Explaining algorithms and automation: A guide for lawyers

Dr. Reuben Binns, Dept. Computer Science, University of Oxford

reuben.binns@cs.ox.ac.uk
Overview

Digitisation, automation, decision-making

Rule-based vs statistical systems

Bias, error, discrimination

Automation bias

Unequal and fettered discretion
World stumbling zombie-like into a digital welfare dystopia, warns UN human rights expert

NEW YORK (17 October 2019) – A UN human rights expert has expressed concerns about the emergence of the “digital welfare state”, saying that all too often the real motives behind such programs are to slash welfare spending, set up intrusive government surveillance systems and generate profits for private corporate interests.

"As humankind moves, perhaps inexorably, towards the digital welfare future it needs to alter course significantly and rapidly to avoid stumbling zombie-like into a digital welfare dystopia," the Special Rapporteur on extreme poverty and human rights, Philip Alston, says in a report to be presented to the General Assembly on Friday.
Digitisation, automation, decision-making

**Digitisation** of paper forms (e.g. tax returns online)

**Automation** of processes (e.g. automatically recurring payments)

Computer-supported /automated **decision-making (ADM)**, e.g.:

- Determining eligibility for benefit
- Risk scoring based on statistical models
- Fraud detection
Rules-based systems

e.g.

IF “years_in_residence” > 5:
   THEN:
      “settled_status_eligibility” = TRUE

Statistical systems

Aim to *classify, predict, or score*

How similar is this benefits application to previously fraudulent ones?

How likely is this person to re-offend (based on statistics from previous cases?)

How risky is the person behind this visa application?
Data -> Salary -> Learning Algorithm -> Model

<table>
<thead>
<tr>
<th>Observation Number</th>
<th>Temperature (X)</th>
<th>Yield (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>122</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>118</td>
</tr>
<tr>
<td>3</td>
<td>54</td>
<td>128</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>121</td>
</tr>
<tr>
<td>5</td>
<td>56</td>
<td>125</td>
</tr>
<tr>
<td>6</td>
<td>59</td>
<td>136</td>
</tr>
<tr>
<td>7</td>
<td>62</td>
<td>144</td>
</tr>
<tr>
<td>8</td>
<td>65</td>
<td>142</td>
</tr>
<tr>
<td>9</td>
<td>67</td>
<td>149</td>
</tr>
<tr>
<td>10</td>
<td>71</td>
<td>161</td>
</tr>
<tr>
<td>11</td>
<td>72</td>
<td>167</td>
</tr>
<tr>
<td>12</td>
<td>74</td>
<td>168</td>
</tr>
<tr>
<td>13</td>
<td>75</td>
<td>162</td>
</tr>
<tr>
<td>14</td>
<td>76</td>
<td>171</td>
</tr>
<tr>
<td>15</td>
<td>79</td>
<td>175</td>
</tr>
<tr>
<td>16</td>
<td>80</td>
<td>182</td>
</tr>
<tr>
<td>17</td>
<td>82</td>
<td>180</td>
</tr>
<tr>
<td>18</td>
<td>85</td>
<td>183</td>
</tr>
<tr>
<td>19</td>
<td>87</td>
<td>188</td>
</tr>
<tr>
<td>20</td>
<td>90</td>
<td>200</td>
</tr>
<tr>
<td>21</td>
<td>93</td>
<td>194</td>
</tr>
<tr>
<td>22</td>
<td>94</td>
<td>206</td>
</tr>
<tr>
<td>23</td>
<td>95</td>
<td>207</td>
</tr>
<tr>
<td>24</td>
<td>97</td>
<td>210</td>
</tr>
<tr>
<td>25</td>
<td>100</td>
<td>219</td>
</tr>
</tbody>
</table>
If $P(\text{default}) > \text{threshold}$, then deny credit.
High-dimensionality
Complexity
- Non-linear
- Non-monotonic
‘Deep learning’

features: \{ 1,1 = black, 1,2 = brown, 1,3 = grey \ldots \}

hidden layers: \{?\}

...labrador?
Bias, error, discrimination in statistical models

- False positives vs false negatives
- Fitting to the majority population
- Reflecting (and compounding) structural discrimination
Bias, error, discrimination in statistical models

- False positives vs false negatives
- Fitting to the majority population
- Reflecting (and compounding) structural discrimination
False Positive:
the boy cried wolf... but no wolf

False Negative:
The villagers thought ‘no wolf’ ... but wolf!
$x = \text{reoffend}$

$o = \text{not reoffend}$
$x = \text{reoffend}$

$0 = \text{not reoffend}$
\( x = \text{reoffend} \)
\( o = \text{not reoffend} \)

