Responsible Deep Learning: Part 1

Danna Gurari
University of Colorado Boulder
Spring 2024

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2024/

Review

- Last lecture on efficient learning:
  - Motivation
  - Curriculum Learning
  - Active Learning
  - Dataset Distillation
  - Few-shot Learning
  - Interview by Lijun Chen
- Assignments (Canvas):
  - Final project presentations due in two weeks
- Questions?

Today’s Topics

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

Today’s Topics

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

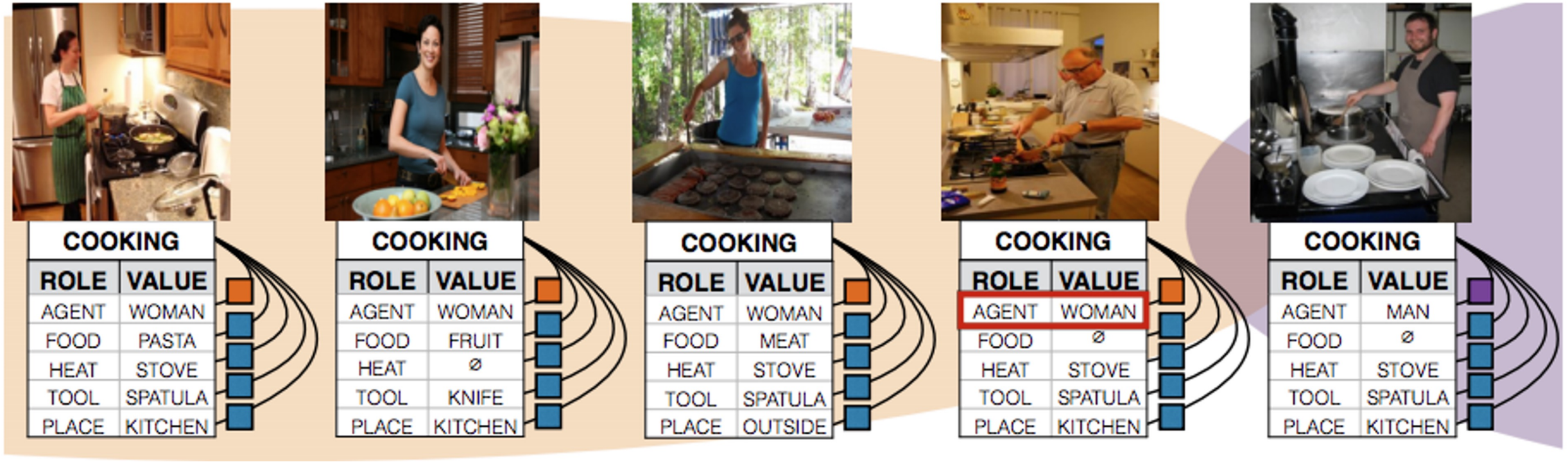
Observation: World Population is Diverse

