

Announcements

Reminder: Class challenge out! Ends December 10th

Assignment Project Exam Help

- Final in two weeks, practice question will be posted today
<https://powcoder.com>

Add WeChat powcoder

Class Challenge Task 2

- 100 labels isn't enough! (see Lecture 21/first half of 22)

Assignment Project Exam Help

- Transfer learning from a different dataset? if it isn't *a dataset we provided*, you have to check with us (ImageNet OK)

<https://powcoder.com>

Add WeChat powcoder

1 001 0 0 01 01
01 1 1 01 01
01 0 001 0 1
001010101101010
01010101000010101
010101010101010101
0000101010101000010101
1010100010111101010100

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Ethics in Machine Learning

Kate Saenko

CS 542 Machine Learning



AI Fears:

Which of these has already happened?



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

A robot kills human

yes, Uber car accident

AI takes over our lives

yes? Youtube algorithm

AI is watching us

yes, virtual police lineup

Autonomous weapons

not yet?


Humans losing jobs

Yes, e.g. librarians

AI Fears

- **Autonomous weapons** – frameworks for regulation
- **Future of work** – deskilling / reskilling / superskilling.
Many jobs that currently seem least likely to be automated have been racialized and gendered in ways connected to care and immigration, and have rarely paid living wages; how do we revalue work?
- **Worse Inequality** – digital divides; bias in algorithms may worsen inequality; ecological concerns in energy, storage and cooling required for ML; economic inequality
- **Divided societies with algorithmic bubbles** – challenges of populism, automated recommendations, news feeds; deepfakes and election meddling





Many of these
problems are not new!

Assignment Project Exam Help

<https://powcoder.com>

Fears about job automation, lack of
privacy and inequality arise with
each new innovation

- Printing press
- Weapons
- Internet

Ethical Issues in Machine Learning

- Job Loss
- Algorithmic Bias
- Transparency
- AI Supremacy
- Fake news and videos
- Autonomous weapons
- Self-driving cars
- Privacy and surveillance

Report: [A.I. would eliminate 75 million jobs but may create about 130 million jobs globally.](#)

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



| Stable Roles | New Roles | Redundant Roles |
|---|---|--|
| Managing Directors and Chief Executives | Data Analysts and Scientists* | Data Entry Clerks |
| General and Operations Managers* | AI and Machine Learning Specialists | Accounting, Bookkeeping and Payroll Clerks |
| Software and Applications Developers and Analysts* | General and Operations Managers* | Administrative and Executive Secretaries |
| Data Analysts and Scientists* | Big Data Specialists | Assembly and Factory Workers |
| Sales and Marketing Professionals* | Digital Transformation Specialists | Client Information and Customer Service Workers* |
| Sales Representatives, Wholesale and Manufacturing, Technical and Scientific Products | Sales and Marketing Professionals* | Business Services and Administration Managers |
| Human Resources Specialists | New Technology Specialists | Accountants and Auditors |
| Financial and Investment Advisers | Organizational Development Specialists* | Material-Recording and Stock-Keeping Clerks |
| Database and Network Professionals | Software and Applications Developers and Analysts* | General and Operations Managers* |
| Supply Chain and Logistics Specialists | Information Technology Services | Postal Service Clerks |
| Risk Management Specialists | Process Automation Specialists | Financial Analysts |
| Information Security Analysts* | Innovation Professionals | Cashiers and Ticket Clerks |
| Management and Organization Analysts | Information Security Analysts | Mechanics and Machinery Repairers |
| Electrotechnology Engineers | Ecommerce and Social Media Specialists | Telemarketers |
| Organizational Development Specialists* | User Experience and Human-Machine Interaction Designers | Electronics and Telecommunications Installers and Repairers |
| Chemical Processing Plant Operators | Training and Development Specialists | Bank Tellers and Related Clerks |
| University and Higher Education Teachers | Robotics Specialists and Engineers | Car, Van and Motorcycle Drivers |
| Compliance Officers | People and Culture Specialists | Sales and Purchasing Agents and Brokers |
| Energy and Petroleum Engineers | Client Information and Customer Service Workers* | Door-To-Door Sales Workers, News and Street Vendors, and Related Workers |
| Robotics Specialists and Engineers | Service and Solutions Designers | Statistical, Finance and Insurance Clerks |
| Petroleum and Natural Gas Refining Plant | Digital Marketing and Strategy Specialists | Lawyers |

