Responsible Computer Vision: Part 1

Danna Gurari

University of Colorado Boulder

Fall 2024

Review

- Last lecture on efficient computer vision:
  - Motivation
  - Model Compression
  - Curriculum Learning
  - Active Learning
- Assignments:
  - Project presentation poster due on Monday
  - Project presentation due in 1 week
  - Peer evaluation due in 1 week (in-class activity)
  - Project report due in 2 weeks
- Questions?

Today’s Topics

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

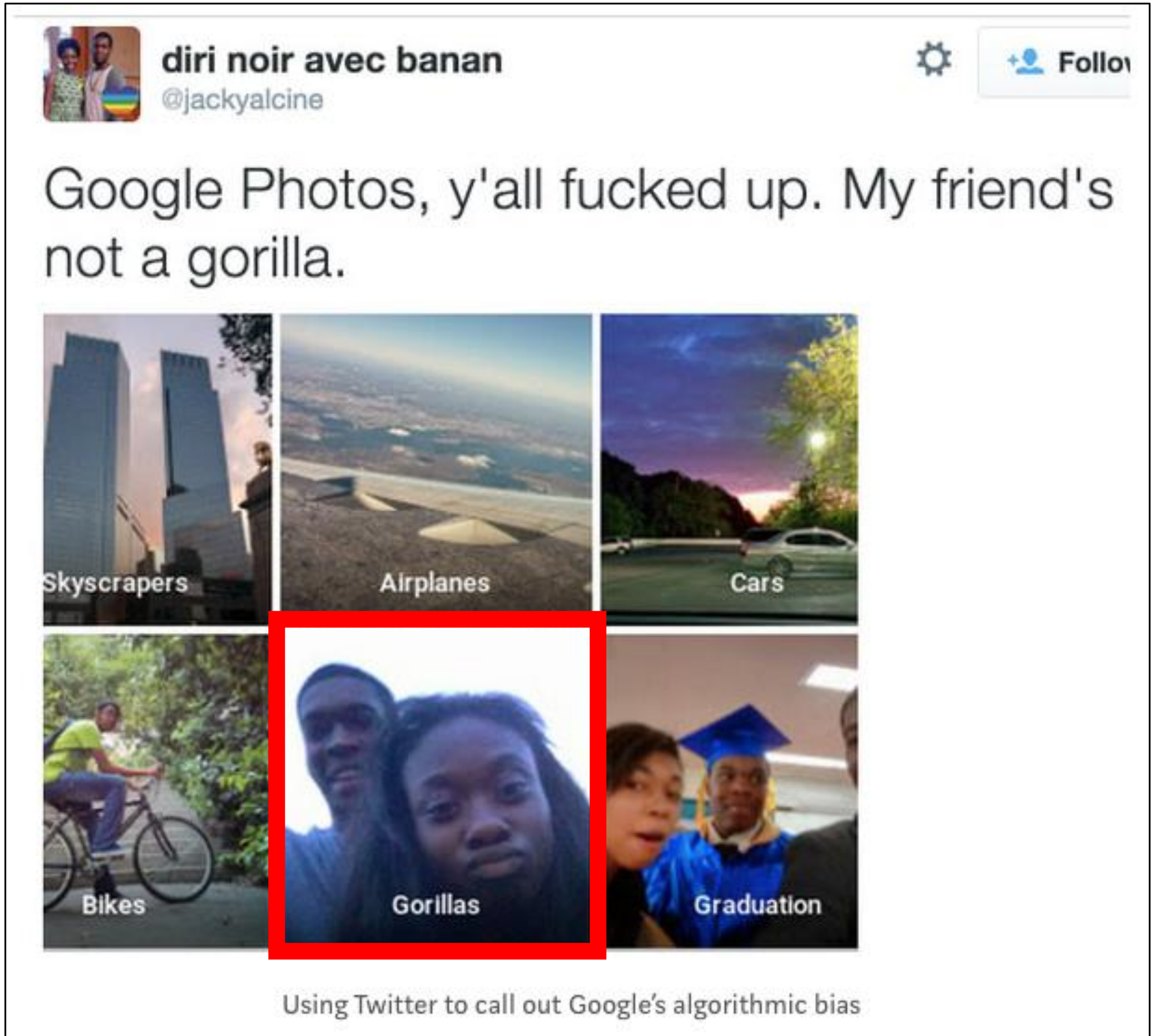
Today’s Topics

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

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



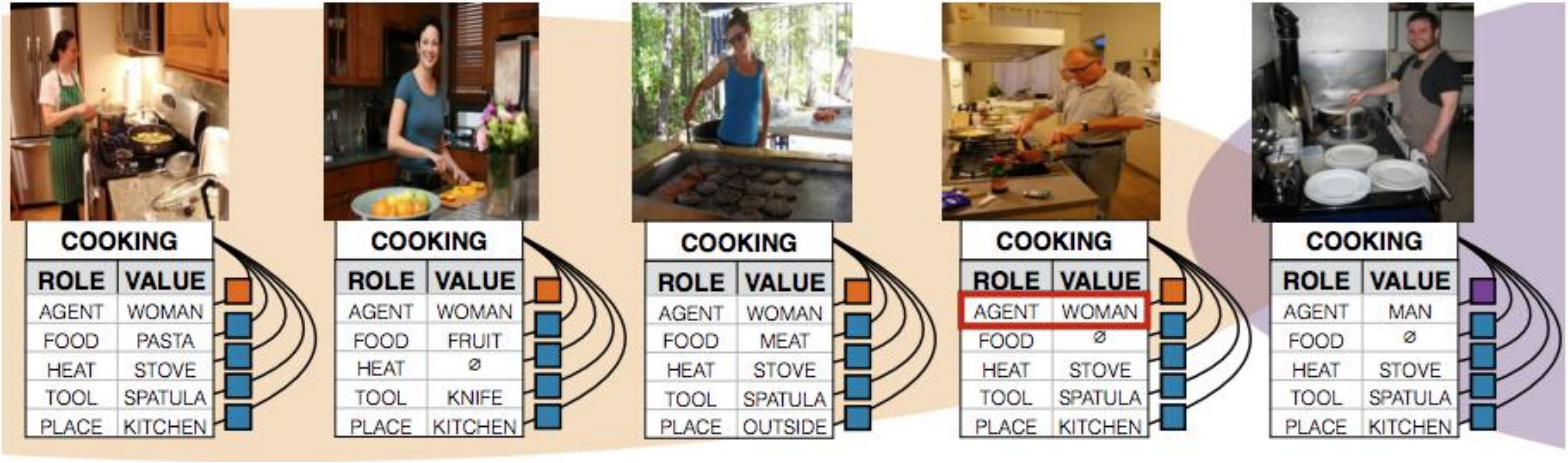
diri noir avec banan
@jackyalcine

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

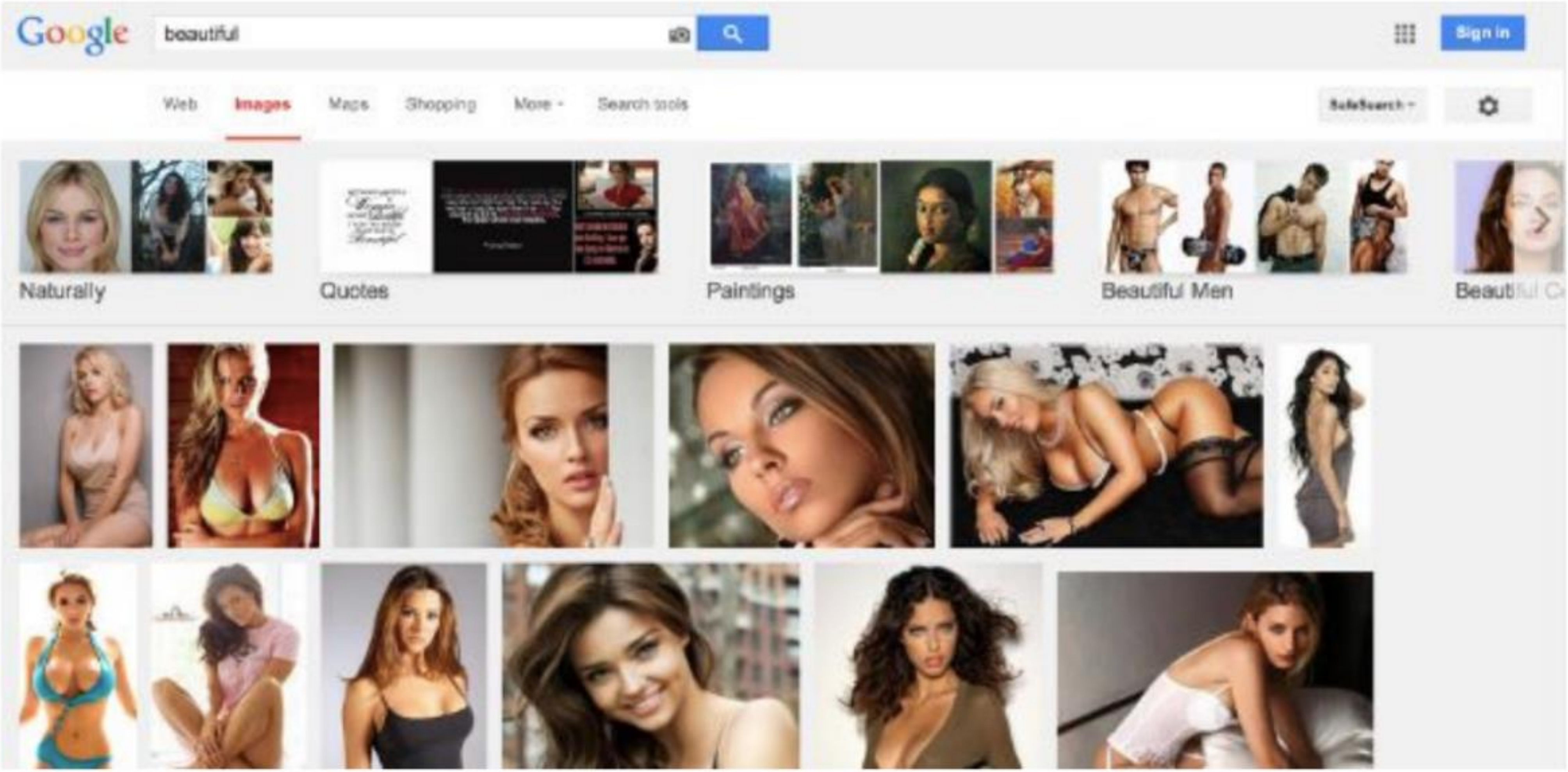
COOKING

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)

