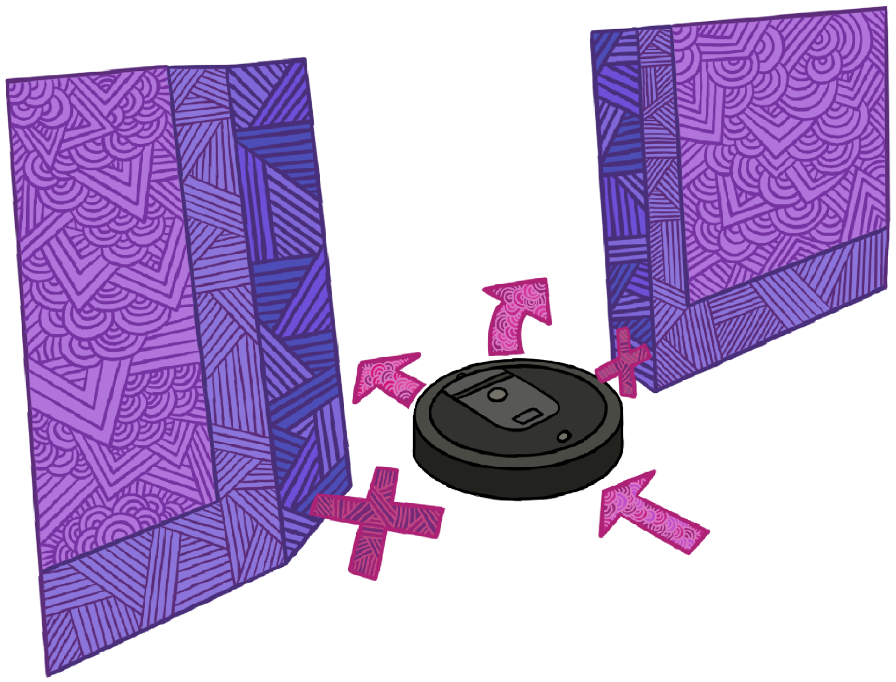
Responsible Data Management

Julia Stoyanovich

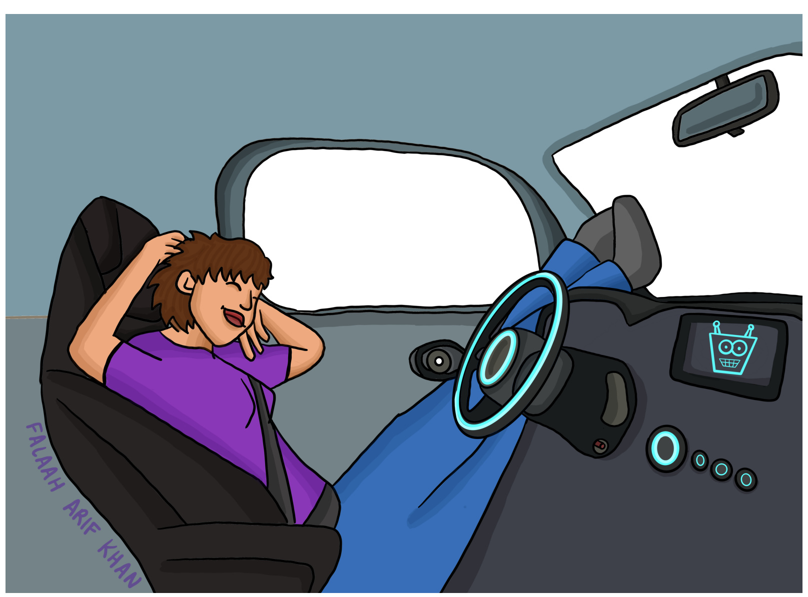
Computer Science and Engineering
Center for Data Science
Center for Responsible AI
Visualization and Data Analytics Center
New York University