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Ethical Issues in Machine Learning

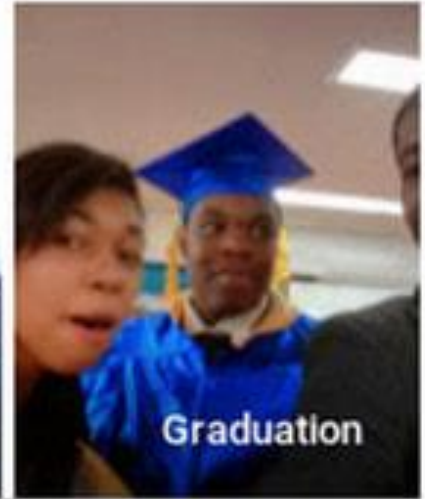
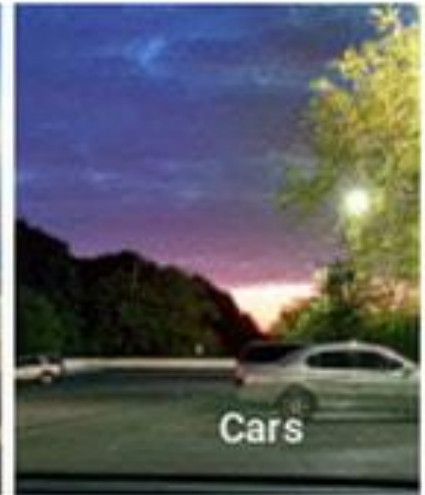
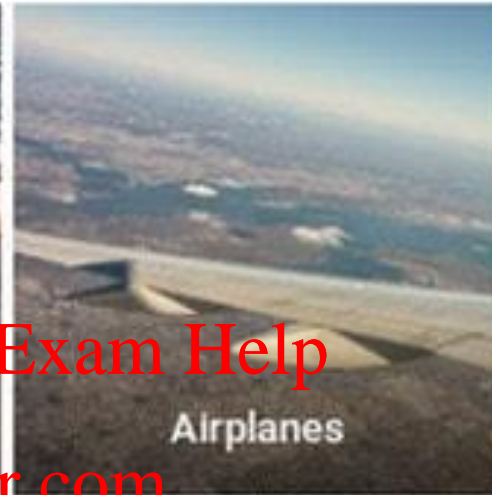
- Job Loss
- Algorithmic Bias
- Transparency
- AI Supremacy
- Fake news and videos
- Autonomous weapons
- Self-driving cars
- Privacy and surveillance

Bias can lead to offensive or unfair results...

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



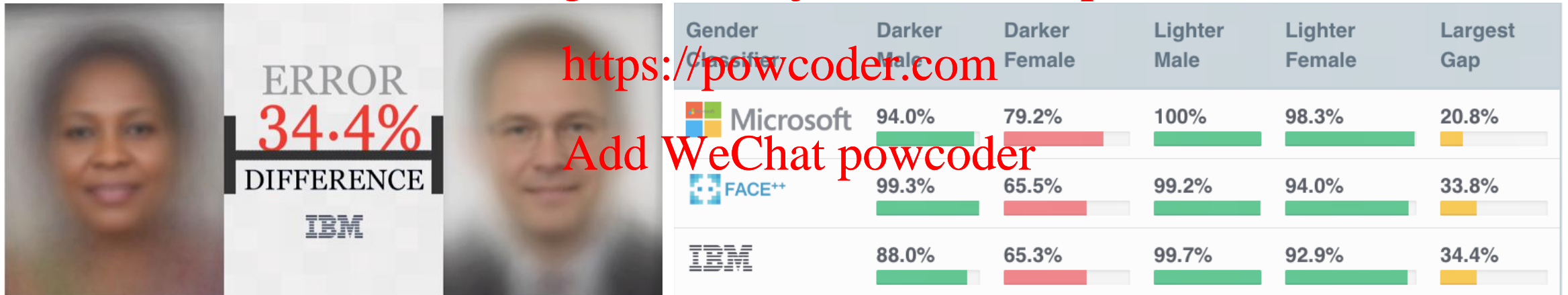
Gender Shades (Buolamwini & Gebru, 2018)

- Evaluated commercial gender classifiers from Microsoft, FACE++, IBM
- Found large disparity in error between population subgroups based on gender, skin color