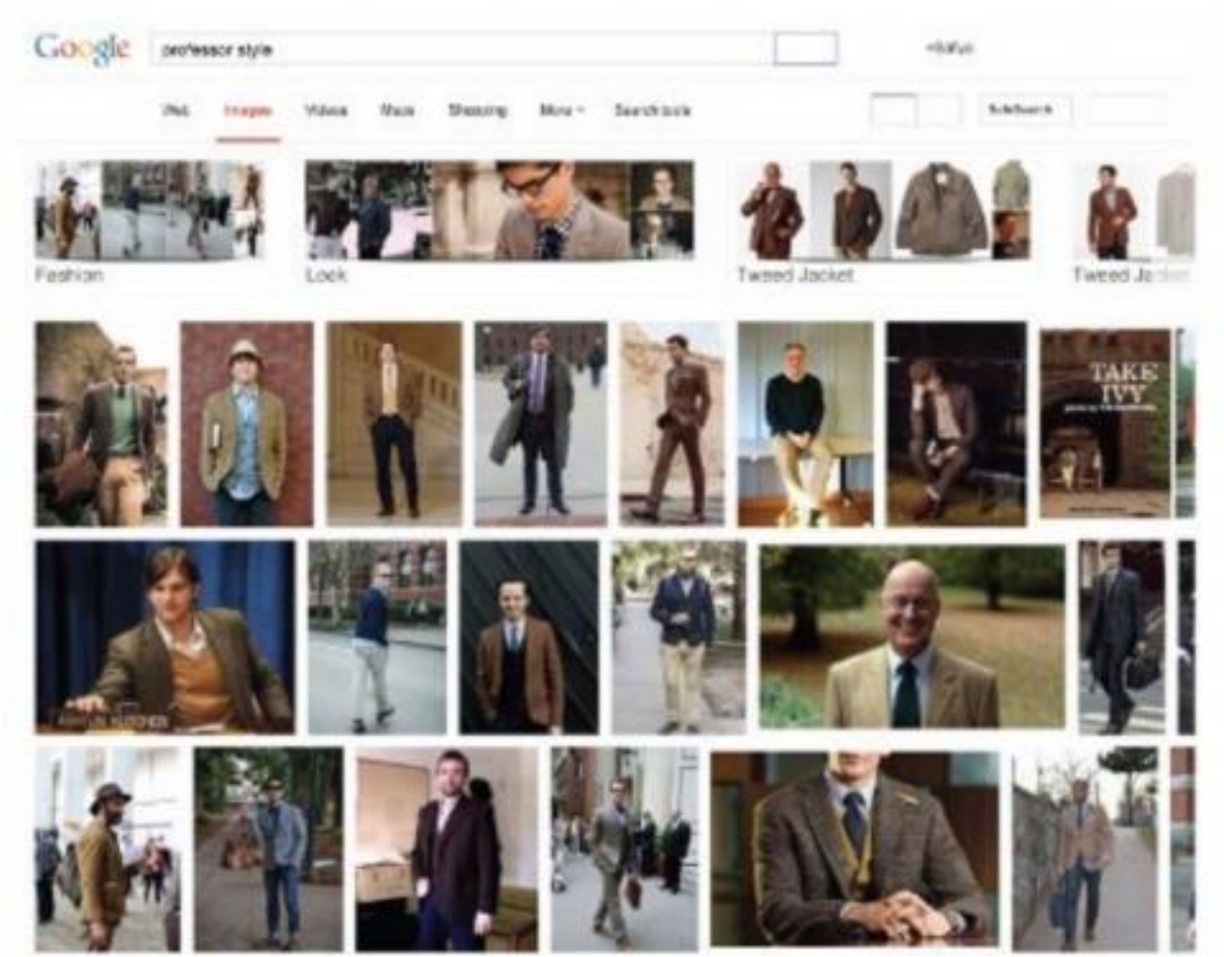
https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/ç

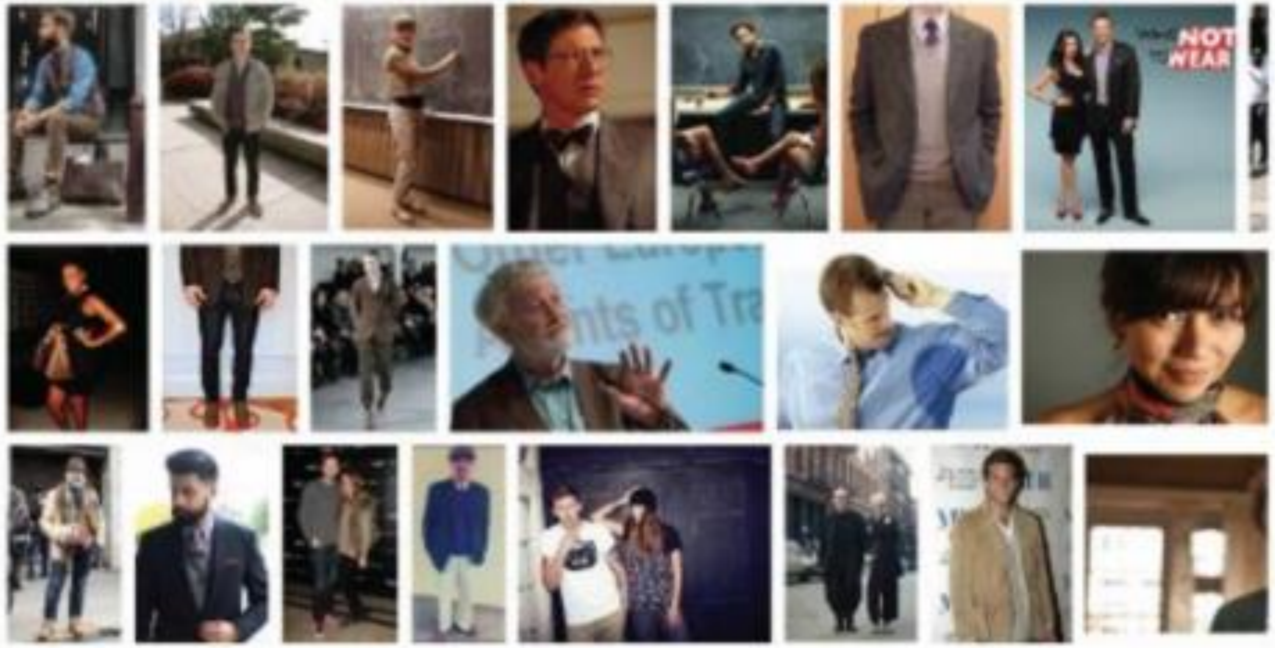
Models Discriminate: Image Tagging (“beautiful”; 2014)



