

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

Danna Gurari

University of Texas at Austin

Spring 2021



Review

- Last week:
 - Machine Learning for Sequential Data
 - Recurrent Neural Networks (RNNs)
 - Training Deep Neural Networks: Hardware & Software
- Assignments (Canvas):
 - Project outline due tonight
 - Prototype of final project ML system due at tomorrow's meeting
 - Final project submission with video due in two weeks
- Questions?

Final Project Video Suggestions

- Video creation/editing resources:
 - https://docs.google.com/document/d/1jBZ1fU1CKDLw1y2ZVM5LvYHjv3iFZoAYPz947_0f2Bs/edit?usp=sharing
- Attributions:
 - Creative commons license generator: <https://creativecommons.org/choose/>

Plagiarism: Definition

- Material from: <https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html>

University of Texas Definition of Plagiarism:

“the appropriation of, buying, receiving as a gift, or obtaining by any means material that is

attributable in whole or in part to another source, including words, ideas, illustrations, structure, computer code, and other expression

or media, and presenting that material as one's own academic work being offered for credit.”

Plagiarism: Definition

- Material from: <https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html>

Plagiarism in Plain English:

Using someone else's work in your own academic work without giving proper credit. Click a button below to see some examples.

Intentional Plagiarism

Unintentional Plagiarism

Plagiarism: Play It Safe, Give Credit Generously

- Material from: <https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html>

Intentional Plagiarism:

- Copying a friend's or classmate's work
- Buying or borrowing papers
- Cutting and pasting blocks of text without providing documentation of the original source
- Borrowing images and other media without documentation of the original source
- Publishing work on the Web without the permission of the creator

Plagiarism: Play It Safe, Give Credit Generously

- Material from: <https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html>

Unintentional Plagiarism:

- Careless paraphrasing
- Poor documentation of sources
- Quoting excessively
- Failure to use your own ideas or words

Plagiarism: Play It Safe, Give Credit Generously

- Material from: <https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html>

During the course of your research, you come across an idea that you use in your paper. You don't use the author's exact words or even paraphrase -- just the idea. Cite it?

Other people's words aren't the only thing you need to cite. You also need to cite ideas. So in this case, you should give the author credit for the idea by citing them.

Plagiarism: Play It Safe, Give Credit Generously

- Material from: <https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html>

You are doing a presentation for your Chemistry class and use an image of the Periodic Table you found on a government web site. Cite it?

You should cite images. Even government websites in the public domain need to be cited.

Plagiarism: Play It Safe, Give Credit Generously

- What can happen if you are accused of plagiarism?
 - Redo assignment
 - Receive a failing grade
 - Be suspended
 - Be expelled
- What resources can help you to avoid plagiarism?
 - Review: <https://legacy.lib.utexas.edu/services/instruction/avoidplagiarism.html>
 - Review: https://legacy.lib.utexas.edu/d7/sites/default/files/services/instruction/AvoidingPlagiarism_guide.pdf
 - Visit writing center: <http://uwc.utexas.edu/>
- Neither you (I believe) nor I have any desire to talk about plagiarism 😊
- Play it safe and give credit generously!!!

Give Credit Generously

- Idea: add credit page to your presentation for resources used
 - e.g., Microsoft Azure
 - e.g., freely-shared code/libraries
 - e.g., links to all images
 - ...

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

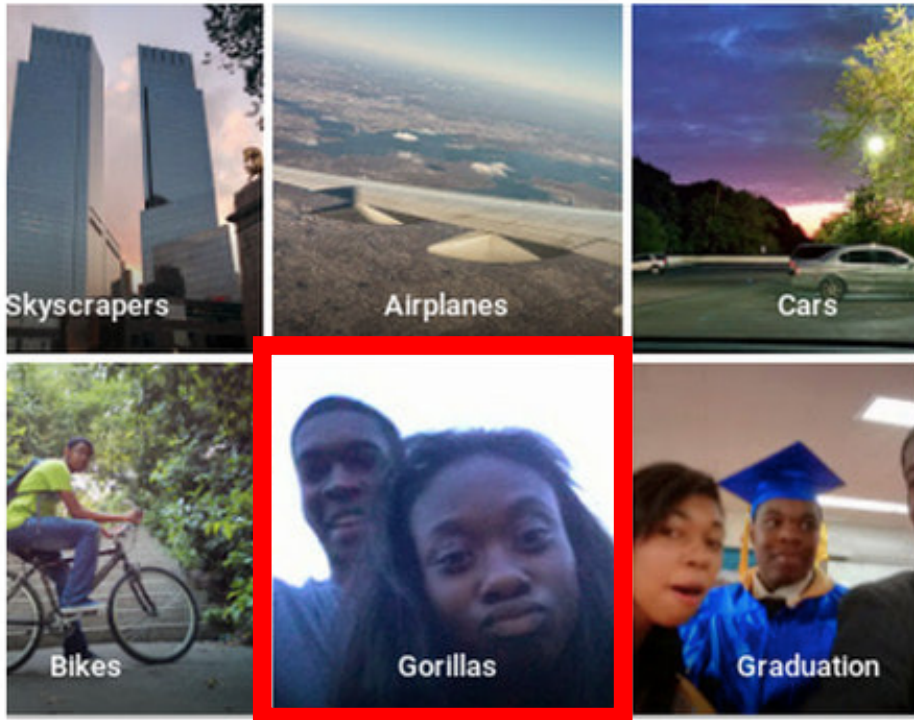


diri noir avec banan
@jackyalcine



Follow

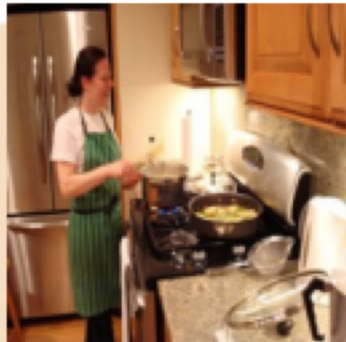
Google Photos, y'all fucked up. My friend's not a gorilla.