AI: algorithms, data, decisions



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

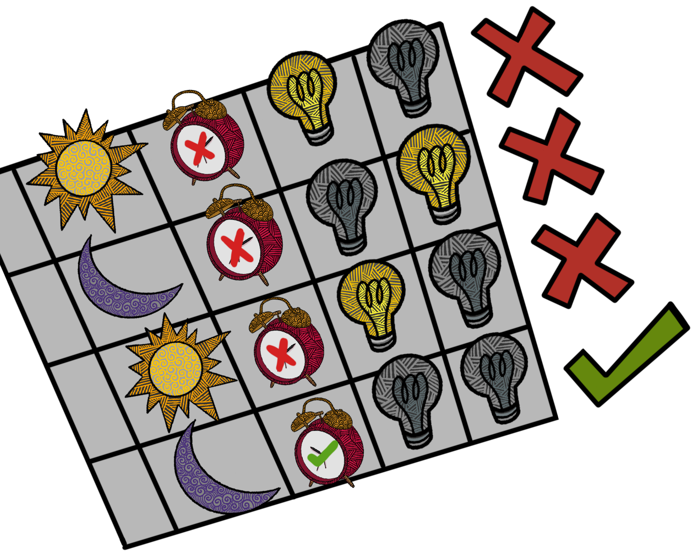
accelerate science

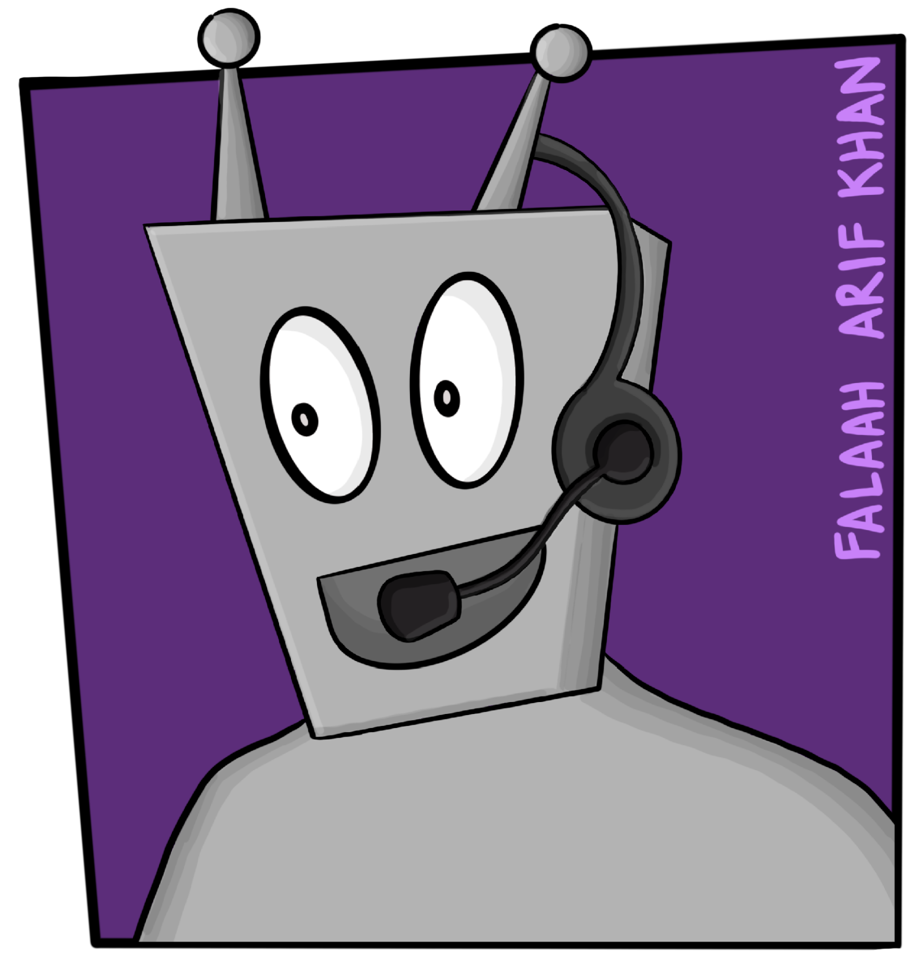
boost innovation

transform government

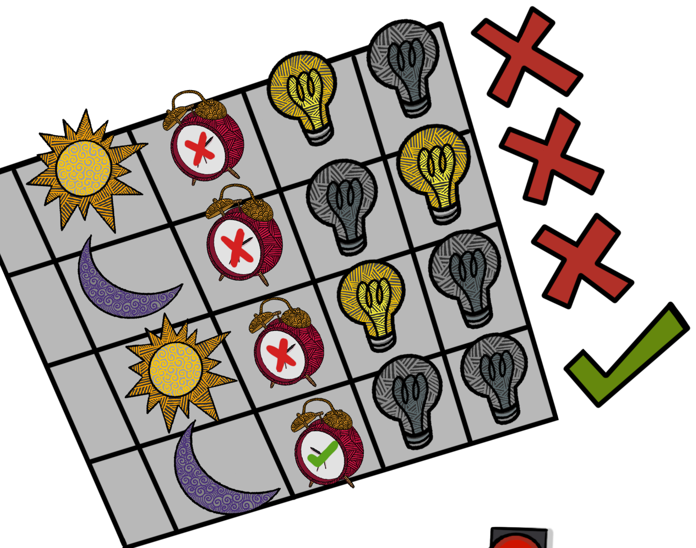
r/ai center for responsible ai

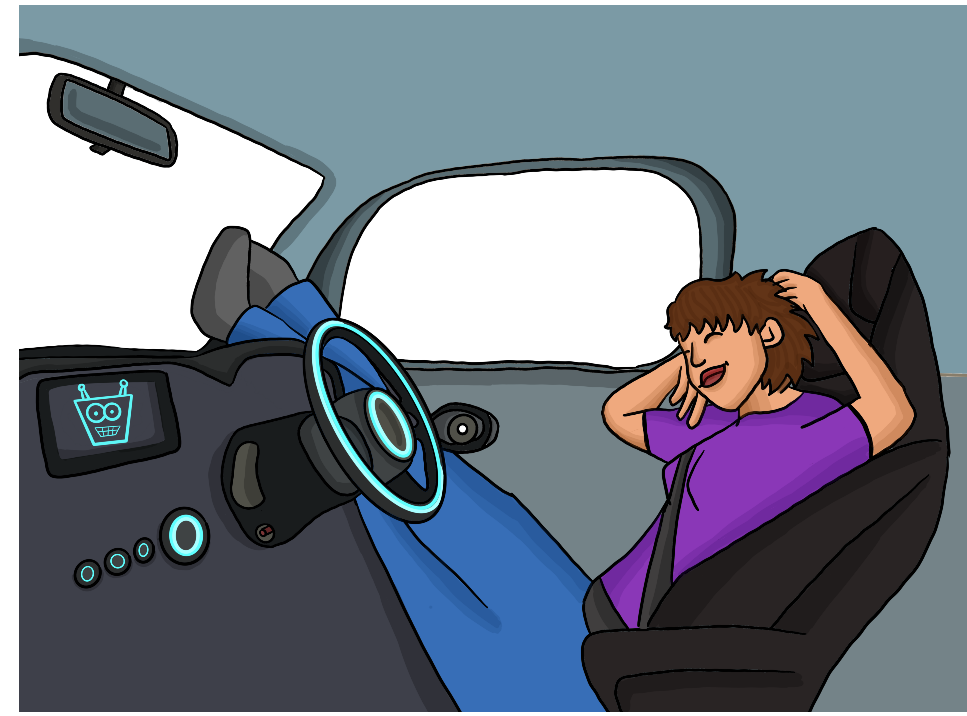
Machines make mistakes



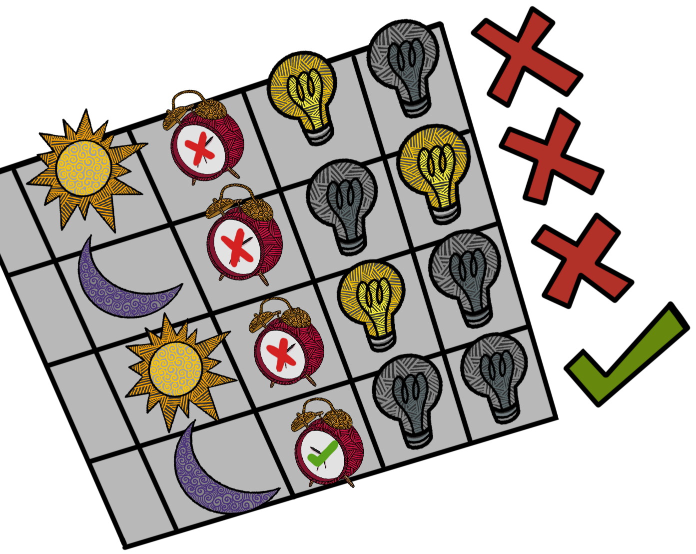


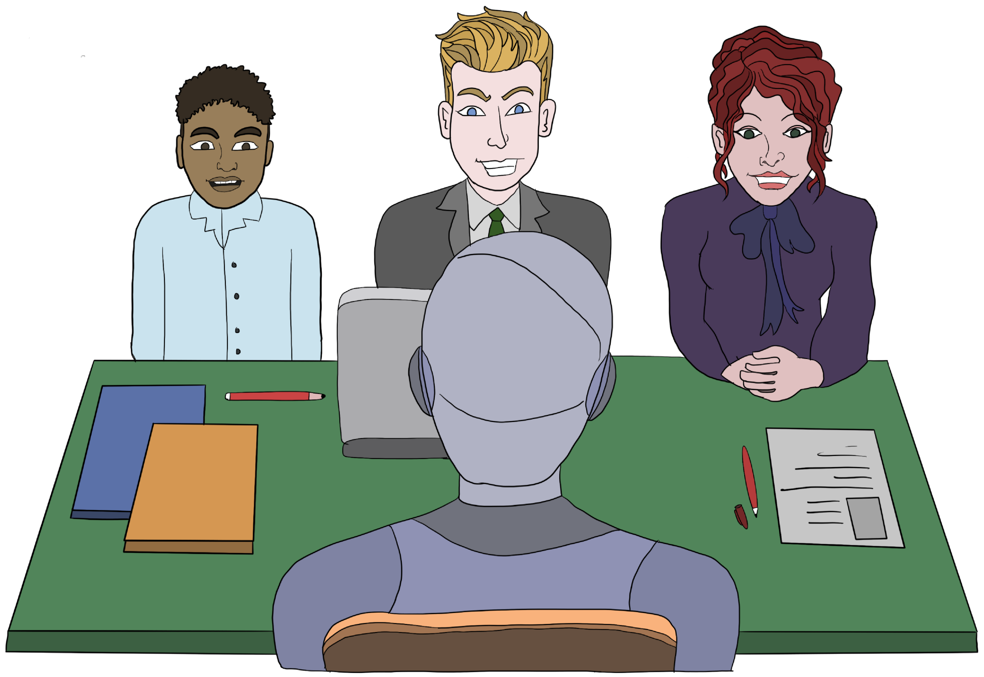
Mistakes lead to harms

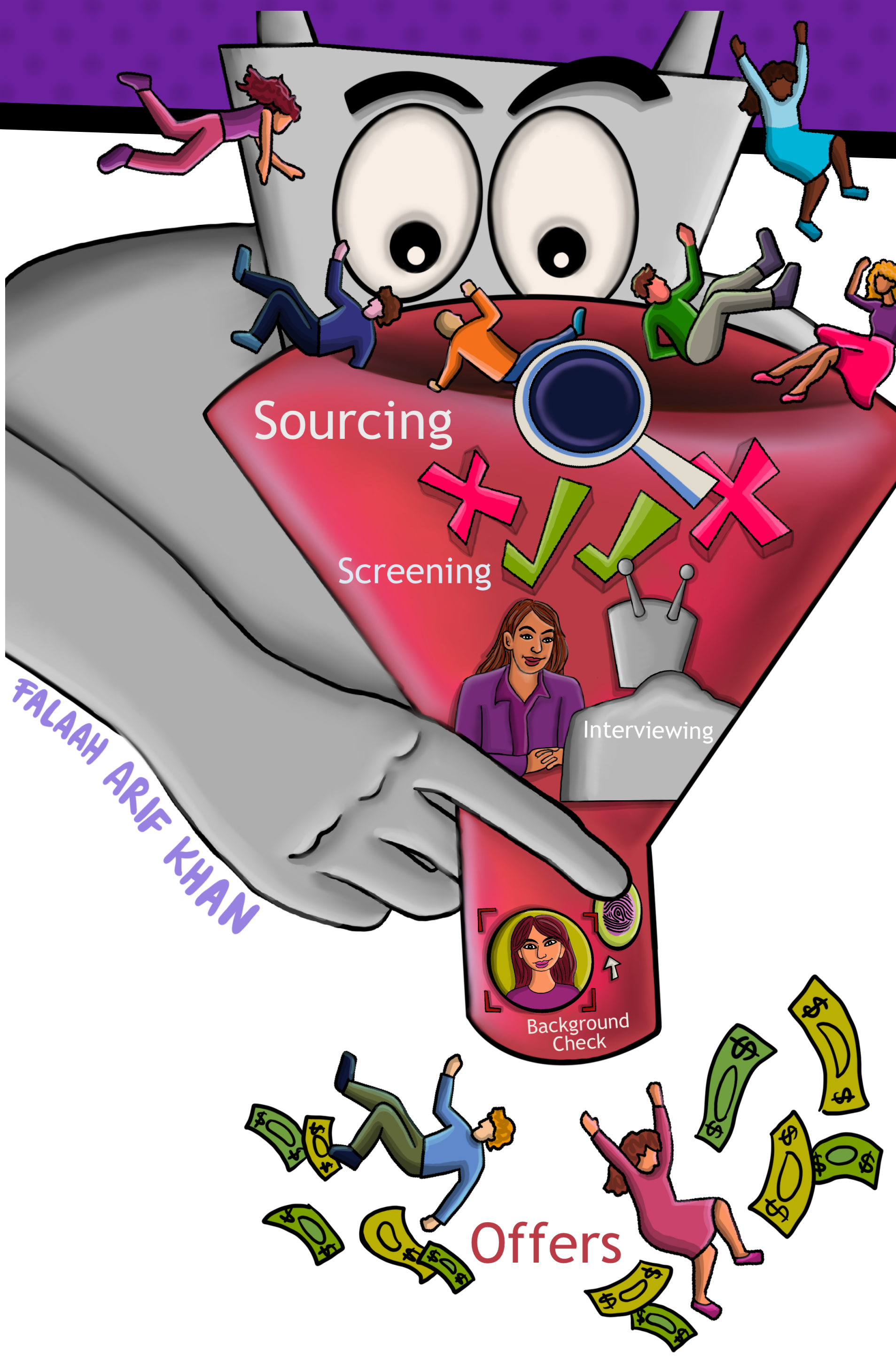




Harms can be cumulative







center for responsible ai

Racial bias in resume screening

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

September 2004

Marianne Bertrand

Sendhil Mullainathan

AMERICAN ECONOMIC REVIEW
VOL. 94, NO. 4, SEPTEMBER 2004
(pp. 991-1013)

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

theguardian **July 2015**

Women less likely to be shown ads for high-paid jobs on Google, study shows

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.

REUTERS **October 2018**

Amazon scraps secret AI recruiting tool that showed bias against women

MIT Technology Review **February 2013**

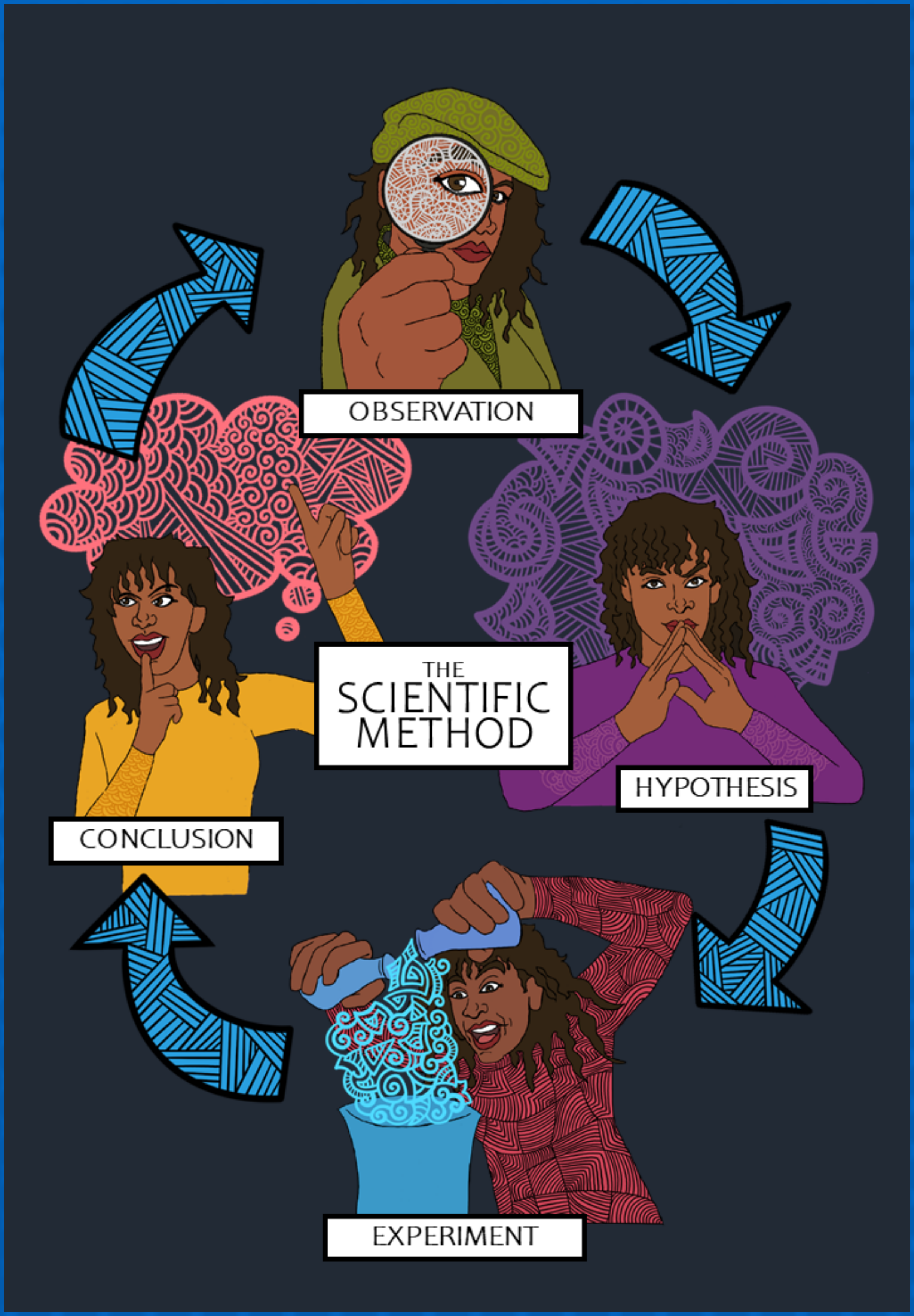
Racism is Poisoning Online Ad Delivery, Says Harvard Professor

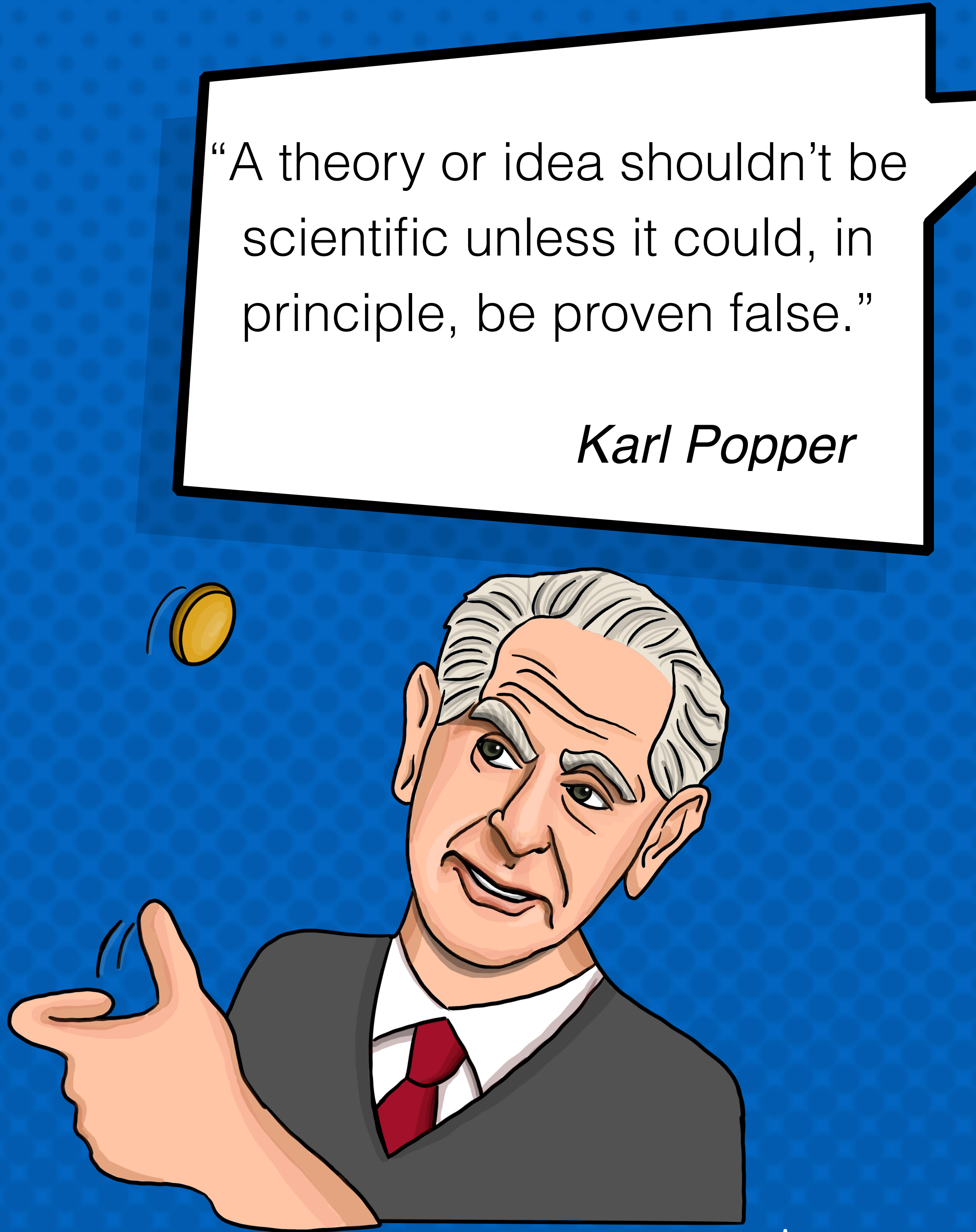
THE WALL STREET JOURNAL. **September 2014**

Are Workplace Personality Tests Fair?

