Responsible Data Management

Julia Stoyanovich

Computer Science and Engineering
Center for Data Science
Center for Responsible AI
Visualization and Data Analytics Center
New York University
Artificial Intelligence (AI)
a system in which algorithms use data and make decisions on our behalf, or help us make decisions
The promise of AI

Opportunity
make our lives convenient
accelerate science
boost innovation
transform government
Machines make mistakes
Mistakes lead to harms
Harms can be cumulative
We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U. S. labor market.
Bias in algorithmic hiring

**The Guardian**  July 2015
Women less likely to be shown ads for high-paid jobs on Google, study shows

**Reuters**  October 2018
Amazon scraps secret AI recruiting tool that showed bias against women

**The New York Times**  March 2021
We Need Laws to Take On Racism and Sexism in Hiring Technology
Artificial intelligence used to evaluate job candidates must not become a tool that exacerbates discrimination.

**MIT Technology Review**  February 2013
Racism is Poisoning Online Ad Delivery, Says Harvard Professor

**The Wall Street Journal**  September 2014
Are Workplace Personality Tests Fair?
Growing Use of Tests Sparks Scrutiny Amid Questions of Effectiveness and Workplace Discrimination
Do these tools actually work?

“A theory or idea shouldn’t be scientific unless it could, in principle, be proven false.”

Karl Popper
a push for regulation
Automated Decision Systems (ADS)

process data about people
help make consequential decisions
combine human & automated decision making
aim to improve efficiency and promote equity
This law requires that a **bias audit** be conducted on an automated employment decision tool prior to the use of said tool. The bill also requires that candidates or employees **be notified about the use of such tools** in the assessment or evaluation for hire or promotion before these tools are used, as well as **be notified about the job qualifications and characteristics that will be used** by the tool. Violations of the provisions of the bill are subject to a civil penalty.
do the tools work?
Personality prediction in hiring
Algorithmic personality tests

**Input:** resume or LinkedIn handle (both systems) or Twitter (Humantic AI)

**Output:** a personality profile + a job fit score (Crystal) or match score (Humantic AI)
An external stability audit framework to test the validity of personality prediction in AI hiring

Alene K. Rhea\textsuperscript{1,2} \cite{1} \@stoyanoj \@stoyanoj - Kelsey Markey\textsuperscript{1,2} \@stoyanoj - Lauren D’Arinzo\textsuperscript{1,2,3} \@stoyanoj - Hilke Schellmann\textsuperscript{4} - Mona Sloane\textsuperscript{2} \@stoyanoj - Paul Squires\textsuperscript{5} - Faalah Arif Khan\textsuperscript{1,2} \@stoyanoj - Julia Stoyanovich\textsuperscript{1,2,6} \\ Received: 6 October 2021 / Accepted: 5 August 2022 / Published online: 17 September 2022 \\ © The Author(s) 2022
Stability audit framework

Key facets across which system assumes its output to be stable

- Facet 1
- Facet 2
- Facet 3
- Facet 4

Input

System Assumptions

Generate Treatments

input\_control

input\_treatment

scores\_control

scores\_treatment

ADS

Auditor

https://link.springer.com/article/10.1007/s10618-022-00861-0
Stability audit framework

<table>
<thead>
<tr>
<th>Facet</th>
<th>Crystal</th>
<th>Humantic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resume file format</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>LinkedIn URL in resume</td>
<td>?</td>
<td>✗</td>
</tr>
<tr>
<td>Source context</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Algorithm-time / immediate</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Algorithm-time / 31 days</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Participant-time / LinkedIn</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Participant-time / Twitter</td>
<td>N/A</td>
<td>✓</td>
</tr>
</tbody>
</table>

https://link.springer.com/article/10.1007/s10618-022-00861-0
all about that bias
Bias in computer systems

**Pre-existing**: exists independently of algorithm, has origins in society

**Technical**: introduced or exacerbated by the technical properties of an ADS

**Emergent**: arises due to context of use

[Friedman & Nissenbaum (1996)]
pre-existing bias
Pre-existing bias has origins in society
Pre-existing bias has origins in society
Pre-existing bias has origins in society
Pre-existing bias has origins in society
Diverse balanced ranking

Goals

diversity: pick $k = 4$ candidates, including 2 of each gender, and at least one per race

utility: maximize the total score of selected candidates

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>A (99)</td>
<td>C (96)</td>
</tr>
<tr>
<td>Black</td>
<td>E (91)</td>
<td>G (90)</td>
</tr>
<tr>
<td>Asian</td>
<td>I (87)</td>
<td>K (86)</td>
</tr>
</tbody>
</table>

Problem

picked the best White and male candidates (A, B) but did not pick the best Black (E, F), Asian (I, J), or female (C, D) candidates

Beliefs

scores are more informative within a group than across groups - effort is relative to circumstance

it is important to reward effort

score = 373

score = 372

[Yang, Gkatzelis, Stoyanovich (2019)]
From beliefs to interventions

Fairness for female candidates

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
<th>G</th>
<th>H</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95</td>
<td>95</td>
<td>90</td>
<td>86</td>
<td>83</td>
<td>83</td>
</tr>
</tbody>
</table>

