Algorithmic Bias in AI Tools

Fall 2020

New York State Permanent Commission on Access to Justice Statewide Civil Legal Aid Technology Conference

Hon. Katherine B. Forrest (fmr.)

Partner
Cravath, Swaine & Moore LLP
kforrest@cravath.com

Algorithms in AI Tools: What Really Happened…

- At the "beginning" of the AI revolution, many assumed AI tools would produce neutral decision-makers.
- The concept was that math = accuracy;
- AI tools would ELIMINATE bias.
  - Not the way it has worked out.
  - AI tools have **human progenitors**.
    - They are made by us,
    - And we teach them (at least initially) what they know.
- Now recognized that AI tools can embed, perpetuate and create bias.
- How and why do AI tools do this?
- What can be done and is being done to change this?
- These are the topics of today's presentation.

Today’s Agenda

- Examples of algorithmic bias
- An ever-so-brief explanation of algorithms (what *are* they, really?)
- The human role in AI tool-creation
  - The potential for embedding, creating and perpetuating bias
- Forms of algorithmic bias:
  - Inputs
  - Weightings / adjustments
  - Output
  - Data sets: training and verification data
    - Historical context: snapshot of time and place
    - Labeled data: who’s doing the labeling and of what
    - Unlabeled data

Today's Agenda (cont'd)

- Judicial challenges to algorithms (bias and other general challenges)
  - What cases have been brought?
  - What is succeeding, what is not?
- Regulatory horizon

Examples of Algorithmic Bias

- **Amazon**: in 2018, Amazon stopped using an AI tool, based on an algorithm, to assist in workforce recruitment
  - 60% of workforce was male
  - 74% of managerial employees were male
  - Algorithm trained on resumes spanning a ten year period
  - Most of those resumes were of “tech experienced” applicants
  - …most of whom were male
  - The AI tool “learned” that men were the preferred applicants
  - The tool penalized resumes that had gendered references – such as “women’s chess club champion”
  - The tool downgraded applicants from all-women colleges
  - Amazon conceded that the tool was faulty and ceased development; it also denied it had ever been used.

Examples of Algorithmic Bias (cont'd)

- **Google**: In 2013, Latanya Sweeney (former Chief Technology Officer of the FTC), while a researcher at Harvard
  - Studied discrimination in online ad delivery
- Found that a Google search for what might be considered African American names (she used "Trevon Jones" as an example), resulted in ad delivery for arrest record searches at a rate disproportionate to use of "white names" (she used "Emma", "Jill" and "Geoffrey")
- Ads such as "criminal background check services" were delivered at a rate 25% higher
- Leading explanation was the use of an algorithm that associated racial categories to names
  - And delivered ads based on racial categories.

Examples of Algorithmic Bias (cont'd)

- **Facial recognition tools**: In 2018, an MIT researcher, Joy Buolamwini, found that three commonly used facial-recognition AI tools demonstrated both race and gender-based biases
- In the experiments, error rates for light-skinned men were never worse than 0.8 percent – that is, 99 percent accurate for white men
  - For darker women, the error rate rose dramatically to between 20-34 percent – that is, only 64-80 percent accurate for darker skinned women
- One of the companies marketed the same facial recognition software as having an accuracy rate of 97 percent
- Buolamwini and her colleagues found that the data set that the AI tool used to learn how to differentiate faces had been 77 percent male and 83 percent white.

Examples of Algorithmic Bias (cont'd)

- **Compas ("Correctional Offender Management Profiling for Alternative Sanctions):** Compas is a widely used suite of software licensed by Northpointe.
- Used in NYS
  - Among the AI tools are programs to assist with classification and housing decisions for prisoners
    - Programs to predict general recidivism
    - Programs to predict violent behavior
- In 2016, ProPublica published the results of a study
  - Examined more than 10,000 criminal defendants in Broward, Florida
  - Compared the predicted recidivism rates with the rate that actually occurred over a two-year period
  - The score generated by Compas correctly predicted general recidivism 61 percent of the time
    - It was correct with regard to predictions of violent recidivism only 20 percent of the time

Examples of Algorithmic Bias (cont'd)

- The ProPublica study also found that:
  - Black defendants who did not recidivate were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent versus 23 percent).
  - White defendants were more often predicted to be less risky than they actually were: white defendants who reoffended had been labeled "low risk" twice as often as black defendants: 48 percent of the time versus 28 percent.
  - When controlling for prior crimes, future recidivism, age and gender, black defendants were 45 percent more likely to be given a higher risk score than their white counterparts.
  - White violent recidivists were 63 percent more likely than black violent recidivists to be misclassified as low risk.
  - Compas's owner, Northpointe, has issued a response stating that the ProPublica study is deficient and refuting its findings.