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

Models Discriminate: Image Tagging

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

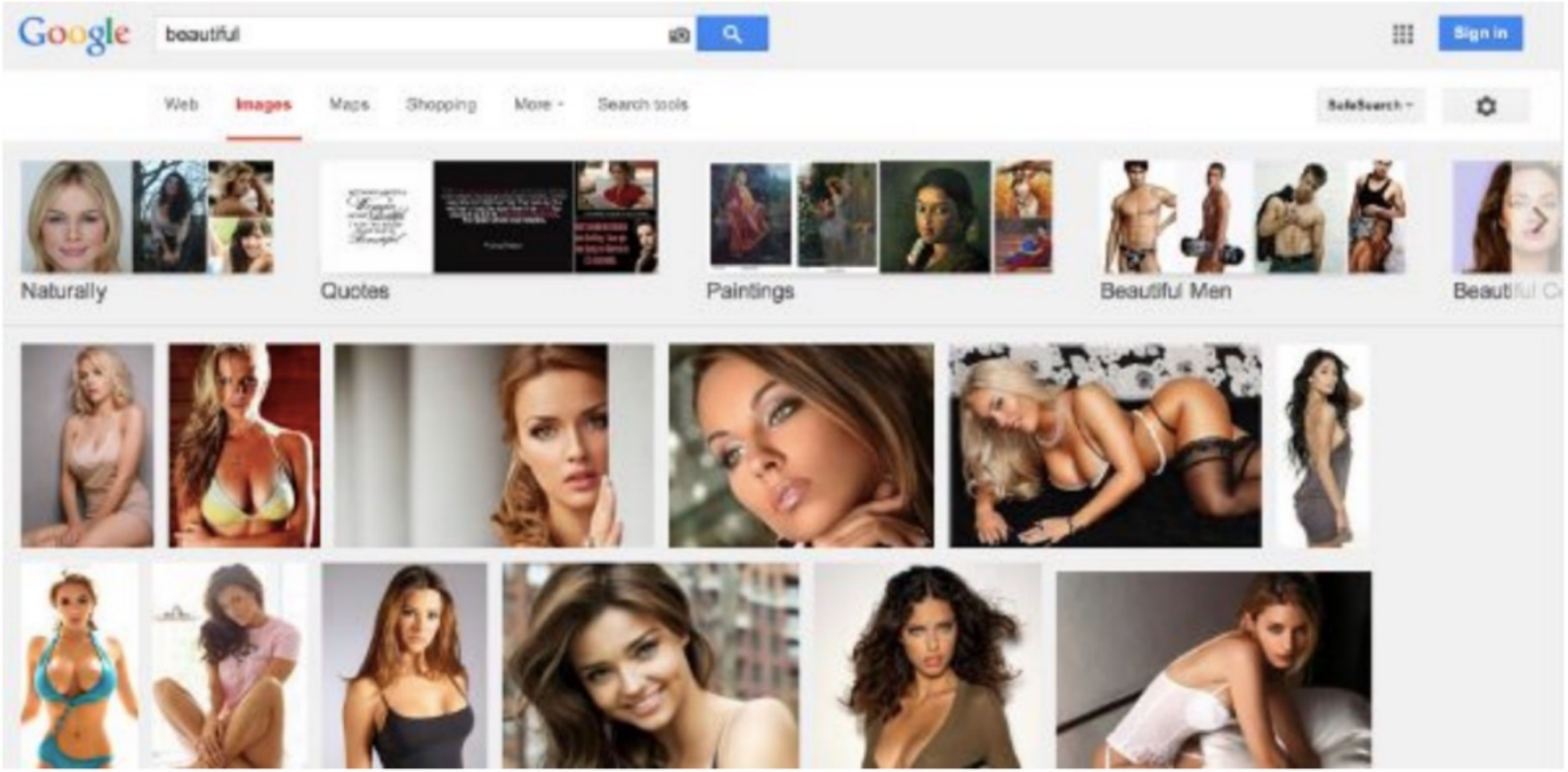
Models Discriminate: Image Tagging



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