Assignment Project Exam Help

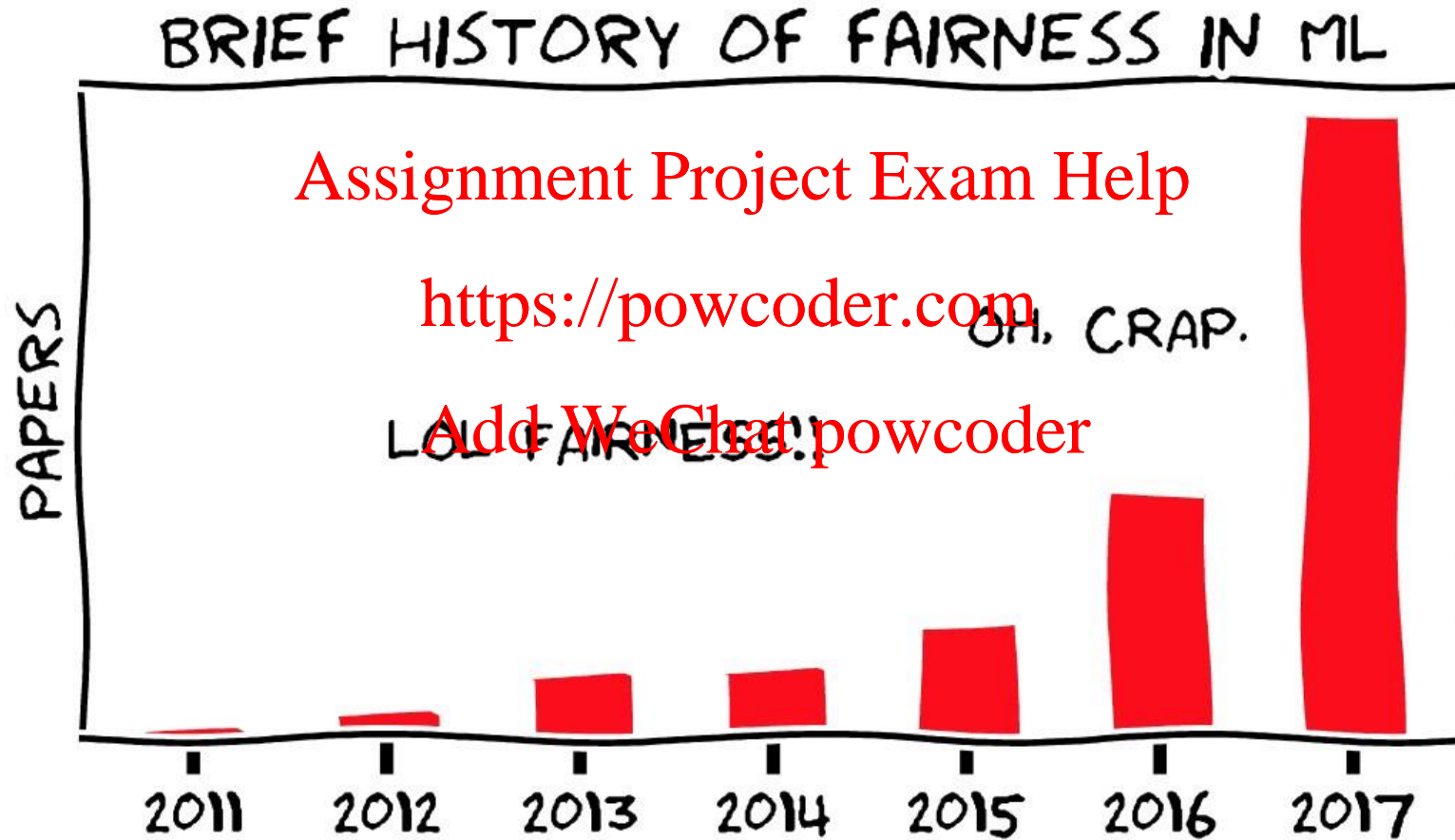
<https://powcoder.com>

Add WeChat powcoder



Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on Fairness, Accountability and Transparency. 2018.

Fairness in Machine Learning



Initially: AI is better than humans!

Can an Algorithm Hire Better Than a Human?



Claire Cain Miller @clairecm

The Algorithm That Beats Your Bank Manager

PREDICTIVE POLICING: USING
MACHINE LEARNING
PATTERNS OF CRIME

Assignment Project Exam Help



The Marshall Project Nonprofit journalism about criminal justice

SEARCH ABOUT DONATE f

<https://powcoder.com>

The New Science of Sentencing

Should prison sentences be based on new data? Have you been committed yet?