What is AI and Why are Algorithms Part of It?

- AI is just that: "artificial" intelligence
  - Machines (software) created by humans; they are artificial, not "alive"
  - Designed to use "intelligence" to carry out tasks
  - "Intelligence" is a broad concept: the ability to acquire knowledge
    - Different from thinking and different from consciousness
- AI must "learn", it needs a "brain" (the algorithm and computing power), and information
  - Think of how human children learn: by interacting with the world and absorbing information
- Various ways of teaching AI:
  - Among them: supervised, unsupervised learning
- AI uses algorithms as a base
  - Many different types of algorithms
  - Type depends on the task.

What are "Algorithms"

- **Algorithms: series of ordered steps**
  - Algorithms = mathematical formulas
- **Steps:**
  - Each step is defined (that's what we call an "input")
    - Each input is "worth" a certain amount, or "weighted"
  - Aiming towards a goal (that's the "output")
- **Like a recipe**
  - Flour, salt, sugar, yeast, water, oil = inputs
  - Amounts of each = weightings
  - Certain steps go first, others follow (activate the yeast, add flour, allow to rise, punch down, rise again, bake)
  - Output (goal) = bread
- **Simple algorithm: 2 (4 + 1) = y**
  - Goal: "To solve for y"
  - Inputs: 2, 4, 1
  - Steps: "First, add 4 + 1", "Next, multiply by 2"

Algorithmic Inputs, Weightings and Output

- **Output**: a first step
  - Formulating the desired goal: what do we want the AI tool to DO?
- Design an algorithm directed at achieving that goal
- **Inputs**:
  - No golden tablet
  - No master set of accepted principles
  - Can be: chosen by humans
    - Chosen by the software (pattern recognition)
  - Importance of WHO decides
    - How decisions are made
    - Is there a validation process
- **Weightings** of inputs:
  - No golden tablet with all the “correct” measurements
  - Who decides: machine or human? adjustments?

How Algorithms "Learn"

- **AI has to have information** – just like humans
  - Information: in data sets
  - No master data set
- **Data set issues**:
  - Who selected?
  - Why this data set and not that one?
  - What's embedded within the data set and did anyone think about it?
    - Historical context
    - Social context
    - Regional variation
    - Biases embedded: not if, just question of what

How Algorithms “Learn” (cont’d)

- Labeled data
  - Who is doing the labeling?
  - Is the data sufficiently diverse in terms of known characteristics?
  - *Where* is the labeling being done and does it matter?
- Unlabeled data
  - Who has chosen?

Human Biases: How They Can Seep In

- Humans have biases – many types
  - Race
  - Gender
  - Age
  - Religion
  - Ability
  - Nationality / origin
  - Sexual orientation
- Explicit biases:
  - Conscious biases
  - May be known, but not acknowledged as incorrect, or ethically or morally wrong
  - May be assumed as "right"
  - May be based on faulty teaching, science, or perceived personal experience

Human Biases: How They Can Seep In

- Implicit biases:
  - Stereotypes
  - Often denied, ignored as "bias"
  - Even if recognized as existing, not necessarily viewed as problematic
  - Conflation of a bias with an accepted world order: this is the way it is
- Implicit biases can impact our views of the world in an unconscious manner
  - A person may exhibit implicit bias without understanding why or even that it has occurred.