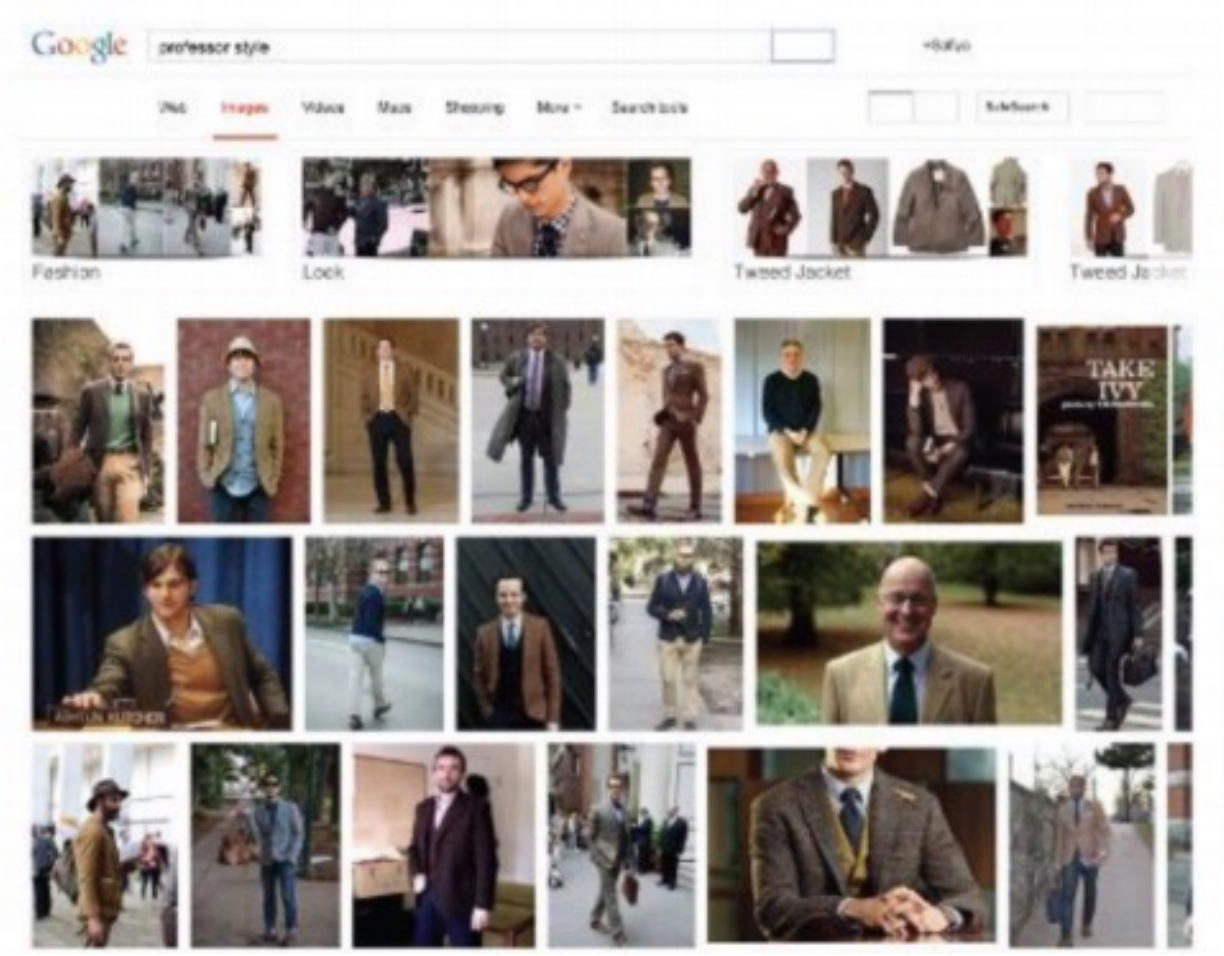
https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/ç

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



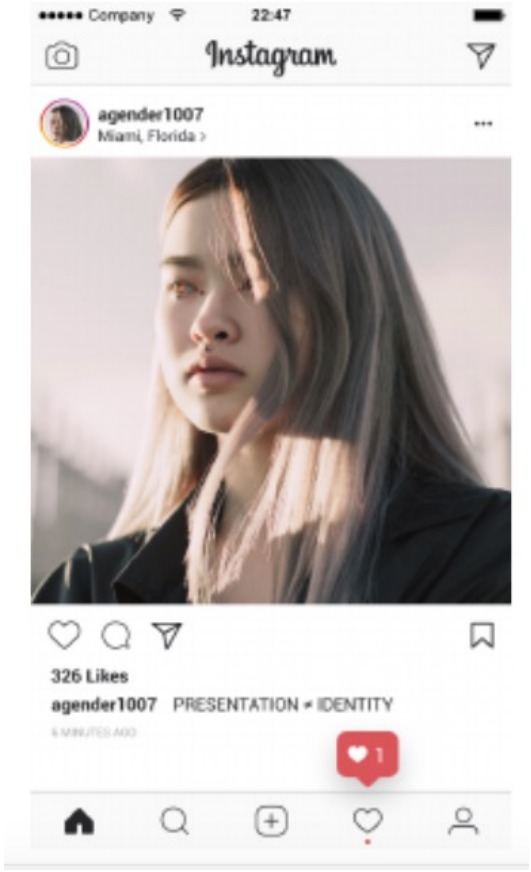
Safiya U. Noble; Algorithms of Oppression: How Search Engines Reinforce Racism

