

Algorithm FATE (Fairness, Accountability, Transparency, & Ethics)

Danna Gurari

University of Texas at Austin

Spring 2020



Review

- Last week:
 - Active Learning
 - Curriculum Learning
 - Reinforcement Learning
- Assignments (Canvas):
 - Project presentation due next week
 - Final project report due in two weeks
- Questions?

Final Project Video Suggestions

- Video creation/editing resources:
 - https://docs.google.com/document/d/1y1AENPLDGi4N1oUmd7g4Z4id_ih31HwUOmrM1jy2Gjg/edit
- Attributions:
 - Creative commons license generator: <https://creativecommons.org/choose/>

Today's Topics

- Machine Learning Algorithms that Discriminate
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Machine Learning
- Guest: Dr. Mehrnoosh Sameki from Microsoft

Today's Topics

- Machine Learning Algorithms that Discriminate
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Machine Learning
- Guest: Dr. Mehrnoosh Sameki from Microsoft

Observation: World Population is Diverse



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

Algorithms Discriminate: Google Search



Algorithms Discriminate: Google Search

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

Ad - Why this

[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

Gooooooooooooo**ogle** >
1 2 3 4 5 6 7 8 9 10 [Next](#)

[Advanced search](#) [Search Help](#) [Give us feedback](#)

[Google Home](#) [Advertising Programs](#) [Business Solutions](#) [Privacy & Terms](#)
[About Google](#)

Algorithms Discriminate: Image Tagging

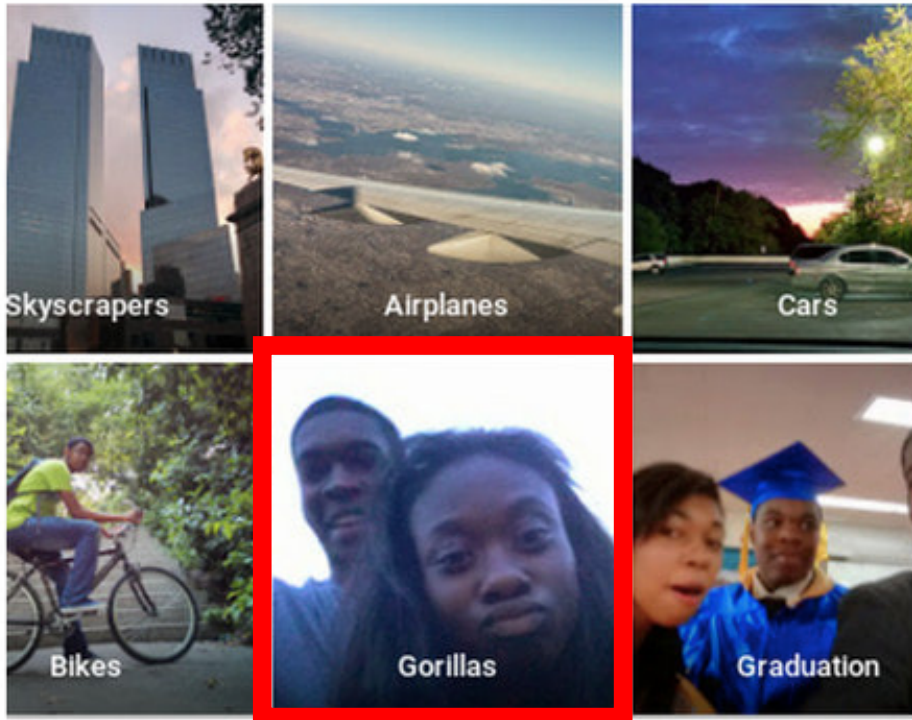


diri noir avec banan
@jackyalcine



+ Follow

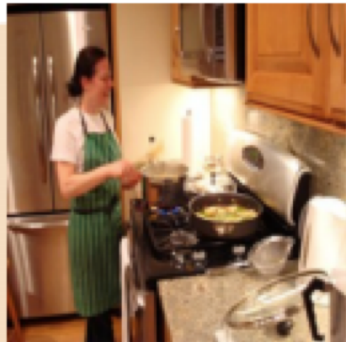
Google Photos, y'all fucked up. My friend's not a gorilla.