Human Biases: How They Can Seep In

- As we discussed above: AI tools have human *progenitors*
- At various steps of the design process, a designer can embed his or her personal biases
  - **Defining the output:** when a task an AI tool is initially defined, bias can creep (or leap) in
  - Examples:
    - An AI tool that assumes an ideal employee must be able-bodied (Tasked with: "identify employees who have the following characteristics", including words like "active and athletic")
    - An AI tool that assumes predicting families needing food support are comprised of citizens (Tasked with: "identify families in this social security database most likely to need food support")
    - An AI tool that assumes crime is correlated with racially segregated neighborhoods (Tasked with: "identify likely crimes in these zip codes")

Human Biases: Input Bias

- As we saw above, inputs are the elements of the recipe:
  - Certain AI tools determine inputs themselves: by identifying patterns within a data set
    - If, for instance, an AI tool is asked to determine the likelihood of recidivism, it might learn its task by using a data set of arrest records, searching for patterns among those who have been arrested
  - From that review, it could come up with its own inputs – for instance:
    - Race
    - Age
    - Gender
    - Education
    - Drug use
    - Residential status and stability
    - Marital status

Human Biases: Input Bias (cont'd)

- What do we think of *those* inputs?
  - Dangers?
- Do we want humans to adjust them?
- To tell the AI tool affirmatively not to use race?
  - Who makes that decision?
  - Are we all in agreement as to which characteristics should be ignored?
- But what about "proxies" for the eliminated input?
  - What if the AI also used zip codes as an indicator of recidivism
  - And the area in question has zip codes that correspond with majorities of one race versus another?
  - Should the AI also be told to ignore zip codes?
  - How about the names of high schools or after school programs?

Human Bias: Weightings / Adjustment Bias

- Every input in an algorithm has a weight
- Most AI tools determine the weights to place on an input based on the **patterns revealed** by the data set
- Humans can also choose weightings initially as part of the algorithm design
- Weights can and sometimes should be readjusted
  - AI tool designed to identify preferred employees, and that had been taught using a data set of current and prior employees, could weight gender highly (e.g. "since all current and prior employees in management have been male, 'being male' shall be weighted highly.")
  - A human may want to readjust the weighting to eliminate gender altogether (weight it at zero? 2% 5% 75%?).

Human Bias: Weightings / Adjustment Bias (cont'd)

- Or: AI reviewing arrest data base to look for patterns, and data base was based on a time when stop and frisk arguably resulted in the over arrest of minorities, could *weight* race higher than other characteristics
  - A human may want to readjust race downward to zero (or 3% or 15%...?)
- Several issues regarding weighting choice and/or adjustment:
  - **Who** is making the decision?
  - **What is the basis** for any adjustment?
    - Research
    - Policy
    - Whim
  - Is anyone supervising, vetting or validating the decision?
- It is highly unlikely that everyone would have the same views as to what weightings are problematic, how they should be adjusted and the basis for any adjustment.

Human Bias: Data Sets

- AI learns from data – that is, information
  - Because it learns from a set of records / photographs, etc., we call these "data sets".
- No "master" data set of anything at all
- How are data sets made?
  - Some AI learns from **labeled** data sets
  - For instance, photographs are labeled "people"
  - If the photographs that are being labeled are disproportionately *white or male*, then what the AI learns is that "people" are mostly white and male
    - Same with age, physical characteristics
    - The AI may then make normative judgments based on that learning
    - But the AI may also make mistakes about groups that it has less experience with (for instance, women and people of color).

Human Bias: Data Sets (cont'd)

- Humans (often in areas of the world that are lower wage and may not be western) *do the labeling*.
  - Who chooses the data that even gets labeled in the first place? who teaches the labelers?
  - Labeling data is labor intensive (millions of photos or portions of photos would need to be labeled, for instance).
  - Accuracy of labeling has clear impacts on what is learned.
- AI can also learn from unlabeled data sets
  - The AI must then know what characteristics to look for, what patterns are within the data.
    - AI could be given photos of crowds of people and be told to find the people within the photo.
      - It would engage in trial and error and recursively learn the portions of the images that are people or not people.

Human Bias: Data Sets (cont'd)

- A human chooses the data set that is used to teach the AI what it needs to know
  - What is that human's background?
  - If the data set requires judgment, who is exercising that judgment and who is reviewing that judgment?
  - Is some random person choosing the data set?
- Examples of data set issues:
  - <u>Recidivism ex.</u>: A data set to predict likelihood of recidivism that uses a data set
    - Of arrest records (not conviction…)
    - Time period may capture particular policing policies
    - Or, capture a crime bubble (e.g. meth flare up in community) that is no longer relevant
    - Maybe from a different part of the country
    - Maybe totally out of date.