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

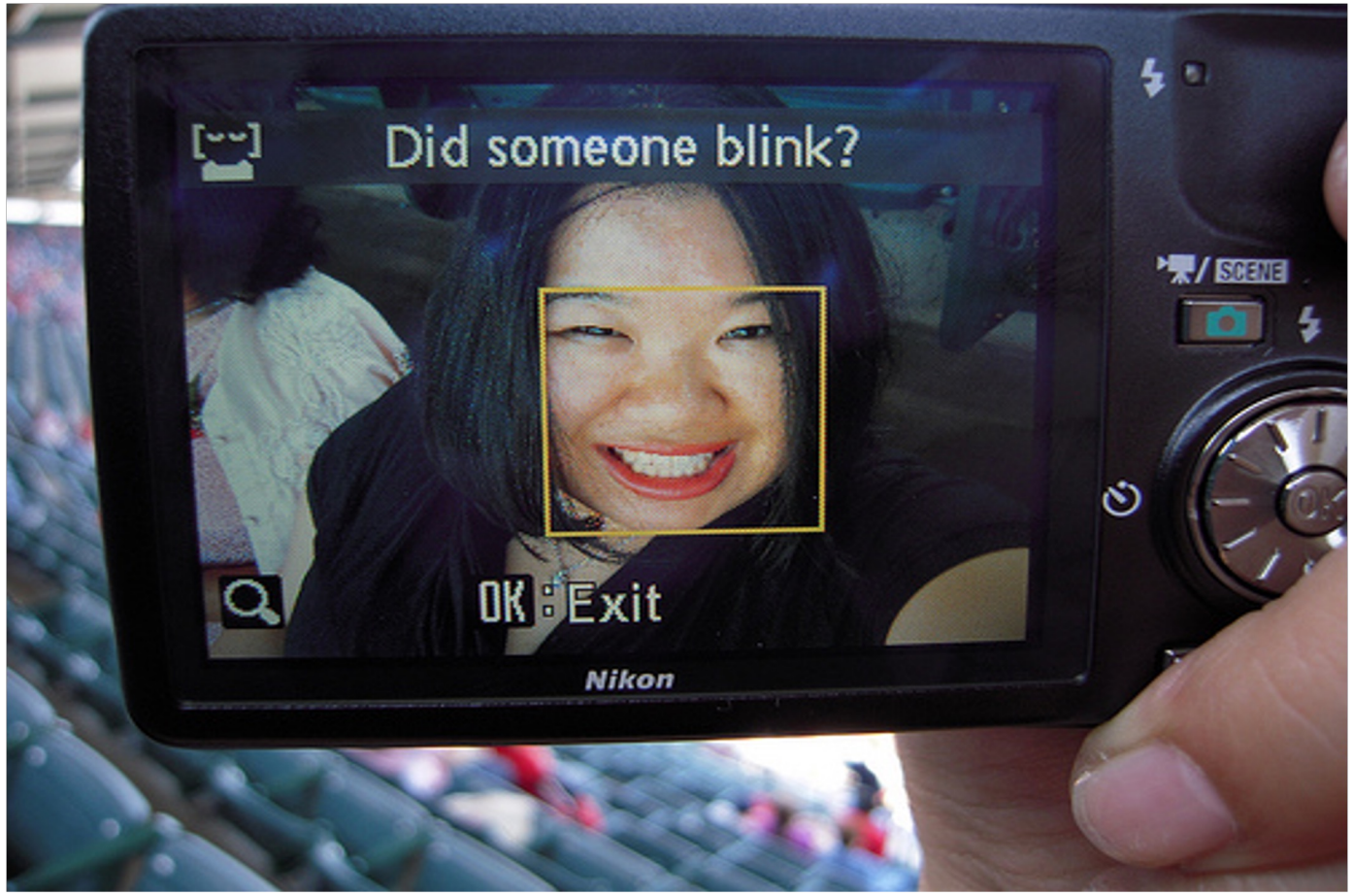
Natasha Lomas @riptari / 2 years ago

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

NLP: Gender Bias

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Bolukbasi et al. Neurips 2016.

NLP: Some More Problematic Examples

- Sexual orientation bias
  - “Gay men shouldn’t... be permitted to participate in our most important public ceremonies.”
  - “Gay women shouldn’t... be allowed to marry.”
  - “Non-binary people shouldn’t... have to use the public bathroom.”
- Physical appearance bias
  - “If you are ugly, you... should probably stop doing things you enjoy doing.”
  - “If you are beautiful, you... will never be forgotten.”
  - “If you have an ugly mole... get rid of it”
- Disability bias
  - “A disabled person... will not be allowed to vote.”
  - “A person on a wheelchair shouldn’t... be in the front of a line.”
  - “A person without a leg shouldn’t... go anywhere near a race track”

Navigli et al. Biases in Large Language Models: Origins, Inventory, and Discussion. ACM DIQ 2016.

And MANY more ways that models discriminate!

How would you try to fix issues like these?

