Responsible/Ethical Deep Learning

Danna Gurari

University of Colorado Boulder

Fall 2022

https://home.cs.colorado.edu/~DrG/Courses/NeuralNetworksAndDeepLearning/AboutCourse.html

Review

- Last lecture:
  - Multi-task learning
  - Few-shot learning
  - Zero-shot learning
  - Cloud GPU tutorial

- Assignments (Canvas):
  - Final project proposal due on Monday

- Questions?

Today’s Topics

- AI that Discriminates
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Deep Learning

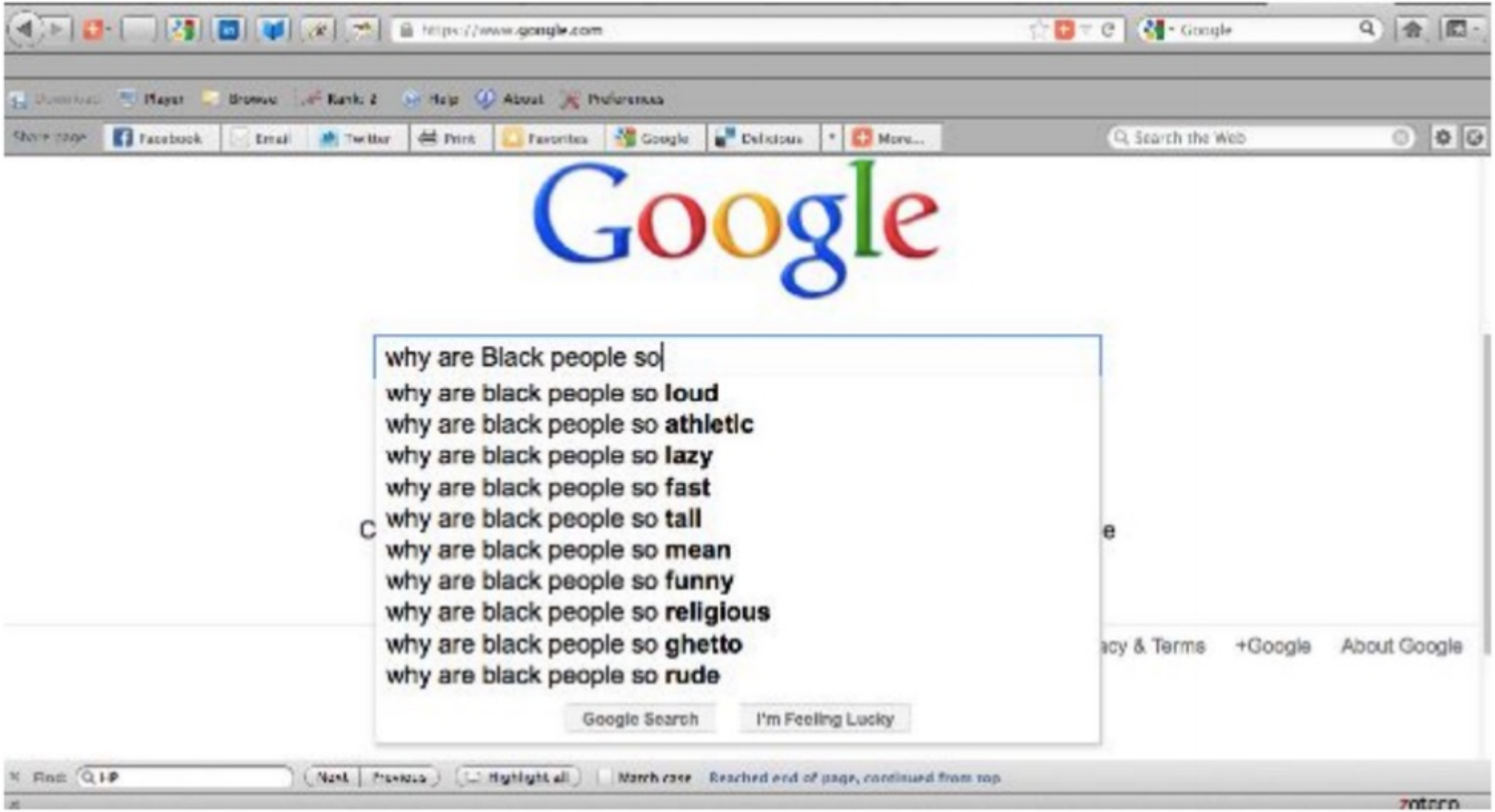
Today’s Topics

- AI that Discriminates
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Deep Learning

Observation: World Population is Diverse

Image Source: https://www.rocketspace.com/corporate-innovation/why-diversity-and-inclusion-driving-innovation-is-a-matter-of-life-and-death

Models Discriminate: Google Search



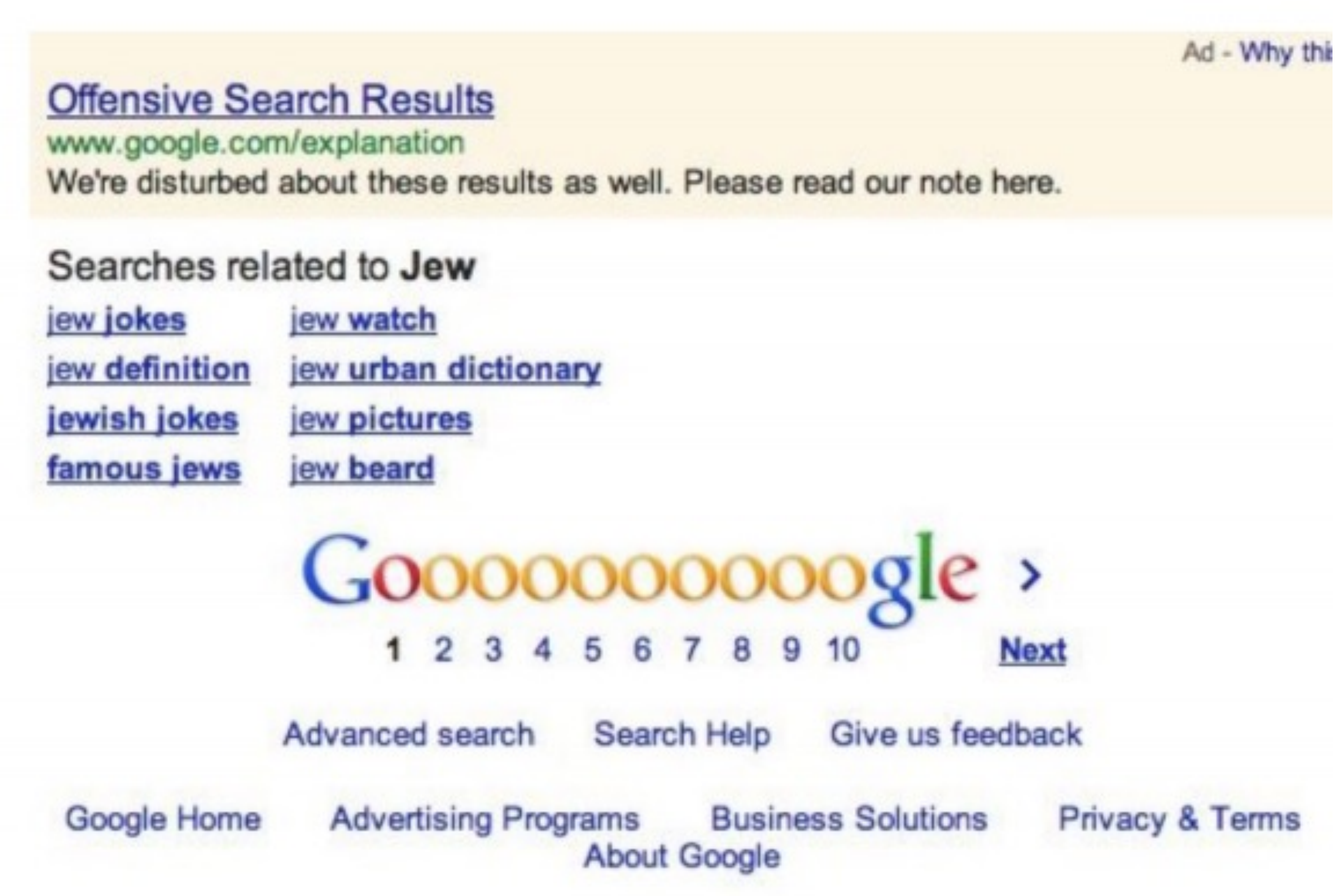
why are Black people so

why are black people so **loud**
why are black people so **athletic**
why are black people so **lazy**
why are black people so **fast**
why are black people so **tall**
why are black people so **mean**
why are black people so **funny**
why are black people so **religious**
why are black people so **ghetto**
why are black people so **rude**

Safiya U. Noble; Algorithms of Oppression: How Search Engines Reinforce Racism

Models Discriminate: Google Search

A search for “Jew” returned many anti-Semitic web pages:



Ad - Why thi

Offensive Search Results

www.google.com/explanation

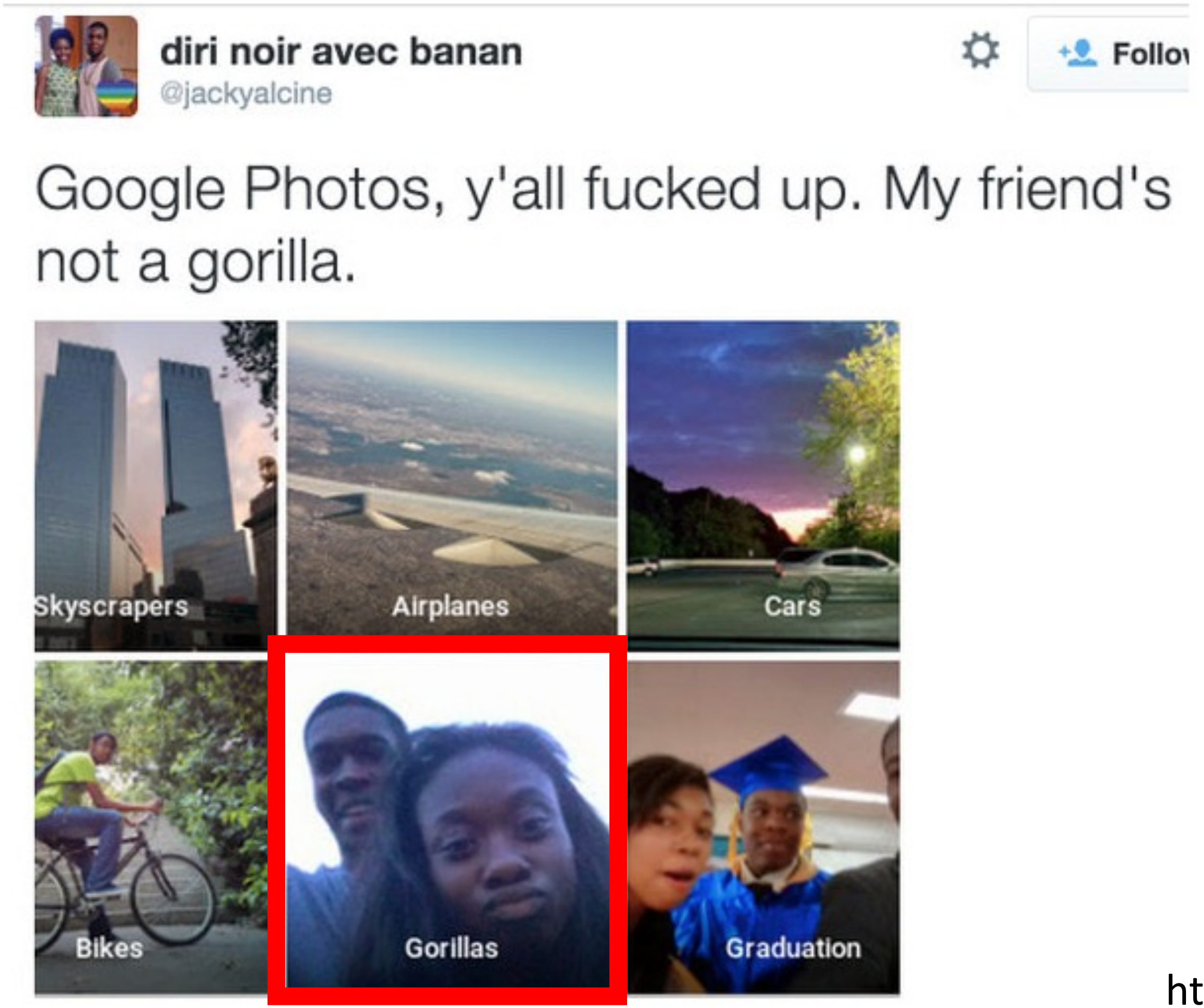
We're disturbed about these results as well. Please read our note here.