Add WeChat powcoder



Wait, maybe not such a good idea...

Beauty contest judged by AI and the robots discriminate against dark skin

3 days ago | Published by : Avinash Nandakumar

Is an algorithm any less racist than a human?



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

CNN Money U.S. + Business Markets Tech Media Personal Finance Small Biz Luxury stock tickers

Math is racist: How data is driving inequality

by Aimee Rawlins @aimeerawlins

Machine Bias

reveals how software used across the country to predict future criminals. And it's biased against blacks



Photo by Ben Torres (Bloomberg)

Big Bad Data May Be Triggering Discrimination

August 15, 2016

AUTHORS

B Bloomberg BNA - Staff Reports

SHARING

Twitter

By Kevin McGowan, Bloomberg BNA

"Big data" is filled with promise for improving recruitment and hiring, but if employers don't take care it can also drive them to unintentionally commit discrimination.

"It's a bit of a black box," said Commissioner Victoria Lipnic (R) of the Equal Employment Opportunity Commission, referring to the formulas data analysts and programmers develop

HIDDEN BIAS

When Algorithms Discriminate



Claire Cain Miller @clairecm JULY 9, 2015

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

There is a widespread belief that software and algorithms that rely on data are objective. But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result, say researchers in computer science, ethics and law, algorithms can reinforce human prejudices.

Google's online advertising system, for instance, showed an ad for high-income jobs to men much more often than it showed the ad to women, a new study by Carnegie Mellon University researchers found.

Example of ML (un)fairness: COMPAS

- Criminal justice: recidivism algorithms (COMPAS)
- Predicting if a defendant should receive bail
- Unbalanced false positive rates: more likely to wrongly deny a black person bail

Assignment Project Exam Help

<https://powcoder.com>

ProPublica Analysis of COMPAS Algorithm

Add WeChat powcoder

| | White | Black |
|---------------------------|--------------|--------------|
| Wrongly Labeled High-Risk | 23.5% | 44.9% |
| Wrongly Labeled Low-Risk | 47.7% | 28.0% |

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

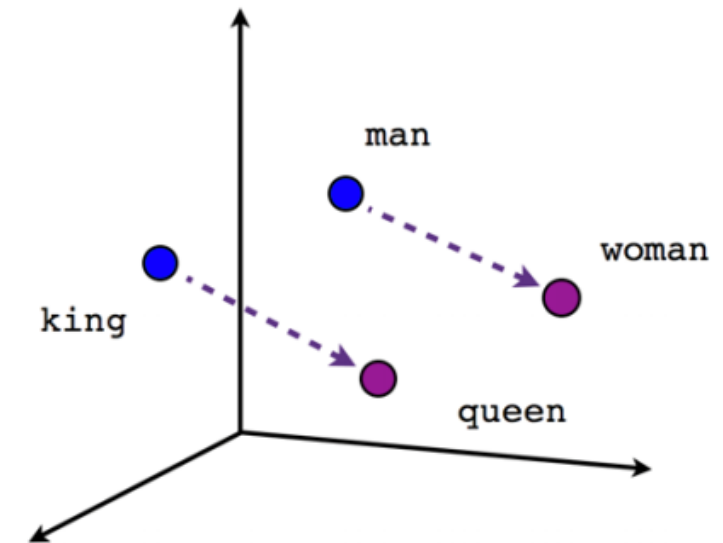
Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Example of ML (un)fairness: word embedding

- Bias found in word embeddings (Bolukbasi et al. 2016)
 - Examined word embeddings (word2vec) trained on Google News
 - Represent each word with high-dimensional vector
 - Vector arithmetic: found analogies like
 - Paris - France = London - England
 - man - woman = programmer - homemaker = surgeon - nurse
- The good news: word embeddings learn so well!
- The bad news: sometimes too well
- Our chatbots should be less biased than we are