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

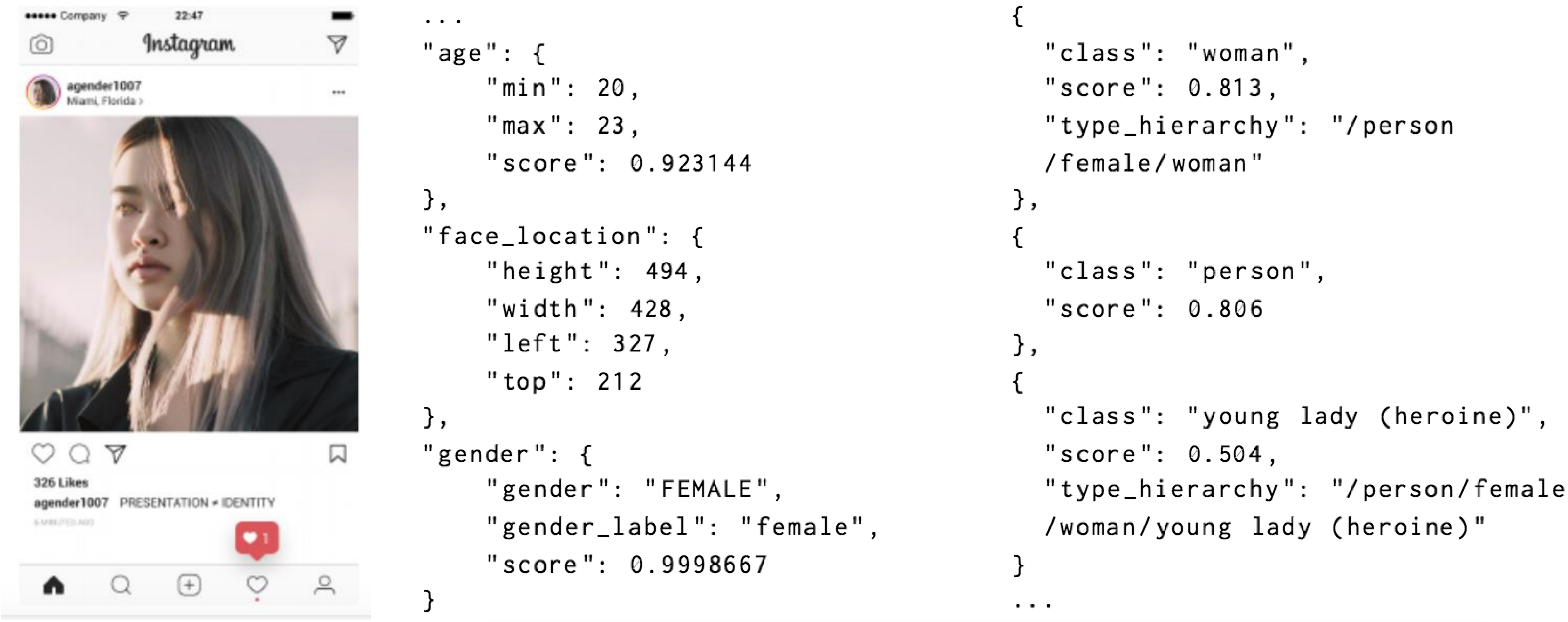
Models Discriminate: Image Tagging ("professor style"; 2014)