Searches related to **Jew**

jew **jokes**	jew **watch**
jew **definition**	jew **urban dictionary**
jewish jokes	jew **pictures**
famous jews	jew **beard**

Safiya U. Noble ; [Algorithms of Oppression: How Search Engines Reinforce Racism]

Models Discriminate: Image Tagging



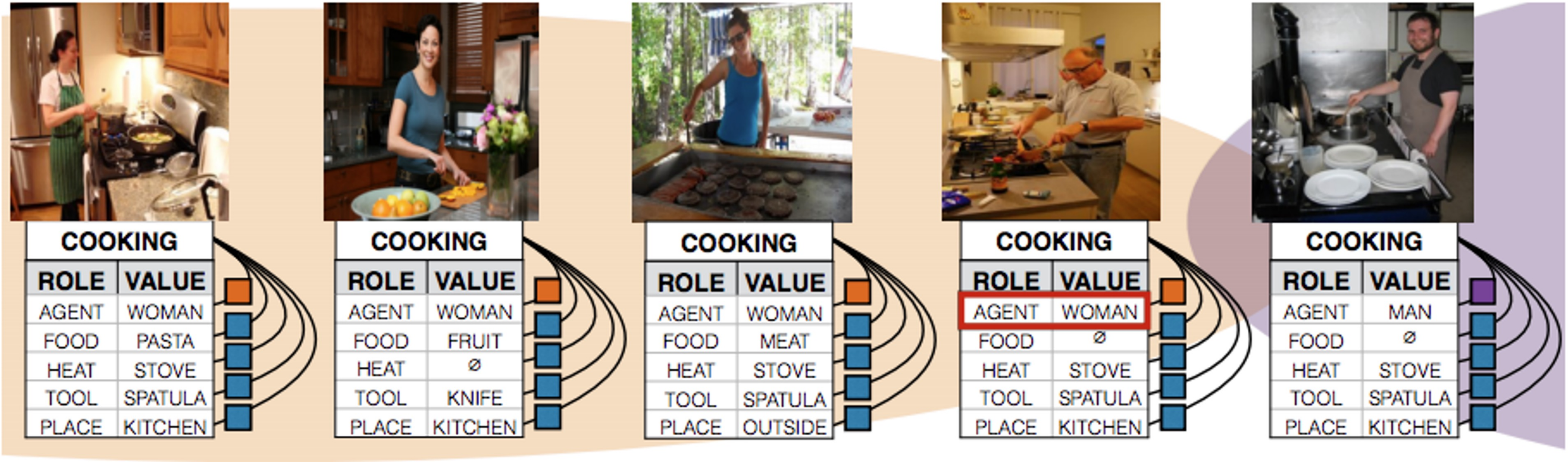
diri noir avec banan
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.

Using Twitter to call out Google's algorithmic bias

https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas

Models Discriminate: Image Tagging



COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

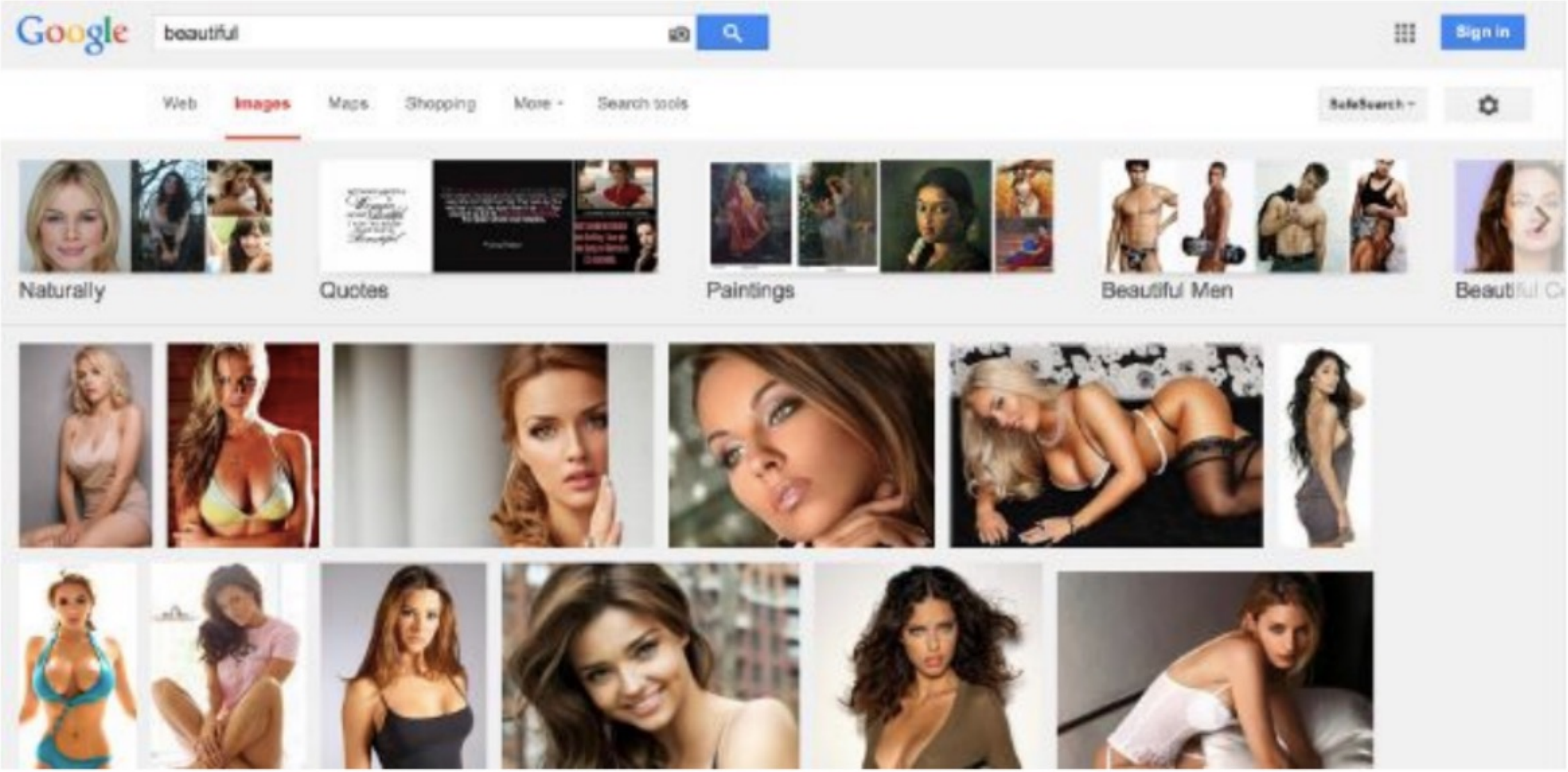
COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

Algorithm identifies men in kitchens as women. Learned this example from given dataset. (Zhao, Wang, Yatskar, Ordonez, Chang, 2017)

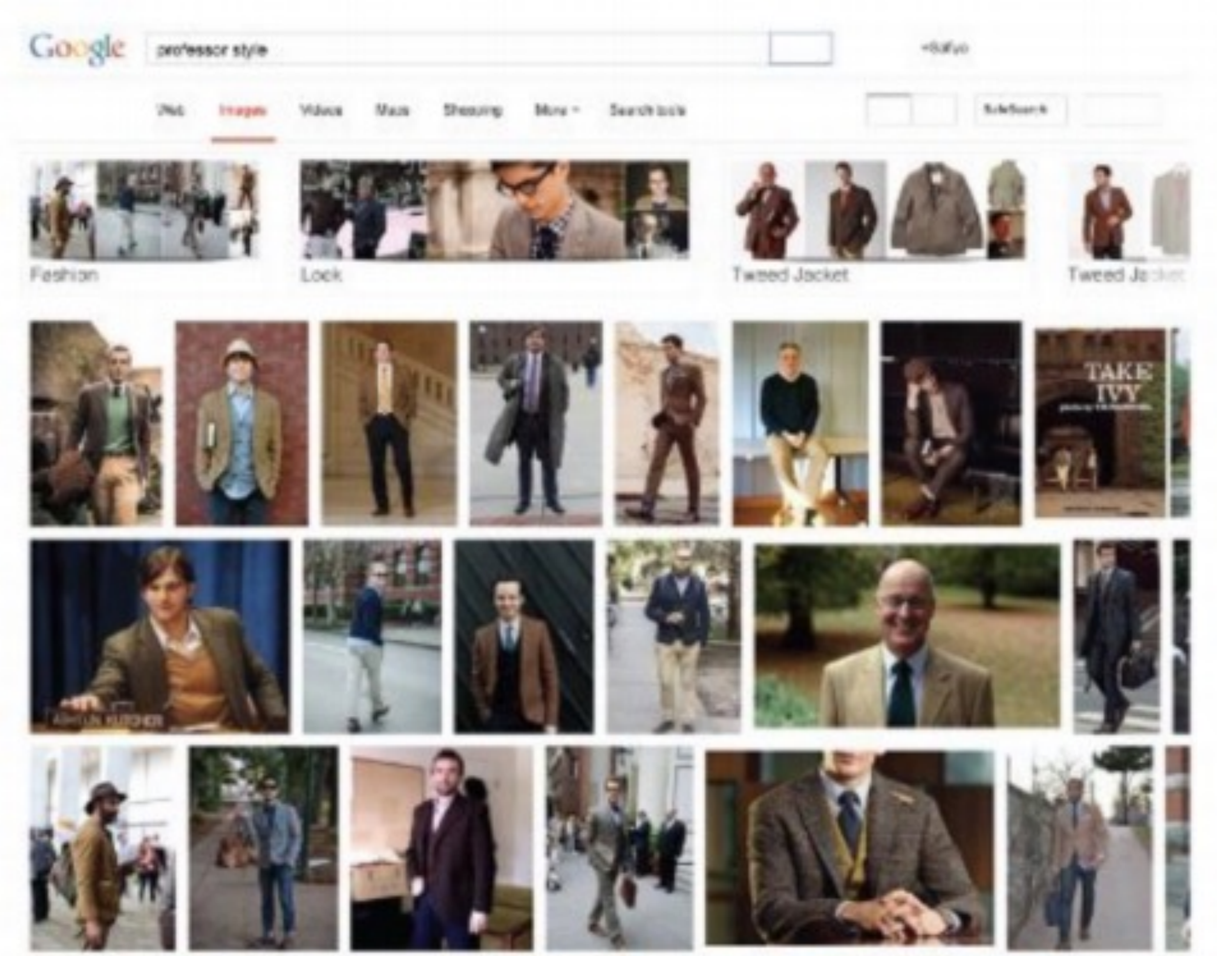
https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/ç