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>

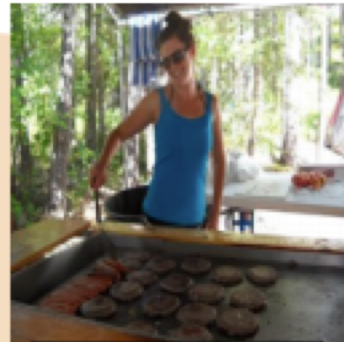
Algorithms Discriminate: Image Tagging



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

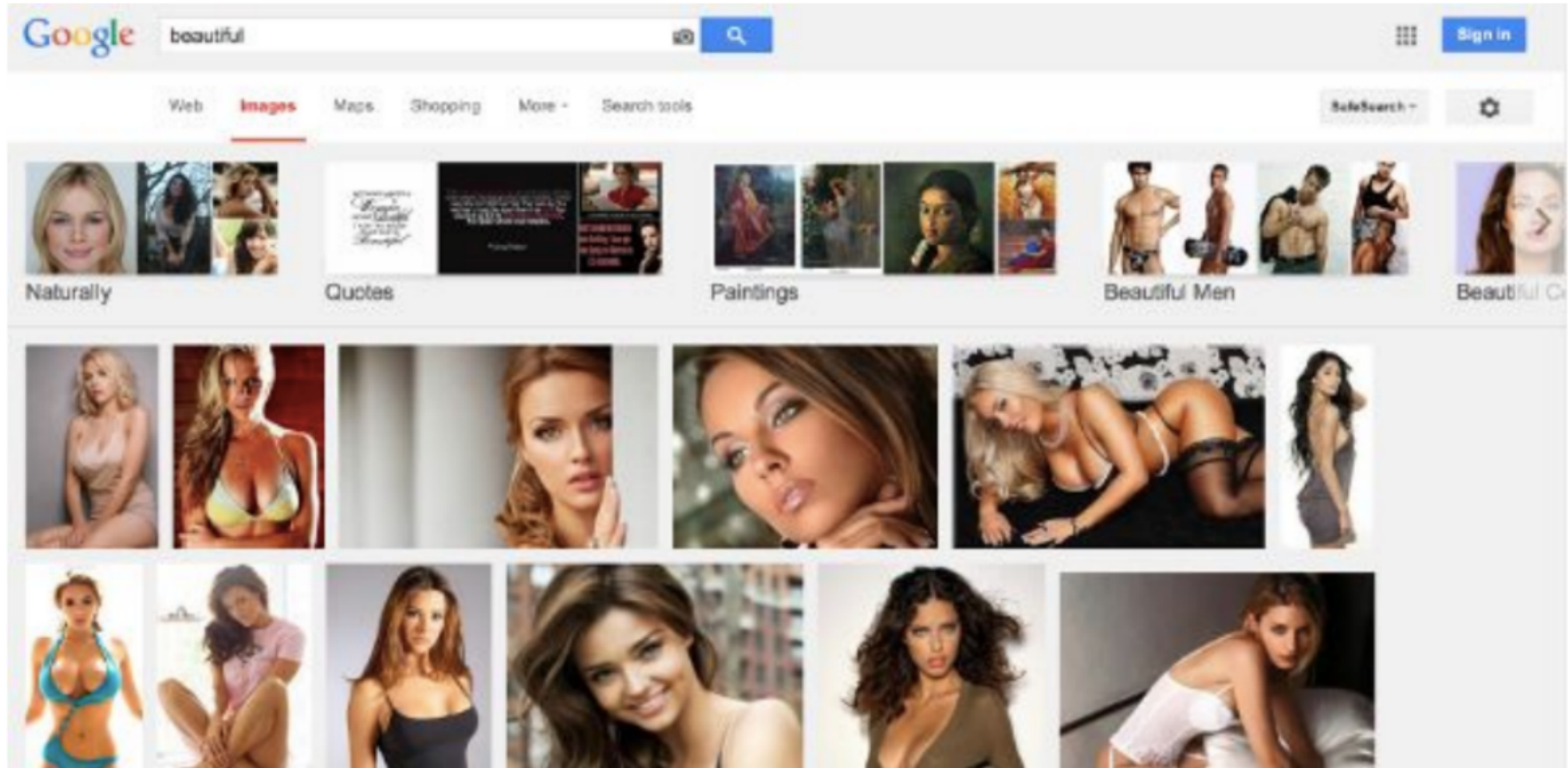


COOKING	
ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

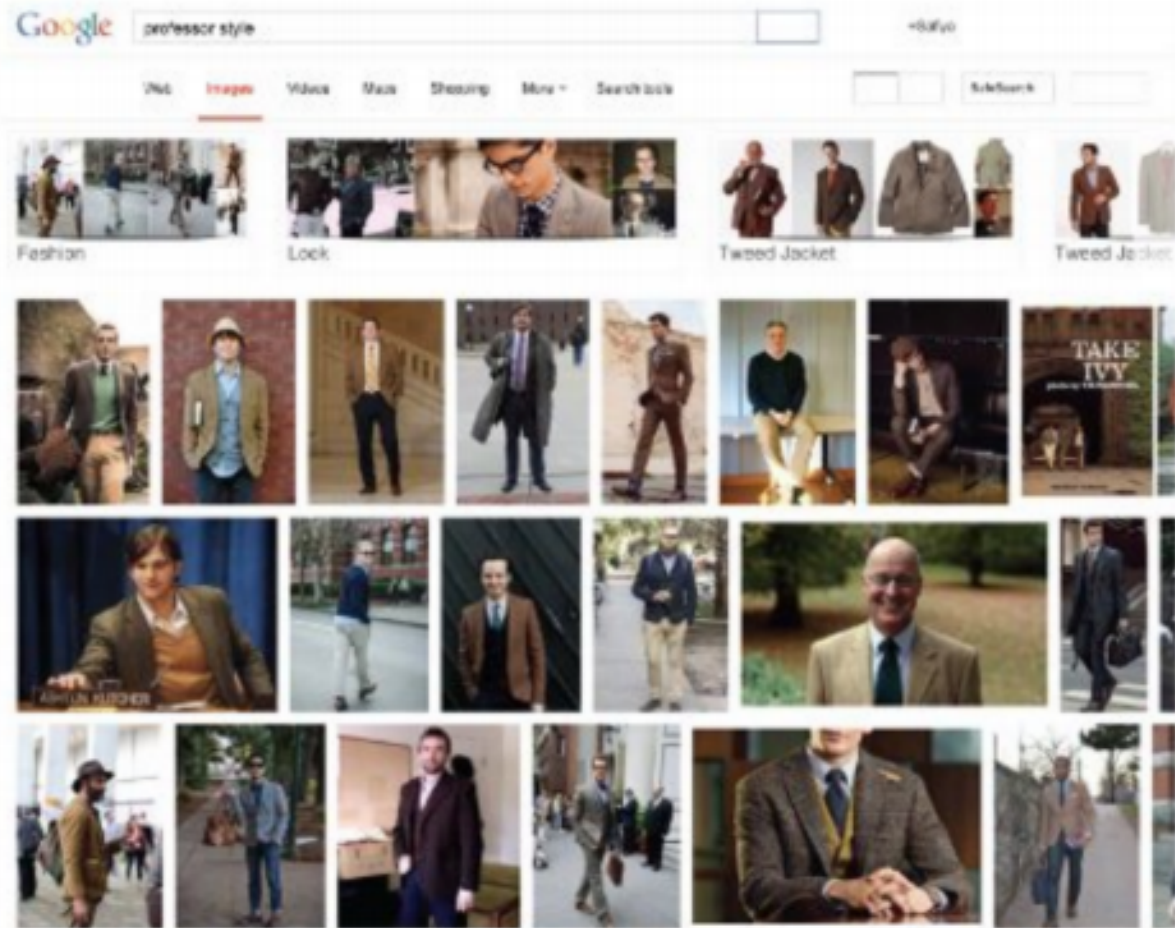
Algorithm identifies men in kitchens as women. Learned this example from given dataset. (Zhao, Wang, Yatskar, Ordonez, Chang, 2017)

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