83 / 95 = 0.91

highest-scoring skipped

lowest-scoring selected

BEFORE: diversity constraints only

AFTER: diversity and fairness constraints

Beliefs

scores are more informative within a group than across groups - effort is relative to circumstance

it is important to reward effort

[Yang, Gkatzelis, Stoyanovich (2019)]
Fairness in Ranking, Part I: Score-based Ranking

MEIKE ZEHLIKE, Humboldt University of Berlin, Max Planck Institute for Software Systems, and Zalando Research, Germany
KE YANG, New York University, NY, and University of Massachusetts, Amherst, MA, USA
JULIA STOYANOVICH, New York University, NY, USA

In the past few years, there has been much work on incorporating fairness requirements into algorithmic rankers, with contributions coming from the data management, algorithms, information retrieval, and recommender systems communities. In this survey we give a systematic overview of this work, offering a broad perspective that connects formalizations and algorithmic approaches across subfields. An important contribution of our work is in developing a common narrative around the value frameworks that motivate specific fairness-enhancing interventions in ranking. This allows us to unify the presentation of mitigation objectives and of algorithmic techniques to help meet those objectives or identify trade-offs.

In this first part of this survey, we describe four classification frameworks for fairness-enhancing interventions, along which we relate the technical methods surveyed in this paper, discuss evaluation datasets, and present technical work on fairness in score-based ranking. In the second part of this survey, we present methods that incorporate fairness in supervised learning, and also give representative examples of recent frameworks for fair score-based ranking methods.

CSC Concepts: • Information system technology policy.

Additional Key Words and Phrases: fairness, algorithms, information retrieval, recommender systems.

ACM Reference Format:

Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems

MEIKE ZEHLIKE, Humboldt University of Berlin, Max Planck Institute for Software Systems, and Zalando Research, Germany
KE YANG, New York University, NY, and University of Massachusetts, Amherst, MA, USA
JULIA STOYANOVICH, New York University, NY, USA

In the past few years, there has been much work on incorporating fairness requirements into algorithmic rankers, with contributions coming from the data management, algorithms, information retrieval, and recommender systems communities. In this survey we give a systematic overview of this work, offering a broad perspective that connects formalizations and algorithmic approaches across subfields. An important contribution of our work is in developing a common narrative around the value frameworks that motivate specific fairness-enhancing interventions in ranking. This allows us to unify the presentation of mitigation objectives and of algorithmic techniques to help meet those objectives or identify trade-offs.

In the first part of this survey, we describe four classification frameworks for fairness-enhancing interventions, along which we relate the technical methods surveyed in this paper, discuss evaluation datasets, and present technical work on fairness in score-based ranking. In this second part of this survey, we present methods that incorporate fairness in supervised learning, and also give representative examples of recent work on fairness in recommendation and matchmaking systems. We also discuss evaluation frameworks for fairness in ranking methods.

topics → Computing /

commender Systems, ACM
technical bias
Technical bias may be introduced or exacerbated by the technical properties of an ADS.
Model development lifecycle

Goal
design a model to predict an appropriate level of compensation for job applicants

Problem
accuracy is lower for middle-aged women - a fairness concern!

now what?

[Schelter, He, Khilnani, Stoyanovich (2020)]
Missing values: Observed data
Missing values: Imputed distribution
Missing values: True distribution
Missing value imputation

are values **missing at random** (e.g., gender, age, years of experience, disability status on job applications)?

are we ever interpolating **rare categories** (e.g., Native American)

are **all categories** represented (e.g., non-binary gender)?
Data filtering

“filtering” operations (like selection and join), can arbitrarily change demographic group proportions

select by zip code, country, years of C++ experience, others?

<table>
<thead>
<tr>
<th>age_group</th>
<th>county</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>CountyA</td>
</tr>
<tr>
<td>60</td>
<td>CountyA</td>
</tr>
<tr>
<td>20</td>
<td>CountyA</td>
</tr>
<tr>
<td>60</td>
<td>CountyB</td>
</tr>
<tr>
<td>20</td>
<td>CountyB</td>
</tr>
<tr>
<td>20</td>
<td>CountyB</td>
</tr>
</tbody>
</table>

50% vs 50%

<table>
<thead>
<tr>
<th>age_group</th>
<th>county</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>CountyA</td>
</tr>
<tr>
<td>60</td>
<td>CountyA</td>
</tr>
<tr>
<td>20</td>
<td>CountyA</td>
</tr>
</tbody>
</table>

66% vs 33%
“filtering” operations (like selection and join), can arbitrarily change demographic group proportions.

select by zip code, country, years of C++ experience, others?