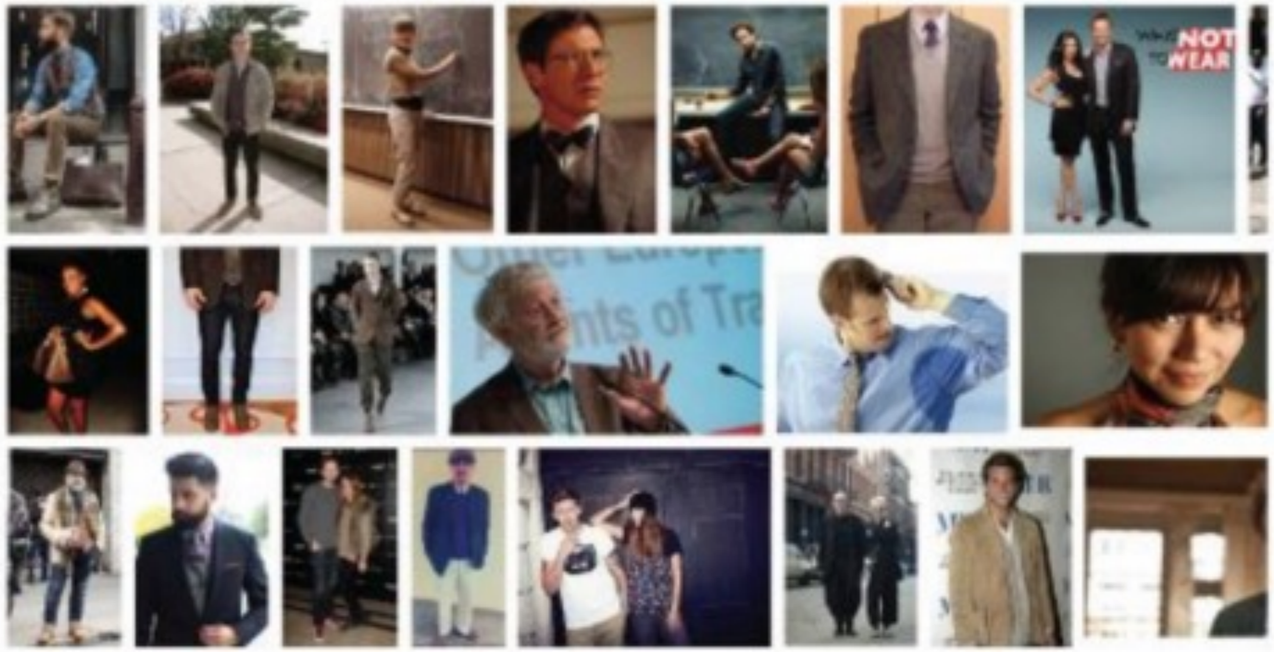
Models Discriminate: Image Tagging ("beautiful"; 2014)



Safiya U. Noble; Algorithms of Oppression: How Search Engines Reinforce Racism

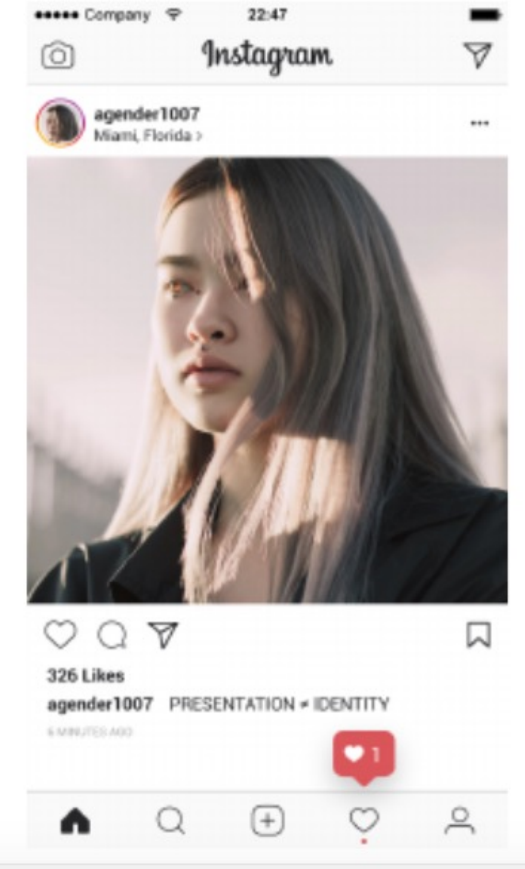
Models Discriminate: Image Tagging ("professor style"; 2014)





Safiya U. Noble; Algorithms of Oppression: How Search Engines Reinforce Racism

Models Discriminate: Image Tagging



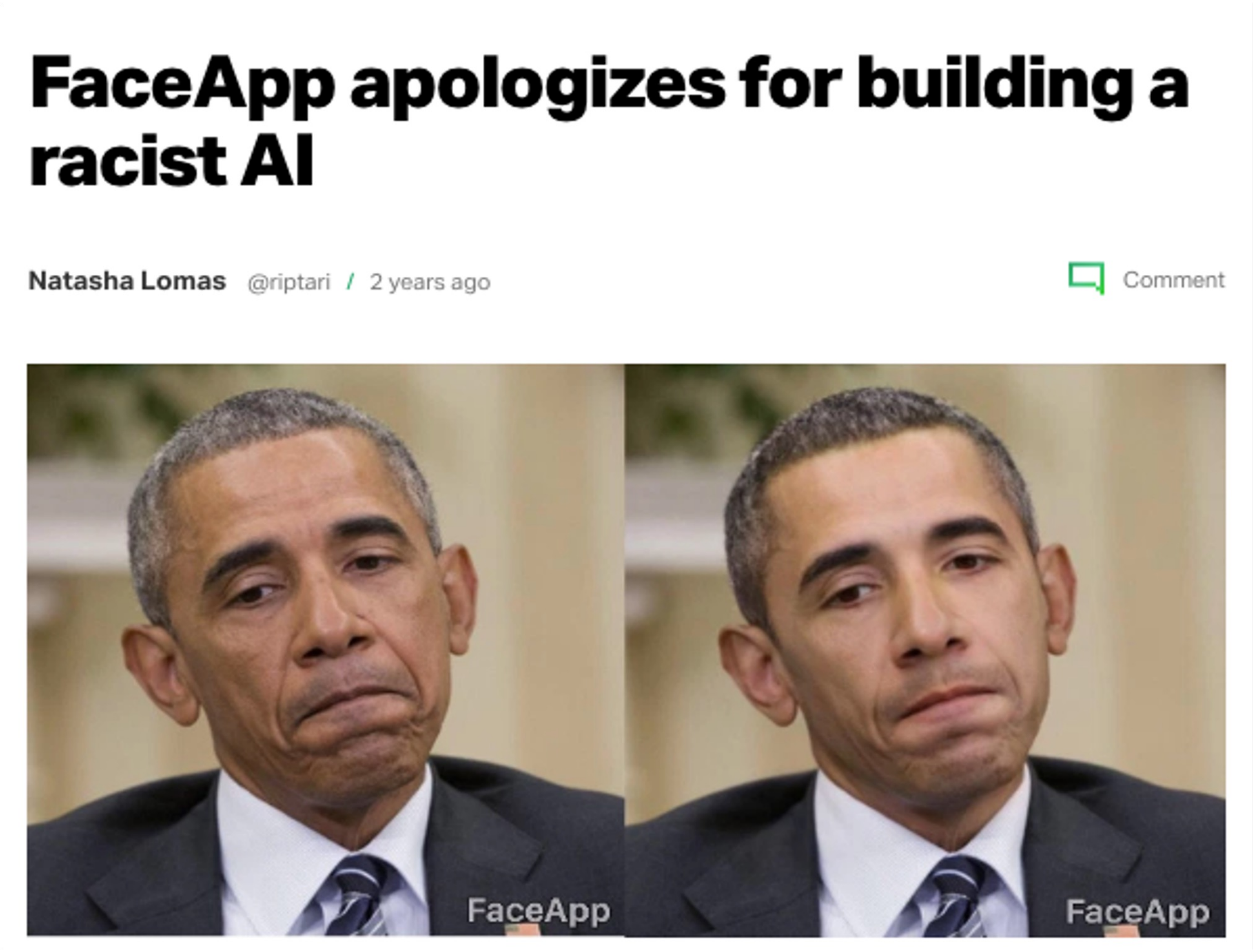
```
...
"age": {
    "min": 20,
    "max": 23,
    "score": 0.923144
},
"face_location": {
    "height": 494,
    "width": 428,
    "left": 327,
    "top": 212
},
"gender": {
    "gender": "FEMALE",
    "gender_label": "female",
    "score": 0.9998667
}
```

```
{
  "class": "woman",
  "score": 0.813,
  "type_hierarchy": "/person
  /female/woman"
},
{
  "class": "person",
  "score": 0.806
},
{
  "class": "young lady (heroine)",
  "score": 0.504,
  "type_hierarchy": "/person/female
  /woman/young lady (heroine)"
}
...
```

Person identifies as agender (gender-less, and so non-binary)

Morgan Klaus Scheurman, Jacob M. Paul, and Jed R. Brubaker, “How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services.” CSCW 2019.