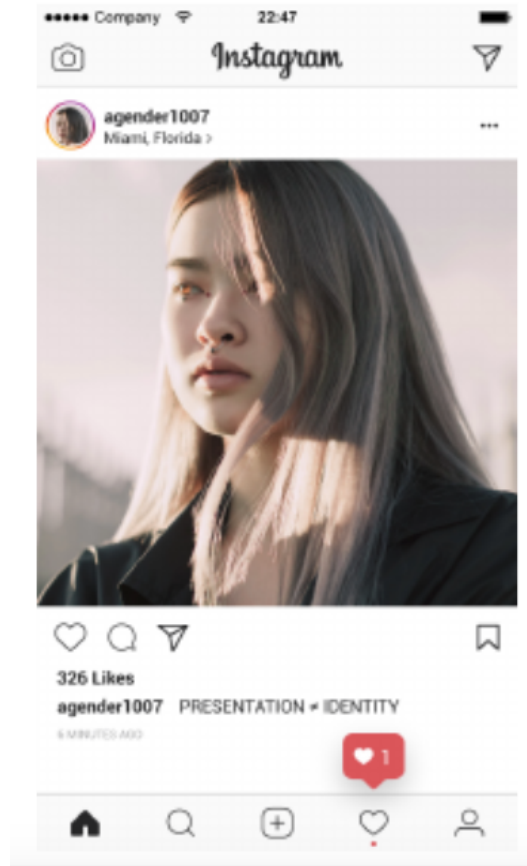
Algorithms Discriminate: Image Tagging (“beautiful”; 2014)



Algorithms Discriminate: Image Tagging (“professor style”; 2014)



Algorithms 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.

Algorithms 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/>

Algorithms 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?"



Algorithms 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/>

Algorithms Discriminate: Book Shopping

Amazon search results for "history of rothschilds".