<table>
<thead>
<tr>
<th>patients</th>
<th>healthcare spending</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssn</td>
<td>race</td>
</tr>
<tr>
<td>000-00-0001</td>
<td>white</td>
</tr>
<tr>
<td>000-00-0002</td>
<td>black</td>
</tr>
<tr>
<td>000-00-0003</td>
<td>white</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ssn</th>
<th>spent</th>
<th>race</th>
</tr>
</thead>
<tbody>
<tr>
<td>000-00-0001</td>
<td>10,000$</td>
<td>white</td>
</tr>
<tr>
<td>000-00-0003</td>
<td>8,000$</td>
<td>white</td>
</tr>
</tbody>
</table>
Data distribution debugging: mlinspect

Python script for preprocessing, written exclusively with native pandas and sklearn constructs

```python
# load input data sources, join to single table
patients = pandas.read_csv(...) histories = pandas.read_csv(...) data = pandas.merge([patients, histories], on=['ssn'])

# compute mean complications per age group, append as column
complications = data.groupby('age_group').agg(mean_complications=('complications', 'mean'))
data = data.merge(complications, on=['age_group'])

# Target variable: people with frequent complications
data['label'] = data['complications'] > 1.2 * data['mean_complications']

# Project data to subset of attributes, filter by counties
data = data[['smoker', 'last_name', 'county', 'num_children', 'race', 'income', 'label']]
data = data[data['county'].isin(counties_of_interest)]

# Define a nested feature encoding pipeline for the data
impute_and_encode = sklearn.Pipeline([sklearn.SimpleImputer(strategy='most_frequent'), sklearn.OneHotEncoder()])
featurisation = sklearn.ColumnTransformer(transformers=[[impute_and_encode, ['smoker', 'county', 'race']], Word2VecTransformer(), ['last_name']])(sklearn.StandardScaler(), ['num_children', 'income'])

# Define the training pipeline for the model
neural_net = sklearn.KerasClassifier(build_fn=create_model())
pipeline = sklearn.Pipeline([['features', featurisation], ['learning_algorithm', neural_net]])

# Train-test split, model training and evaluation
train_data, test_data = train_test_split(data)
model = pipeline.fit(train_data, train_data.label)
print(model.score(test_data, test_data.label))
```

Declarative inspection of preprocessing pipeline

mlinspect
.PipelineInspector.on_pipeline('health.py').no_bias_introduced_for(["age_group", "race"]).no_illegal_features().no_missing_embeddings().verify()
Automated Data Cleaning Can Hurt Fairness in ML-based Decision Making

<table>
<thead>
<tr>
<th>Model</th>
<th>Fairness Worse</th>
<th>Auto-cleaning Makes Fairness Better</th>
<th>Fairness &amp; Accuracy Better</th>
</tr>
</thead>
<tbody>
<tr>
<td>xgboost</td>
<td>21.2% (45)</td>
<td>10.8% (23)</td>
<td>6.6% (14)</td>
</tr>
<tr>
<td>knn</td>
<td>24.5% (52)</td>
<td>13.7% (29)</td>
<td>11.8% (25)</td>
</tr>
<tr>
<td>log-reg</td>
<td>19.8% (42)</td>
<td>12.3% (26)</td>
<td>7.5% (16)</td>
</tr>
</tbody>
</table>

TABLE V

Impact of auto-cleaning on accuracy and fairness for different ML models on 212 configurations in total. We list cases where fairness gets worse, fairness gets better, and where both fairness and accuracy get better. Auto-cleaning is more likely to worsen than to improve fairness across all models.
emergent bias
Emergent bias arises in the context of use of a technical system
Hiring and AI: Let Job Candidates Know Why They Were Rejected

Artificial-intelligence tools are seeing ever broader use in hiring. But this practice is also hotly criticized because we rarely understand how these tools select candidates, and whether the candidates they select are, in fact, better qualified than those who are rejected.

To help answer these crucial questions, we should give job seekers more information about the hiring process and the decisions. The solution I propose is a twist on something we see every day: nutritional labels. Specifically, job candidates would see simple, standardized labels that show the factors that go into the AI’s decision.

https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313
Anatomy of a job posting label

Qualifications:
- Knowledge of financial systems
- Team player
- BS in Accounting
- GPA > 3

Data:
- Resume
- LinkedIn profile
- Credit score
- Other social media (optional)

Assessment:
- AI-assisted personality prediction
- Personal interview (accommodations upon request)

https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313
wrapping up
We are AI comics

dataresponsibly.github.io/we-are-ai/comics
We are AI comics: in Spanish

dataresponsibly.github.io/we-are-ai/comics
Thank you!

Julia Stoyanovich  
New York University  
USA

Serge Abiteboul  
Inria & ENS Paris  
France

Bill Howe  
University of Washington  
USA

H.V. Jagadish  
University of Michigan  
USA

Sebastian Schelter  
University of Amsterdam  
The Netherlands

Research supported in part by NSF Grants No. 1934464, 1934565, 1934405, 1926250, 1741022, 1740996, 1916505, by Microsoft, and by Ahold Delhaize