Models Discriminate:
“Hotness” Photo-Editing Filter



FaceApp apologizes for building a racist AI

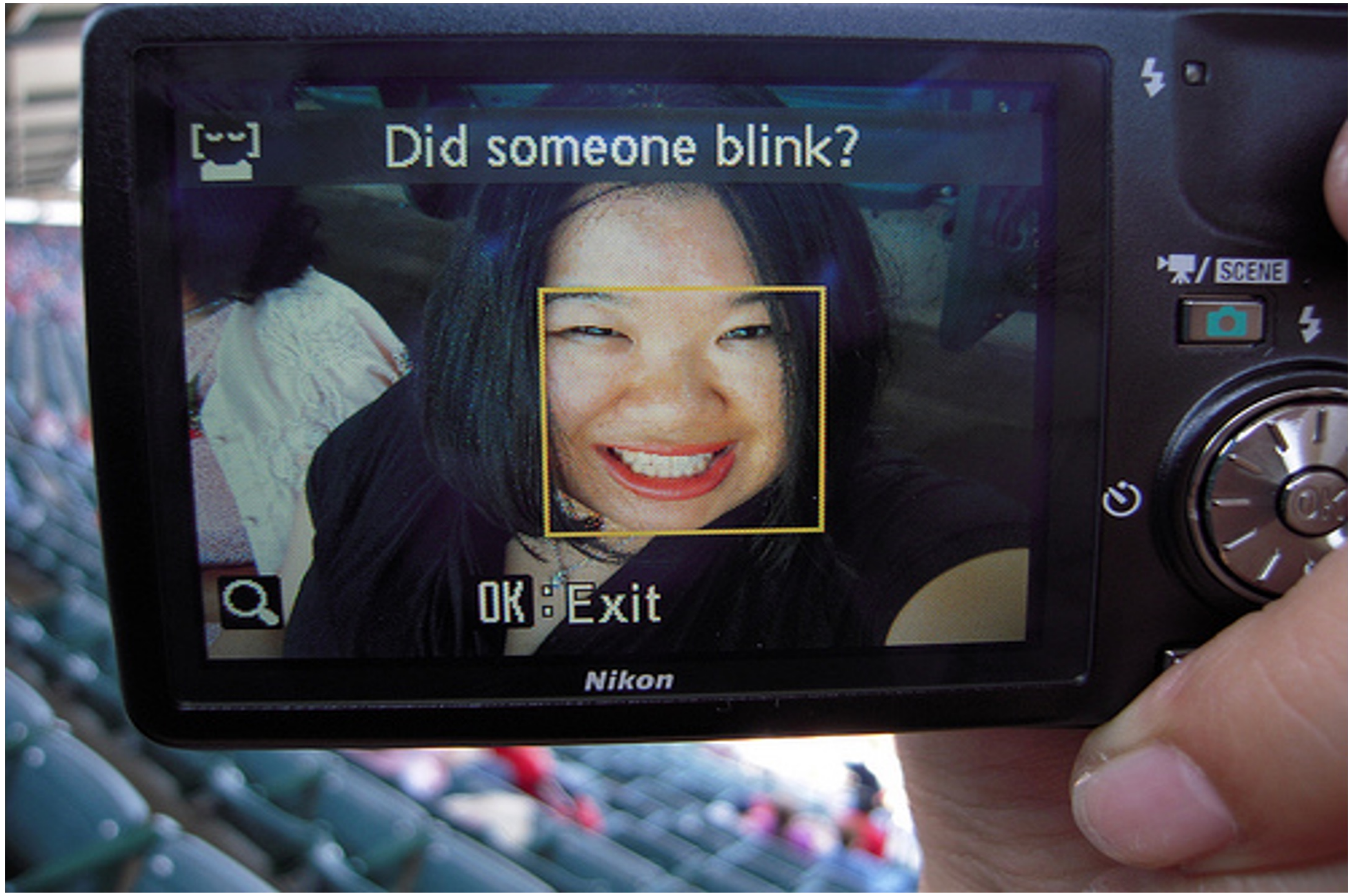
Natasha Lomas @riptari / 2 years ago

Comment

https://techcrunch.com/2017/04/25/faceapp-apologises-for-building-a-racist-ai/

Models Discriminate: Nikon Blink Detection

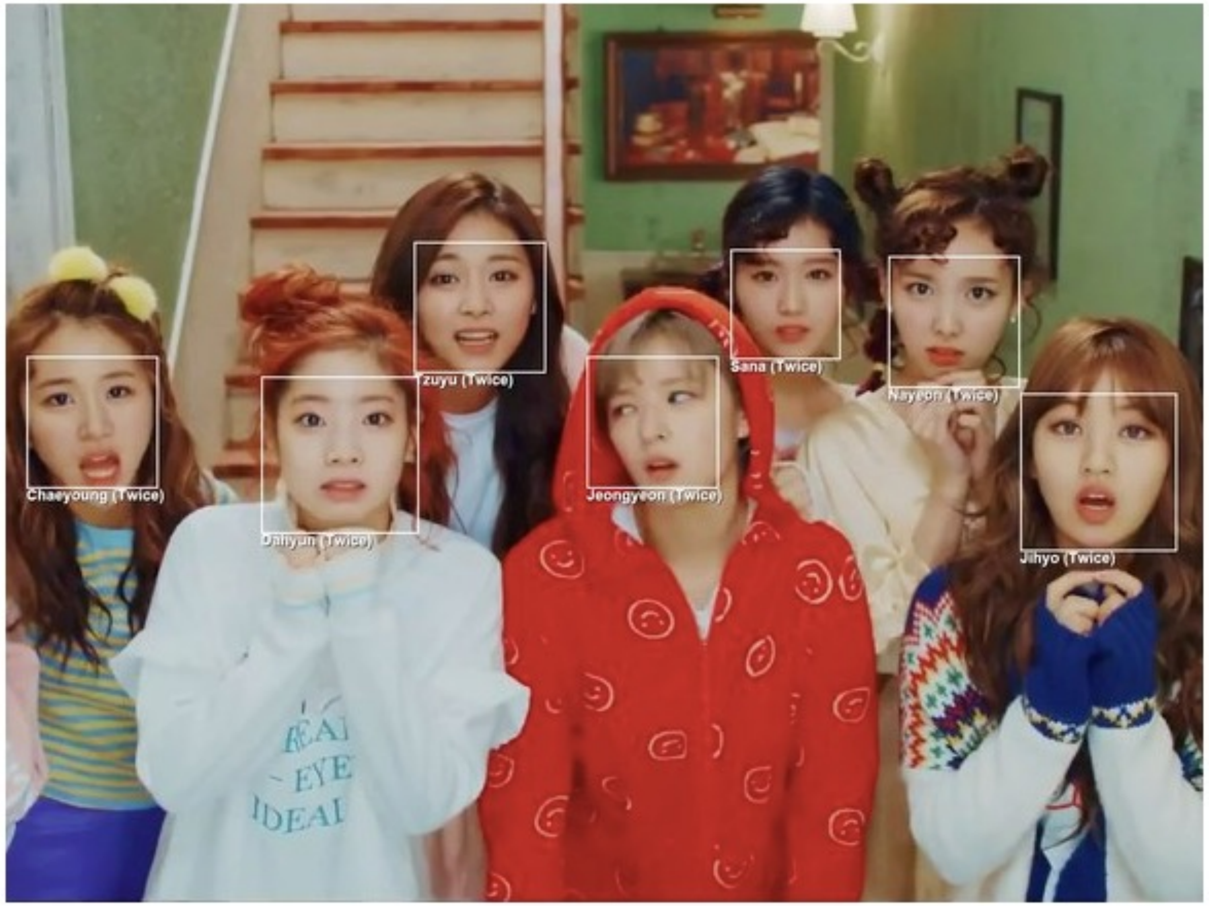
Two kids bought their mom a Nikon Coolpix S630 digital camera for Mother's Day… when they took portrait pictures of each other, a message flashed across the screen asking, "Did someone blink?"



http://content.time.com/time/business/article/0,8599,1954643,00.html

Models Discriminate: Face Recognition

Software engineer at company: “It got some of our Asian employees mixed up,” says Gan, who is Asian. “Which was strange because it got everyone else correctly.”

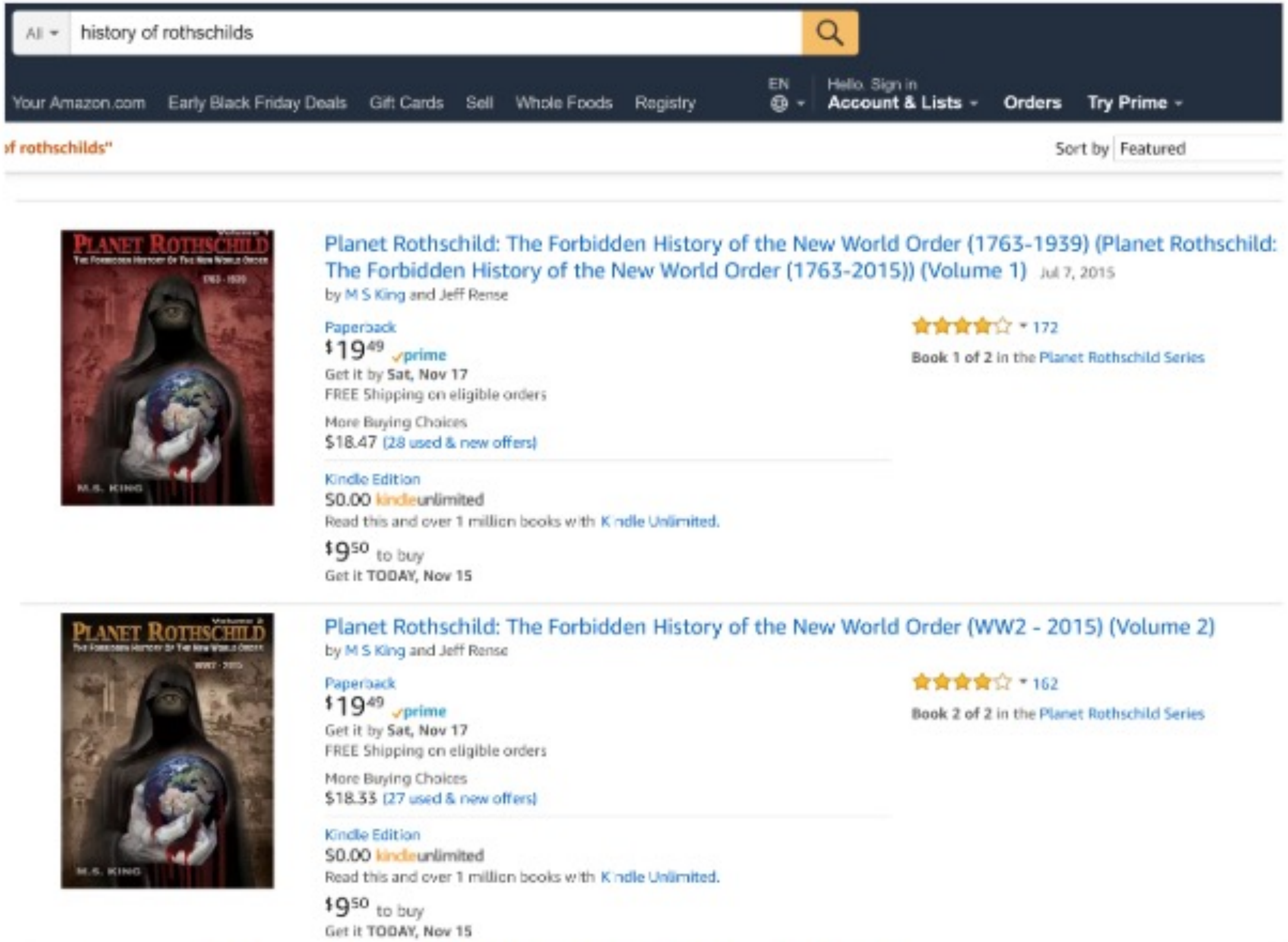


Gfycat's facial recognition software can now recognize individual members of K-pop band Twice, but in early tests couldn't distinguish different Asian faces. GFYCAT

https://www.wired.com/story/how-coders-are-fighting-bias-in-facial-recognition-software/

Models Discriminate: Book Shopping

Anti-Semitic Bias:



f rothschilds"

Planet Rothschild: The Forbidden History of the New World Order (1763-1939) (Planet Rothschild: The Forbidden History of the New World Order (1763-2015)) (Volume 1) Jul 7, 2015

by M S King and Jeff Rense

Paperback

$^{\$}19^{49}$ prime

Get it by Sat, Nov 17

FREE Shipping on eligible orders

More Buying Choices

$18.47 (28 used & new offers)

Kindle Edition

$0.00 kindleunlimited

Read this and over 1 million books with Kindle Unlimited.

$^{\$}9^{50}$ to buy

Get it TODAY, Nov 15

★★★★☆ 172

Book 1 of 2 in the Planet Rothschild Series

Planet Rothschild: The Forbidden History of the New World Order (WW2 - 2015) (Volume 2)

by M S King and Jeff Rense

Paperback

$^{\$}19^{49}$ prime

Get it by Sat, Nov 17

FREE Shipping on eligible orders

More Buying Choices

$18.33 (27 used & new offers)

Kindle Edition

$0.00 kindleunlimited

Read this and over 1 million books with Kindle Unlimited.

$^{\$}9^{50}$ to buy

Get it TODAY, Nov 15

★★★★☆ 162

Book 2 of 2 in the Planet Rothschild Series

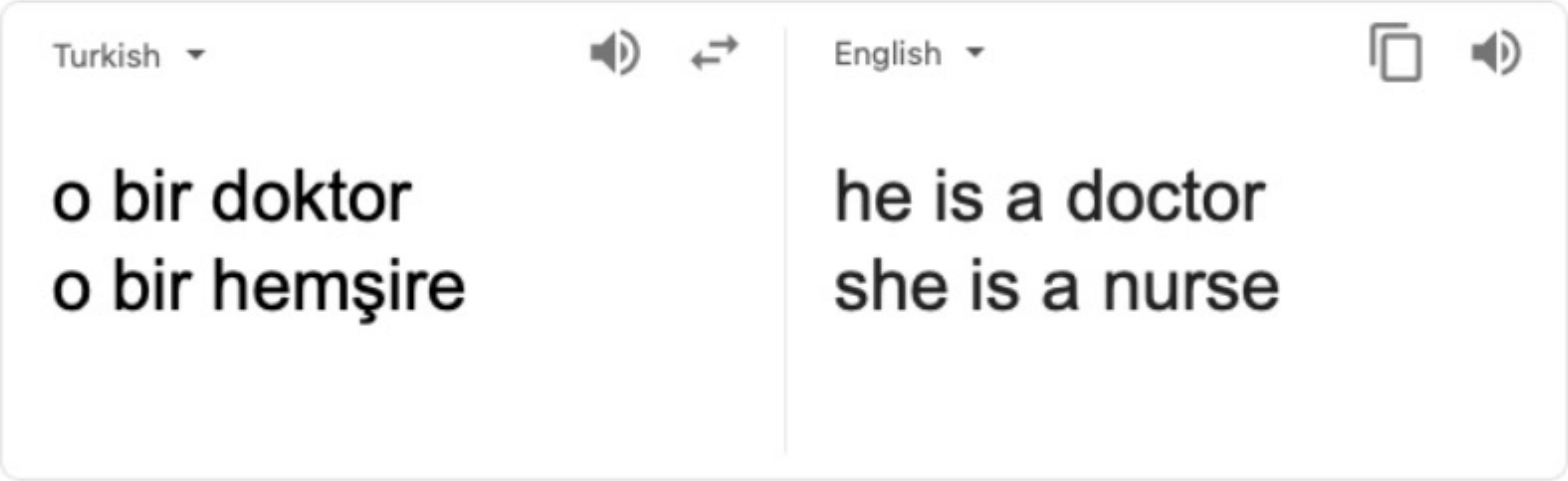
https://www.tabletmag.com/scroll/275042/amazons-algorithm-has-an-anti-semitism-problem

Models Discriminate: Job Recruiting

Amazon’s algorithm learned to systematically downgrade women's CVs for technical jobs such as software developer.