Growing Use of Tests Sparks Scrutiny Amid Questions of Effectiveness and Workplace Discrimination

Do these tools actually work?





r/ai center for responsible ai

a push for regulation

center for responsible ai

Automated Decision Systems (ADS)

Automated Decision Systems (ADS)

process data about people

help make consequential decisions

combine human & automated decision making

aim to improve **efficiency** and promote **equity**

New York City Local Law 144 of 2021

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

December 11, 2021

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?

center for responsible ai

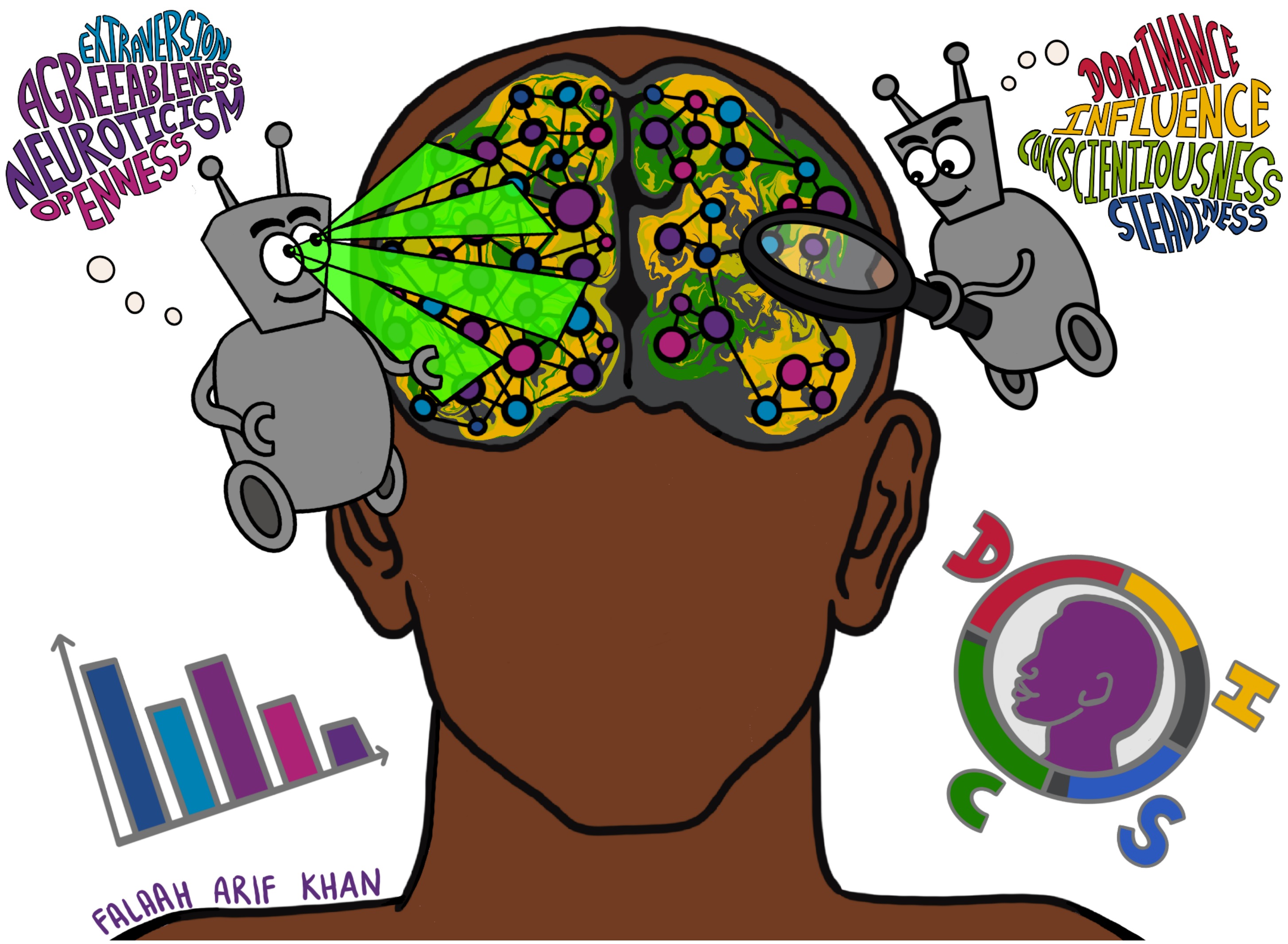
Personality prediction in hiring

center for responsible ai

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)



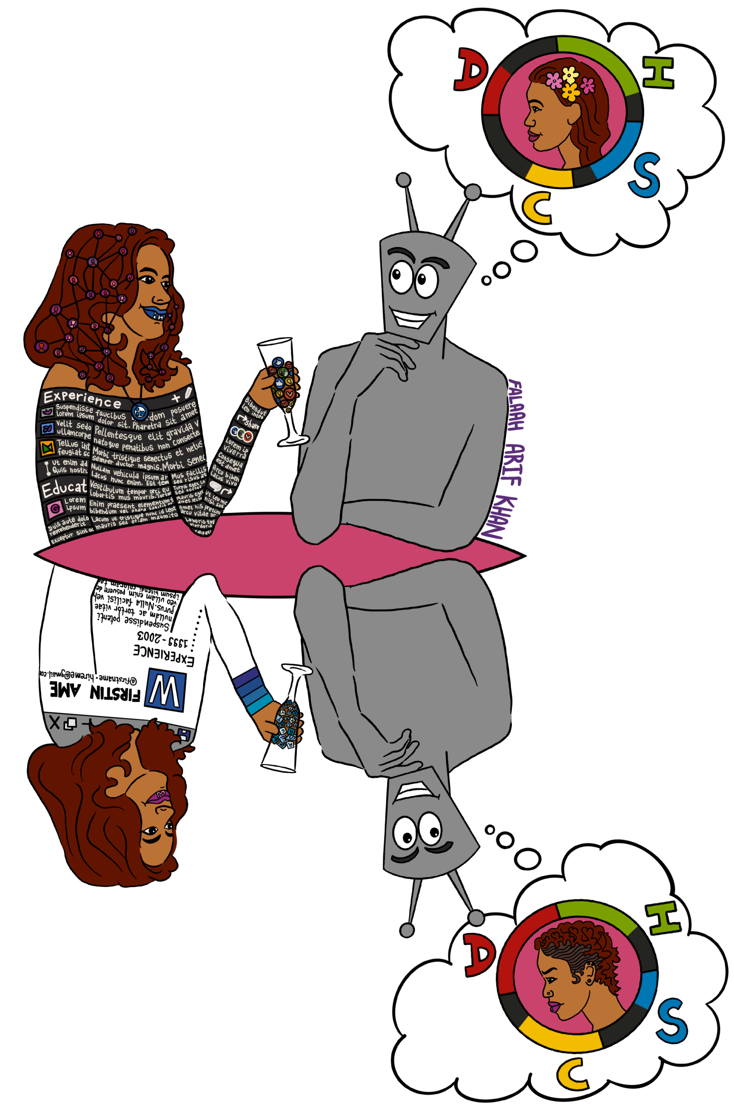
Stability audit framework

Data Mining and Knowledge Discovery (2022) 36:2153–2193
https://doi.org/10.1007/s10618-022-00861-0

An external stability audit framework to test the validity of personality prediction in AI hiring

Alene K. Rhea[1,2] · Kelsey Markey[1,2] · Lauren D'Arinzo[1,2,3] · Hilke Schellmann[4] · Mona Sloane[2] · Paul Squires[5] · Falaah Arif Khan[1,2] · Julia Stoyanovich[1,2,6]

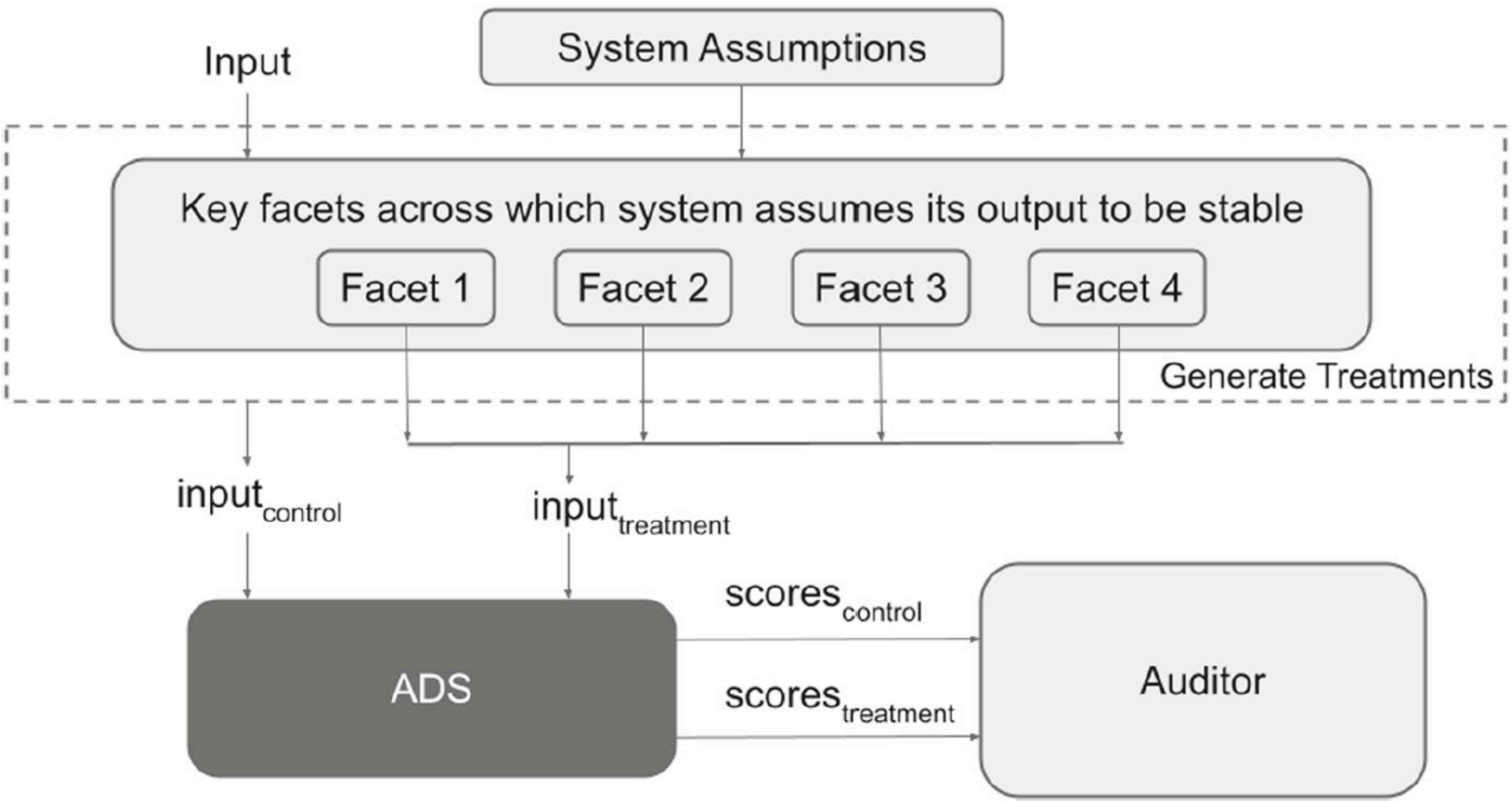
Received: 6 October 2021 / Accepted: 5 August 2022 / Published online: 17 September 2022