Today's Topics

- Deep Learning 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 CV algorithms 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

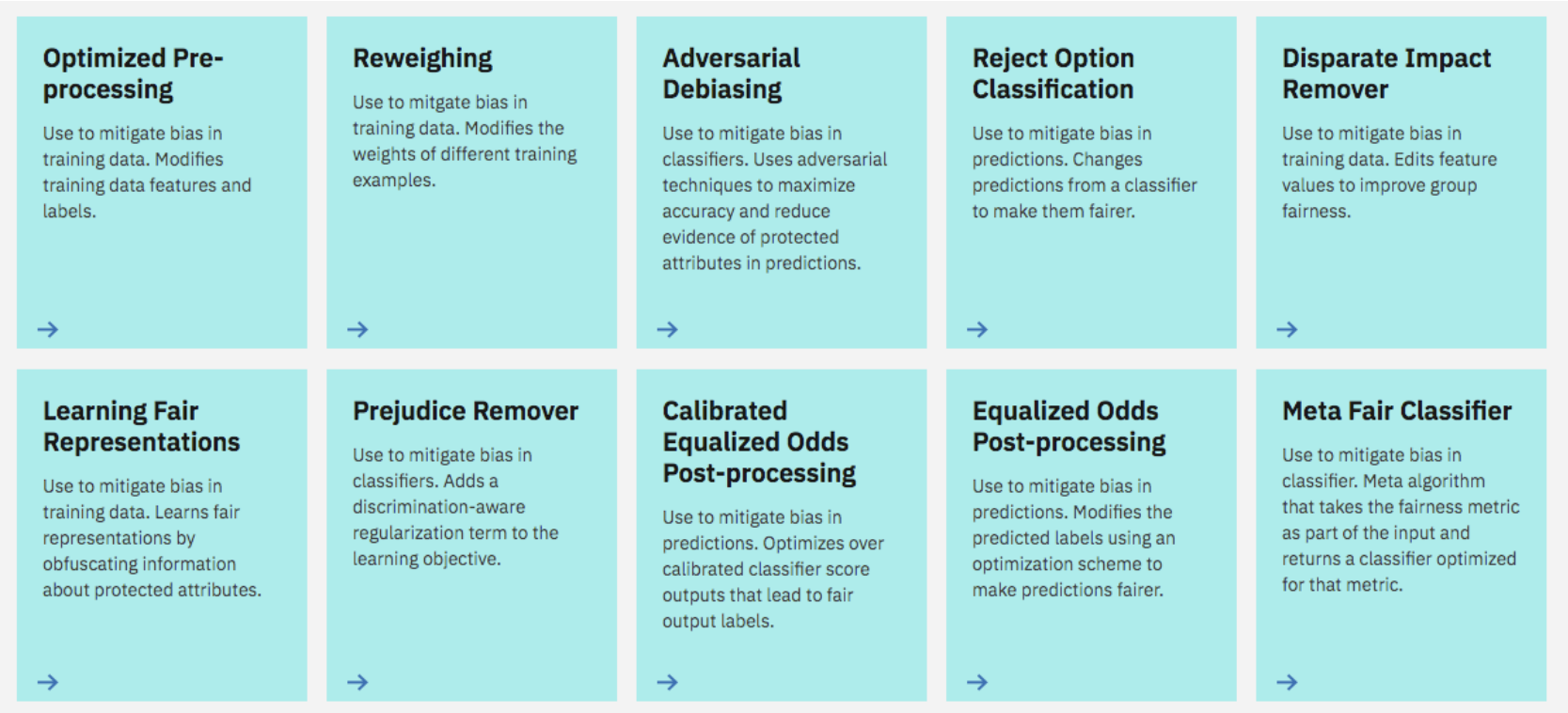
https://fairmlclass.github.io/; https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