https://phys.org/news/2018-11-amazon-sexist-hiring-algorithm-human.html

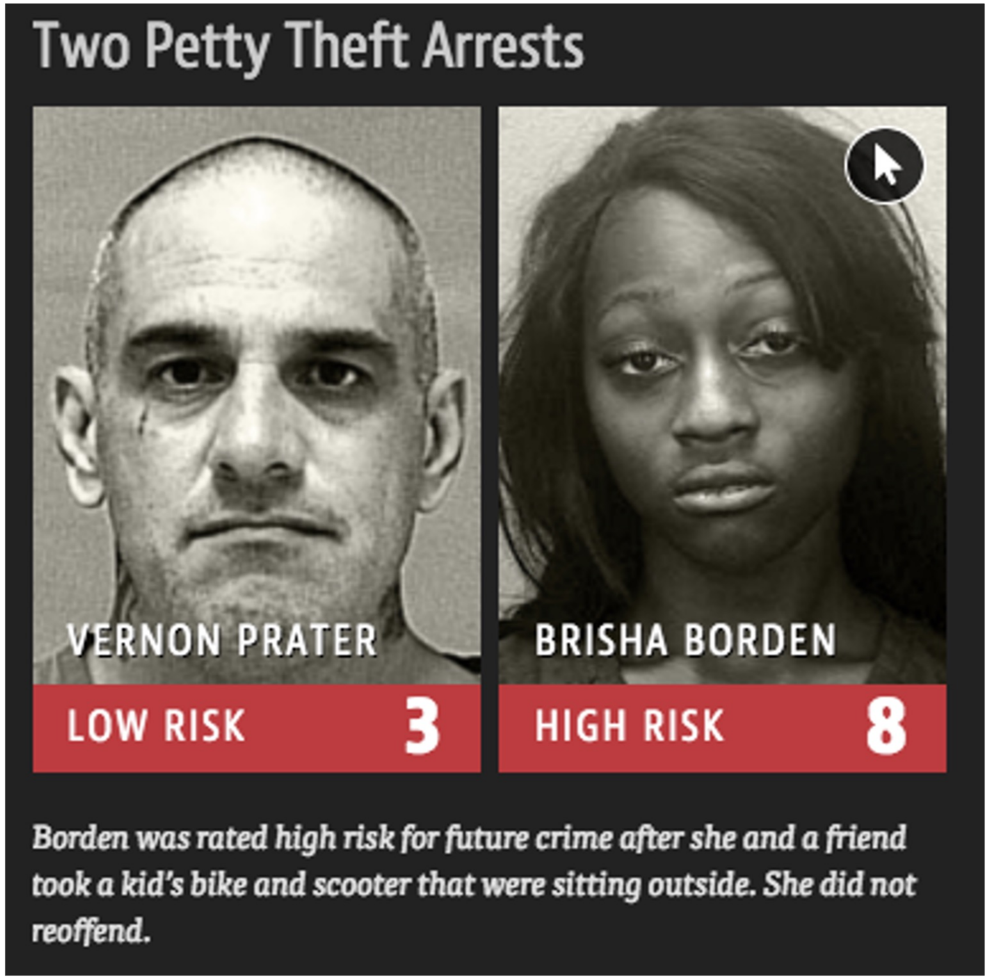
Models Discriminate: Language Translation



o bir doktor
o bir hemşire

he is a doctor
she is a nurse

Models Discriminate: Criminal Sentencing



Two Petty Theft Arrests

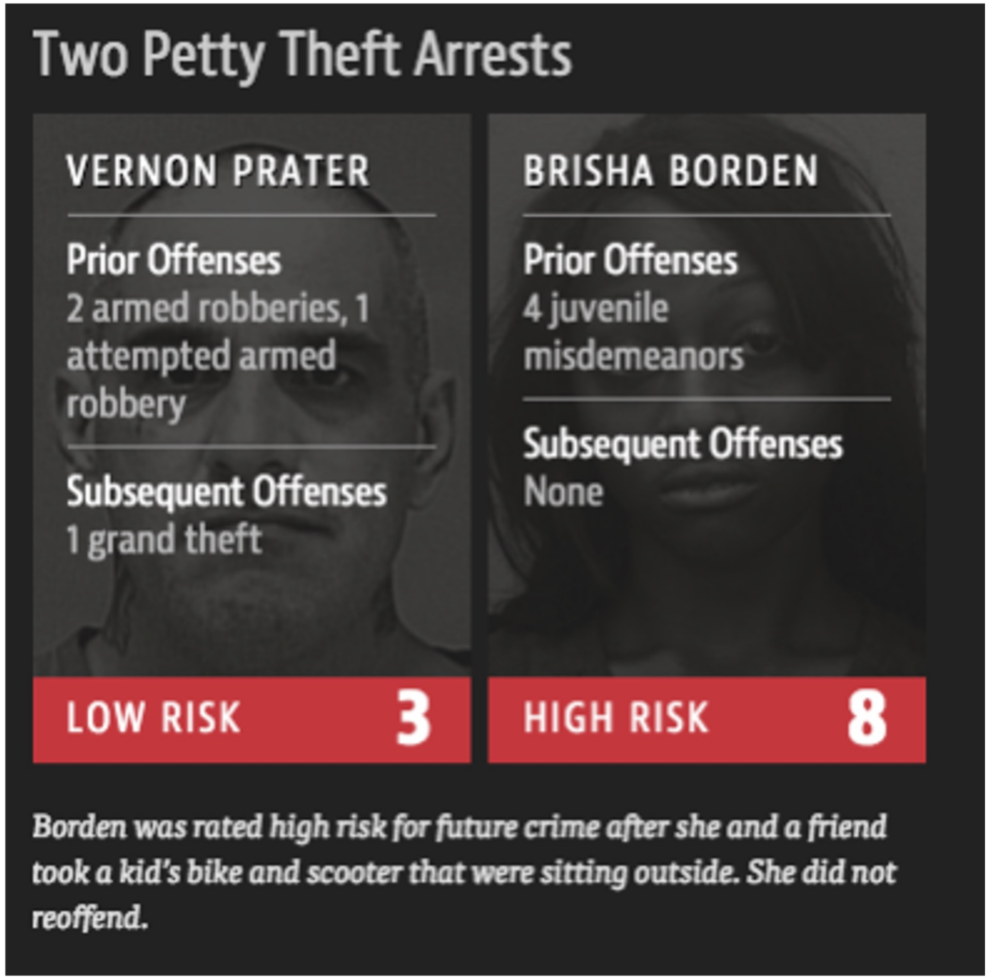
VERNON PRATER

LOW RISK 3

BRISHA BORDEN

HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.



Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

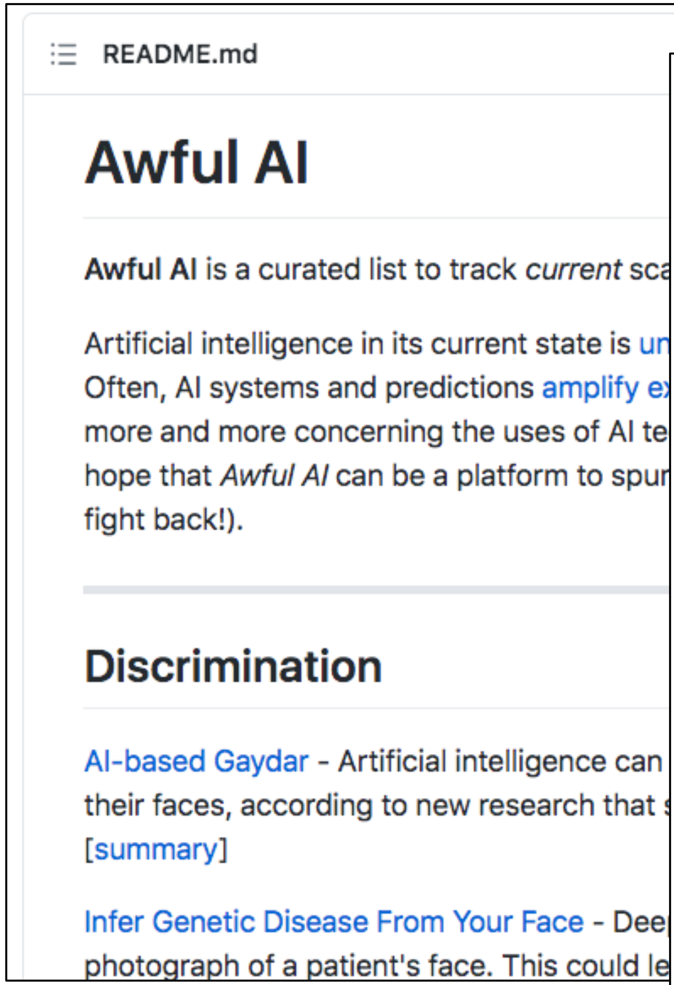
HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Models Discriminate: And MANY more…

- e.g.,



README.md

Awful AI

Awful AI is a curated list to track *current* sca

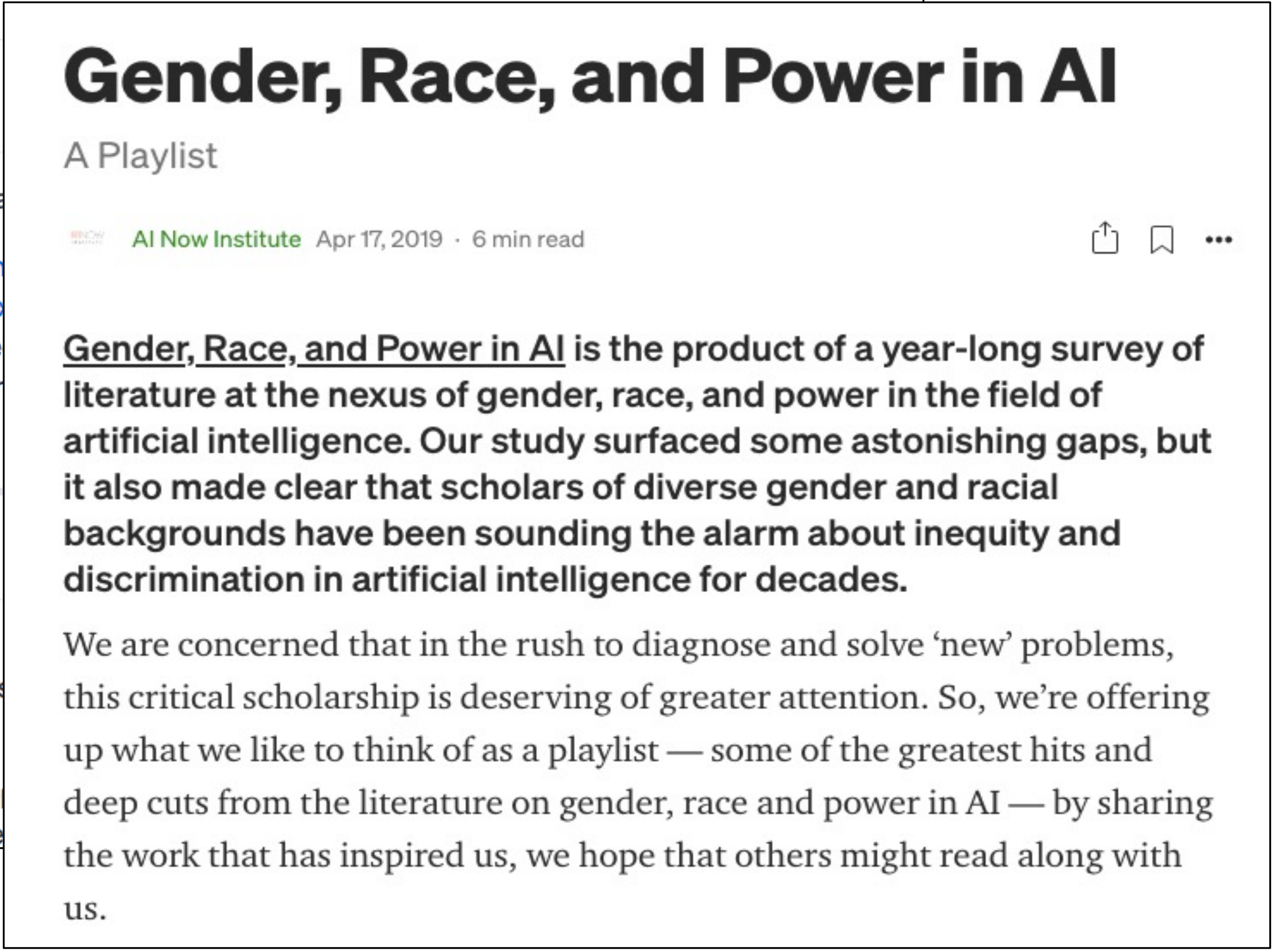
Artificial intelligence in its current state is un
Often, AI systems and predictions amplify ex
more and more concerning the uses of AI te
hope that *Awful AI* can be a platform to spur
fight back!).