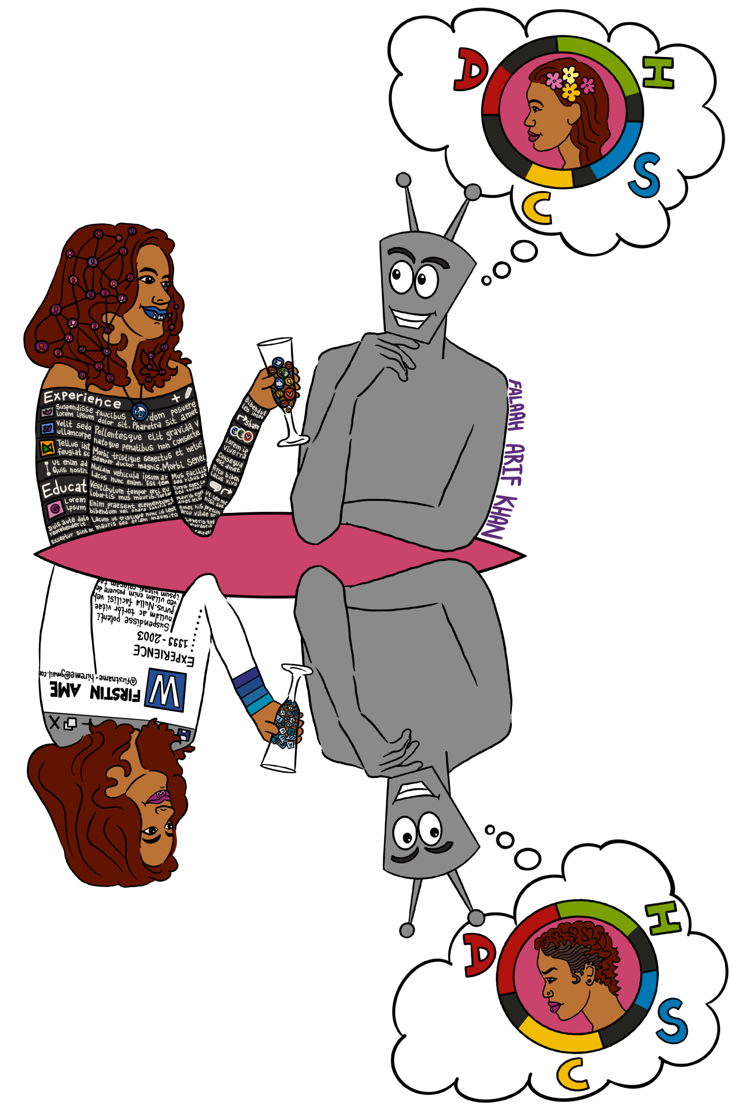


https://link.springer.com/article/10.1007/s10618-022-00861-0

center for responsible ai

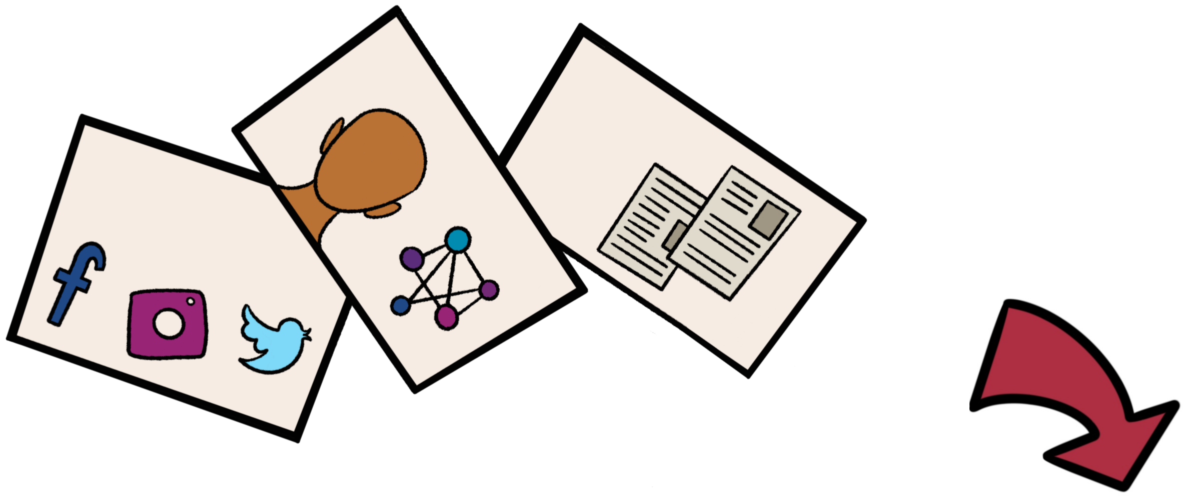
Stability audit framework



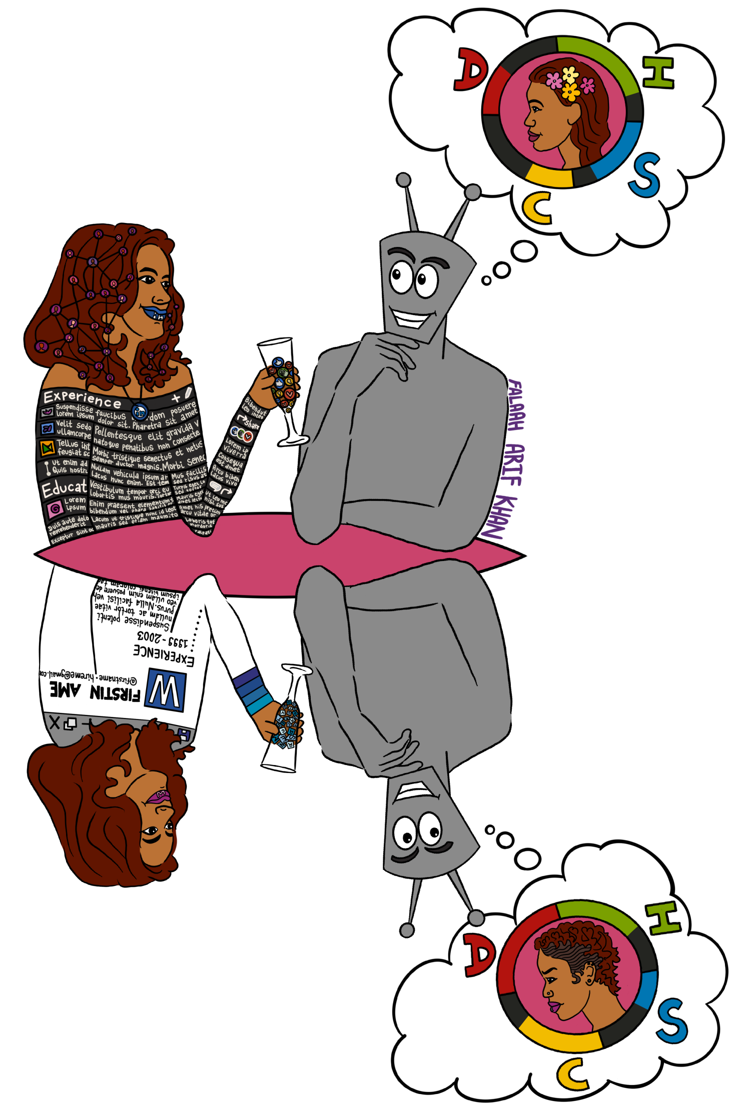


https://link.springer.com/article/10.1007/s10618-022-00861-0

Stability audit framework



Facet	Crystal	Humantic
Resume file format	✗	✓
LinkedIn URL in resume	?	✗
Source context	✗	✗
Algorithm-time / immediate	✓	✓
Algorithm-time / 31 days	✓	✗
Participant-time / LinkedIn	✗	✗
Participant-time / Twitter	N/A	✓



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

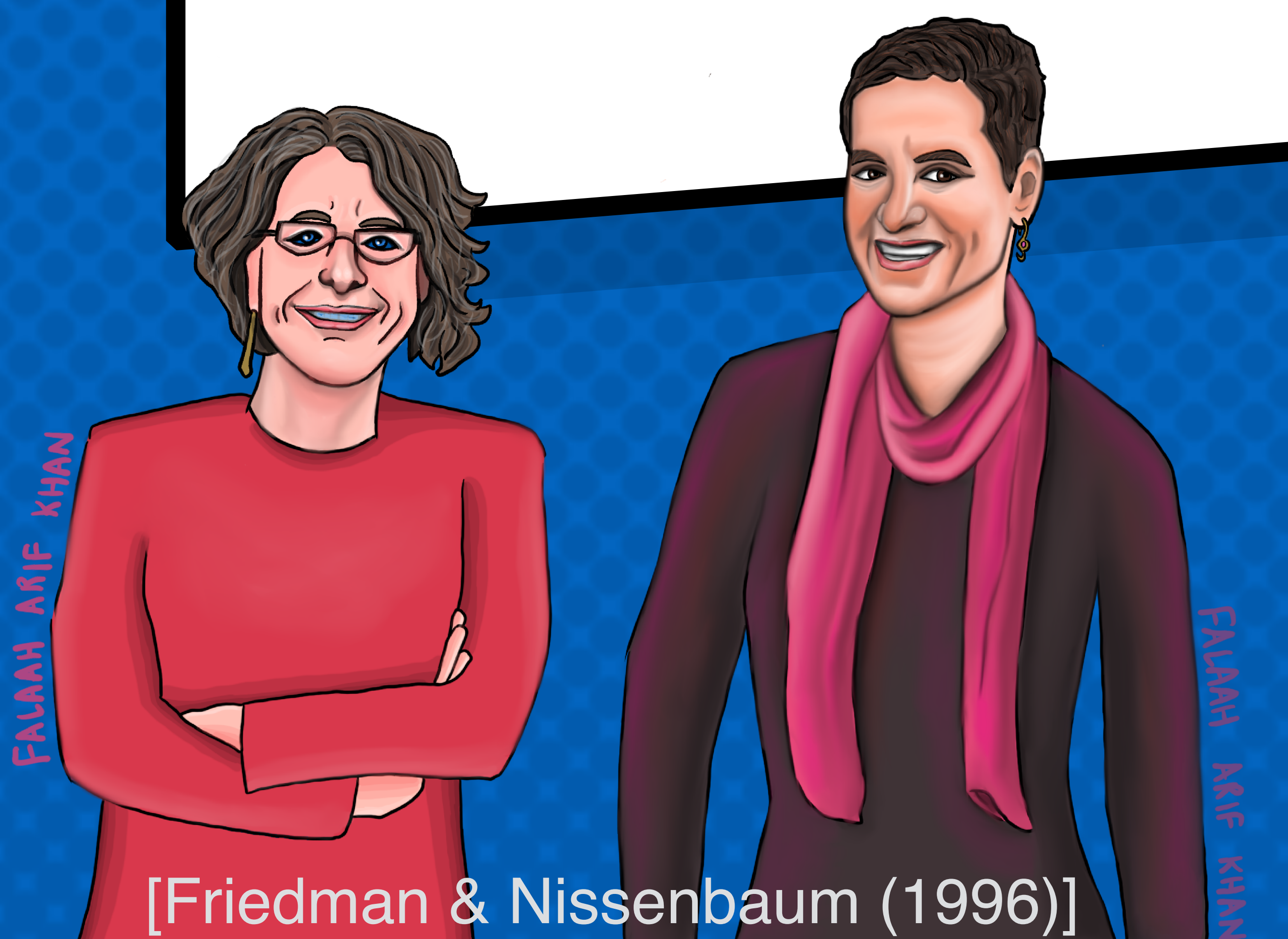
r/ai center for responsible ai

Bias in computer systems

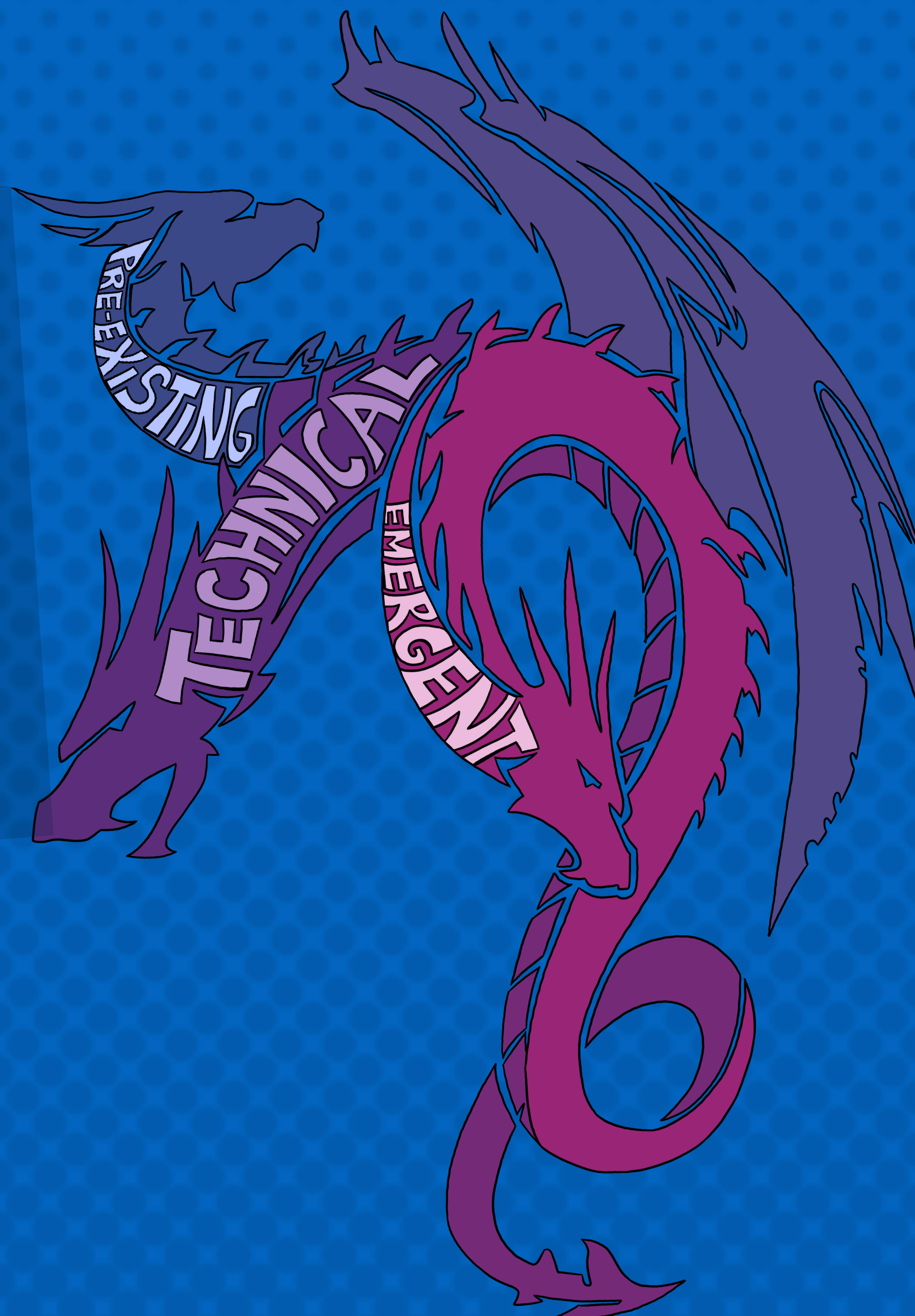
Pre-existing: exists independently of algorithm, has origins in society