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

Models Discriminate: Image Tagging



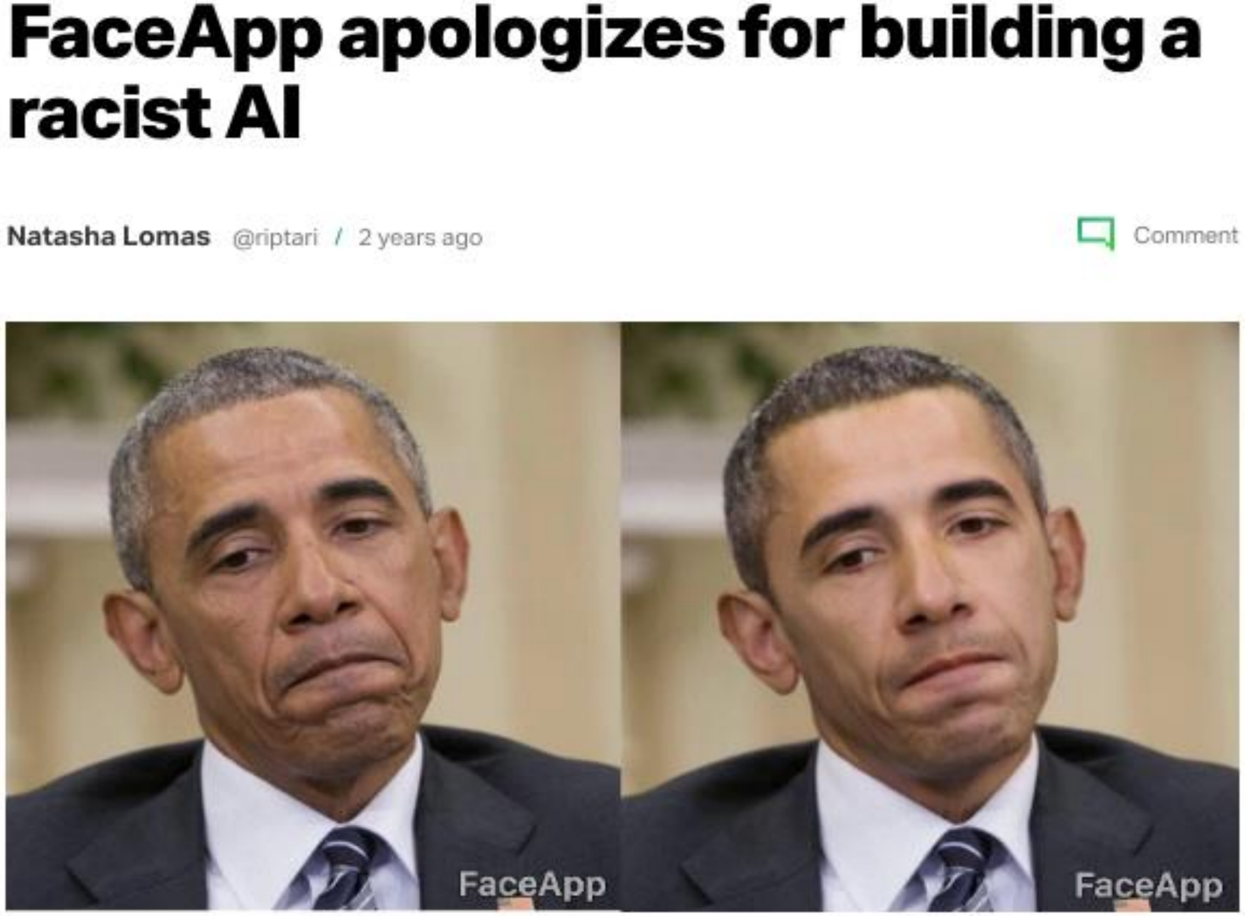
```
...
"age": {
    "min": 20,
    "max": 23,
    "score": 0.923144
},
"face_location": {
    "height": 494,
    "width": 428,
    "left": 327,
    "top": 212
},
"gender": {
    "gender": "FEMALE",
    "gender_label": "female",
    "score": 0.9998667
}
```

```
{
  "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

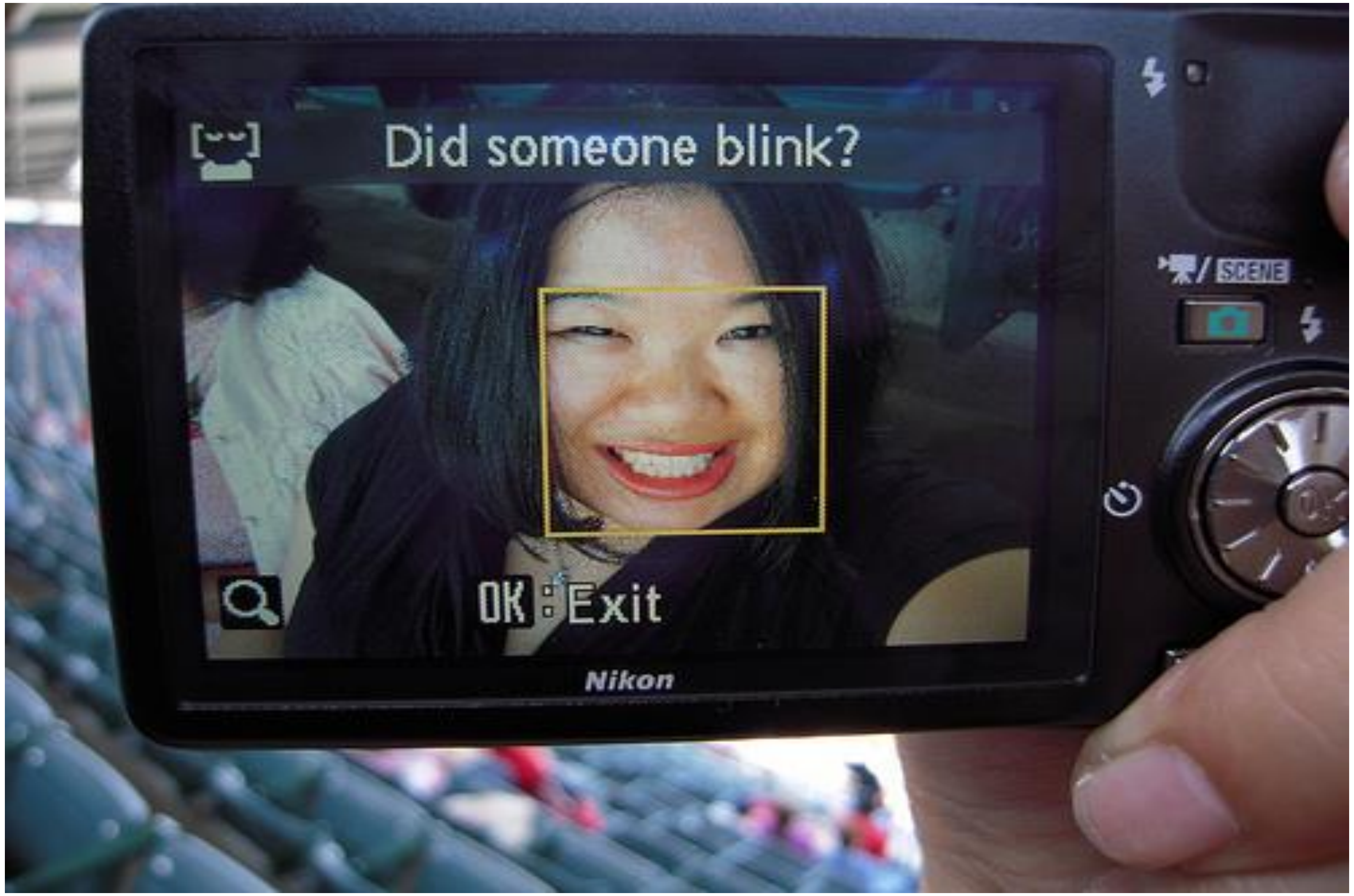
Natasha Lomas @riptari / 2 years ago

Comment

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

Models Discriminate: Nikon Blink Detection

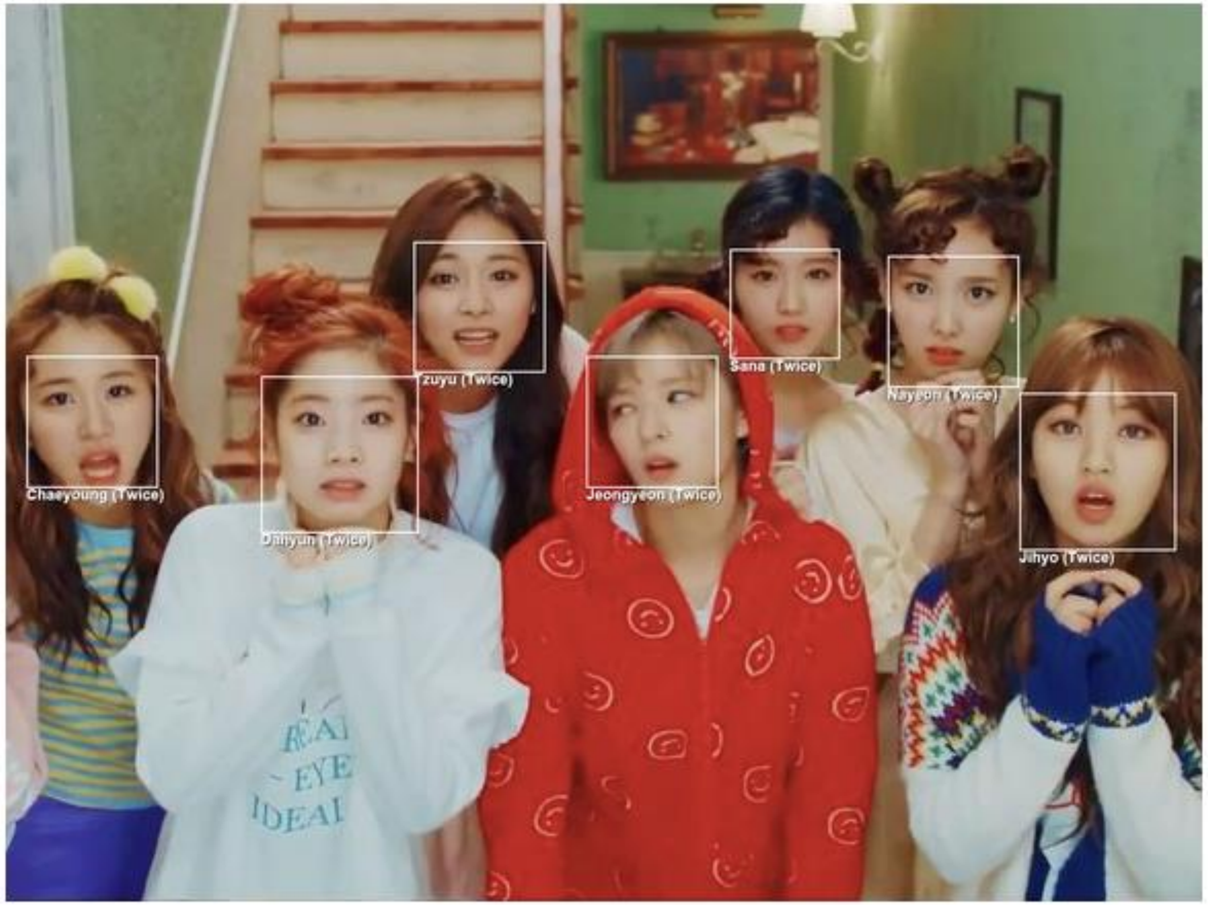
Two kids bought their mom a Nikon Coolpix S630 digital camera for Mother's Day… when they took portrait pictures of each other, a message flashed across the screen asking, "Did someone blink?"



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/

And MANY more ways that models discriminate!