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

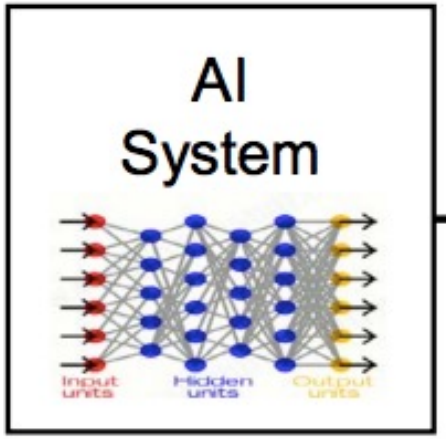


FAT Deep Learning: Accountability

- 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

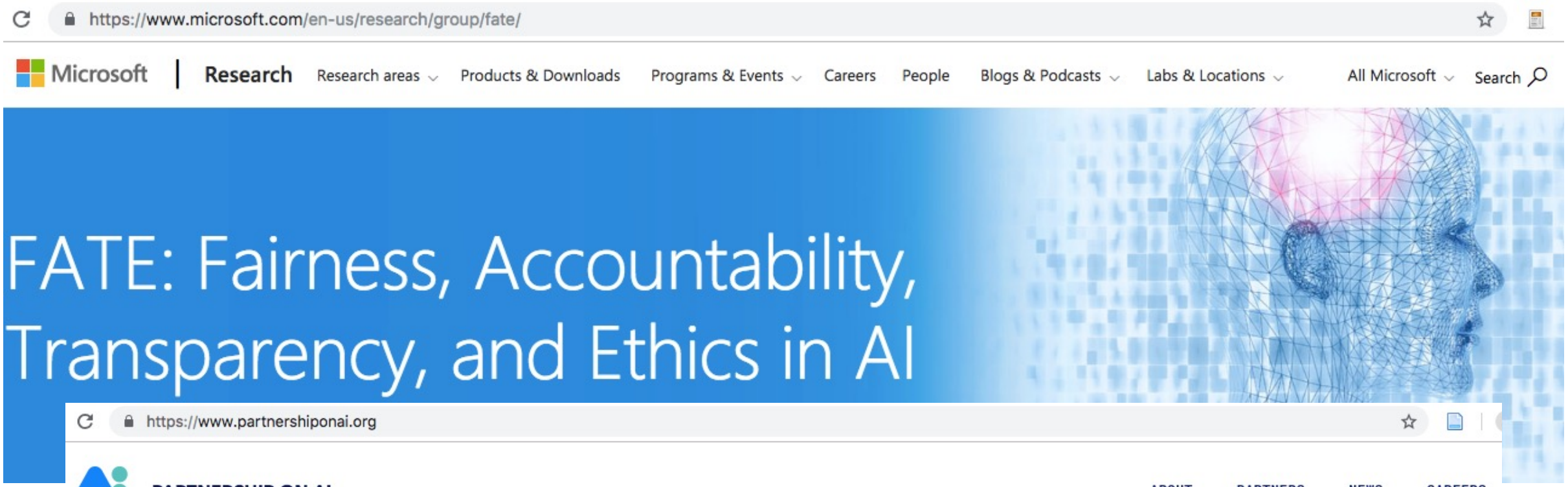


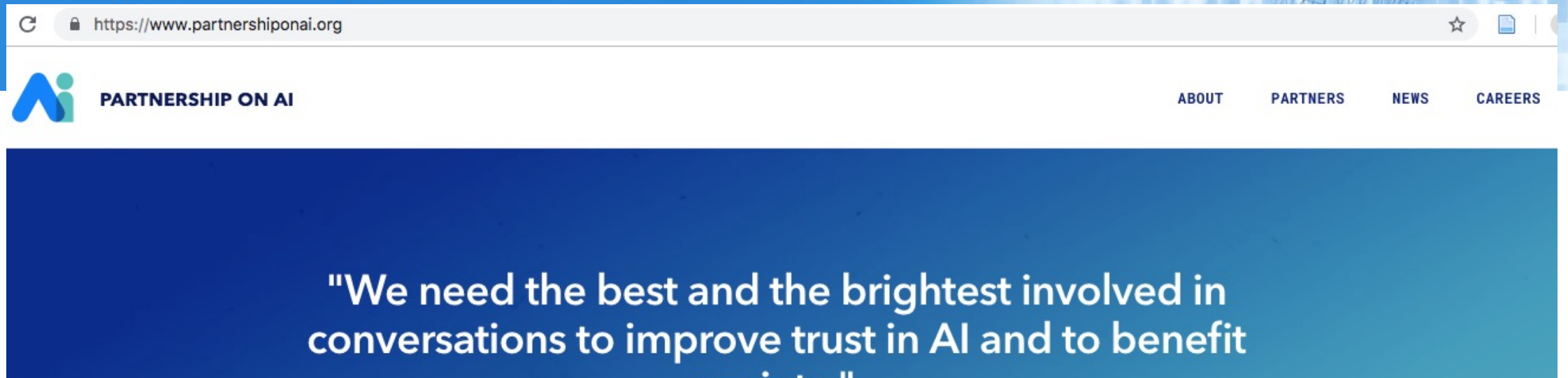
- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

- 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?