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>

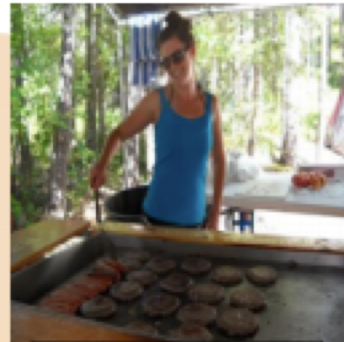
Algorithms Discriminate: Image Tagging



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

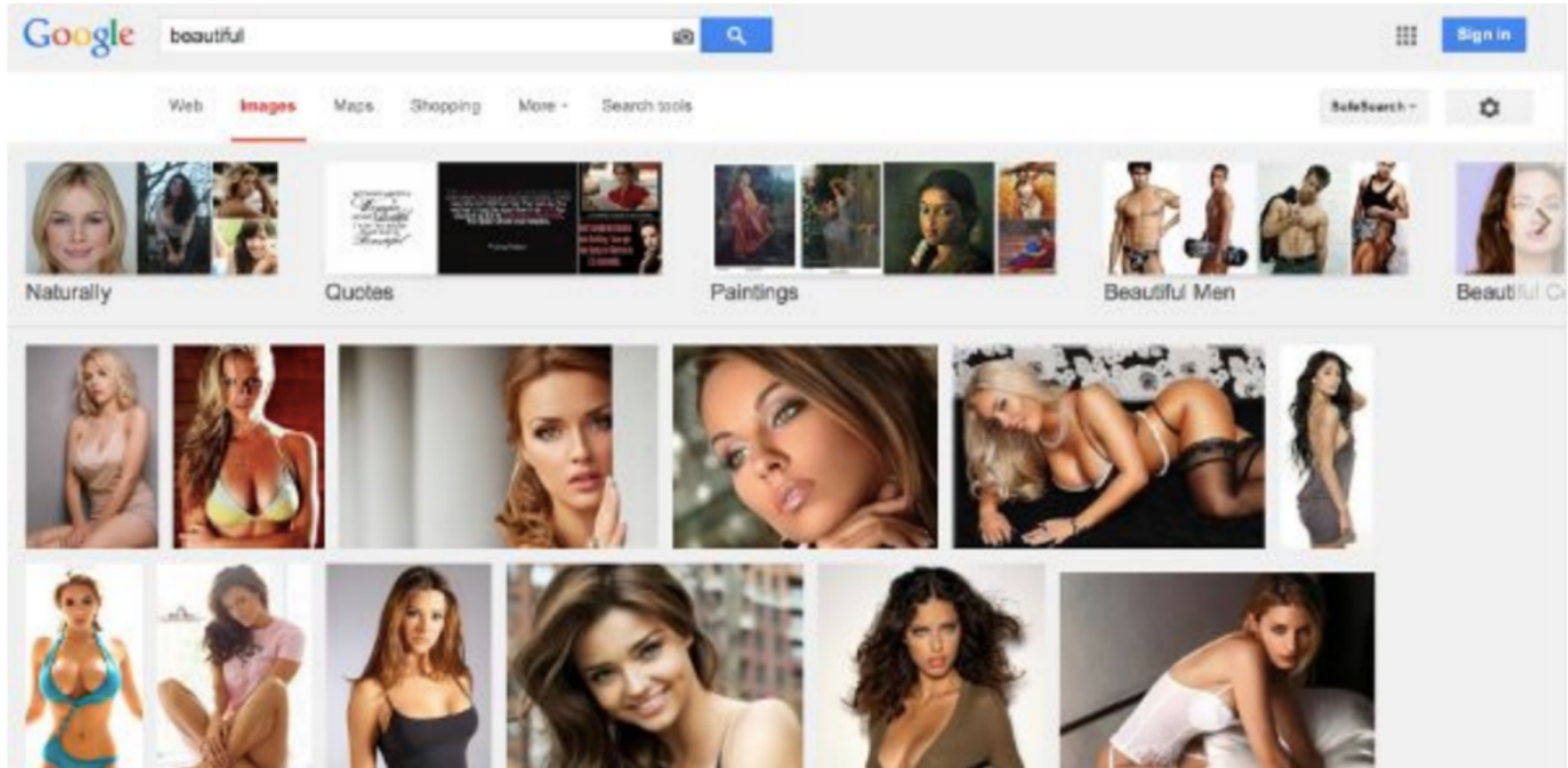


COOKING	
ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

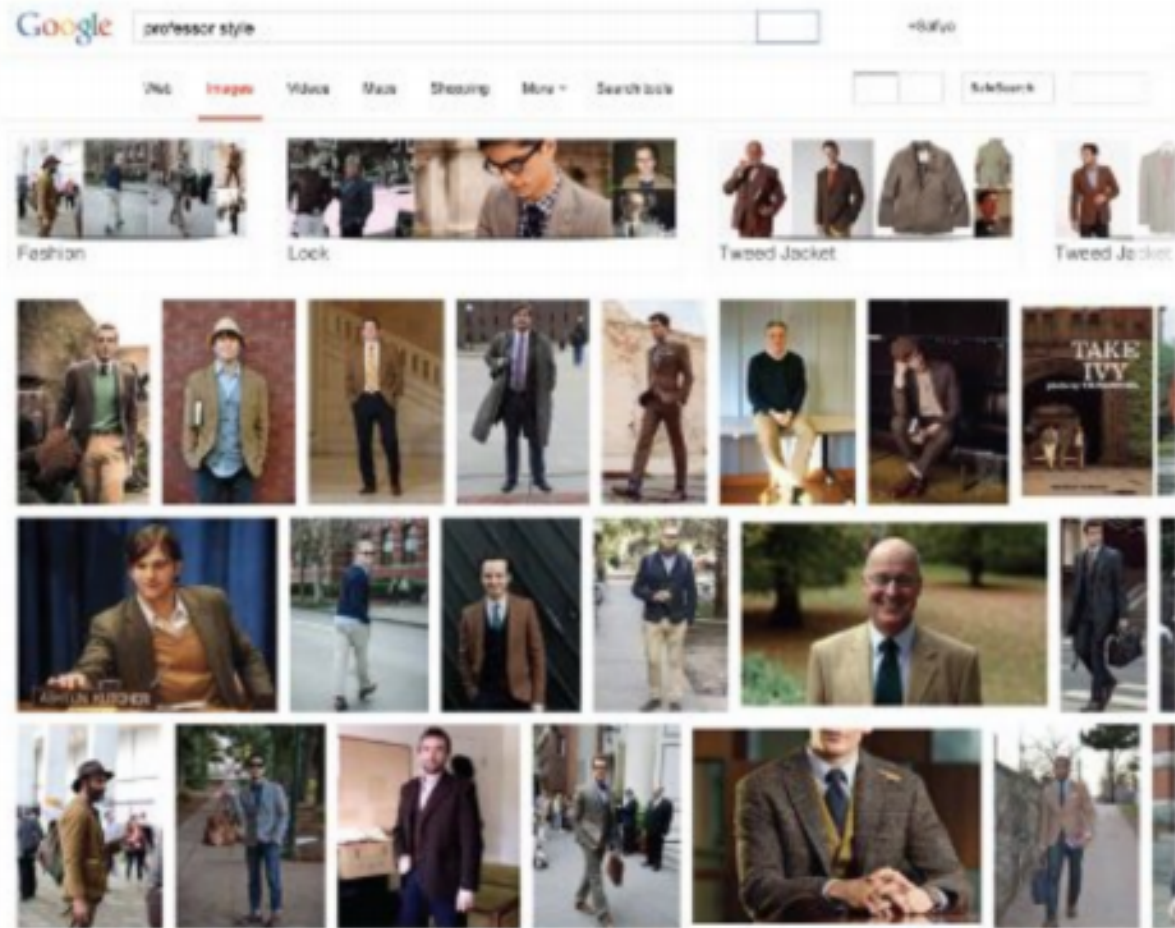
Algorithm identifies men in kitchens as women. Learned this example from given dataset. (Zhao, Wang, Yatskar, Ordonez, Chang, 2017)