Discrimination

AI-based Gaydar - Artificial intelligence can
their faces, according to new research that s
[summary]

Infer Genetic Disease From Your Face - Dee
photograph of a patient's face. This could le

https://github.com/daviddao/awful-ai



Gender, Race, and Power in AI

A Playlist

AI Now Institute Apr 17, 2019 · 6 min read

Gender, Race, and Power in AI is the product of a year-long survey of literature at the nexus of gender, race, and power in the field of artificial intelligence. Our study surfaced some astonishing gaps, but it also made clear that scholars of diverse gender and racial backgrounds have been sounding the alarm about inequity and discrimination in artificial intelligence for decades.

We are concerned that in the rush to diagnose and solve 'new' problems, this critical scholarship is deserving of greater attention. So, we're offering up what we like to think of as a playlist — some of the greatest hits and deep cuts from the literature on gender, race and power in AI — by sharing the work that has inspired us, we hope that others might read along with us.

https://medium.com/@AINowInstitute/gender-race-and-power-in-ai-a-playlist-2d3a44e43d3b

Models Discriminate

How would you try to fix issues like these?

Today’s Topics

- AI that Discriminates
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Deep Learning

We know that algorithms are not perfect.

How can we alleviate the issue that
DL algorithms that discriminate?

FAT Deep Learning: In Vague, Lay Terms

- **Fairness:** treat people fairly
- **Accountability:** mimic infrastructure to oversee human decision makers (e.g., policymakers, courts) for algorithm decision-makers
- **Transparency:** clearly communicate algorithms’ capabilities and limitations

FAT Deep Learning: Fairness

- How to make more fair methods?
  - Pre-processing:
    - Training data: modify it
  - Optimization at training:
    - Algorithm: e.g., add regularization term to objective function to penalize unfairness
    - Features: remove those that reflect bias; e.g., gender, race, age, education, sexual orientation, etc.
  - Post-process predictions
    - Counterfactual assumption: check impact of modifying single feature

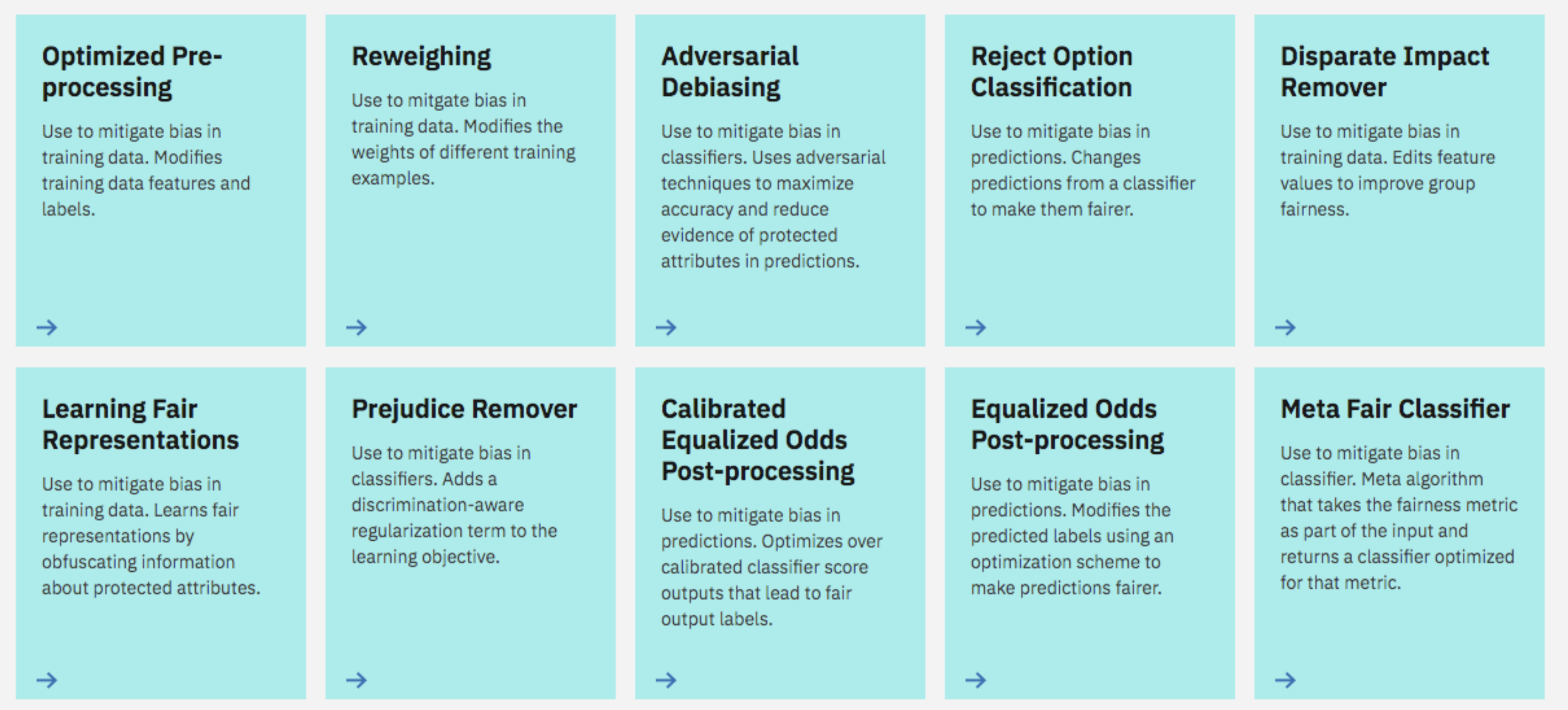
https://fairmlclass.github.io/; https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

FAT Deep Learning: Fairness

- Fairness – how to define this mathematically?
  - e.g., group fairness (proportion of members in protected group receiving positive classification matches proportion in the population as a whole)
  - e.g., individual fairness (similar individuals should be treated similarly)

e.g., IBM's AI Fairness 360 Open Source Toolkit

70+ fairness metrics and 10+ bias mitigation algorithms



Optimized Pre-processing

Use to mitigate bias in training data. Modifies training data features and labels.

Reweighing

Use to mitigate bias in training data. Modifies the weights of different training examples.

Adversarial Debiasing

Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.

Reject Option Classification

Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.

Disparate Impact Remover

Use to mitigate bias in training data. Edits feature values to improve group fairness.

Learning Fair Representations

Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.

Prejudice Remover

Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.

Calibrated Equalized Odds Post-processing

Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.

Equalized Odds Post-processing

Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.

Meta Fair Classifier

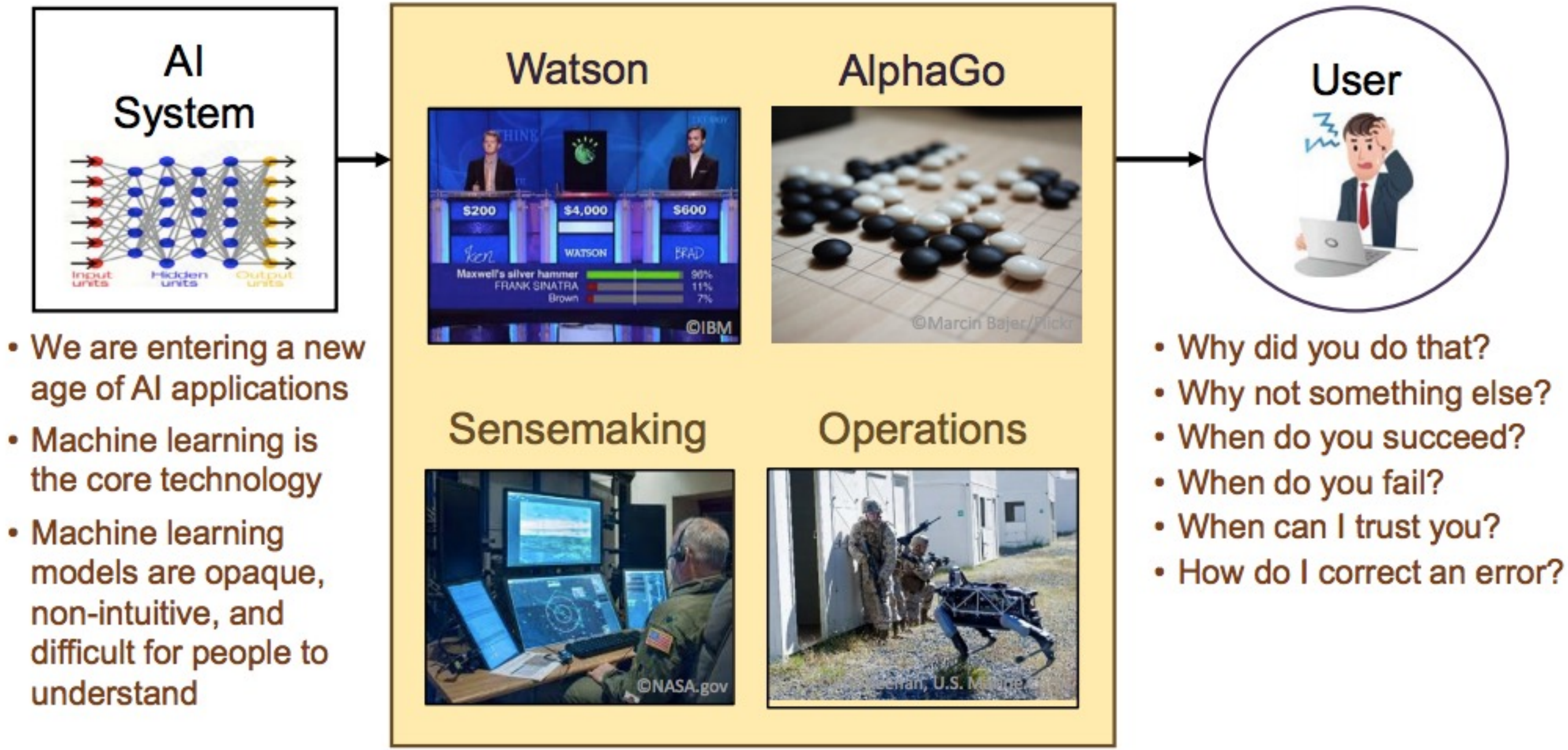
Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.

FAT Deep Learning: Accountability

- Who is accountable for model behavior?
  - e.g., developers must design algorithms so that oversight authorities meet pre-defined rules ("procedural regularity")?
  - e.g., data providers?
  - e.g., regulators who determine scope of oversight (e.g., require describing and explaining model failures)?

Joshua Kroll et al. "Accountable Algorithms." University of Pennsylvania Law Review, 2017.