Search bar: All history of rothschilds

Navigation: Your Amazon.com, Early Black Friday Deals, Gift Cards, Sell, Whole Foods, Registry, EN, Hello, Sign in, Account & Lists, Orders, Try Prime

Sort by: Featured

Search results for "of 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
4.5 stars (172 reviews)
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
4.5 stars (162 reviews)
Book 2 of 2 in the Planet Rothschild Series



Anti-Semitic Bias:



Algorithms Discriminate: Job Recruiting

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



Algorithms Discriminate: Language Translation

Turkish ▾  

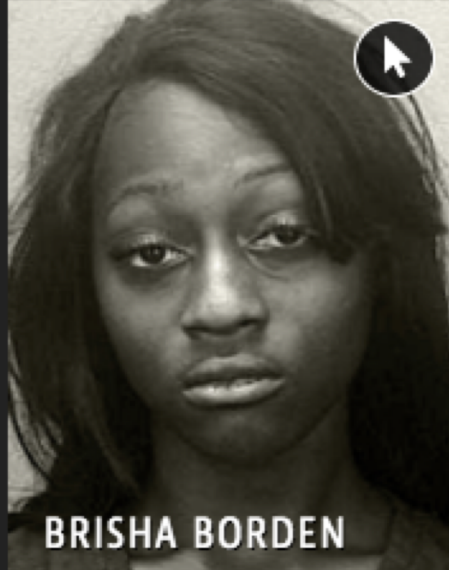

English ▾  

o bir doktor
o bir hemşire

he is a doctor
she is a nurse

Algorithms Discriminate: Criminal Sentencing

Two Petty Theft Arrests



VERNON PRATER	BRISHA BORDEN
LOW RISK	HIGH RISK
3	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	BRISHA BORDEN
Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK	HIGH RISK
3	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.

Group Discussion

How would you try to fix issues like these?

Today's Topics

- Biased Machine Learning Algorithms
- **FAT (Fair, Accountable, & Transparent) Algorithms**
- Ethics in Machine Learning
- Guest: Dr. Mehrnoosh Sameki from Microsoft

We know that algorithms are not perfect.

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

FAT Machine 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 Machine Learning: Fairness

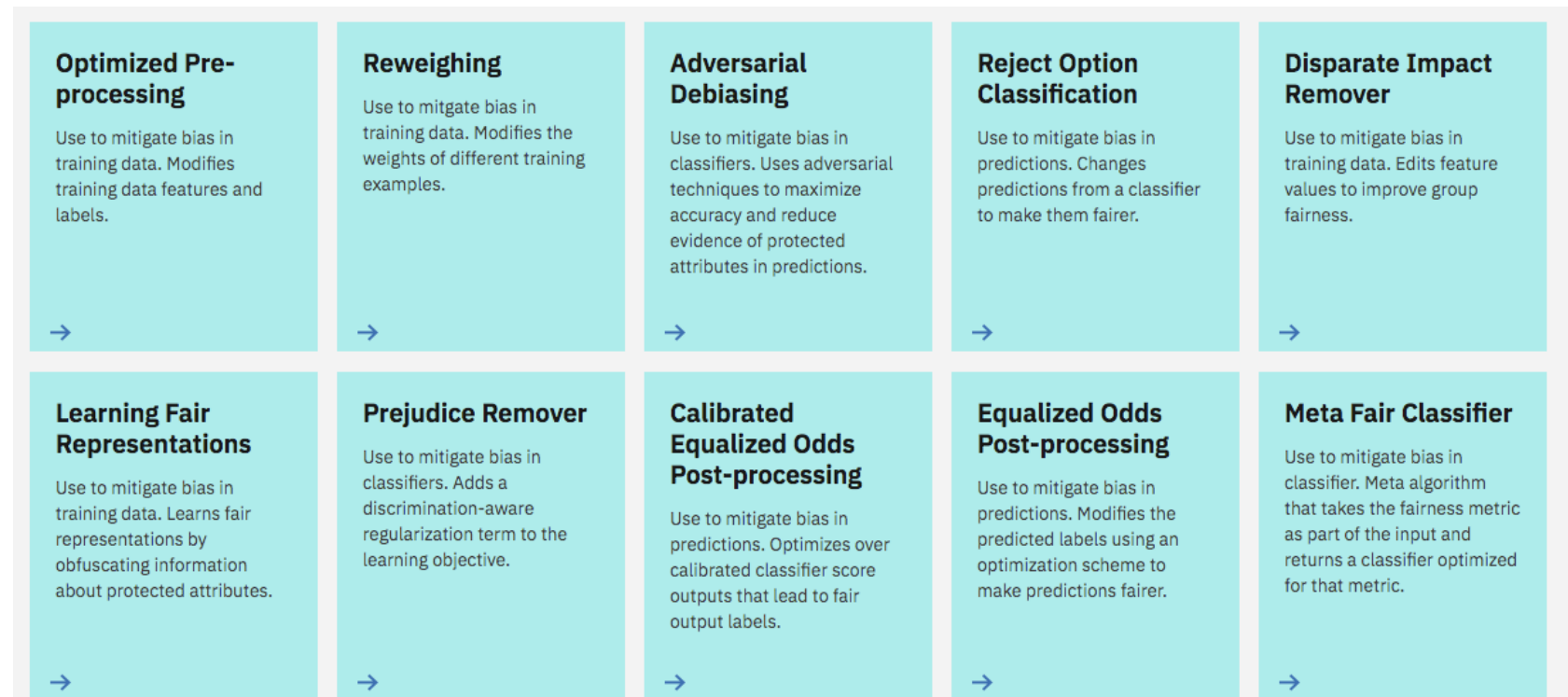
- How to make more fair methods?
 - Pre-processing:
 - Training data: modify it
 - Optimization at training:
 - Algorithm: e.g., add regularization term to objective function to penalize unfairness
 - Features: remove those that reflect bias; e.g., gender, race, age, education, sexual orientation, etc.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