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

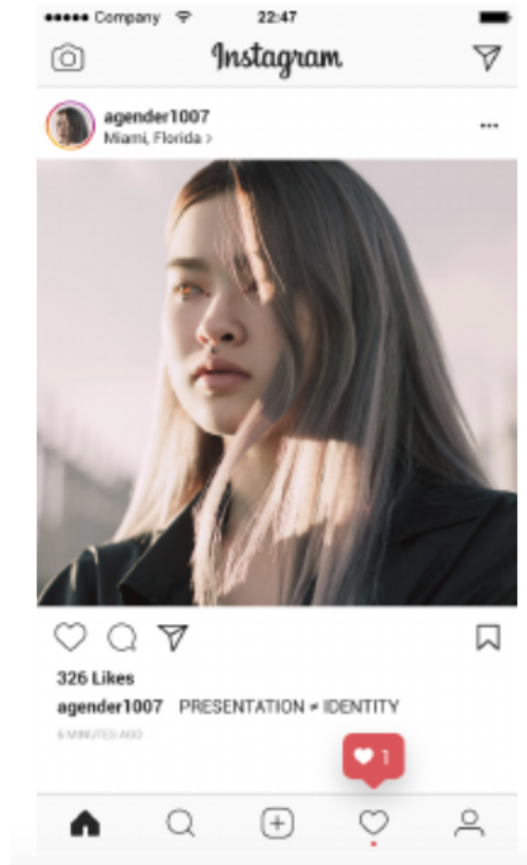
Algorithms Discriminate: Image Tagging (“beautiful”; 2014)



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

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 CVs for technical jobs such as software developer.



Algorithms Discriminate: Language Translation

Turkish   

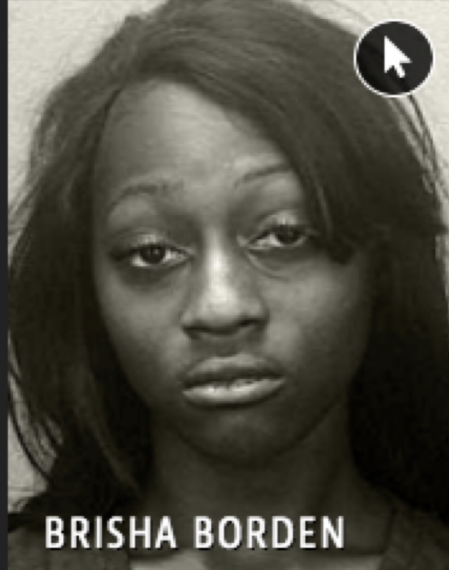

English   

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.