How would you try to fix issues like these?

Today’s Topics

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

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

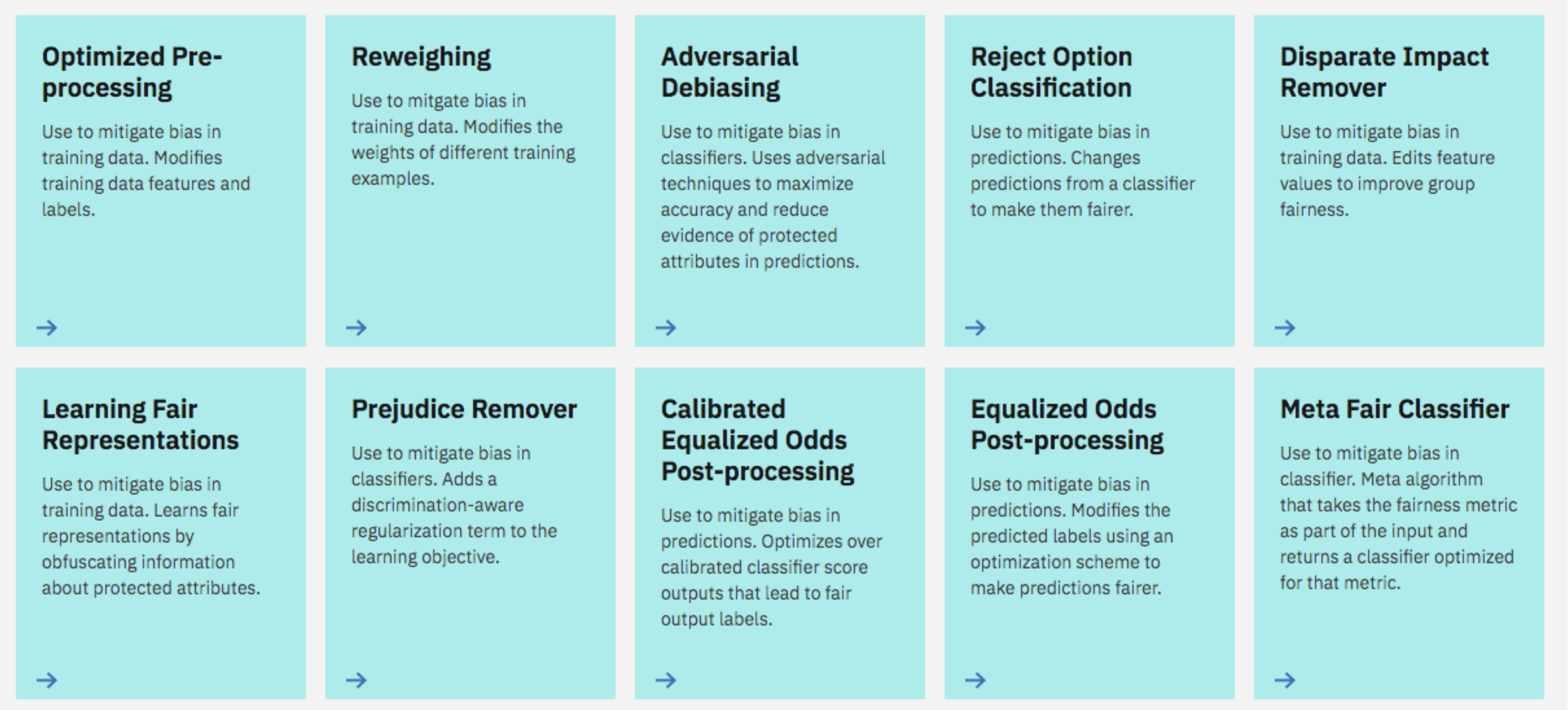
https://fairmlclass.github.io/; https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

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



Optimized Pre-processing

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

Reweighing

Use to mitgate 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

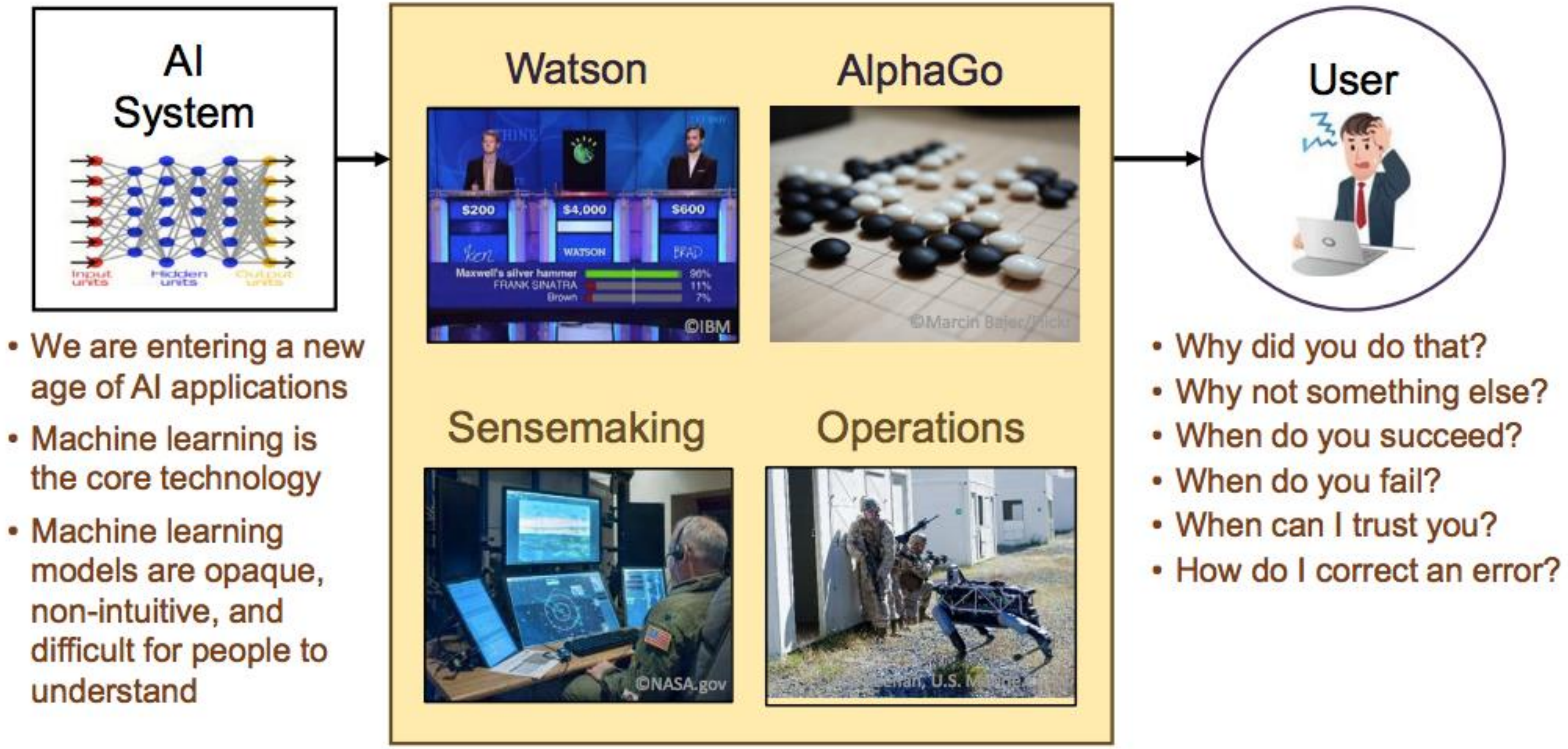
Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.

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

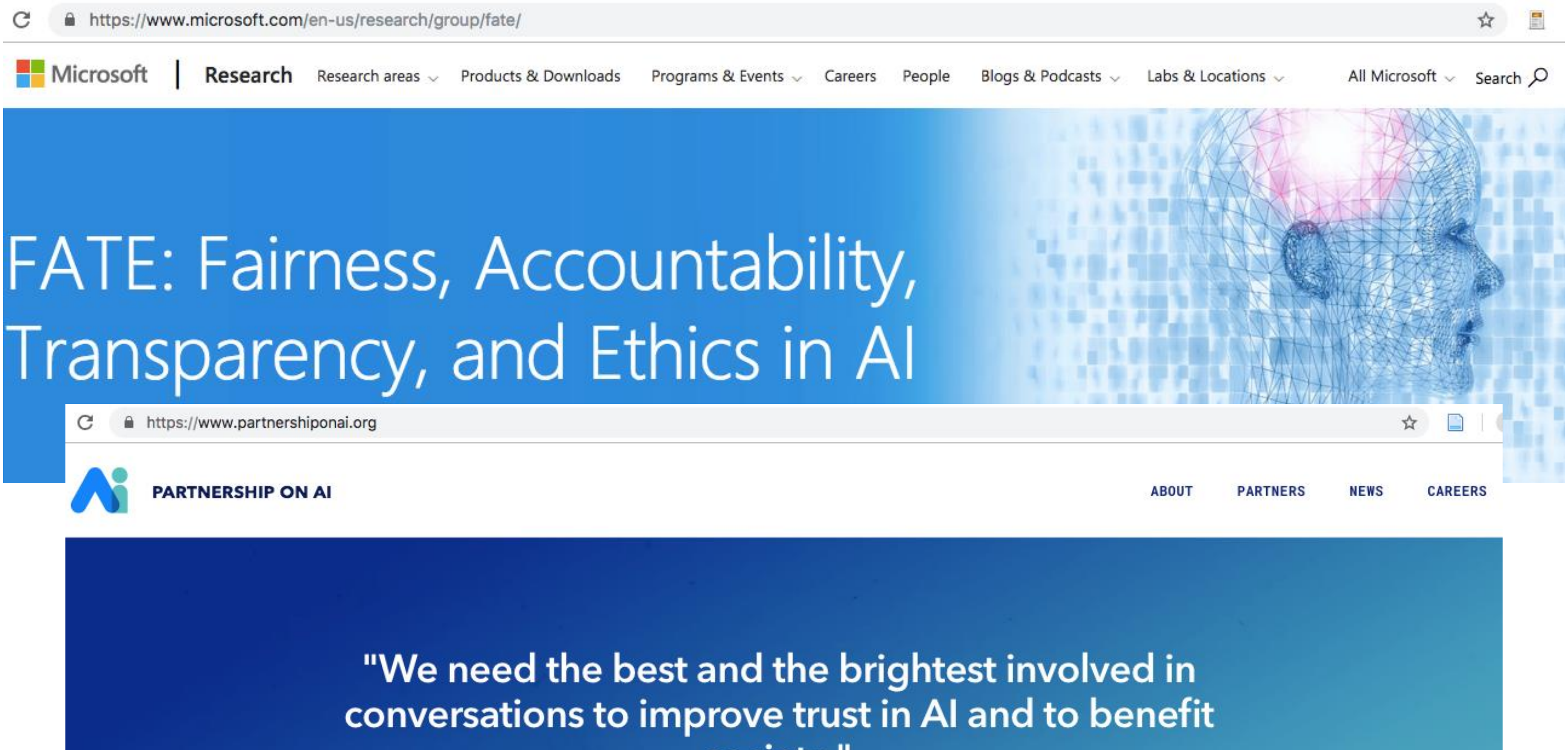


- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

- 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…)



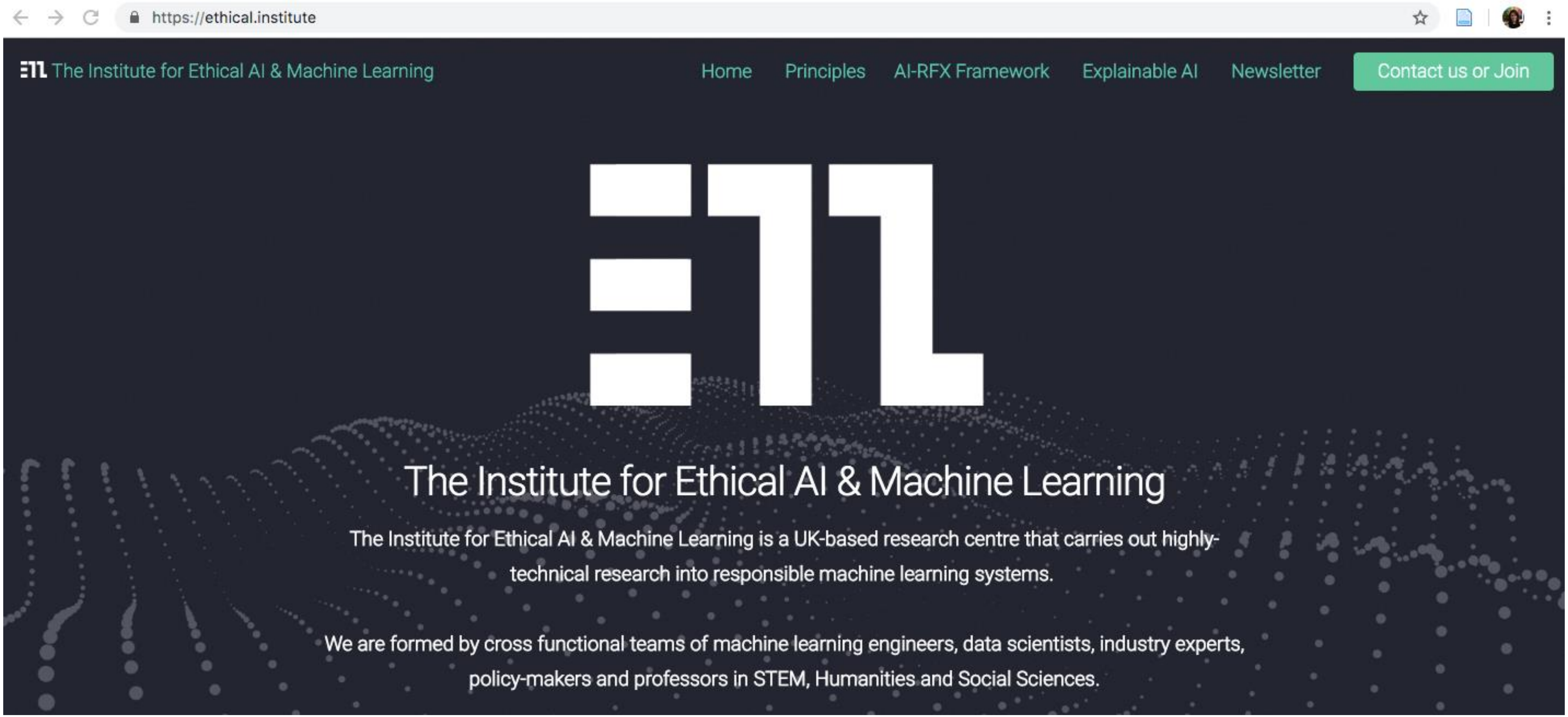
https://www.microsoft.com/en-us/research/group/fate/

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



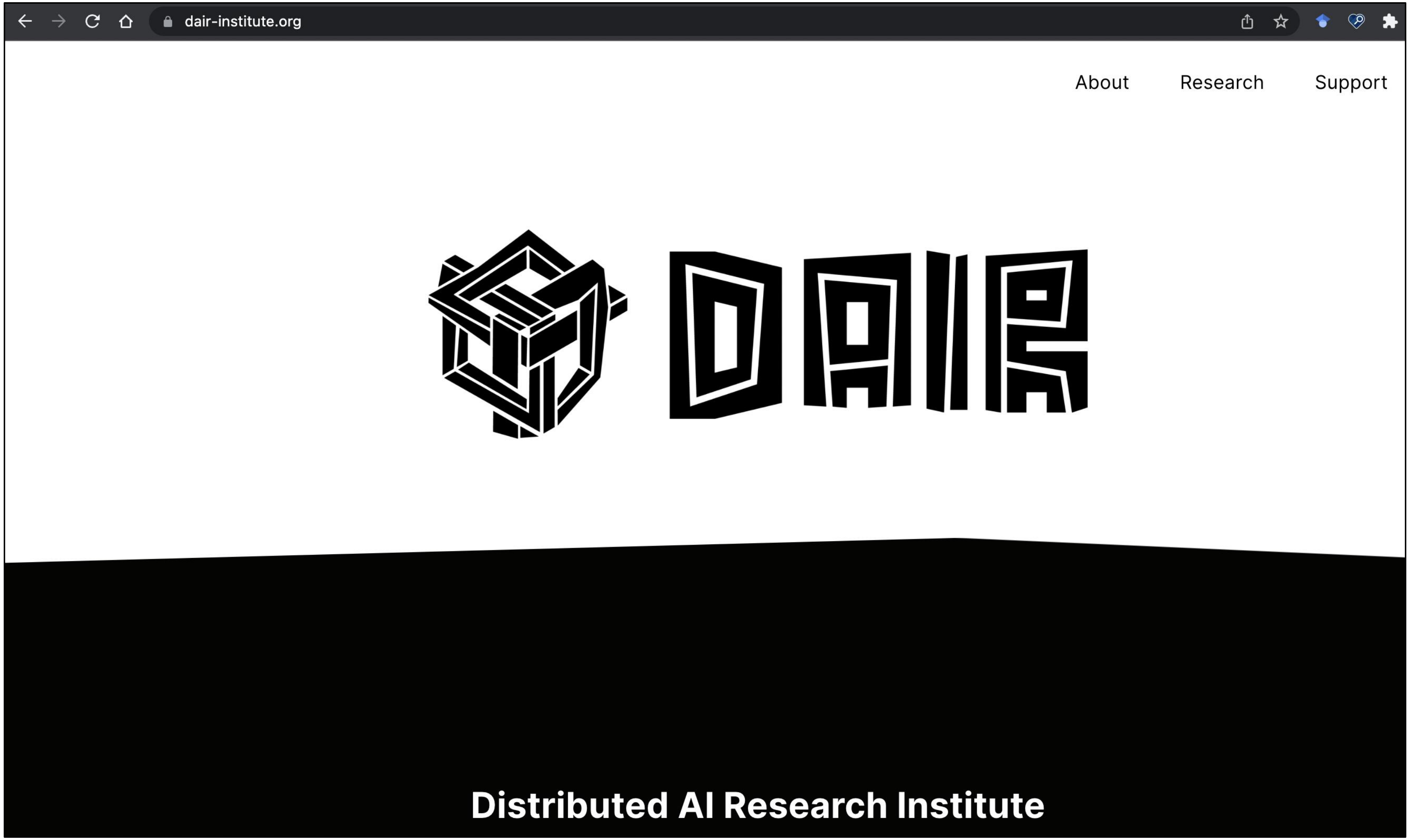