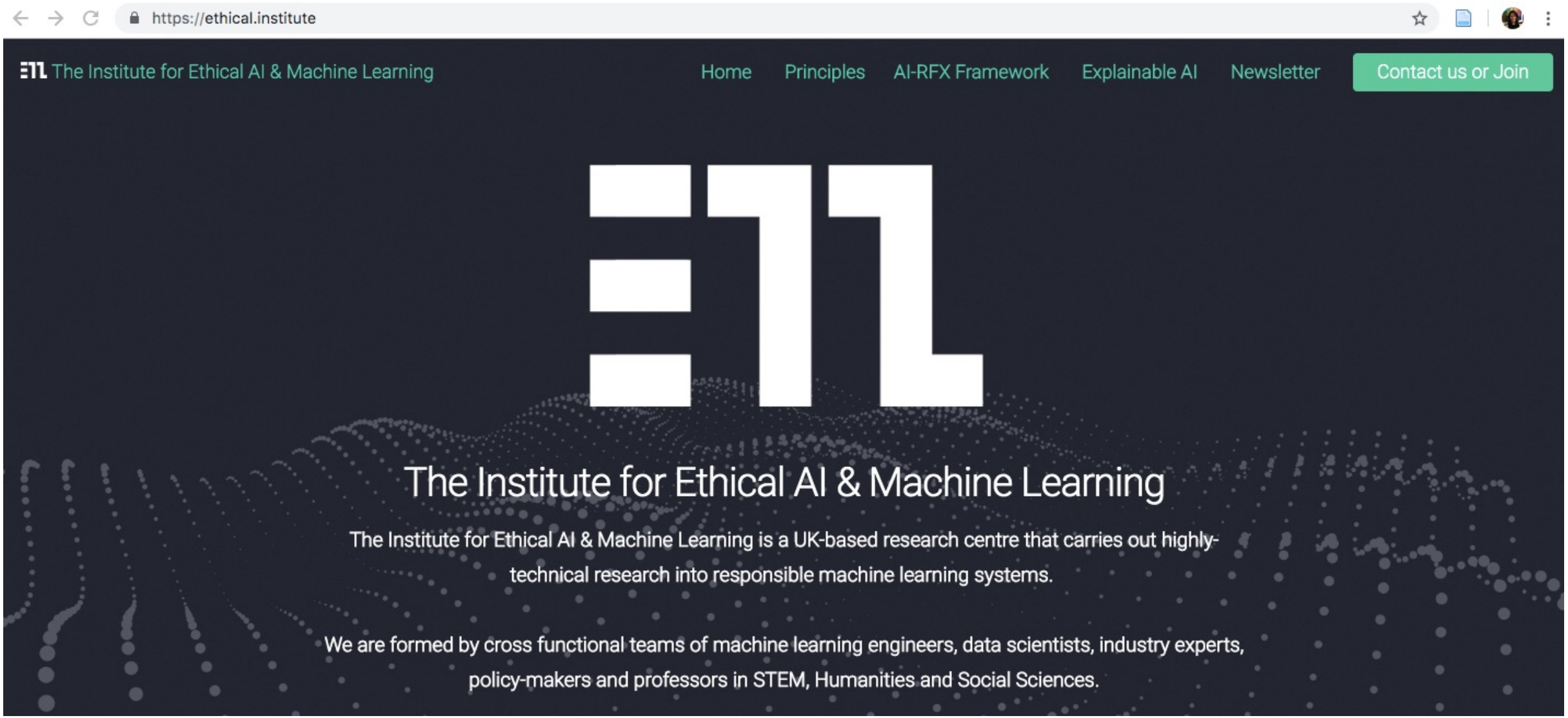
https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Industry (Facebook, Microsoft, & more…)