FAT Machine 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

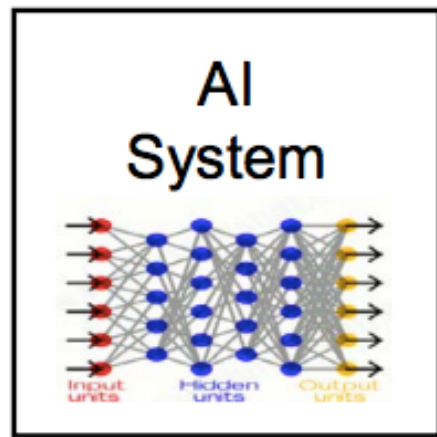
70+ fairness metrics and 10+
bias mitigation algorithms



FAT Machine Learning: Accountability

- Accountability: who is accountable for ML algorithm behavior?
 - e.g., developers who 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 failures in ML systems)?

FAT Machine 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

Watson

A screenshot from the game show Jeopardy! showing the AI system Watson competing against human contestants. The screen displays scores of \$200, \$4,000, and \$600. Below the scores, there are clues and progress bars for 'Maxwell's silver hammer', 'FRANK SINATRA', and 'Brown'. The IBM logo is visible in the bottom right corner.

AlphaGo

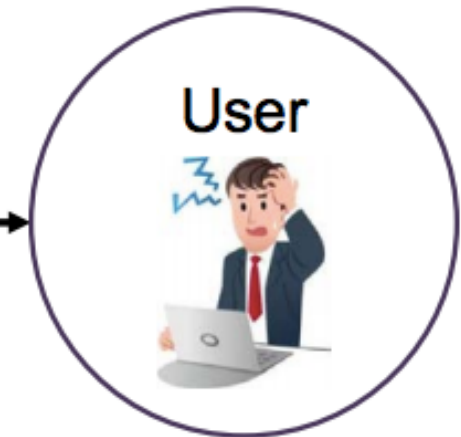
A close-up photograph of black and white Go stones scattered on a wooden Go board.

Sensemaking

A photograph of a person in a military-style uniform sitting at a desk in a control room, looking at several computer monitors displaying data and maps. The NASA.gov logo is visible in the bottom right corner.

Operations

A photograph of a soldier in camouflage gear operating a small, four-wheeled robot in an outdoor setting. The robot is carrying a rifle. The text 'Sennan, U.S. Marine' is visible in the bottom right corner.



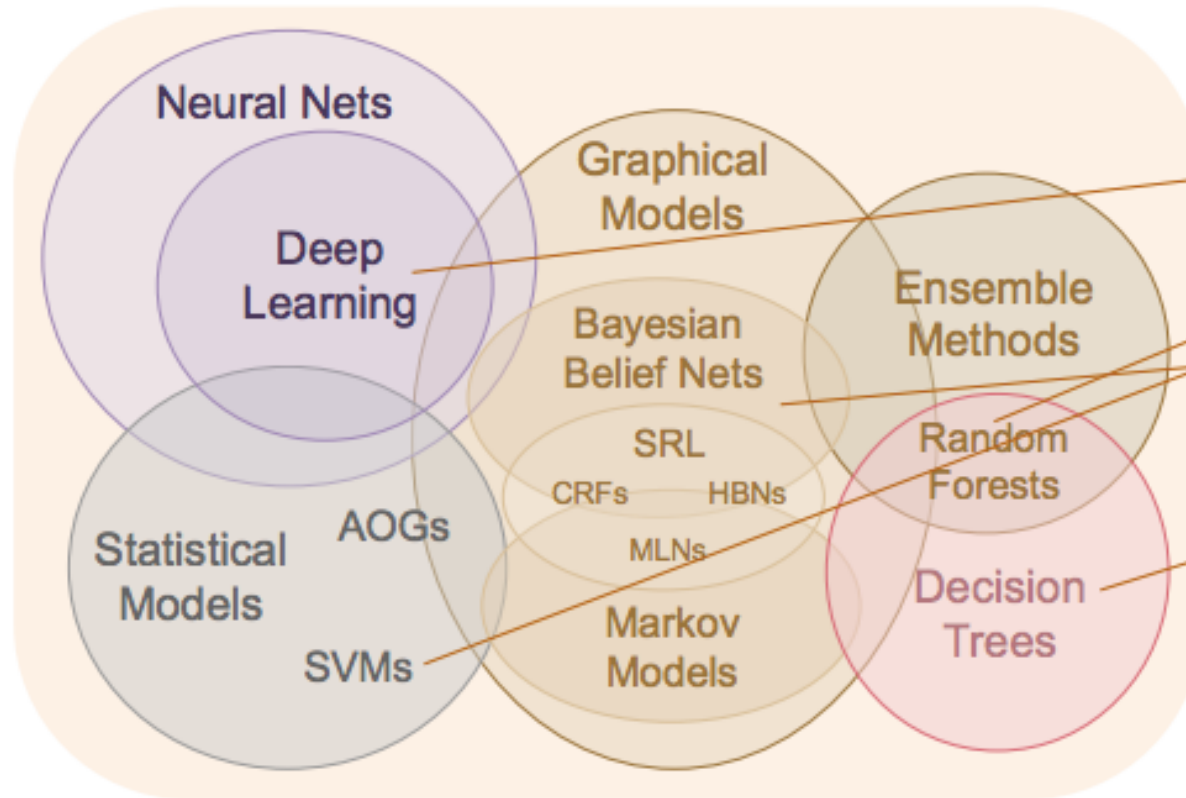
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