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

Emergent: arises due to context of use



[Friedman & Nissenbaum (1996)]

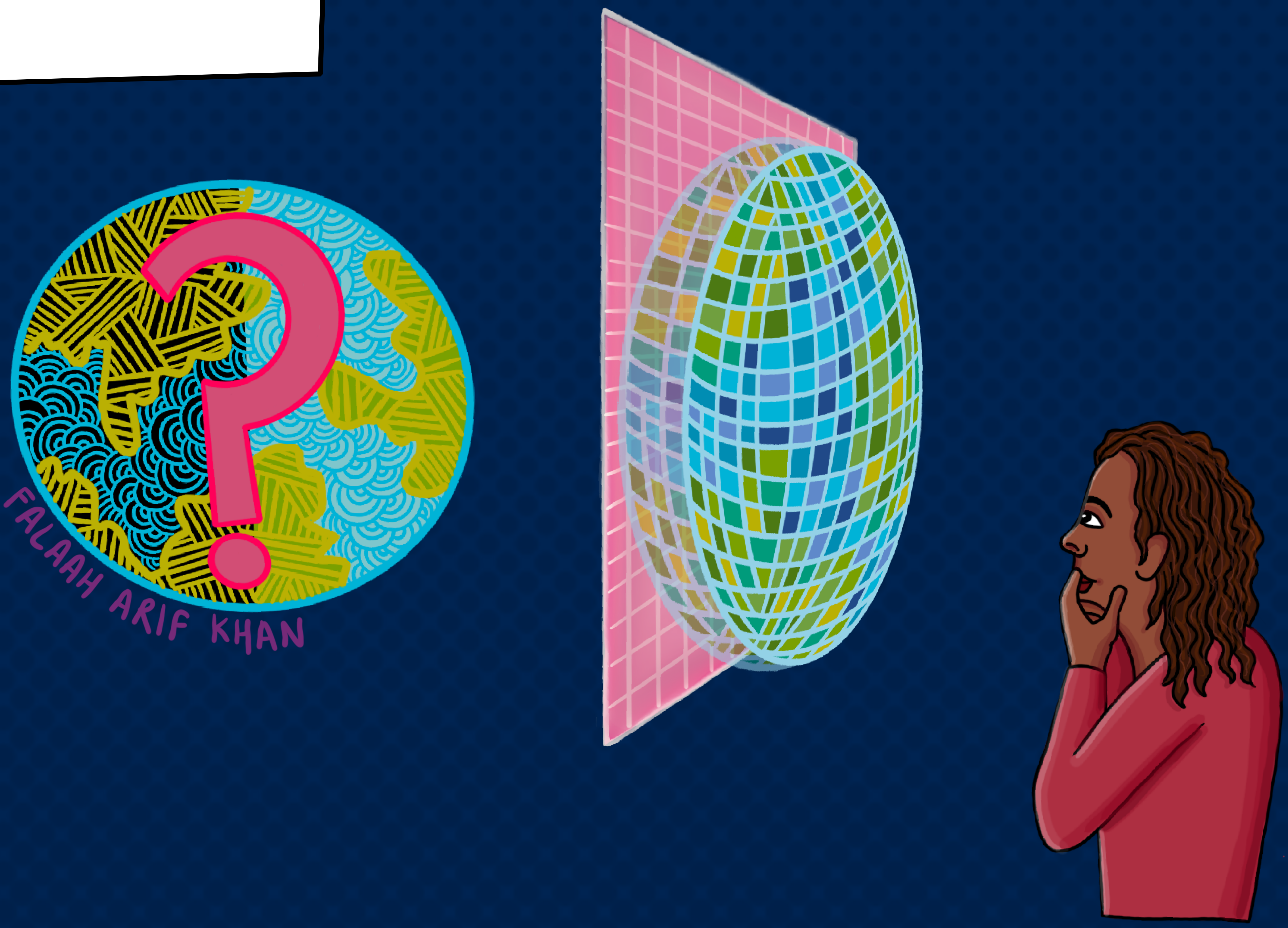


r/ai center for responsible ai

pre-existing bias

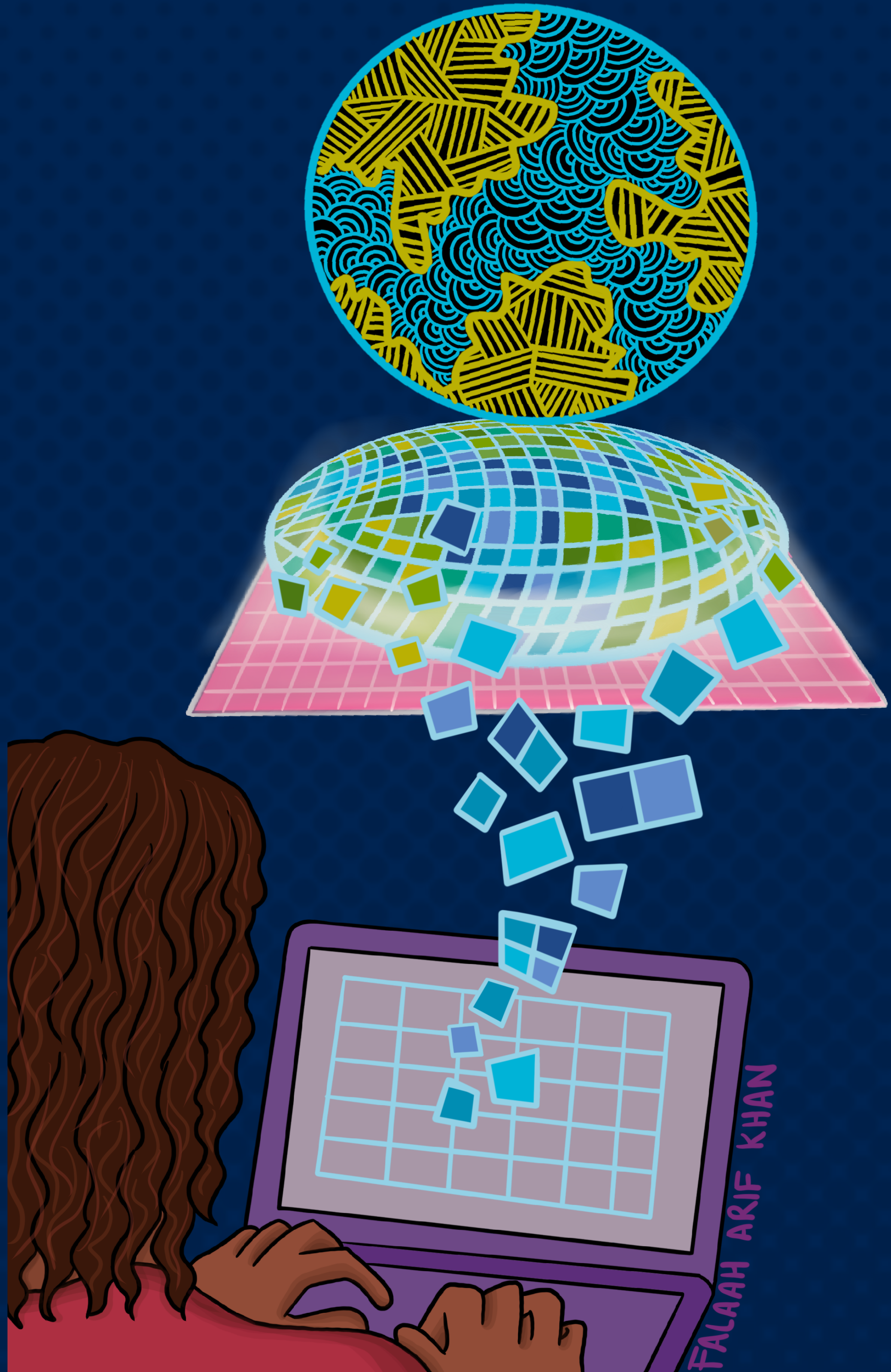
r/ai center for responsible ai

Pre-existing bias has origins in society



r/ai center for responsible ai

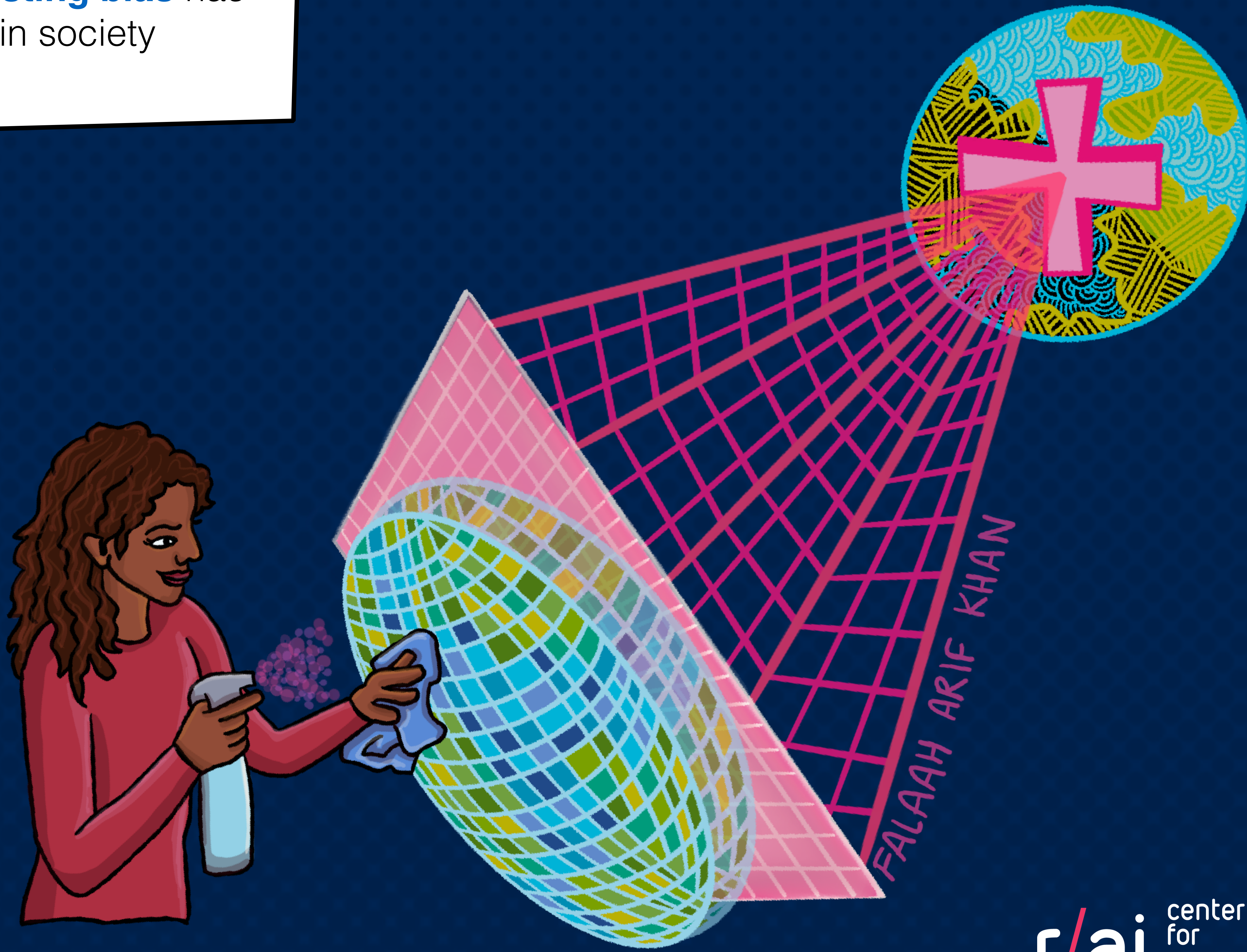
Pre-existing bias has origins in society



r/ai center for responsible ai

Pre-existing bias has origins in society

Pre-existing bias has origins in society



r/ai center for responsible ai

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

score = 373

	Male		Female	
White	A (99)	B (98)	C (96)	D (95)
Black	E (91)	F (91)	G (90)	H (89)
Asian	I (87)	J (87)	K (86)	L (83)

score = 372

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**

[Yang, Gkatzelis, Stoyanovich (2019)]

From beliefs to interventions

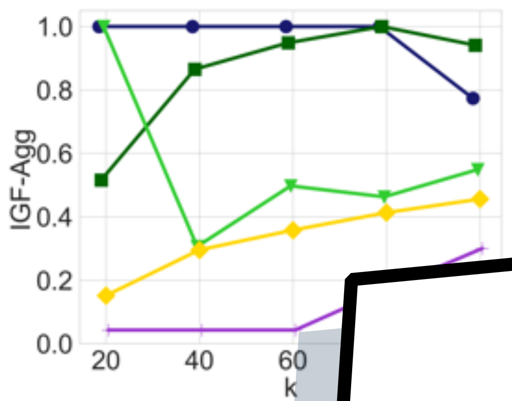
Fairness for female candidates **83 / 95 = 0.91**

C	D	G	H	K	L
95	95	90	86	83	83

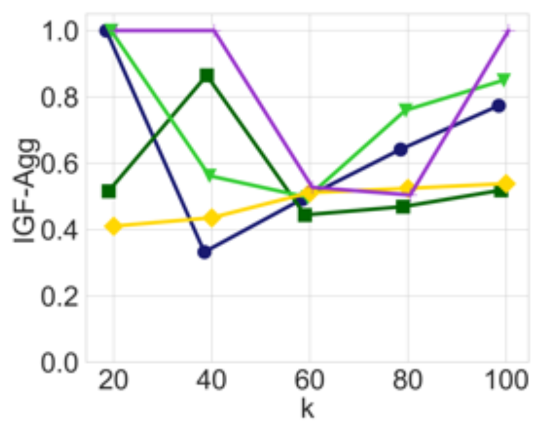
highest-scoring skipped

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)]

r/ai center for responsible ai

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 rankin
ranking methods.

