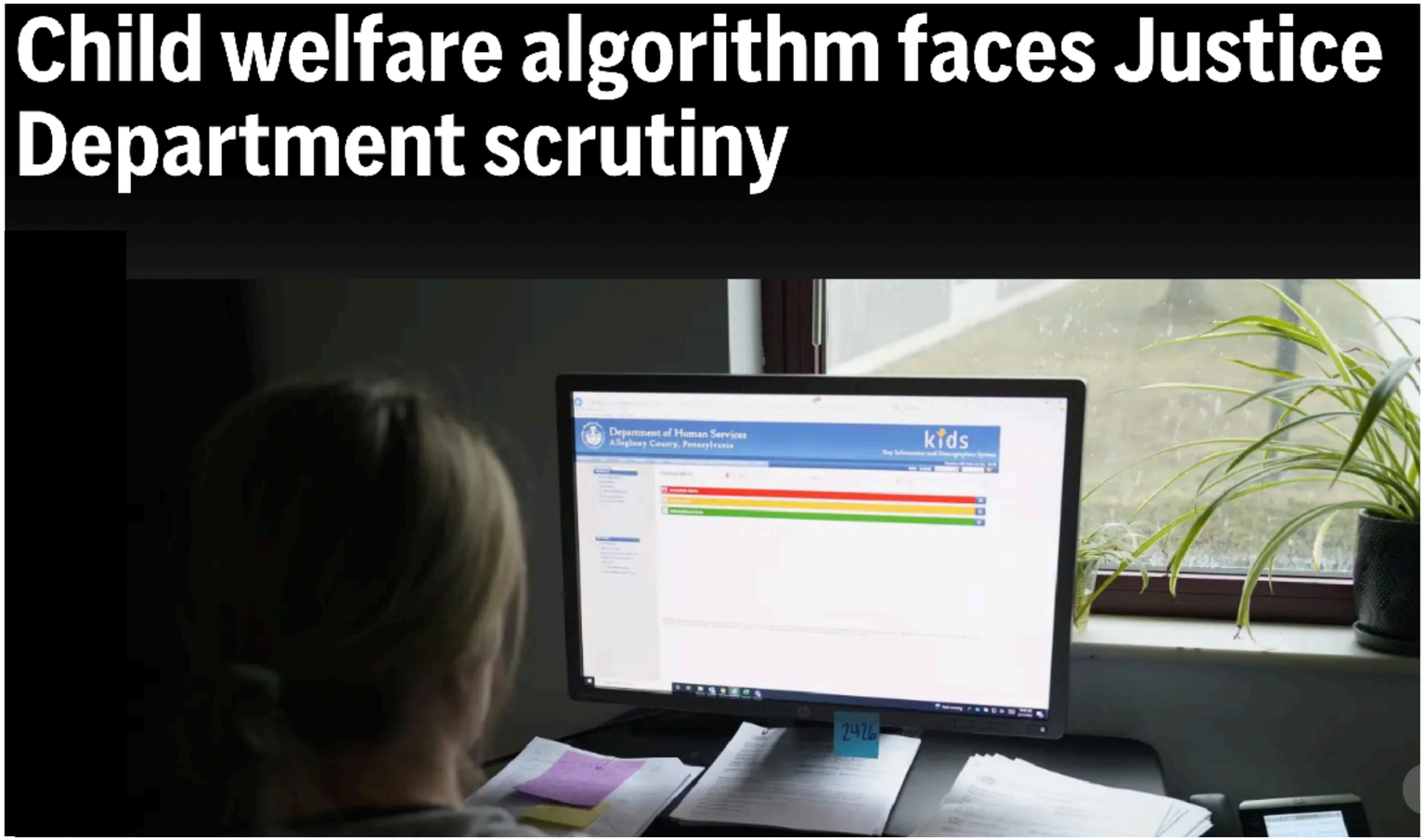
The information displayed represents a de-identified listing of arrest data from August 1, 2012 to July 31, 2016, that is used by the Chicago Police Department's Strategic Subject Algorithm, created by the Illinois Institute of Technology and funded through a Department of Justice Bureau of Justice Assistance grant, to create a risk assessment score known as the Strategic Subject List or "SSL." These scores reflect an individual's probability of being involved in a shooting incident either as a victim or an offender. Scores are calculated and placed on a scale ranging from 0 (extremely low risk) to 500 (extremely high risk).

Based on this time frame's version of the Strategic Subject Algorithm, individuals with criminal records are ranked using eight attributes, not including race or sex. These attributes are: number of times being the victim of a shooting incident, age during the latest arrest, number of times being the victim of aggravated battery or assault, number of prior arrests for violent offenses, gang affiliation, number of prior narcotic arrests, trend in recent criminal activity and number of prior unlawful use of weapon arrests.

"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."

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

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."

Algorithms can help inform high-stakes decisions *if*...

People affected by them (e.g. applicants, defendants) understand what data about them is used and how their scores are derived

Decision makers advised by them (e.g. judges) understand what they mean and what mistakes they can make

Policymakers understand their strengths and weaknesses

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

All this requires transparency!

Vermont (and Connecticut, California, …)

“Automated Decision System”: an algorithm that uses data-based analytics to make or support government decisions or judgments

An agency will inventory the use of such systems in state government, including:

- their intended benefits
- what data the AI uses, and how this data is collected, processed, **weighted and combined**
- how data is securely stored and processed to protect privacy
- whether the AI has been audited by an independent third party **using local data** for bias **and accuracy**
- whether its decisions can be explained to impacted individuals **and decision makers**
- whether its decisions are contestable and reversible by a human decision maker

No state agency shall enter into any contract to purchase, lease, or use a tool unless the vendor discloses enough about the algorithm to make these independent audits possible

Questions?