FAT Machine Learning: Transparency

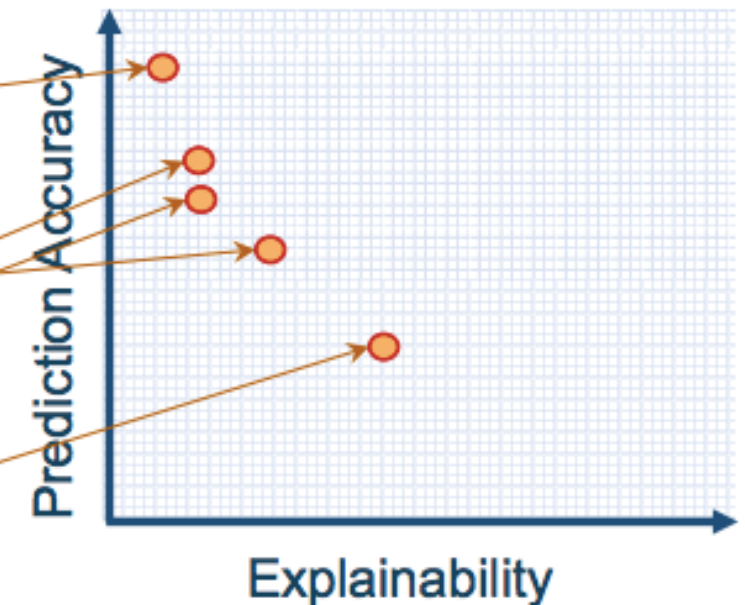
New Approach

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

Learning Techniques (today)

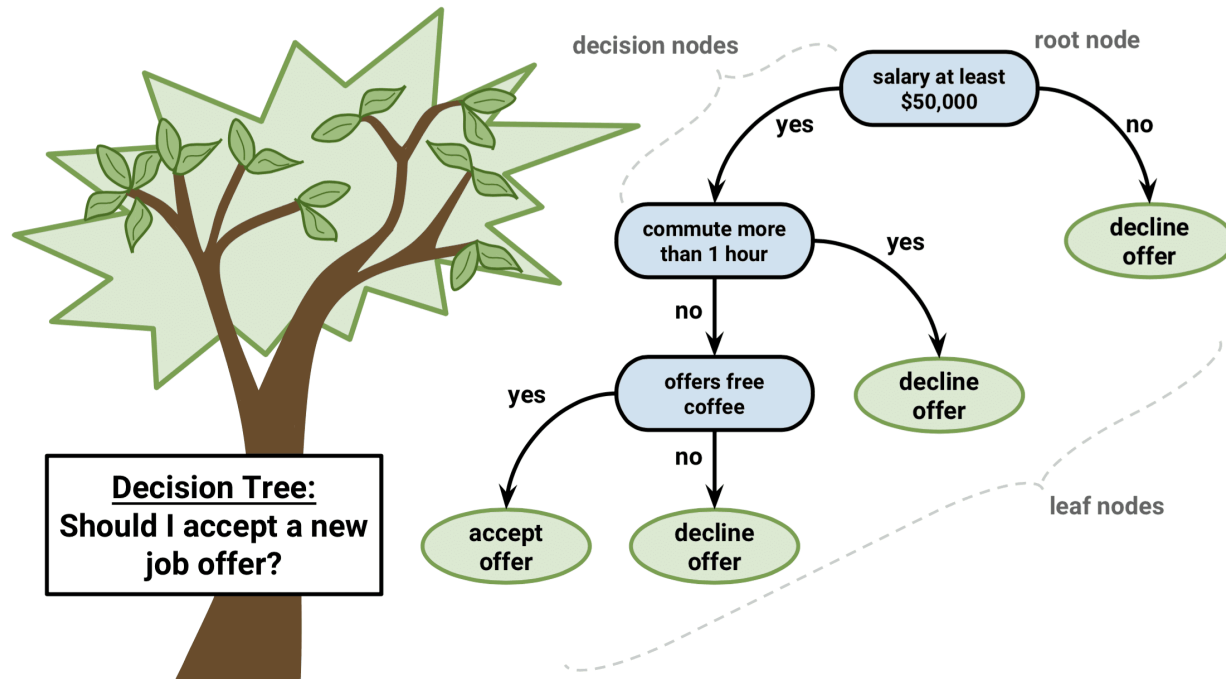


Explainability (notional)