CCS Concepts: • **Information systen**
technology policy.

Additional Key Words and Phrases: fairı

ACM Reference Format:
Meike Zehlike, Ke Yang, and Julia Stoy
Article 1 (January 2022), 35 pages. https

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

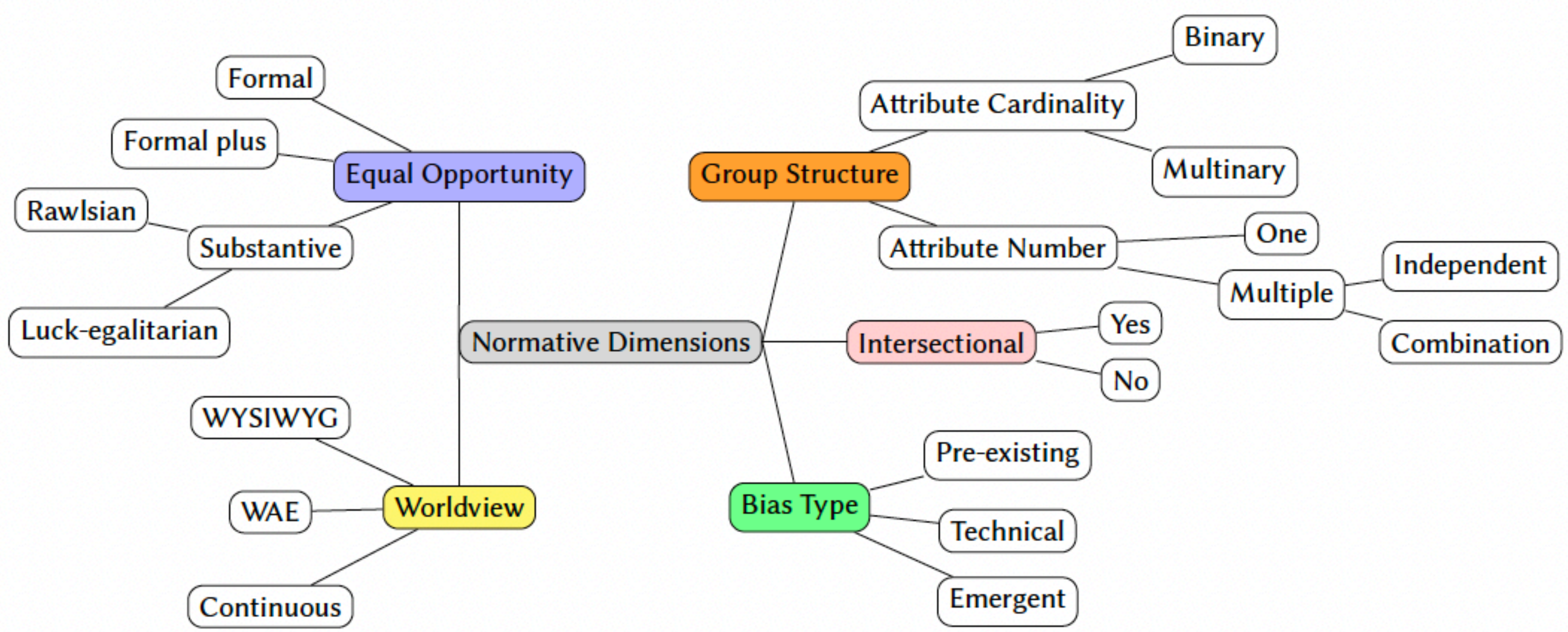
MEIKE ZEHLIKE, Humboldt University of Berlin, Max Planck Institute for Software Systems, and Zalando Research, Germany
KE YANG, New York University, NY, and University of Massachusetts, Amherst, MA, USA
JULIA STOYANOVICH, New York University, NY, USA

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 fair
air ranking methods.

l **topics** → **Computing** /

commender Systems. *ACM*

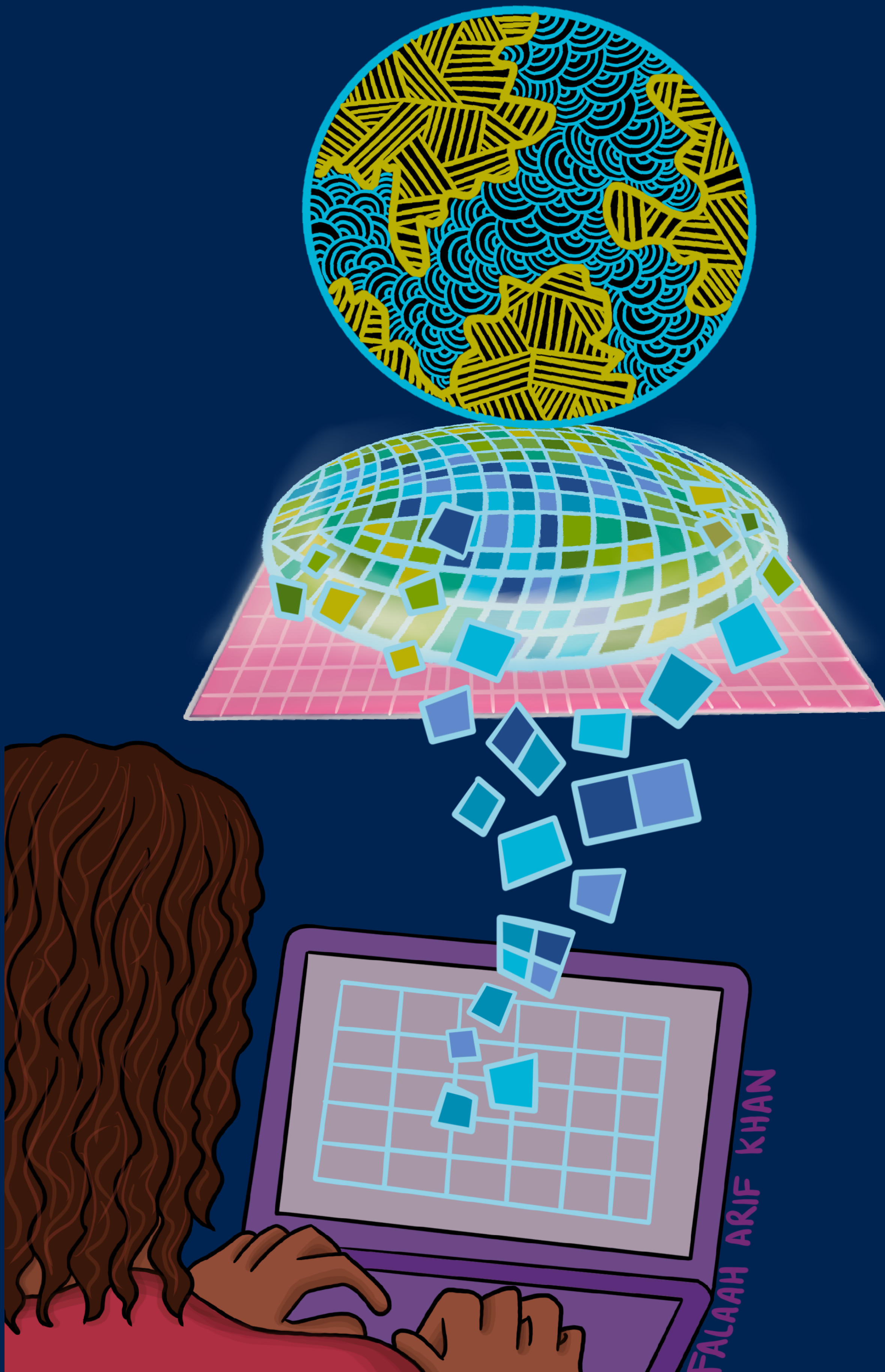


[Zehlike, Yang, Stoyanovich (2022)]

center for responsible ai

technical bias

r/ai center for responsible ai



Technical bias may be introduced or exacerbated by the technical properties of an ADS

r/ai center for responsible ai

Model development lifecycle

Goal

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

Problem

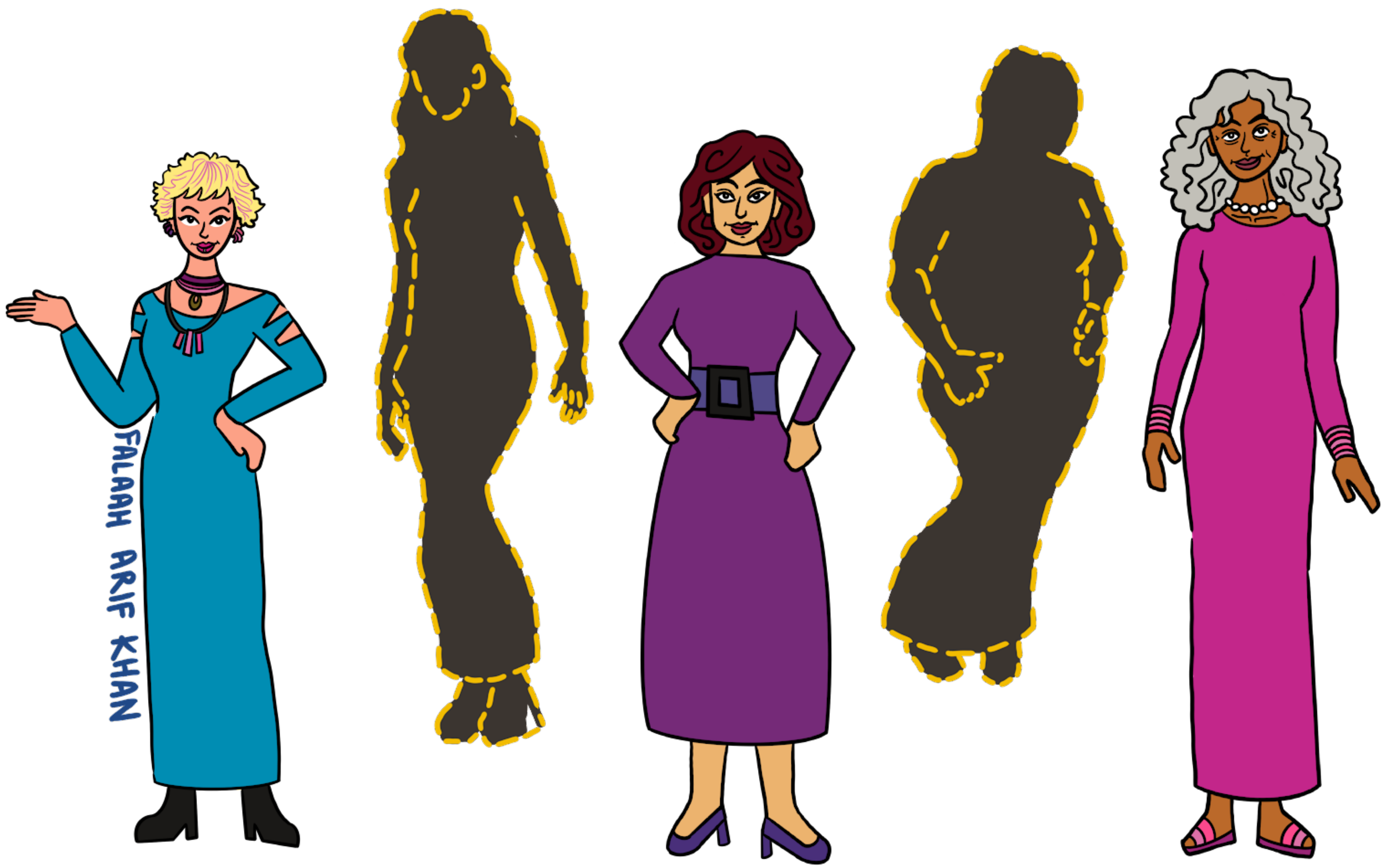
accuracy is lower for middle-aged women - **a fairness concern!**

now what?