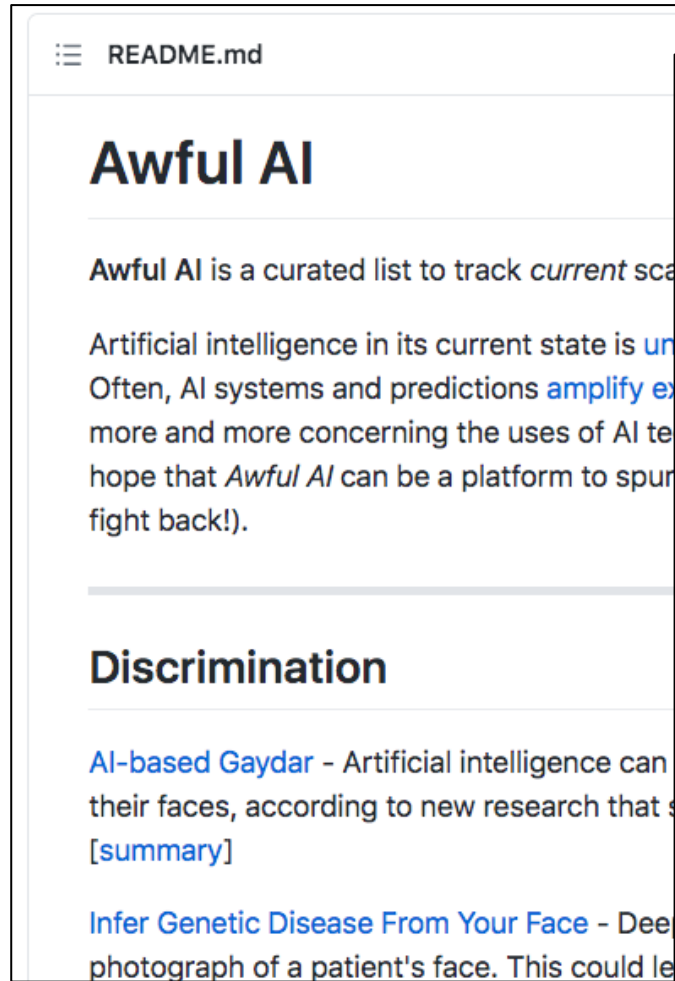
Two Petty Theft Arrests

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.

Algorithms Discriminate: And MANY more...

• e.g.,



☰ README.md

Awful AI

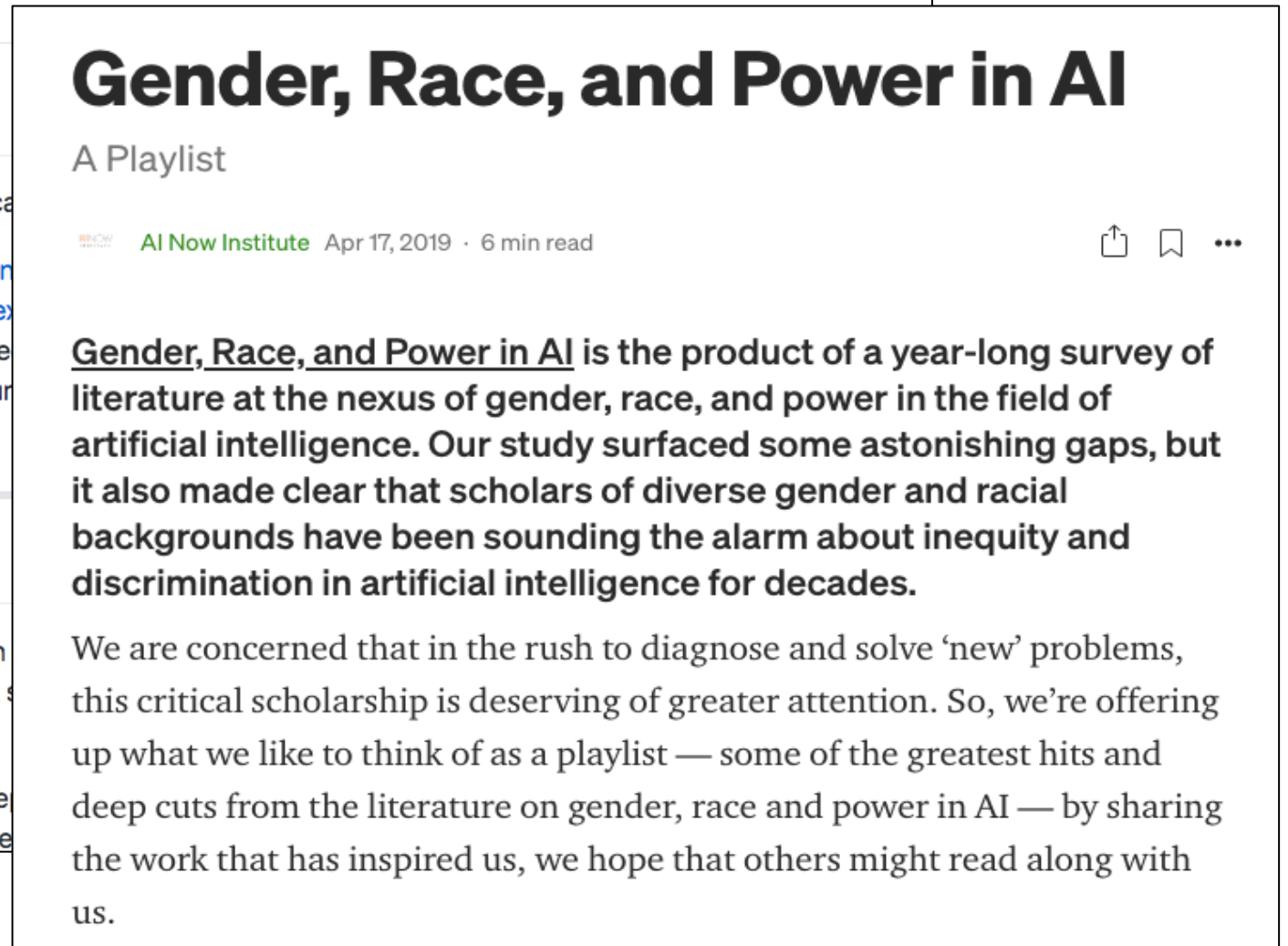
Awful AI is a curated list to track *current* scary AI news. Artificial intelligence in its current state is *unfair*. Often, AI systems and predictions *amplify existing biases* more and more concerning the uses of AI technology. We hope that *Awful AI* can be a platform to spur people to fight back!).