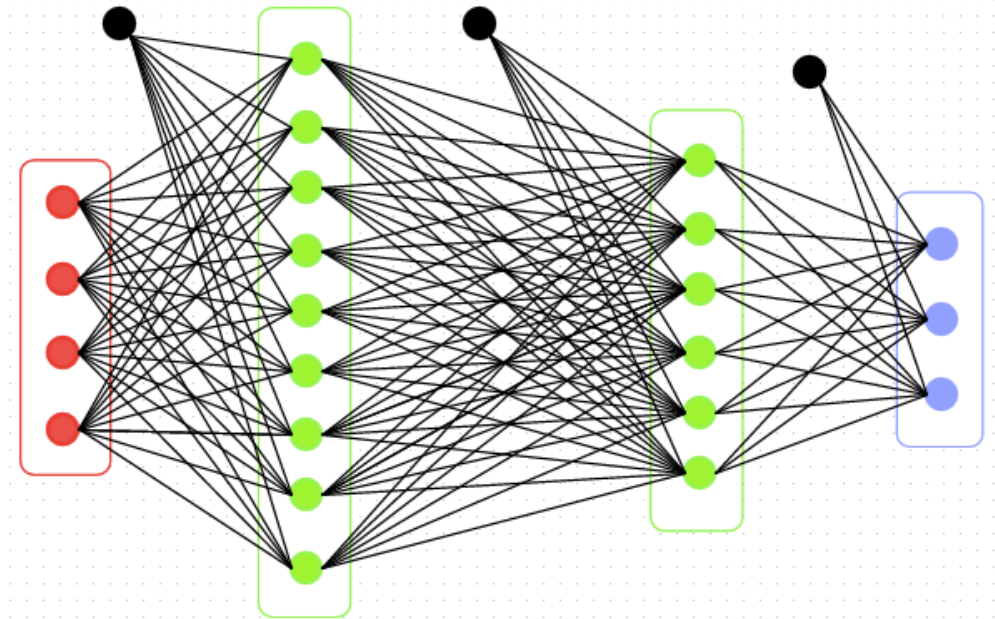


FAT Machine Learning: Transparency

- Transparency: how are predictions made by black box ML algorithms?
 - e.g.,



VS



Source: <http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/>

Source: <https://towardsdatascience.com/build-your-first-deep-learning-classifier-using-tensorflow-dog-breed-example-964ed0689430>

Industry (Facebook, Google, Uber, & more...)

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

Microsoft | Research Research areas Products & Downloads Programs & Events Careers People Blogs & Podcasts Labs & Locations All Microsoft Search