Human Bias: Data Sets (cont'd)

- Financial services ex.: A data set to predict who "should" get a loan, or where loans with certain rates should be marketed
  - Based on zip codes of profitable loans
  - Those zip codes happen to correspond to white neighborhoods
- Employment ex.: A data set used with a tool to predict who should get a job
  - Taken from records of former employees when the organization hired few women or minorities.

Understanding the Algorithm

- The biggest challenge facing those impacted by algorithmic decision-making is obtaining sufficient information about the algorithm.
  - Two general possibilities:
    - (1) those which humans can follow and understand the AI tool's choices
      - How the inputs, weightings and data set were chosen
    - (2) those that are considered "black box": humans cannot (easily) understand
  - Where is the evidence of what's in the algorithm?
    - Source code
    - Often quite difficult to obtain
  - Some companies assert "trade secret" / proprietary
    - Protective orders and judicial restrictions on access are a vehicle to address (or, court-appointed expert)
  - Transparency and fairness issues.

Judicial Challenges to Algorithmic Bias

- An Initial question: is the discrimination resulting from the AI tool qualitatively different from humans undertaking the very same task?
  - Or, do we just want and expect more from AI tools?
- **Legal Theories**:
  - **Discrimination**: employment, housing, lending, etc.
  - *Intentional discrimination* presents issues:
    - Absence of adequate considerations may be “un”-intentional
    - If the AI tool has itself learned inputs and weightings based on the data set, difficult to show that the human designer has engaged in *intentional* conduct
    - Difficult to show that data set was *chosen* to convey bias
  - *Disparate Impact* presents fewer issues:
    - Showing the disparate impact may be the easiest element to meet
    - Harder for a business to show “necessity” of using a particular data set, inputs of weightings.

Judicial Challenges: Cases

- **Constitutional**
  - **Due process**:
    - Loomis v. Wisconsin: lost the challenge (new pro se challenge proceeding: Henderson v. Stensberg, Wisconsin, 2020; but see People v. Younglove ,MI,rejected claim as not preserved; State v. Gordon, IA, not preserved)
    - Algorithm used to predict a defendant's recidivism; judge used
    - Raised question of whether lack of access to the source code prevented the defendant from fully understanding basis of decision
    - Company that made the software refused to turn over source code, arguing it was "proprietary" and trade secret; Court agreed
    - The inputs, weightings and data set choices were not known
      - General information on what was included in the algorithm was provided
    - Appellate court: no due process violation.

Judicial Challenges: Cases (cont'd)

- **Public Employment / teaching assessment tools**
  - "Value-added Assessment Tools"
  - Algorithms used to try and capture value of a teacher to student outcomes
  - Series of cases:
    - Wagner v. Haslam: challenge to TN tool
      - Challenge to the inputs
      - Equal protection and due process claims; court said passed muster
    - Trout v. Knox County Board of Ed.
      - Focused on statistical inaccuracy
      - Court denied challenge
    - Houston Federation of Teachers
      - Argued that procedural due process violated
      - Court denied SJ (results could not be replicated)

Judicial Challenges: Cases (cont'd)

- **Public benefits**
  - <u>KW v. Armstrong</u> (Kentucky)
  - Concerned reduction of benefits to developmentally disabled
  - Algorithm assigned weights to variables used to calculate a budget for a potential recipient
    - Based on assessment of personal characteristics and prediction of participant's needs
    - *Adjustment of weight of input variables* relating to "needing assistance with mobility" and "living situation" caused significant decrease in budgeted amounts
  - Budgets decreased based on algorithmic findings and letters sent out
  - No prior notice or opportunity for a hearing
  - Procedural **due process** challenge
  - Court found the tool was arbitrary and unreliable.

Judicial Challenges: Cases (cont'd)

- **<u>Confrontation Clause</u>**:
- When the AI tool is asked to answer a specific question, is it providing "testimony" for purposes of the confrontation clause?
- <u>People v. Wakefield</u> (2019): (Justice Pritzker):
  - DNA analysis software program(TrueAllele) is an AI tool that "automates the interpretation of the data signals generated in the lab"
  - Uses an algorithm for that gives probabilities
  - Has additional AI capability to make inferences when other information is available
  - The AI tool then answers a specific question: "how much more the suspect matches the evidence [than] a random person would"
  - Answer is in the form of a likelihood ratio
  - One issue in the case was whether the source code itself was a declarant within the meaning of the Confrontation Clause
  - Court found under facts here, code was not a "declarant".