https://ethical.institute
ΞΊL The Institute for Ethical AI & Machine Learning
Home
Principles
AI-RFX Framework
Explainable AI
Newsletter
Contact us or Join

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

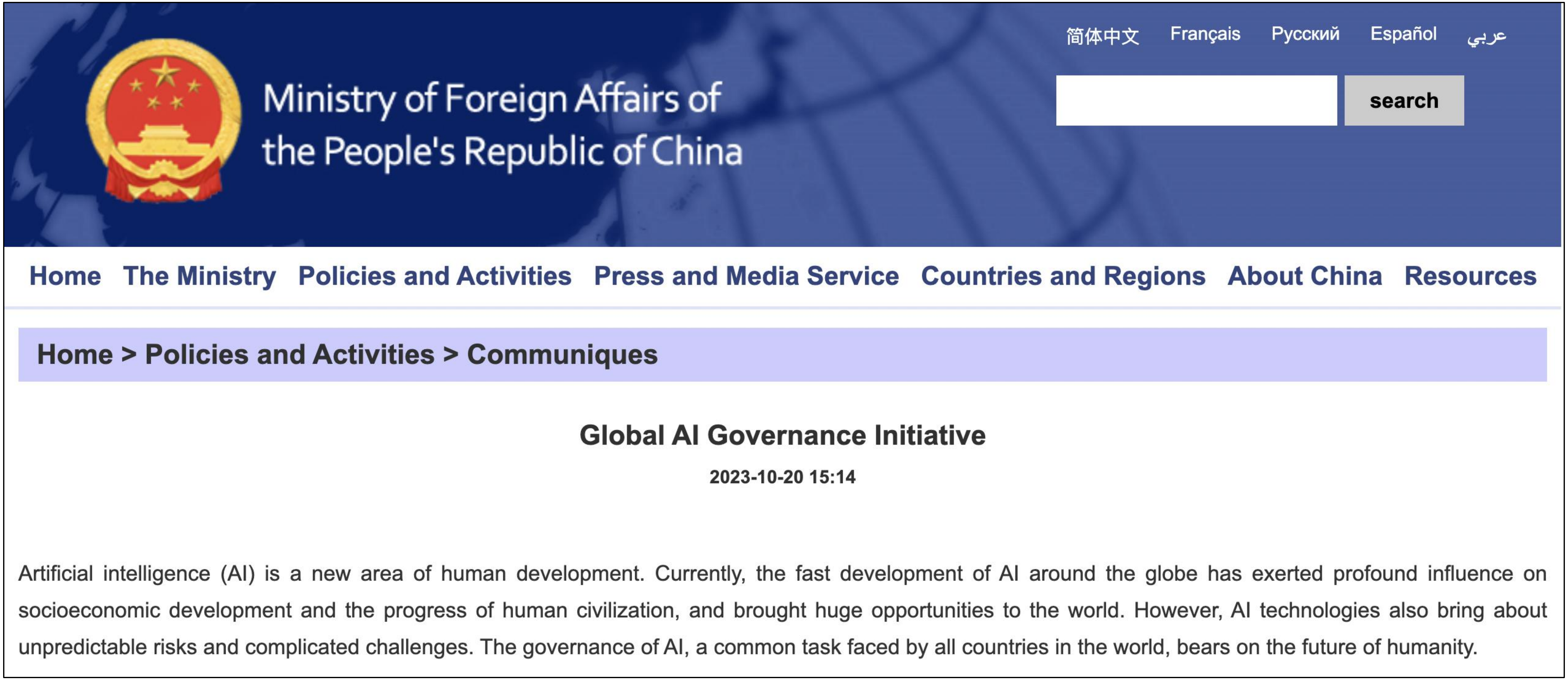
About Research Support

DAIR

Distributed AI Research Institute

Governments

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

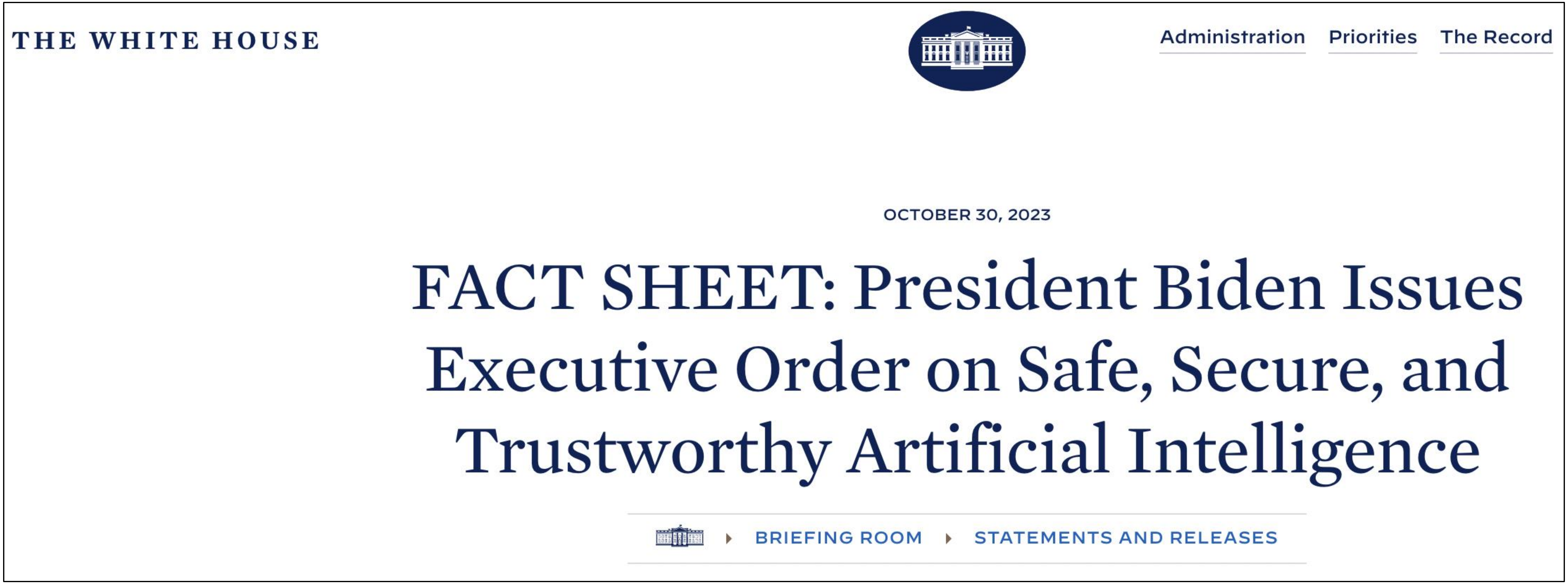


Global AI Governance Initiative

2023-10-20 15:14

Artificial intelligence (AI) is a new area of human development. Currently, the fast development of AI around the globe has exerted profound influence on socioeconomic development and the progress of human civilization, and brought huge opportunities to the world. However, AI technologies also bring about unpredictable risks and complicated challenges. The governance of AI, a common task faced by all countries in the world, bears on the future of humanity.

Governments

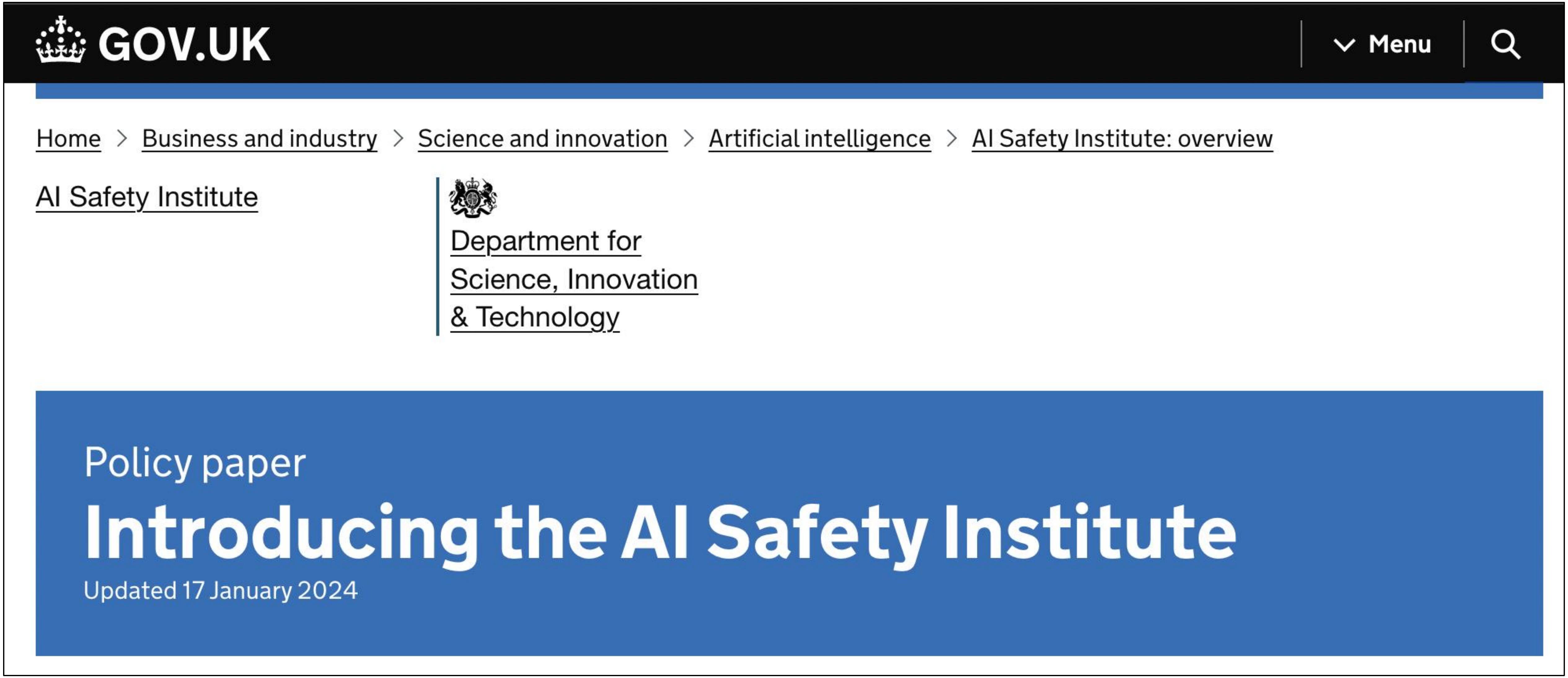