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



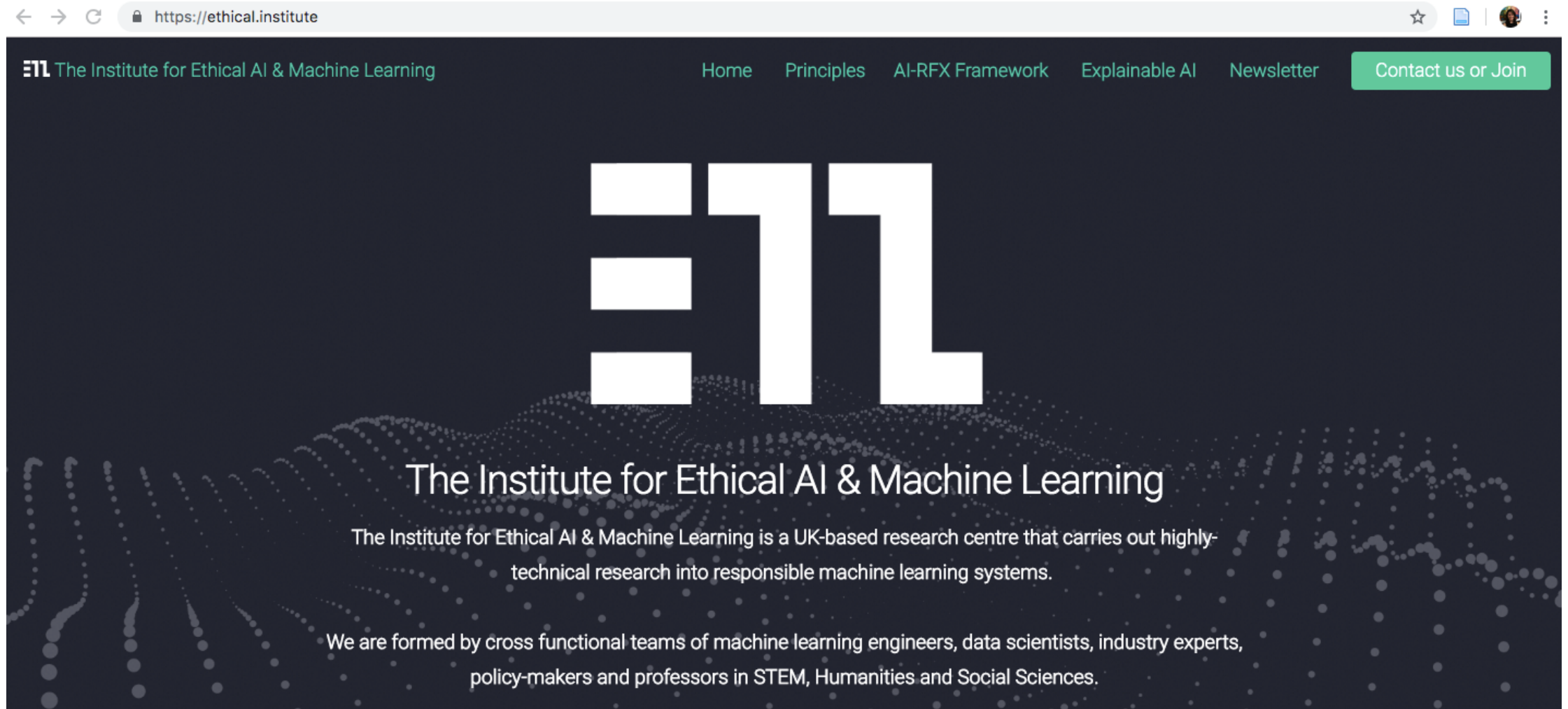
https://www.partnershiponai.org

 PARTNERSHIP ON AI

ABOUT PARTNERS NEWS CAREERS

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

Institutes



The screenshot shows a web browser window with the URL <https://ethical.institute>. The page features a dark blue background with a large, stylized white 'EML' logo in the center. Below the logo, the text reads: 'The Institute for Ethical AI & Machine Learning'. A descriptive paragraph follows: 'The Institute for Ethical AI & Machine Learning is a UK-based research centre that carries out highly-technical research into responsible machine learning systems.' At the bottom, another paragraph states: '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.' The browser's address bar and navigation icons are visible at the top. The website's navigation menu includes 'Home', 'Principles', 'AI-RFX Framework', 'Explainable AI', and 'Newsletter', along with a green 'Contact us or Join' button.

← → ↻ <https://ethical.institute> ☆ | 🌐 | 👤 | ⋮

EML The Institute for Ethical AI & Machine Learning

Home Principles AI-RFX Framework Explainable AI Newsletter [Contact us or Join](#)

EML

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.

<https://www.fast.ai/2018/09/24/ai-ethics-resources/>

Academia: Workshops

Not Secure | ethicsinnlp.org

Ethics in NLP 2018

NAACL 2018
New Orleans, Louisiana
June 5th

Academia: Workshops

https://fatconference.org

ACM FAT* Conference 2019 ▾ 2018 ▾

Organization Resources ▾

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

Not Secure | fairware.cs.umass.edu/agenda.html



[Home](#)

[Agenda](#)

[Keynote](#)

[Call for Papers](#)

[Organization](#)

Academia: Annual Workshop for 5 Years Now...



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.

Academia: Annual Workshop Scope...

Questions to the machine learning community include:

- How can we achieve high classification accuracy while eliminating discriminatory biases? What are meaningful formal fairness properties?
- How can we design expressive yet easily interpretable classifiers?
- Can we ensure that a classifier remains accurate even if the statistical signal it relies on is exposed to public scrutiny?
- Are there practical methods to test existing classifiers for compliance with a policy?

Academia: And Many More Resources...

<https://fatconference.org/resources.html>

Today's Topics

- Biased Machine Learning Algorithms
- FAT (Fair, Accountable, & Transparent) Algorithms
- **Ethics in Machine Learning**
- Guest: Dr. Mehrnoosh Sameki from Microsoft

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 ML to sentence people for a crime

Unacceptable to acceptable:
Using ML to diagnose diseases

Unacceptable to acceptable:
Using ML to filter resumes for jobs

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

Google Form: Guest Speaker & Class Feedback

- Google form:
 - Guest: Dr. Mehrnoosh Sameki, Technical Program Manager at Microsoft (<http://cs-people.bu.edu/sameki/>); list one question for her for today's visit
- Then, take a short break.
- Class resumes at 4:50pm CST.

Today's Topics

- Biased Machine Learning Algorithms
- FAT (Fair, Accountable, & Transparent) Algorithms
- Ethics in Machine Learning
- **Guest: Dr. Mehrnoosh Sameki from Microsoft**