Discrimination

[AI-based Gaydar](#) - Artificial intelligence can identify their faces, according to new research that says [summary]


[Infer Genetic Disease From Your Face](#) - Deep learning can identify a photograph of a patient's face. This could lead to

<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>

Algorithms Discriminate

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

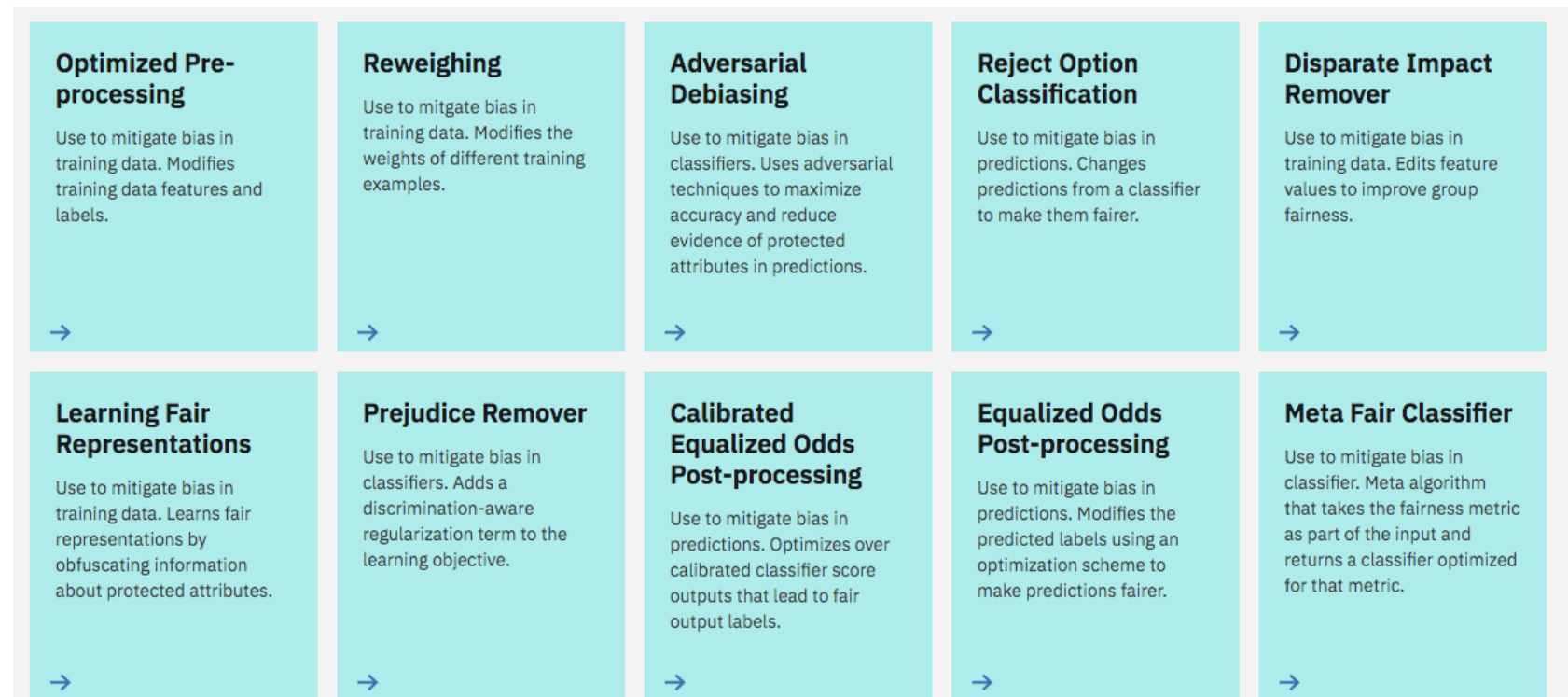
- How to make more fair methods?
 - Pre-processing:
 - Training data: modify it
 - Optimization at training:
 - Algorithm: e.g., add regularization term to objective function to penalize unfairness
 - Features: remove those that reflect bias; e.g., gender, race, age, education, sexual orientation, etc.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