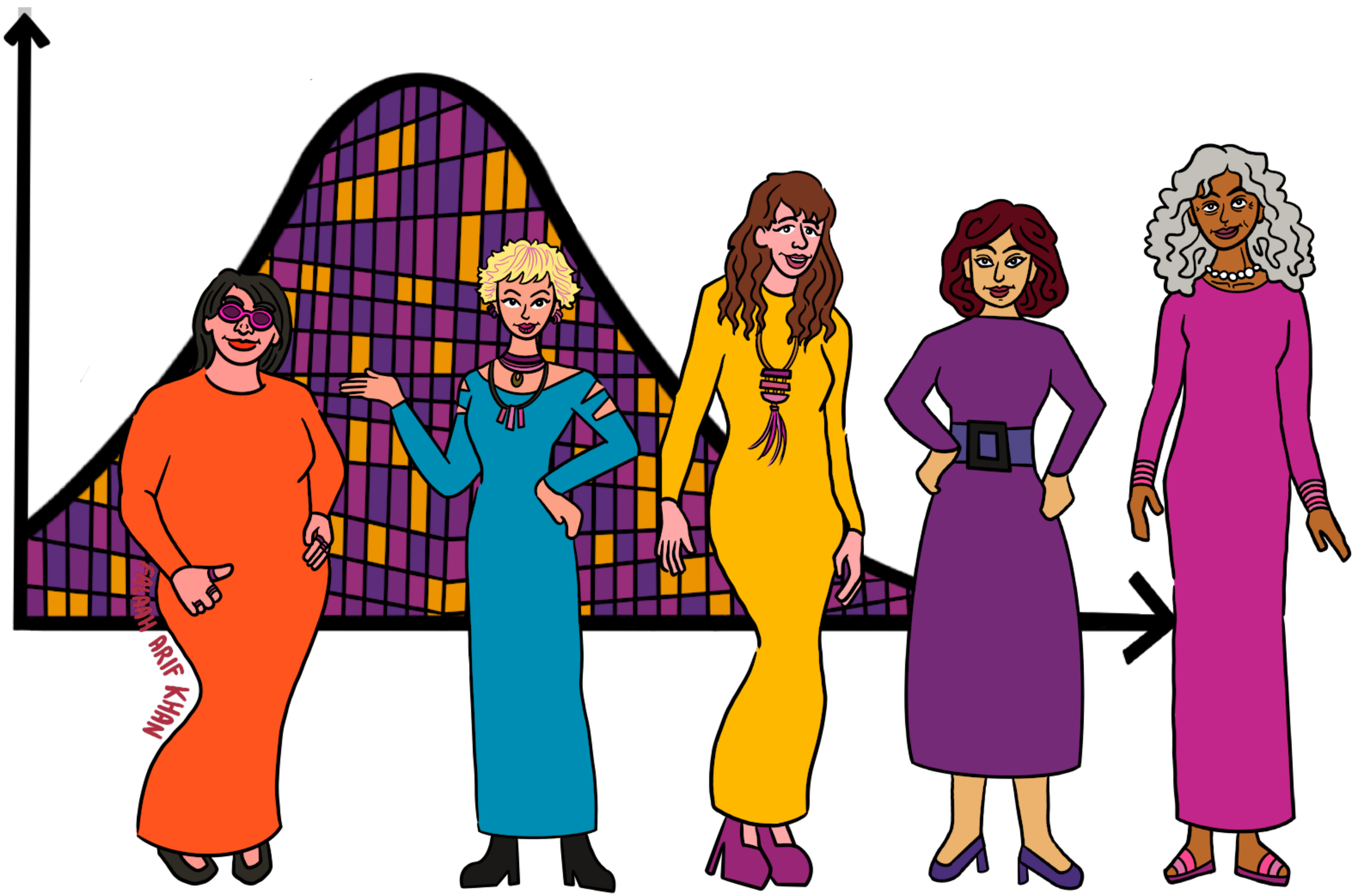
[Schelter, He, Khilnani, Stoyanovich (2020)]

r/ai center for responsible ai

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?

age_group	county
60	CountyA
60	CountyA
20	CountyA
60	CountyB
20	CountyB
20	CountyB

50% vs 50%

→

age_group	county
60	CountyA
60	CountyA
20	CountyA

66% vs 33%

Data filtering

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

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

patients

ssn	race
000-00-0001	white
000-00-0002	black
000-00-0003	white

healthcare spending

ssn	spent
000-00-0001	10,000$
000-00-0003	8,000$

Join
on *ssn*

result

ssn	spent	race
000-00-0001	10,000$	white
000-00-0003	8,000$	white

Data distribution debugging: mlinspect

Potential issues in preprocessing pipeline:

1. Join might change proportions of groups in data
2. Column 'age_group' projected out, but required for fairness
3. Selection might change proportions of groups in data
4. Imputation might change proportions of groups in data
5. 'race' as a feature might be illegal!
6. Embedding vectors may not be available for rare names!

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

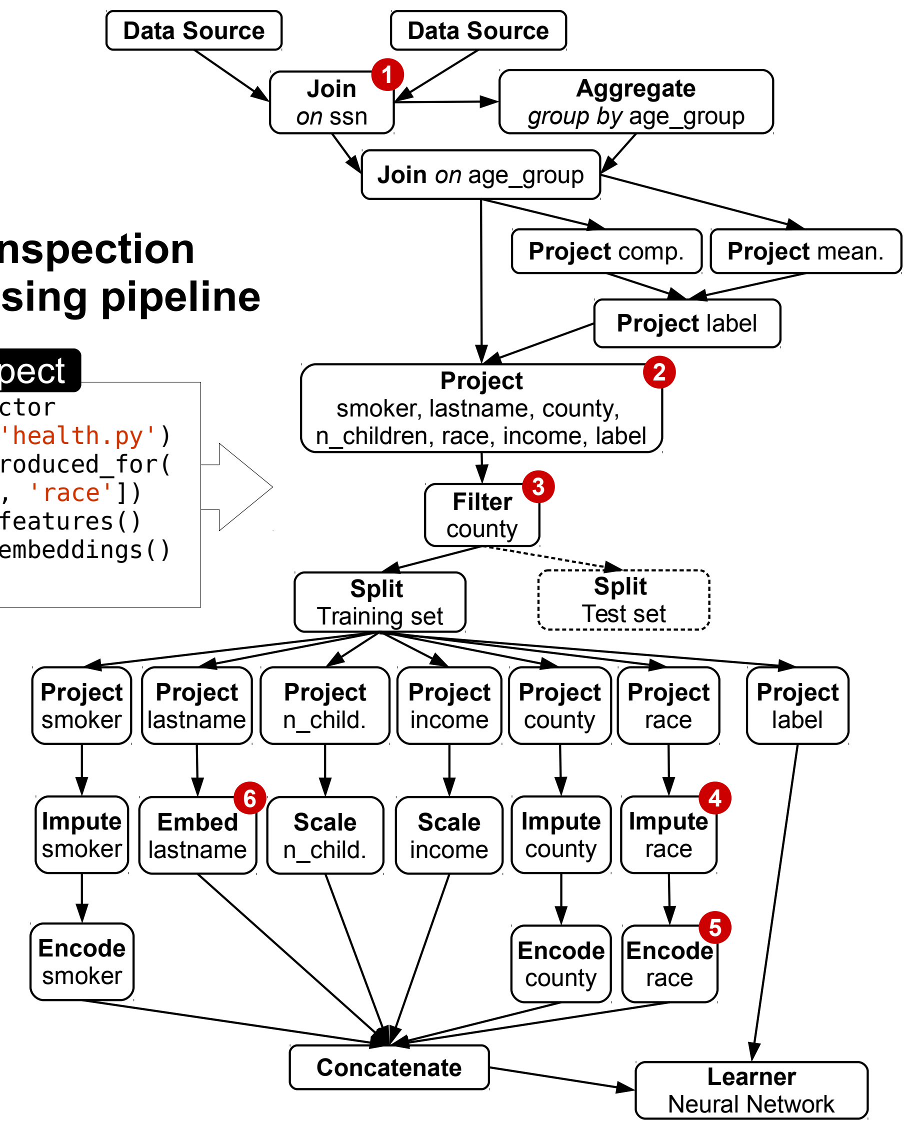
```
# load input data sources, join to single table
patients = pandas.read_csv(…)
histories = pandas.read_csv(…)
data = pandas.merge([patients, histories], on=['ssn'])

# 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()
```

Corresponding dataflow DAG for instrumentation, extracted by *mlinspect*



[Grafberger, Stoyanovich, Schelter (2021)]

center for responsible ai

Impact of automated data cleaning

Automated Data Cleaning Can Hurt Fairness in ML-based Decision Making

Shubha Guha
s.guha@uva.nl
University of Amsterdam

Falaah Arif Khan
fa2161@nyu.edu
New York University

Julia Stoyanovich
stoyanovich@nyu.edu
New York University

Sebastian Schelter
s.schelter@uva.nl
University of Amsterdam

ongoing work

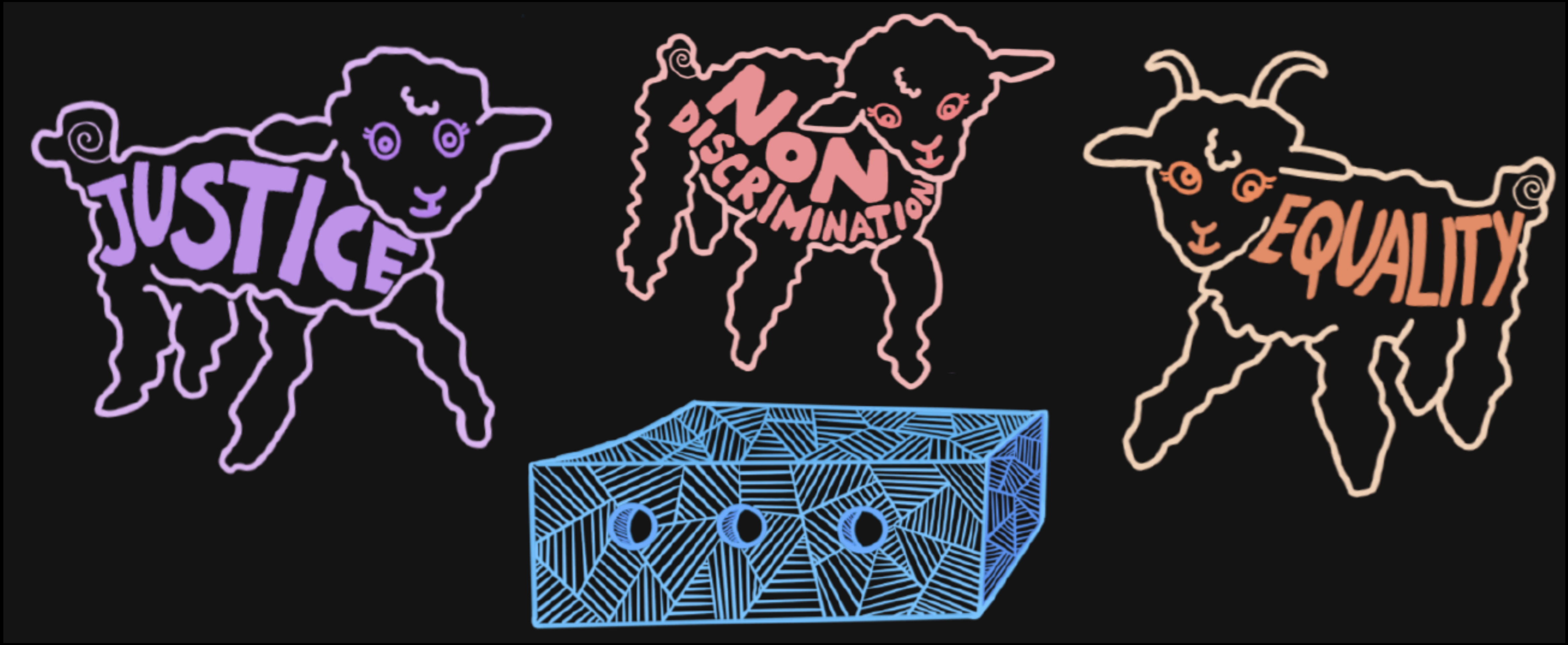
model	auto-cleaning makes fairness worse	auto-cleaning makes fairness better	auto-cleaning makes fairness & accuracy better
`xgboost`	21.2% (45)	10.8% (23)	6.6% (14)
`knn`	24.5% (52)	13.7% (29)	11.8% (25)
`log-reg`	19.8% (42)	12.3% (26)	7.5% (16)

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

center
for
responsible
ai

Emergent bias arises in the context of use of a technical system



center for responsible ai

Nutritional labels for job seekers

THE WALL STREET JOURNAL.

September 22, 2021

Hiring and AI: Let Job Candidates Know Why They Were Rejected

Labels that explain a hiring process that uses AI could allow job seekers to opt out if they object to the employer's data practices.