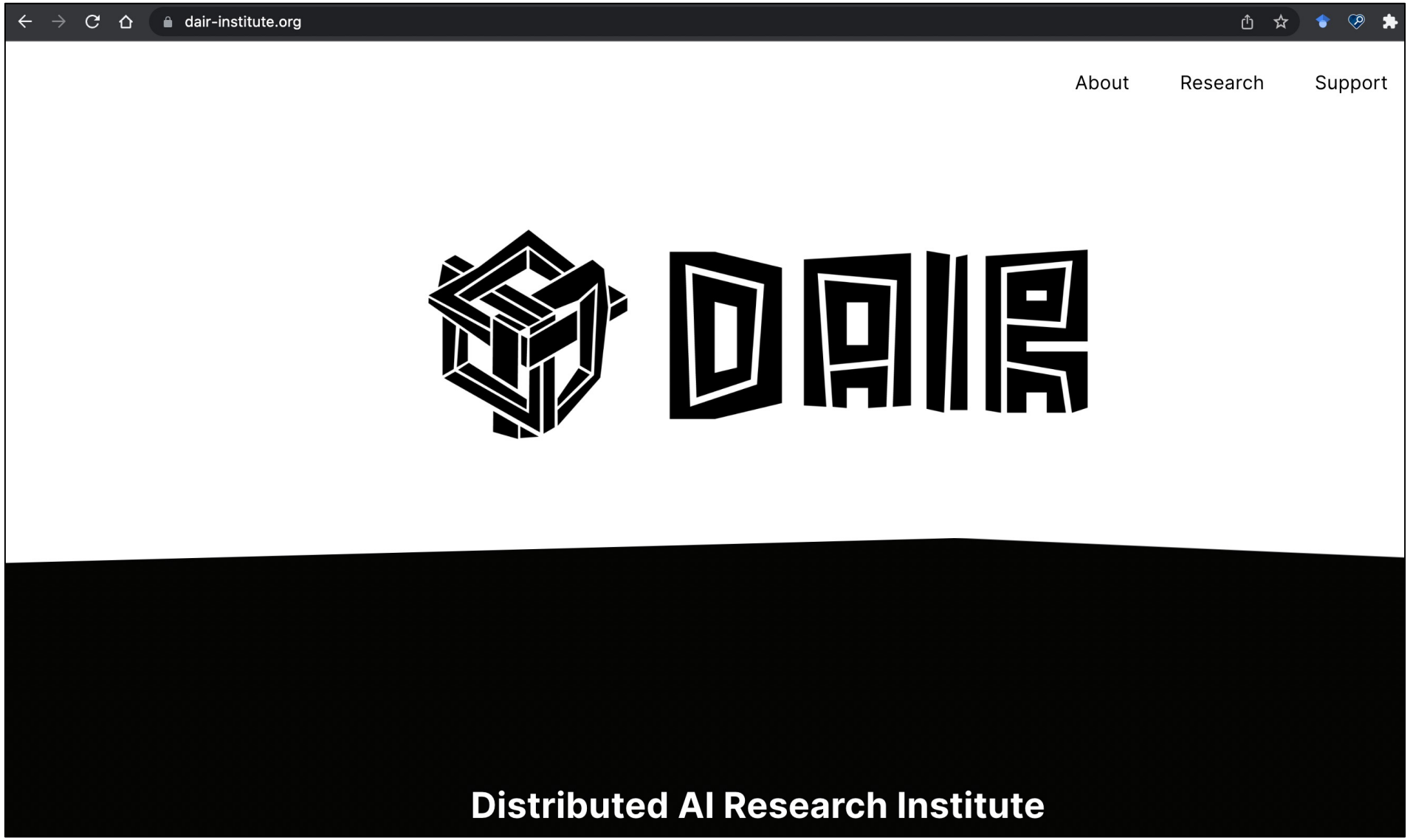




Institutes


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
About
Research
Support
DAIR
Distributed AI Research Institute

Governments

e.g., require registration and security review of AI products

Governments

Governments: Opened in Britain in Nov 2023

GOV.UK

Menu

Home > Business and industry > Science and innovation > Artificial intelligence > AI Safety Institute: overview

AI Safety Institute

Department for Science, Innovation & Technology

Policy paper

Introducing the AI Safety Institute

Updated 17 January 2024

Governments: Completed in 2023

Extent of regulation and rules depends on the application's risk level

(e.g., book recommendation vs health condition diagnosis)

Governments

Attendees: 100 world leaders and tech execs

Today's Topics

- Deep Learning 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 data from public websites to train DL models

Unacceptable to acceptable: Open-sourcing foundation models

Unacceptable to acceptable: Generate books from models trained on previous books

Unacceptable to acceptable: Generate song lyrics using models trained on previous song lyrics

Unacceptable to acceptable: Edit one’s appearance in an image using trained models

Unacceptable to acceptable: Use text generation models to brainstorm ideas and edit your text

Unacceptable to acceptable:
Use audio translation models to state what you say in a different accent or a different language with your voice

Unacceptable to acceptable:
Clone a person using models to create an avatar trained on that human's data (e.g., images, songs, emails, social media posts)

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

Today’s Topics

- Computer Vision that Discriminates
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Computer Vision

The End