RIGOROUS ANALYSIS OF RACIAL BIAS IN GENDER CLASSIFICATION

Vidya Muthukumar

joint work with

Tejaswini Pedapati, Prasanna Sattigeri, Nalini Ratha, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, Kush Varshney

Gender classification across skin type: Pilot Parliaments’ Benchmark dataset

<table>
<thead>
<tr>
<th>Skin Type</th>
<th>Female Accuracy</th>
<th>Male Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dark</td>
<td>75.35%</td>
<td>97.59%</td>
</tr>
<tr>
<td>Light</td>
<td>95.47%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Key question: which *underrepresented* features in dark-skinned females contribute to disparate accuracy?

EXPERIMENT #1: STABILITY TO SKIN TONE CHANGE

classification baseline: "state-of-the-art" IBM Watson classifier

Effect on dark females

Effect on light females

observation: skin tone reflected primarily by Y-component!

conclusion: influence of skin type *alone* on classifier is minimal*

*null hypothesis verified through a one sample t-test

EXPERIMENT #2: INFLUENCE OF HAIR INFORMATION

classification baseline: "state-of-the-art" VGGFace2-trained features + SVM on face-cropped images

Hair information removed

conclusion: disparate accuracy persists in the absence of hair information

IMPLICATIONS OF STUDY

- underrepresented skin type, hair patterns not significant contributors to disparate accuracy
- facial features most relevant —> gender stereotypes also reinforced
- quick-fix algorithms will not solve the problem — balanced dataset creation required!

REFERENCES