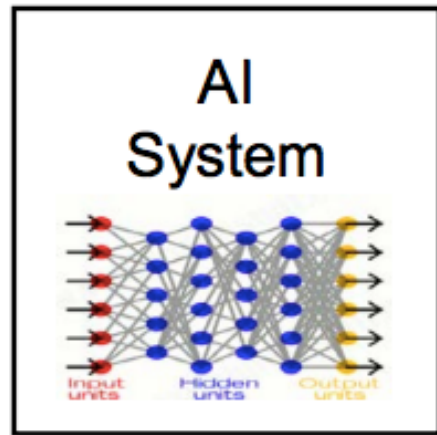
70+ fairness metrics and 10+
bias mitigation algorithms



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! featuring the AI system Watson. The screen shows the game board with scores of \$200, \$4,000, and \$600. Watson is playing against Brad. The current question is 'Maxwell's silver hammer' with a 90% confidence level. Other options are 'FRANK SINATRA' (11%) and 'Brown' (7%).

AlphaGo

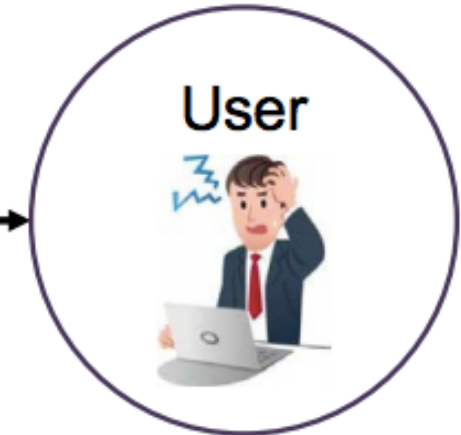
A close-up photograph of black and white Go stones scattered on a wooden board.

Sensemaking

A person in a military uniform is seated at a workstation with multiple computer monitors displaying data and maps. The person is wearing a headset and appears to be focused on the screens.

Operations

A soldier in full combat gear is operating a small, four-wheeled robot in an outdoor setting. The robot is equipped with a camera and other sensors.



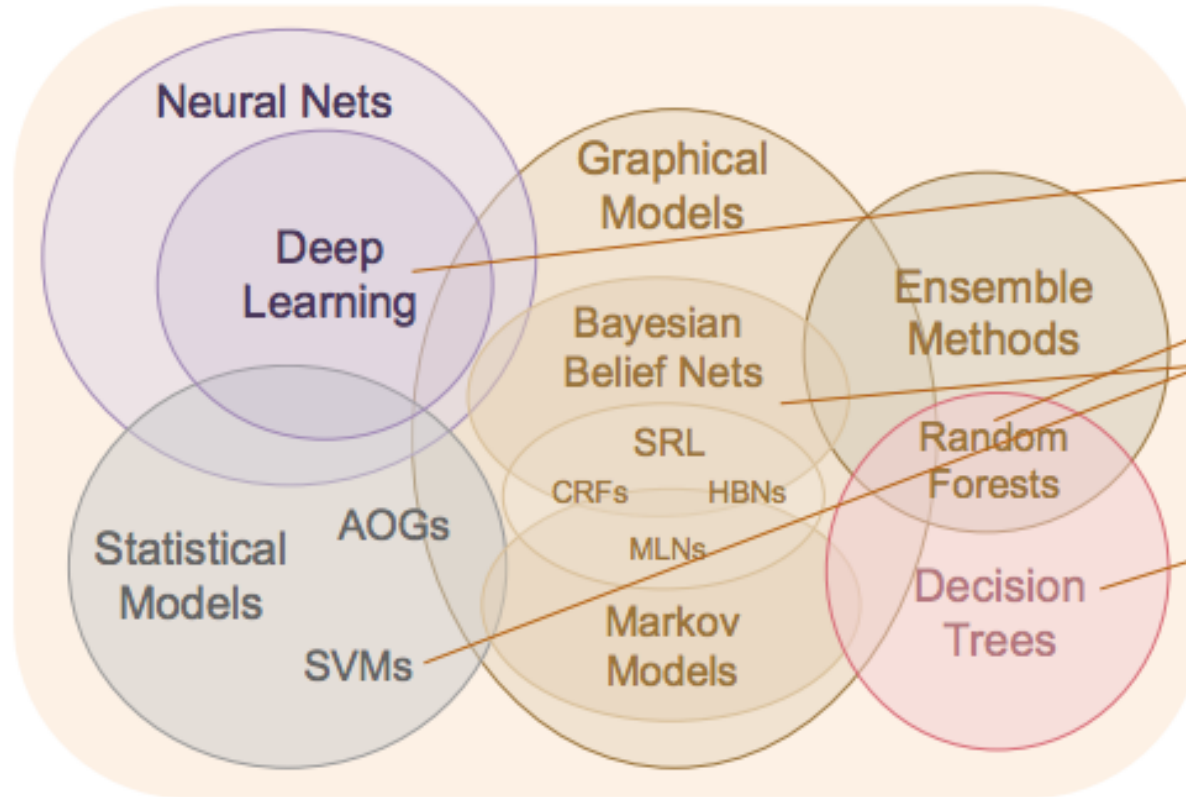
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