THE WHITE HOUSE

OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

Governments: Opened in Britain in Nov 2023



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

AI Safety Institute

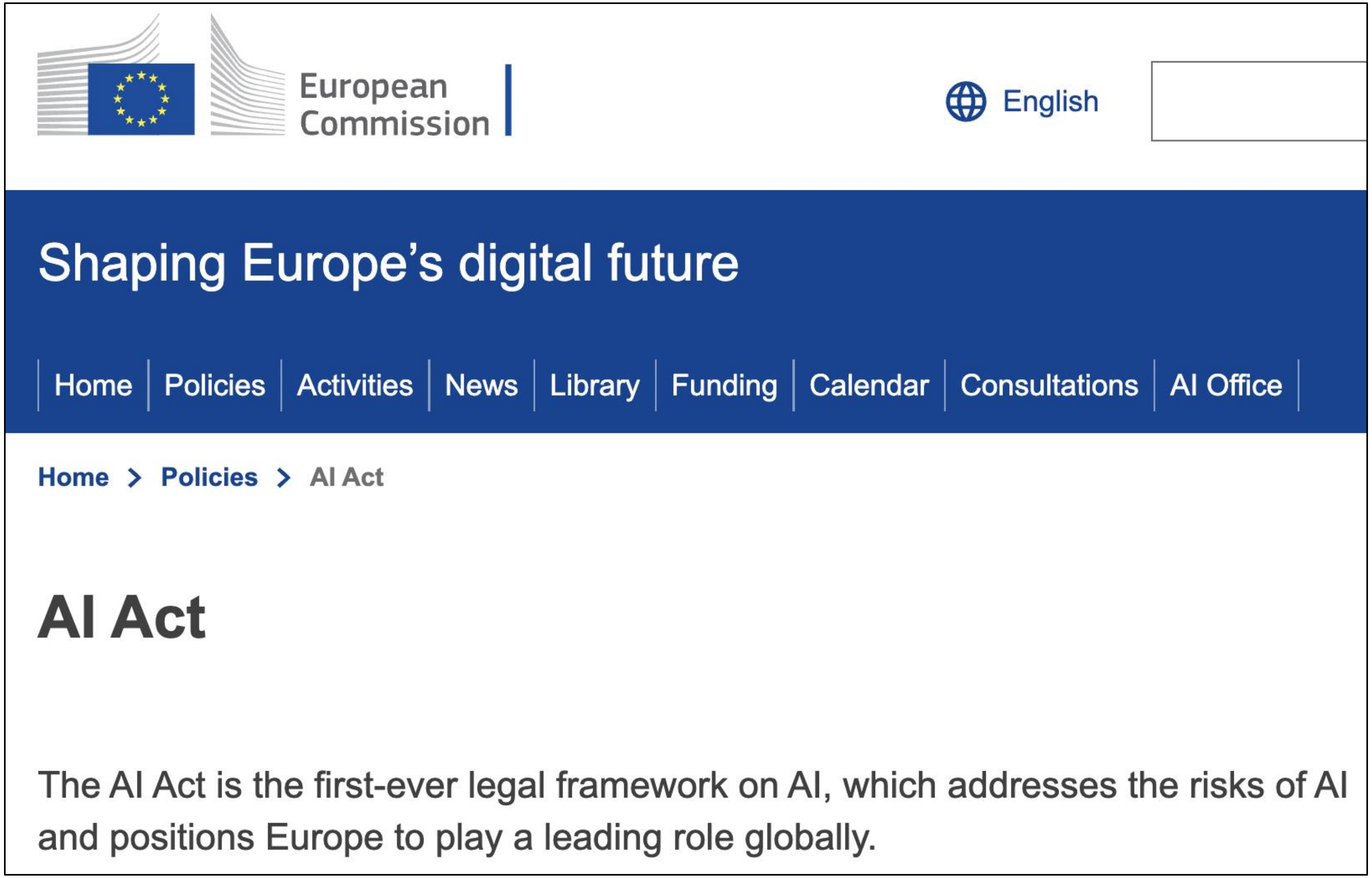
Department for Science, Innovation & Technology

Policy paper

Introducing the AI Safety Institute

Updated 17 January 2024

Governments: Completed in 2023



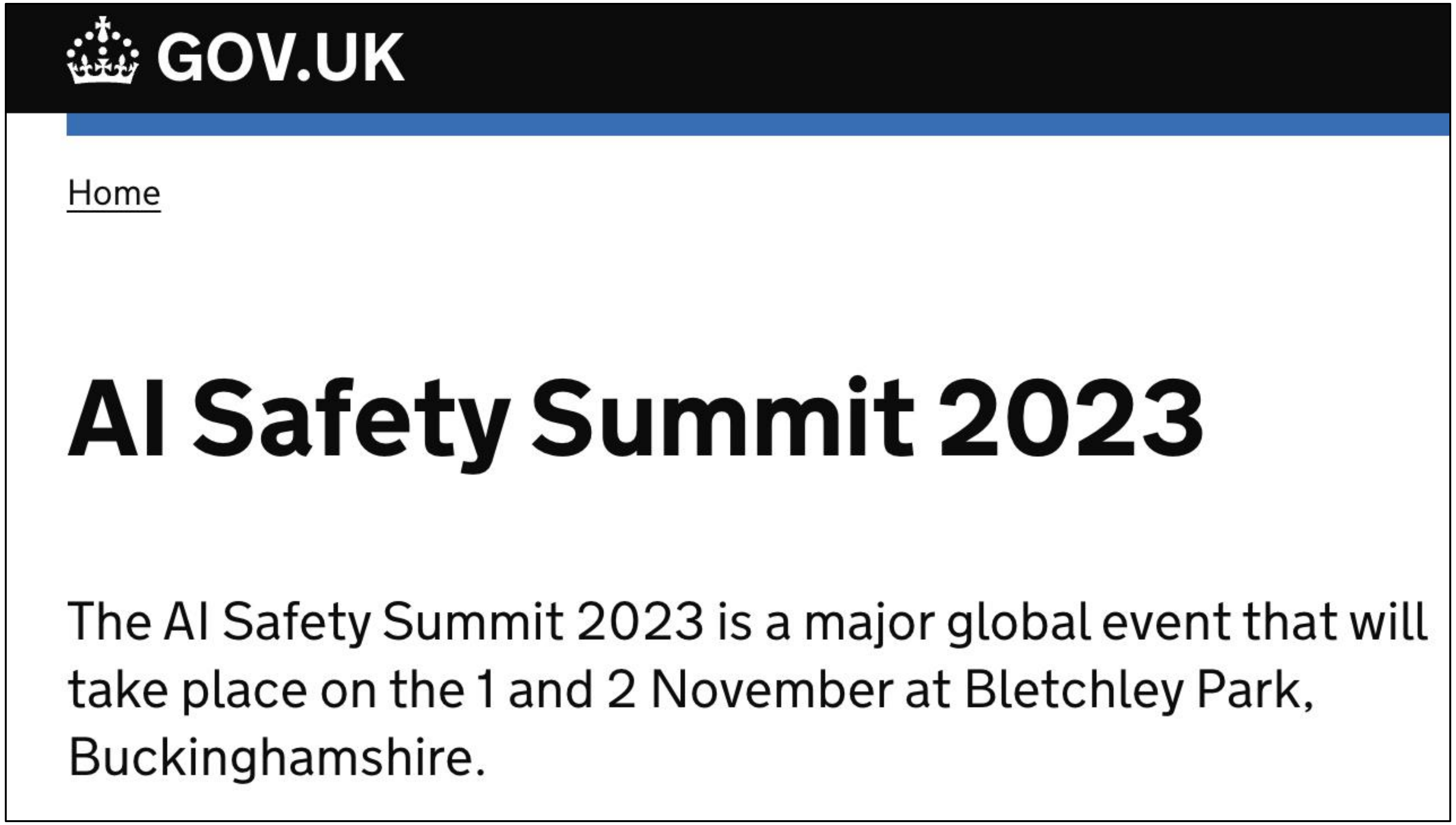
AI Act

The AI Act is the first-ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.

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

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

Governments

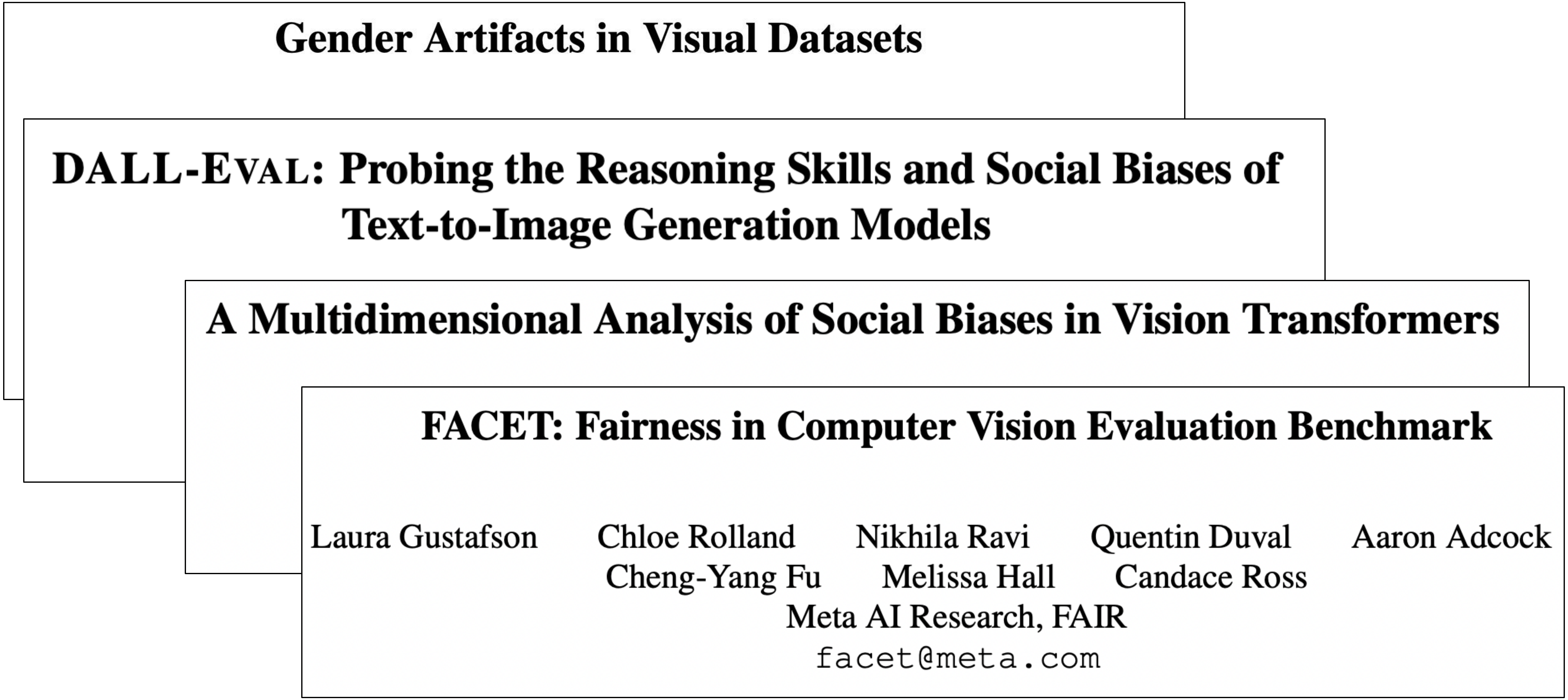


AI Safety Summit 2023

The AI Safety Summit 2023 is a major global event that will take place on the 1 and 2 November at Bletchley Park, Buckinghamshire.

Attendees: 100 world leaders and tech execs

Recent Work: Highlights from ICCV 2023



Gender Artifacts in Visual Datasets

DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models

A Multidimensional Analysis of Social Biases in Vision Transformers

FACET: Fairness in Computer Vision Evaluation Benchmark

Laura Gustafson Chloe Rolland Nikhila Ravi Quentin Duval Aaron Adcock
Cheng-Yang Fu Melissa Hall Candace Ross
Meta AI Research, FAIR
facet@meta.com

Today’s Topics

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

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 CV to diagnose diseases

Unacceptable to acceptable:
Using CV to tag names to people’s faces

Unacceptable to acceptable: Using CV to describe someone’s body shape/size

Unacceptable to acceptable:
Using CV to edit publicly-shared images

Unacceptable to acceptable: Using data from public websites to train CV models

Unacceptable to acceptable:
Open-sourcing vision foundation models

What other ethical issues can you think of around using computer vision algorithms?

Today’s Topics

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

The End