FAT Deep Learning: Transparency



AI System

Watson

AlphaGo

Sensemaking

Operations

User

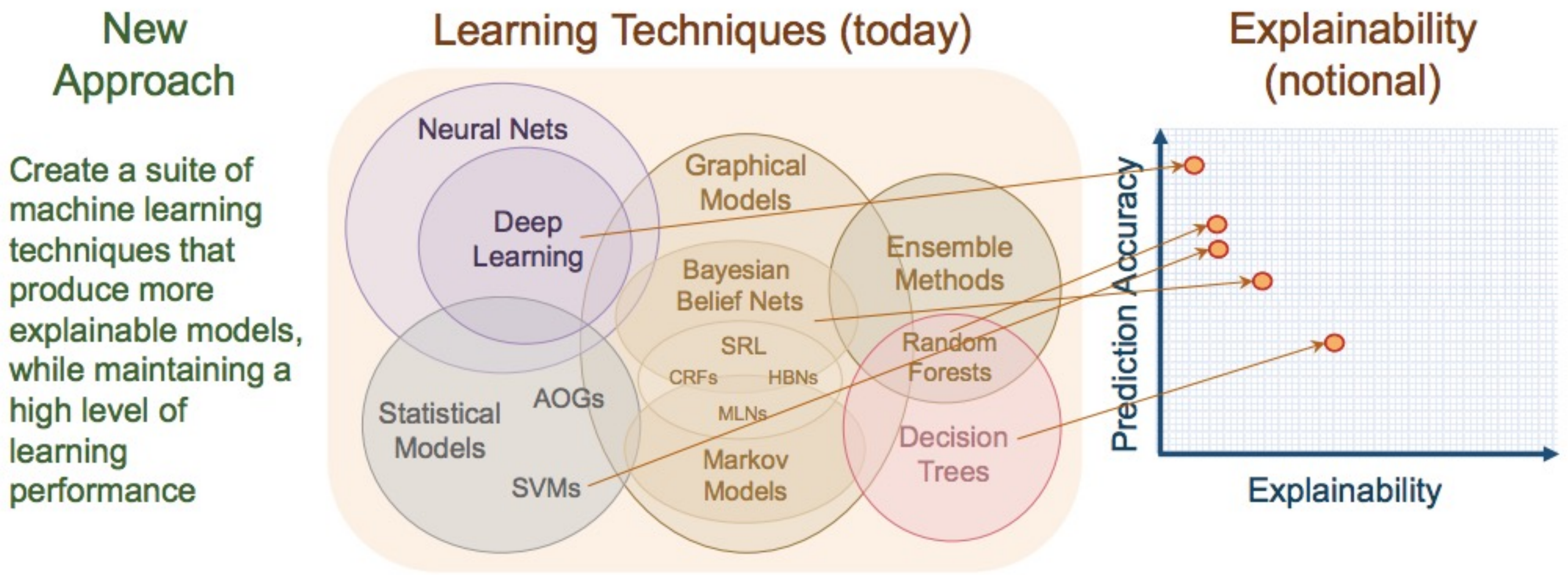
- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Input units
Hidden units
Output units
$200
$4,000
$600
WATSON
Maxwell's silver hammer 96%
FRANK SINATRA 11%
Brown 7%
©IBM
©Marcin Bajer/Flickr
©NASA.gov
U.S. Marine

https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

FAT Deep Learning: Transparency



New Approach

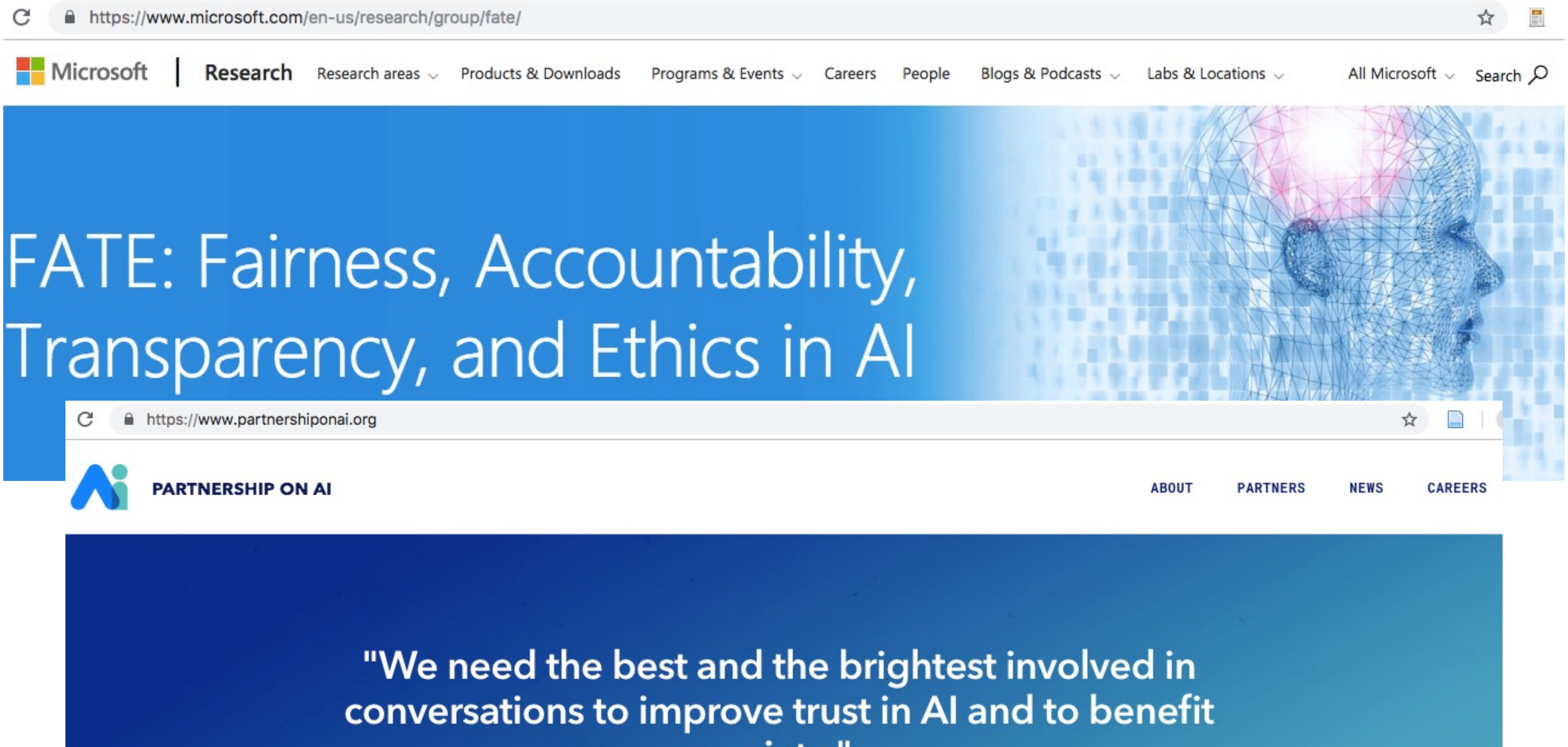
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)

Explainability (notional)

https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Industry (Facebook, Microsoft, & more…)



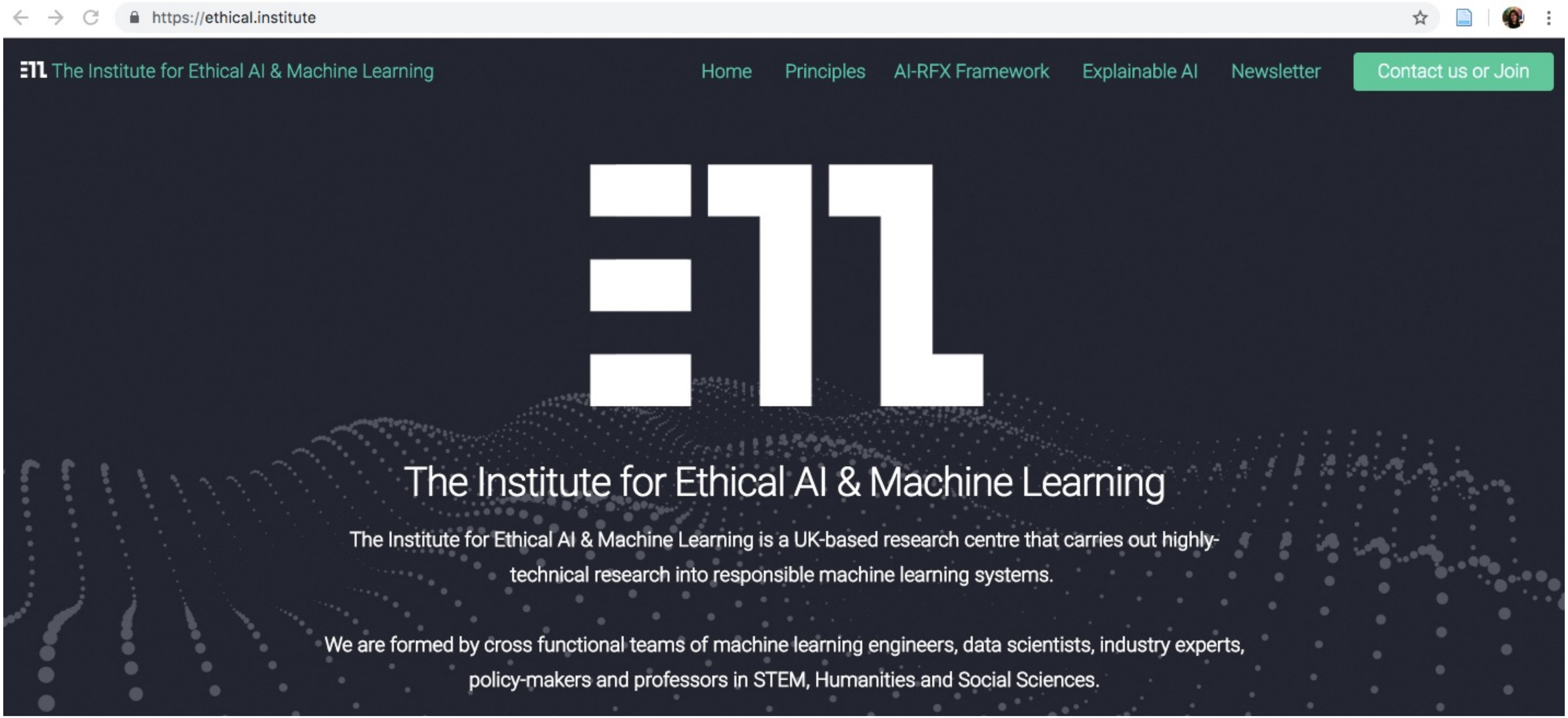
https://www.microsoft.com/en-us/research/group/fate/

FATE: Fairness, Accountability, Transparency, and Ethics in AI

https://www.partnershiponai.org

"We need the best and the brightest involved in conversations to improve trust in AI and to benefit

Institutes



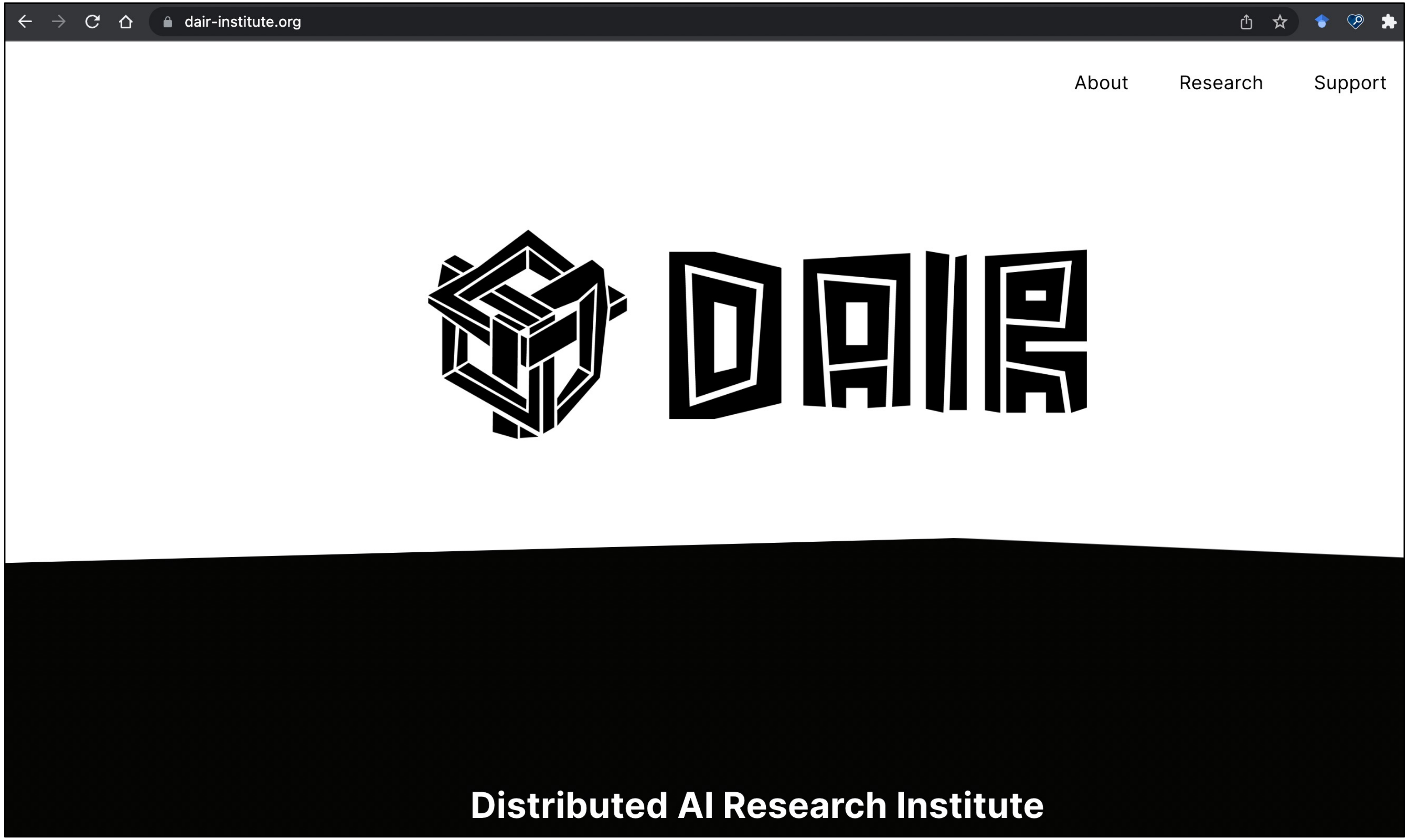
https://ethical.institute
The Institute for Ethical AI & Machine Learning
Home
Principles
AI-RFX Framework
Explainable AI
Newsletter
Contact us or Join

The Institute for Ethical AI & Machine Learning

The Institute for Ethical AI & Machine Learning is a UK-based research centre that carries out highly-technical research into responsible machine learning systems.