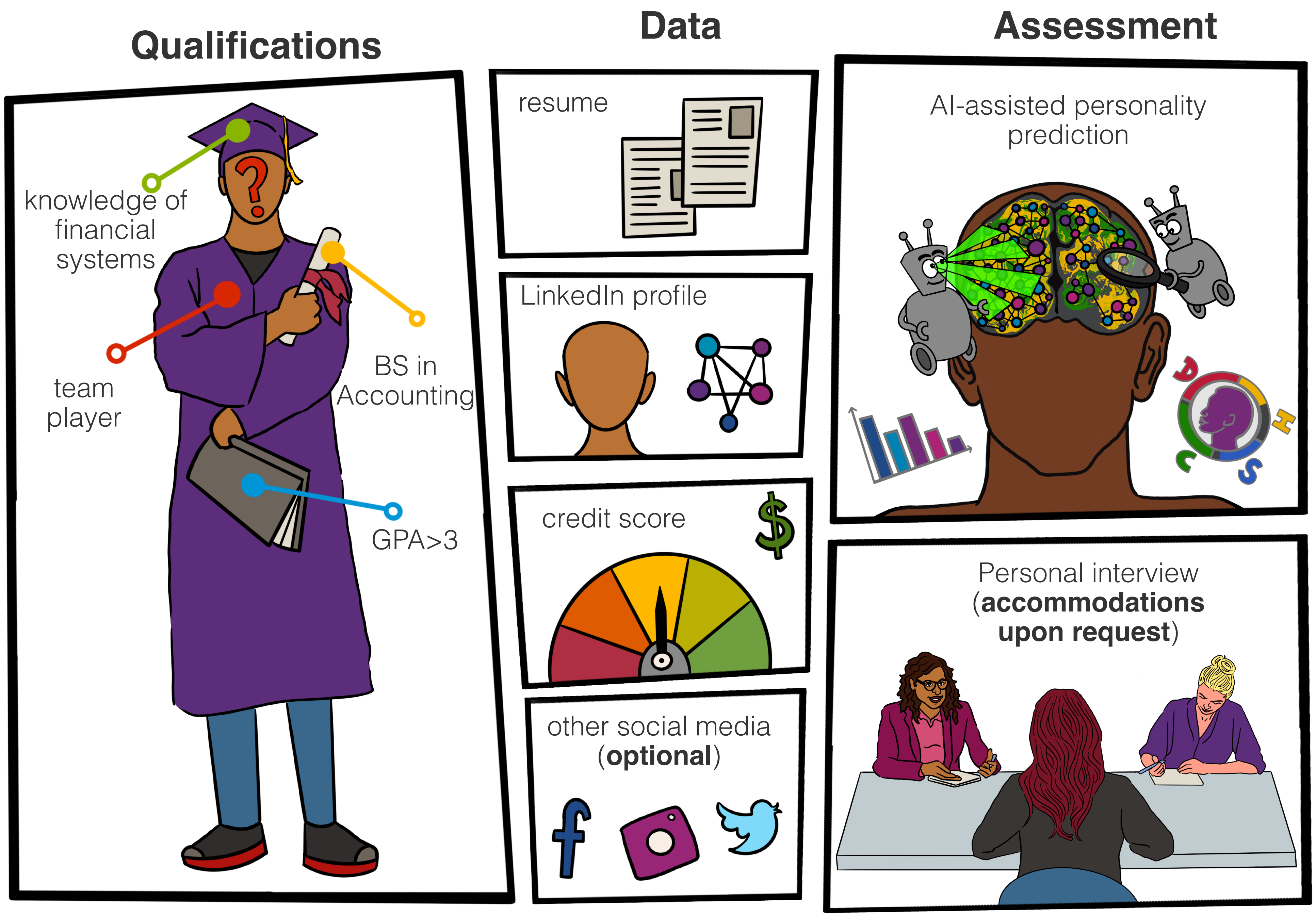
PHOTO: ISTOCKPHOTO/GETTY IMAGES

By Julia Stoyanovich
Updated Sept. 22, 2021 11:00 am ET

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.

Anatomy of a job posting label

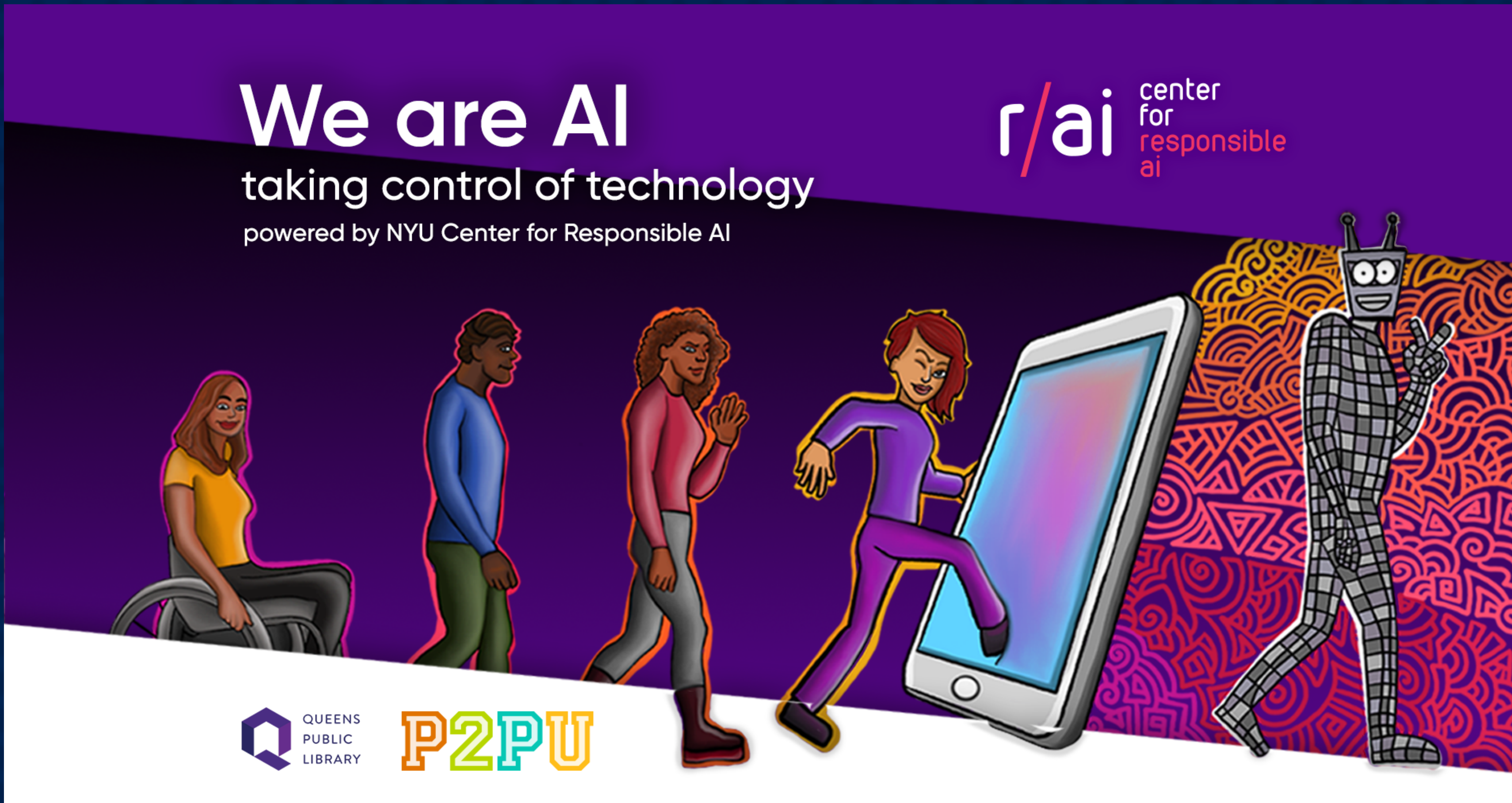


https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313

r/ai center for responsible ai

wrapping up

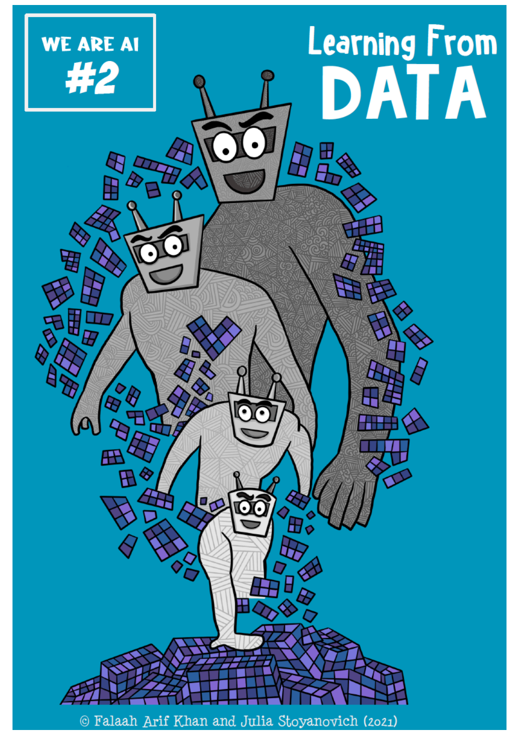
r/ai center for responsible ai

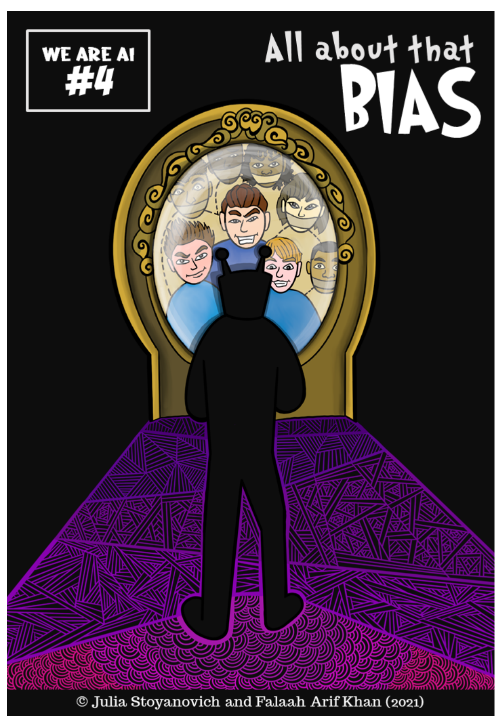


https://dataresponsibly.github.io/we-are-ai/

r/ai center for responsible ai

We are AI comics

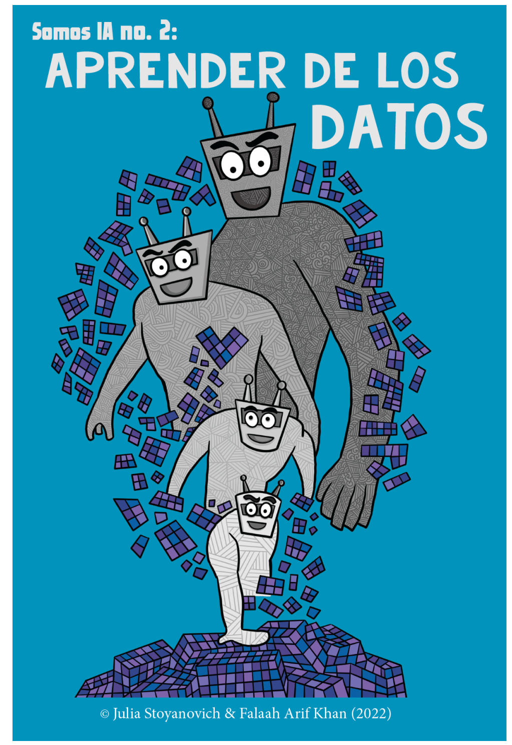


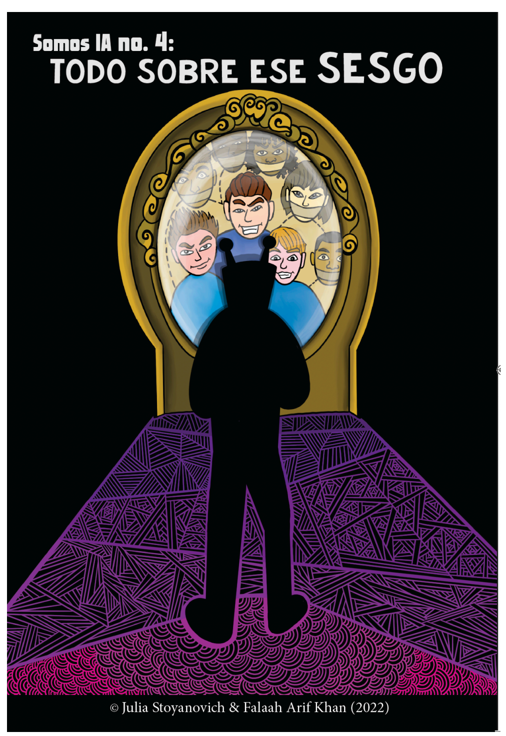


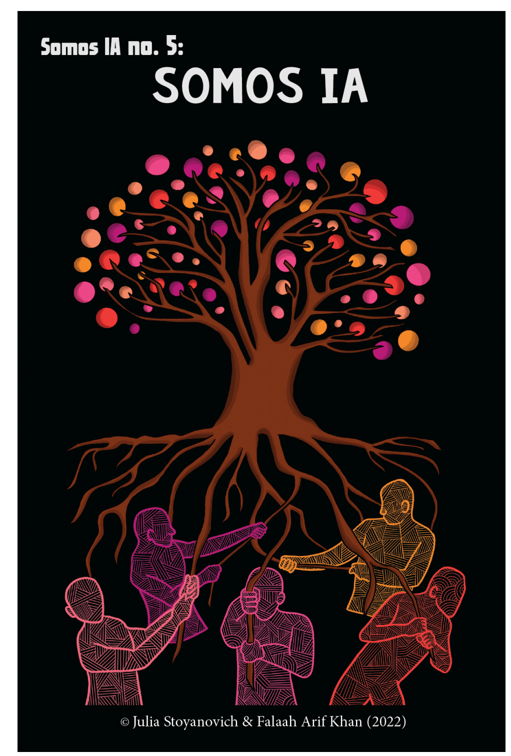
dataresponsibly.github.io/we-are-ai/comics

r/ai center for responsible ai

We are AI comics: in Spanish







dataresponsibly.github.io/we-are-ai/comics

r/ai center for responsible ai

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