FAT Machine Learning: Transparency

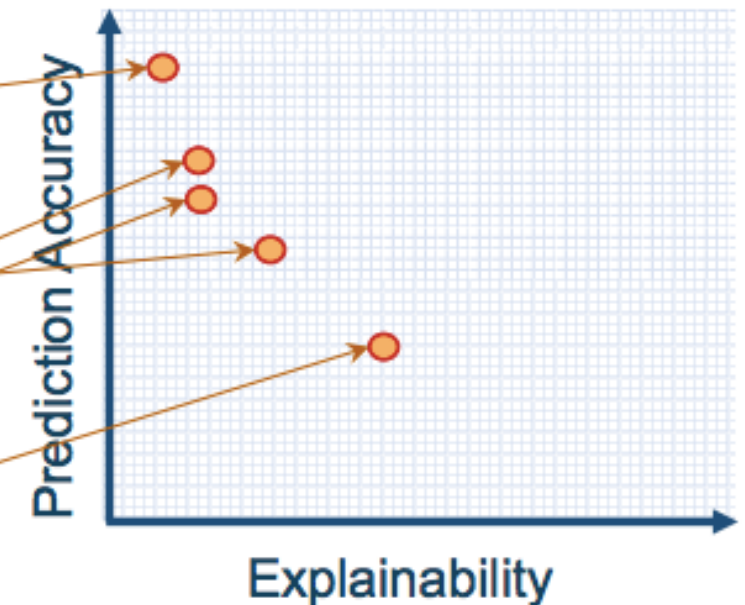
New Approach

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

Learning Techniques (today)

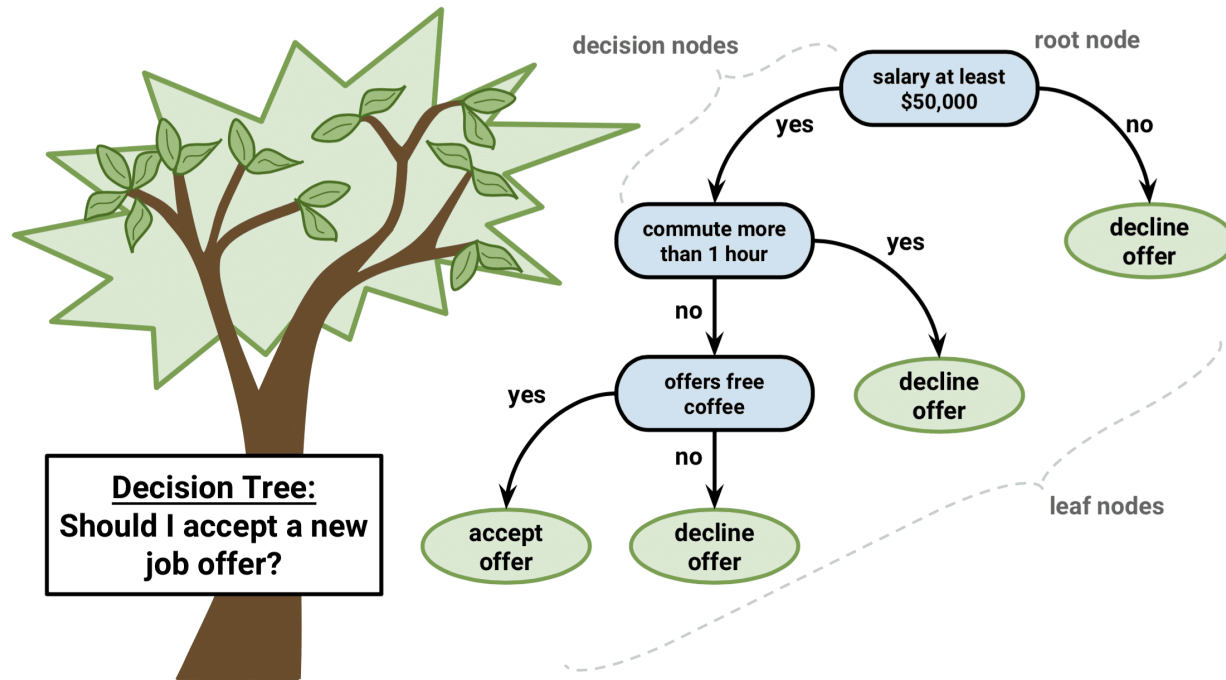


Explainability (notional)