Judicial Challenges: Cases (cont'd)

- **FOIL** requests:
  - Miller v. N.Y. State Dept. of Financial Services:
    - Sought materials including the algorithm relating to creation of database of payday lenders compiled by DFS, and which used an algorithm to predict violations of usury laws
    - DFS objected that it was proprietary and would damage effectiveness, Court agreed (but required more information be provided)
  - NYU (Brennan Center) v. NYPD:
    - Art. 78 petition relating to the algorithms used for predictive policing (use of "Palantir Gotham" – analyzing data to predict crime locations.)
    - Court agreed there was an insufficient showing of trade secret status
    - Request for source code dropped and not primary issue
    - Court required disclosure of host of information.

Judicial Challenges: Cases (cont'd)

- **Frye**/ **Daubert**: Just too new and too untested (using it for ESI)
  - Moore v. Publicis Groupe, S.D.N.Y. 2012 (Peck):
  - Challenge to defendant employer's use of an algorithmic AI tool to assist with document discovery
  - Predictive coding technology
  - Over 3 million documents for review
  - Algorithms determine relevance based on interactions and iterative process with human reviewer
  - Uses a "seed set" of relevant documents – the algorithm looks for patterns and applies those patterns
  - Court allowed the computer assisted review.

Judicial Challenges: Cases (cont'd)

- **Consumer protection litigation: various areas**
  - Force v. Facebook: 2d Cir.
    - Material support of terrorism claim
    - Issue was whether Facebook's use of its algorithms constituted development of content, thereby removing it from the protections of the Communications Decency Act
    - Argument was that the "match-making" algorithms that paired content with users (alleged terrorists) crossed the line
    - Court found that Facebook was not a content creator
    - Found algorithms were neutral
  - Jiminez v. Credit One Bank NA (S.D.N.Y. 2019)
    - Issue was whether Credit One's dialing system, generated by an algorithm, was "predictive" for purposes of Tel. Prot. Act
    - Defense expert opined without obtaining access to the algorithm despite the centrality of the algorithm's capabilities to the issue.
    - Court said it was predictive; algorithm changed with information.

Judicial Challenges: Cases (cont'd)

- **Housing**
  - <u>Conn. Fair Housing Center v. Corelogic Rental Property Solutions</u>
    - Use of AI tool for background checks by landlords
    - Alleged violation of Fair Housing Act and disparate impact claims against screening company (not housing provider)
    - Used public criminal records and a predictive "matching" algorithm
    - Applicant denied on basis of a record marked as "disqualifying"
    - Disqualification based on "unidentified records"; landlord did not know nature of the issue
    - Turned out to be an arrest for shoplifting that had been dropped
    - Defendant moved to dismiss on basis that could not be intentional discrimination
    - Court denied motion – on basis (inter alia) that defendant designed the form the algorithm used as input.

Regulatory Horizon

- Little regulation specific to algorithmic bias in place in the USA
  - Efforts underway
- April 10, 2019: Algorithmic Accountability Act ("AAA") introduced to the U.S. Congress (referred to committee)
  - Would authorize the FTC to create regulatory scheme
  - Scheme oriented around concept of "impact assessments"
  - Requires companies to assess impacts of algorithms for automated decision-making on fairness, bias, discrimination, privacy and security
- In Europe: General Data Privacy Regulation ("GDPR") – in effect April 2018
  - Provisions on automated decision-making (a person has a right not to be subject to a purely automated decision)
  - Has a meaningful right to information about the "logic" involved.

Regulatory Horizon

- New York State
  - Nothing solely on algorithms
  - Bills on:
    - autonomous vehicles
    - biometric technology
  - July 2019: Gov. Cuomo created commission to study how to regulate AI
- New York City
  - 2018 task force on Automated Decisions
  - Issued recommendations in 2019: recommends a structure within the City to review use of automated decisions, recommends focusing on setting out principles for information sharing, channeling public inquiry and assessment
  - Nov. 2019: de Blasio signed an executive order establishing "Algorithm Management and Policy Officer".

Thank you!