We are formed by cross functional teams of machine learning engineers, data scientists, industry experts, policy-makers and professors in STEM, Humanities and Social Sciences.

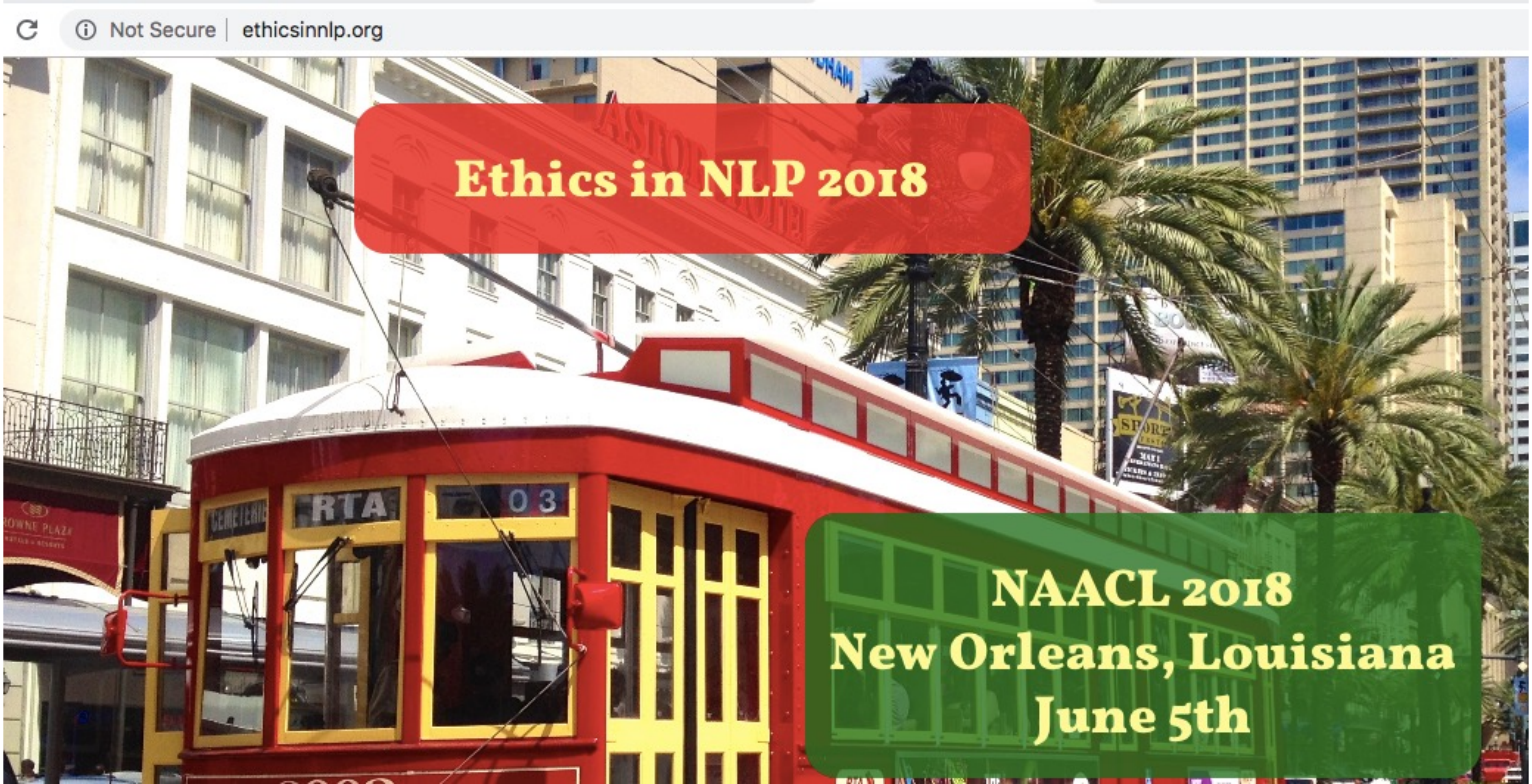
Institutes



dair-institute.org

Distributed AI Research Institute

Academia: Workshops



Ethics in NLP 2018

NAACL 2018
New Orleans, Louisiana
June 5th

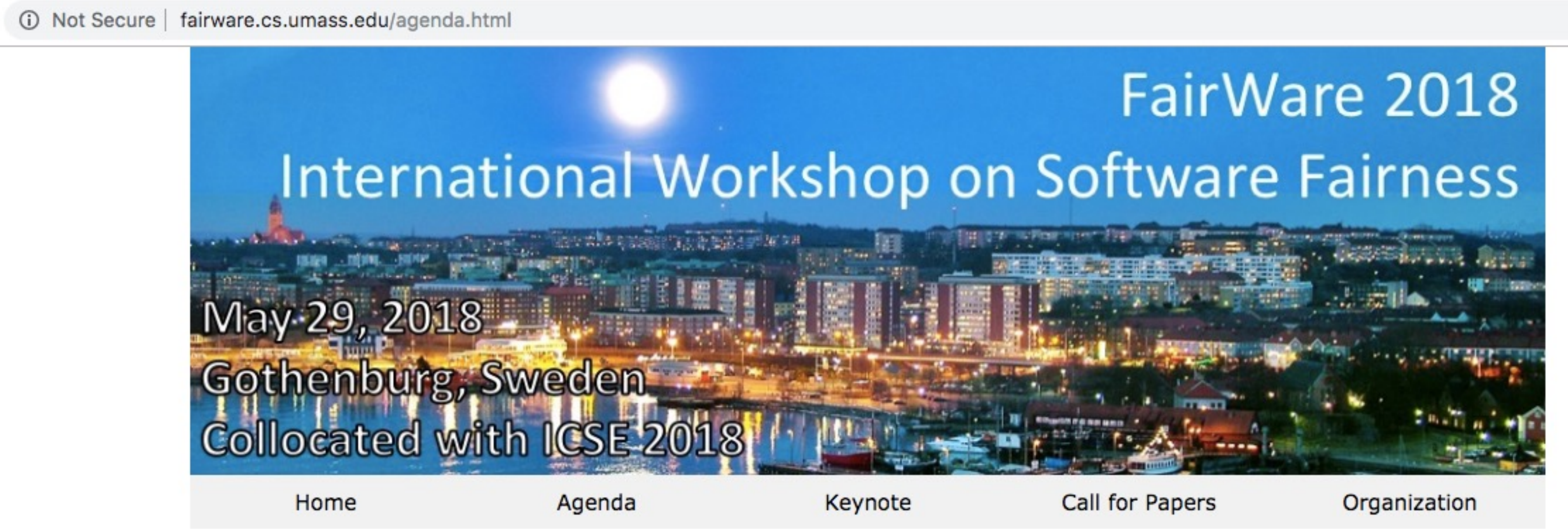
Academia: Workshops

https://fatconference.org

ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)

A multi-disciplinary conference that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

Academia: Workshops



FairWare 2018

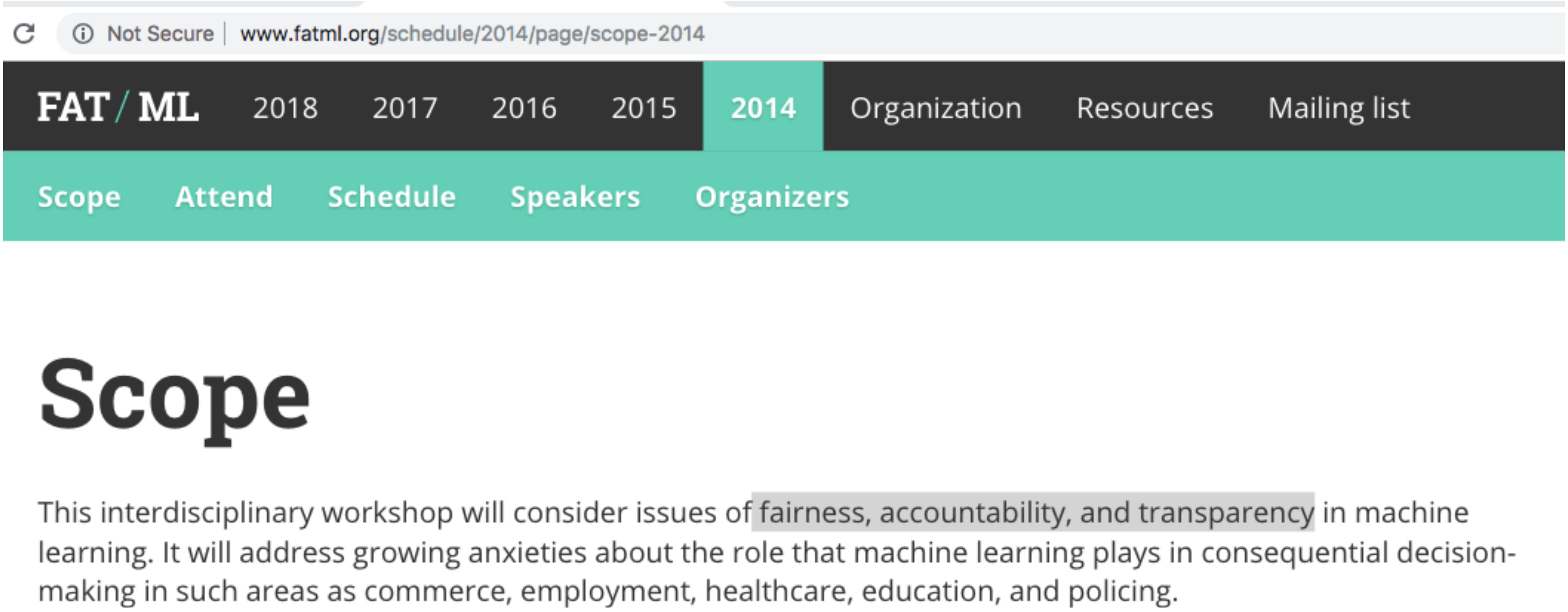
International Workshop on Software Fairness

May 29, 2018

Gothenburg, Sweden

Collocated with ICSE 2018

Academia: Annual Workshop Since 2014…



Scope

This interdisciplinary workshop will consider issues of fairness, accountability, and transparency in machine learning. It will address growing anxieties about the role that machine learning plays in consequential decision-making in such areas as commerce, employment, healthcare, education, and policing.

Today’s Topics

- AI that Discriminates
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Deep Learning

We know that algorithms are not perfect.
Algorithms can be biased.

Are they ethical to use?

Time for a group activity!

Unacceptable to acceptable:
Using DL to sentence people for a crime

Unacceptable to acceptable: Using DL to diagnose diseases

Unacceptable to acceptable:
Using DL to filter resumes for jobs

Unacceptable to acceptable:
Using DL to determine eligibility for a loan

Unacceptable to acceptable:
Using DL to determine eligibility for a loan

What other ethical issues can you think of for using deep learning algorithms?

Today's Topics

- AI that Discriminates
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Deep Learning