False positive rate = \( \frac{3}{14} = 21\% \)

False negative rate = \( \frac{6}{24} = 25\% \)
Better that ten guilty persons escape than that one innocent suffer

— Sir William Blackstone (1765)
Bias, error, discrimination in statistical models

- False positives vs false negatives
- Fitting to the majority population
- Reflecting (and compounding) structural discrimination
\( x = \text{reoffend} \)
\( o = \text{not reoffend} \)
$x = \text{reoffend}$

$o = \text{not reoffend}$
Bias, error, discrimination in statistical models

- False positives vs false negatives
- Fitting to the majority population
- Reflecting (and compounding) structural discrimination
prior offence

Model

re-offend
prior arrests

Model

re-arrest
Usage rates

1.3
Blacks used marijuana at 1.3 times the rate of whites.

Arrest rates

3.7
Blacks were arrested for marijuana possession at 3.7 times the rate of whites.

prior arrests

Model

re-arrest

Report: The War on Marijuana in Black and White, ACLU
Usage rates

1.3

Blacks used marijuana at 1.3 times the rate of whites.

Arrest rates

3.7

Blacks were arrested for marijuana possession at 3.7 times the rate of whites.

prior arrests

Model

re-arrest

Report: The War on Marijuana in Black and White, ACLU
“black people are now nine times more likely to be stopped and searched for drugs despite using drugs at a lower rate than white people”
features: { qualifications, postcode, place of birth, occupation, behavioural data, ... }
Parity of errors between protected classes

protected groups receive equal proportion of errors

Model performance on male applicants

Model performance on female applicants
Parity of calibration between protected classes

Calibration: of those given a particular risk score $S$, $S\%$ should result in the predicted outcome.

Calibration should be equal between protected groups.
Roles for automated decision-making

Decision **support** vs **full automation**

- **Decision support:** providing additional information, informed by statistical or rules-based systems, to aid a human decision-maker in their decision.
  - E.g. a risk score presented to a parole officer to inform their assessment of an offender
- **Fully automated:** the system takes a decision and action in relation to a person or group without human input.
  - E.g. a visa application is automatically assessed and approved

*NB: implications for data protection (GDPR Article 22 ‘solely automated’ decisions)*
Fully automated

Decision support

New Case

automated processing

Outcome 1  Outcome 2

New Case

automated processing

Assigned a Label/Category

Human

manual analysis

Outcome 1  Outcome 2
Automation bias

Human decision-makers may either systematically:

*Under-rely* on computer outputs, ignoring good information

*Over-rely* on computer outputs, ignoring their own judgement and supplemental information from other sources

Daniel Schwen / Wikimedia Commons. Boeing 787 cockpit at the Museum of Flight near Seattle
Unequal application of discretion

Under-reliance and over-reliance might be applied unequally between different groups.

Even if the algorithm is not biased, the way that human decision-makers use it may interact with existing prejudice / bias


Unequal application of discretion

An initial ADM stage may determine *which* human decision makers make the assessment.

Even if no decision is taken without a human, the algorithmic step determines the type and quality of human judgement.

Independent Chief Inspector of Borders and Immigration, ‘An inspection of entry clearance processing operations in Croydon and Istanbul’
Unequal application of discretion

Figure 4: Decisions and streaming tool ratings for Croydon visit visa applications
1 January to 28 February 2017

<table>
<thead>
<tr>
<th>Streaming rating</th>
<th>Applications</th>
<th>Percentage issued</th>
<th>Percentage refused</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>13,560</td>
<td>96.36%</td>
<td>3.64%</td>
</tr>
<tr>
<td>Amber</td>
<td>3,662</td>
<td>81.08%</td>
<td>18.92%</td>
</tr>
<tr>
<td>Red</td>
<td>6,421</td>
<td>48.59%</td>
<td>51.41%</td>
</tr>
</tbody>
</table>
Upstream automation may fetter downstream discretion

Figure 5: Daily benchmarks for deciding visit applications

<table>
<thead>
<tr>
<th>Location</th>
<th>Super Green</th>
<th>Green</th>
<th>Amber</th>
<th>Red</th>
</tr>
</thead>
<tbody>
<tr>
<td>Croydon</td>
<td>N/A</td>
<td>75</td>
<td>35</td>
<td>25</td>
</tr>
<tr>
<td>Istanbul</td>
<td>100</td>
<td>70</td>
<td>35</td>
<td>30</td>
</tr>
</tbody>
</table>
Where is the decision? Who / what made it?
Thanks!

Reuben Binns

reuben.binns@cs.ox.ac.uk

Twitter: @RDBinns