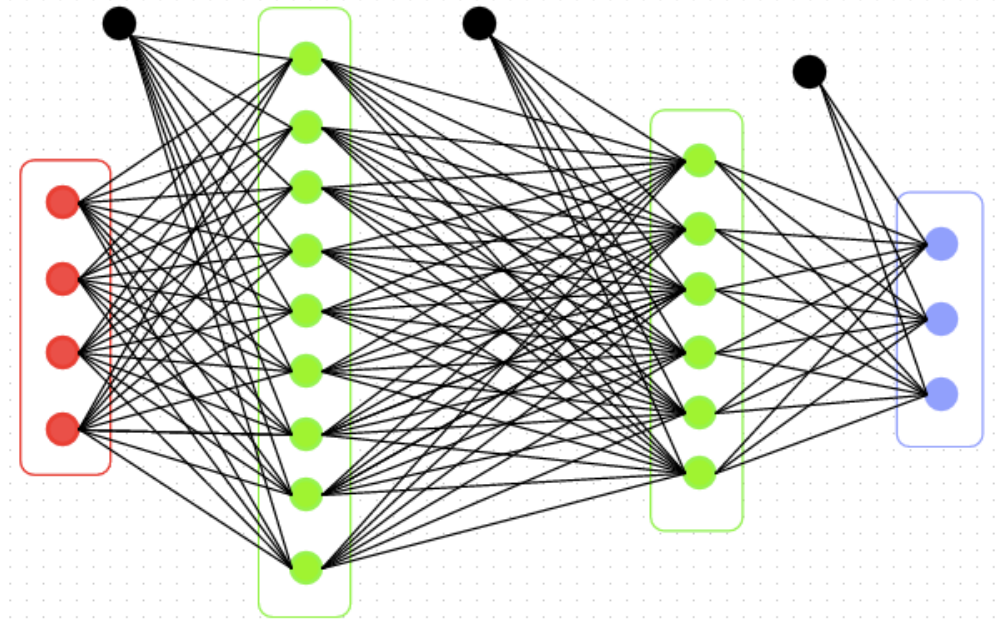


FAT Machine Learning: Transparency

- 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



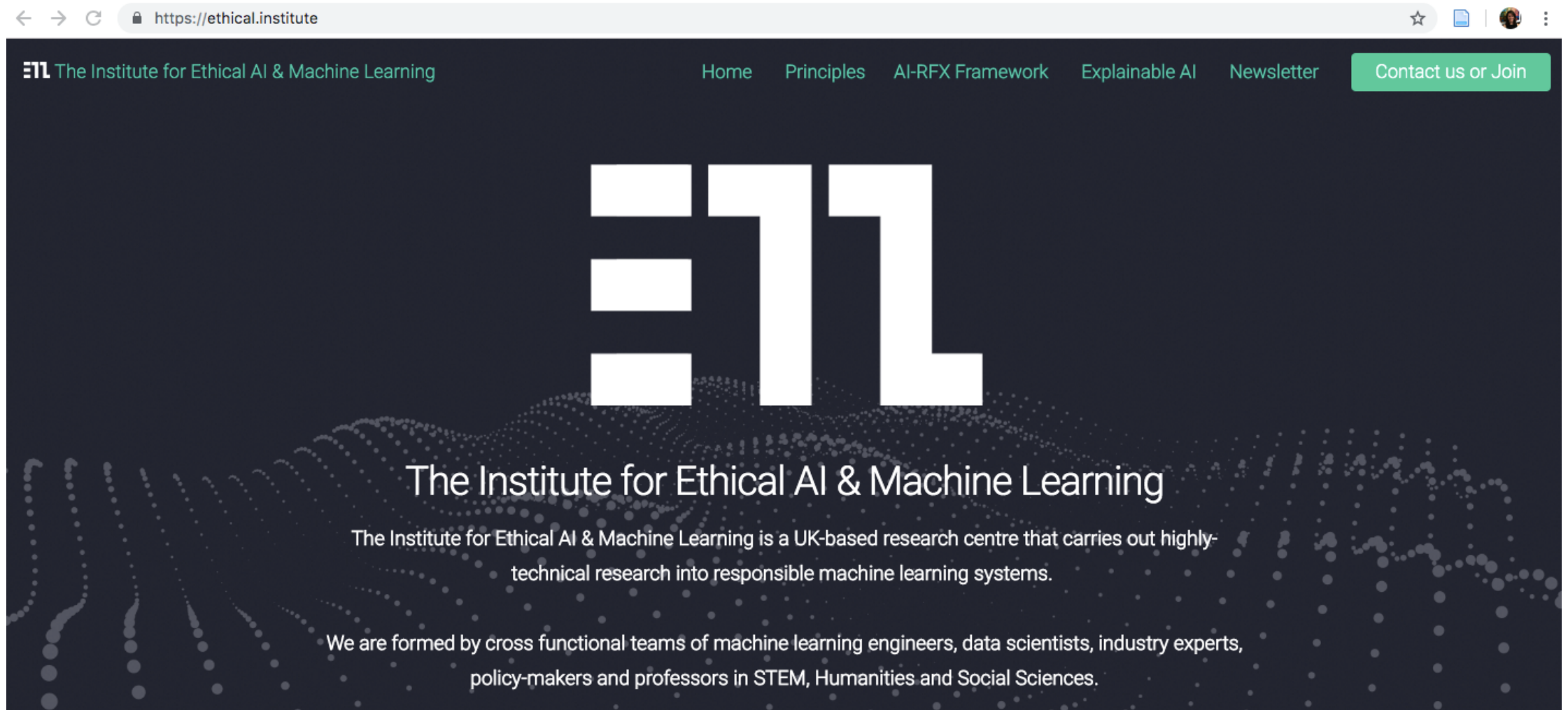
https://www.partnershiponai.org

 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 pattern of white dots and lines. At the top left, the logo "EML The Institute for Ethical AI & Machine Learning" is displayed. The navigation menu includes "Home", "Principles", "AI-RFX Framework", "Explainable AI", "Newsletter", and a green button labeled "Contact us or Join". The main content area features the large white letters "EML" in the center. Below this, the text reads: "The Institute for Ethical AI & Machine Learning" followed by a paragraph: "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."

<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



May 29, 2018

Gothenburg, Sweden

Collocated with ICSE 2018

FairWare 2018

International Workshop on Software Fairness

[Home](#)

[Agenda](#)

[Keynote](#)

[Call for Papers](#)

[Organization](#)

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.

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

Today's Topics

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

Google Form: Guest Speaker

- Guest: Dr. Mehrnoosh Sameki, Senior Technical Program Manager at Microsoft (<https://www.linkedin.com/in/mehrnoosh-sameki-a2a02245/>)
 - Share one question for her for tomorrow's visit