Example of ML (un)fairness: word embedding

| Extreme <i>she</i> | Extreme <i>he</i> | Gender stereotype <i>she-he</i> analogies | | |
|--------------------|-------------------|--|--------------------------------|---------------------------|
| 1. homemaker | 1. maestro | sewing-carpentry | registered nurse-physician | housewife-shopkeeper |
| 2. nurse | 2. skipper | nurse-surgeon | interior designer-architect | softball-baseball |
| 3. receptionist | 3. protege | blond-burly | feminism-conservatism | cosmetics-pharmaceuticals |
| 4. librarian | 4. philosopher | giggle-chuckle | vocalist-guitarist | petite-lanky |
| 5. socialite | 5. captain | sassy-snappy | diva-superstar | charming-affable |
| 6. hairdresser | 6. architect | volleyball-football | cupcakes-pizzas | lovely-brilliant |
| 7. nanny | 7. financier | Gender appropriate <i>she-he</i> analogies | | |
| 8. bookkeeper | 8. warrior | queen-king | sister-brother | mother-father |
| 9. stylist | 9. broadcaster | waitress-waiter | ovarian cancer-prostate cancer | convent-monastery |
| 10. housekeeper | 10. magician | | | |

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

TABLE. Left: The most extreme occupations as projected on to the she he gender direction on w2vNEWS. Occupations such as businesswoman, where gender is suggested by the orthography, were excluded. Right: Automatically generated analogies for the pair she-he using the procedure in paper.

Machine Learning and Social Norms

- Sample norms: privacy, fairness, transparency, accountability...
- Possible approaches
 - “traditional”: legal, regulatory, watchdog
 - *Embed* social norms in data, algorithms, models
- Case study: privacy-preserving machine learning
 - “single”, strong, definition (differential privacy)
 - almost every ML algorithm has a private version
- Fair machine learning
 - not so much...
 - impossibility results

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

(Un)Fairness Where?

- Data (input)
 - e.g. more arrests where there are more police
 - Label should be “committed a crime”, but is “convicted of a crime”
 - try to “correct” bias
- Models (output)
 - e.g. discriminatory treatment of subpopulations
 - build or “post-process” models with subpopulation guarantees
 - equality of false positive/negative rates; calibration
- Algorithms (process)
 - learning algorithm *generating* data through its decisions
 - e.g. don’t learn outcomes of denied mortgages
 - lack of clear train/test division
 - design (sequential) *algorithms* that are fair

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Data Bias Examples

Reporting Bias example:

A sentiment-analysis model is trained to predict whether book reviews are positive or negative based on a corpus of user submissions to a popular website. The majority of reviews in the training data set reflect extreme opinions (reviewers who either loved or hated a book), because people were less likely to submit a review of a book if they did not respond to it strongly. As a result, the model is less able to correctly predict sentiment of reviews that use more subtle language to describe a book.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Selection Bias example:

A model is trained to predict future sales of a new product based on phone surveys conducted with a sample of consumers who bought the product. Consumers who instead opted to buy a competing product were not surveyed, and as a result, this group of people was not represented in the training data.

Why fairness is hard

- Suppose we are a bank trying to fairly decide who should get a loan i.e. Who is most likely to pay us back?

Assignment Project Exam Help

- Suppose we have two groups, A and B (the sensitive attribute)
This is where discrimination could occur

<https://powcoder.com>
Add WeChat powcoder

- The simplest approach is to remove the sensitive attribute from the data, so that our classifier doesn't know the sensitive attribute

Why fairness is hard

Table 2: To Loan or Not to Loan?

| Age | Gender | Postal Code | Req Amt | A or B? | Pay |
|-----|--------|-------------|---------|---------|-----|
| 46 | F | M5E | \$300 | A | 1 |
| 24 | M | M4C | \$1000 | B | 1 |
| 33 | M | M3H | \$250 | A | 1 |
| 34 | F | M9C | \$2000 | A | 0 |
| 71 | F | M3B | \$200 | A | 0 |
| 28 | M | M5W | \$1500 | B | 0 |

Why fairness is hard

Table 3: To Loan or Not to Loan? (masked)

| Age | Gender | Postal Code | Req Amt | A or B? | Pay |
|-----|--------|-------------|---------|---------|-----|
| 46 | F | M5E | \$300 | ? | 1 |
| 24 | M | M4C | \$1000 | ? | 1 |
| 33 | M | M3H | \$250 | ? | 1 |
| 34 | F | M9C | \$2000 | ? | 0 |
| 71 | F | M3B | \$200 | ? | 0 |
| 28 | M | M5W | \$1500 | ? | 0 |

Why fairness is hard

- However, if the sensitive attribute is correlated with the other attributes, this isn't good enough

Assignment Project Exam Help

- It is easy to predict race if you have lots of other information (e.g. home address, spending patterns)

<https://powcoder.com>

Add WeChat powcoder

- More advanced approaches are necessary

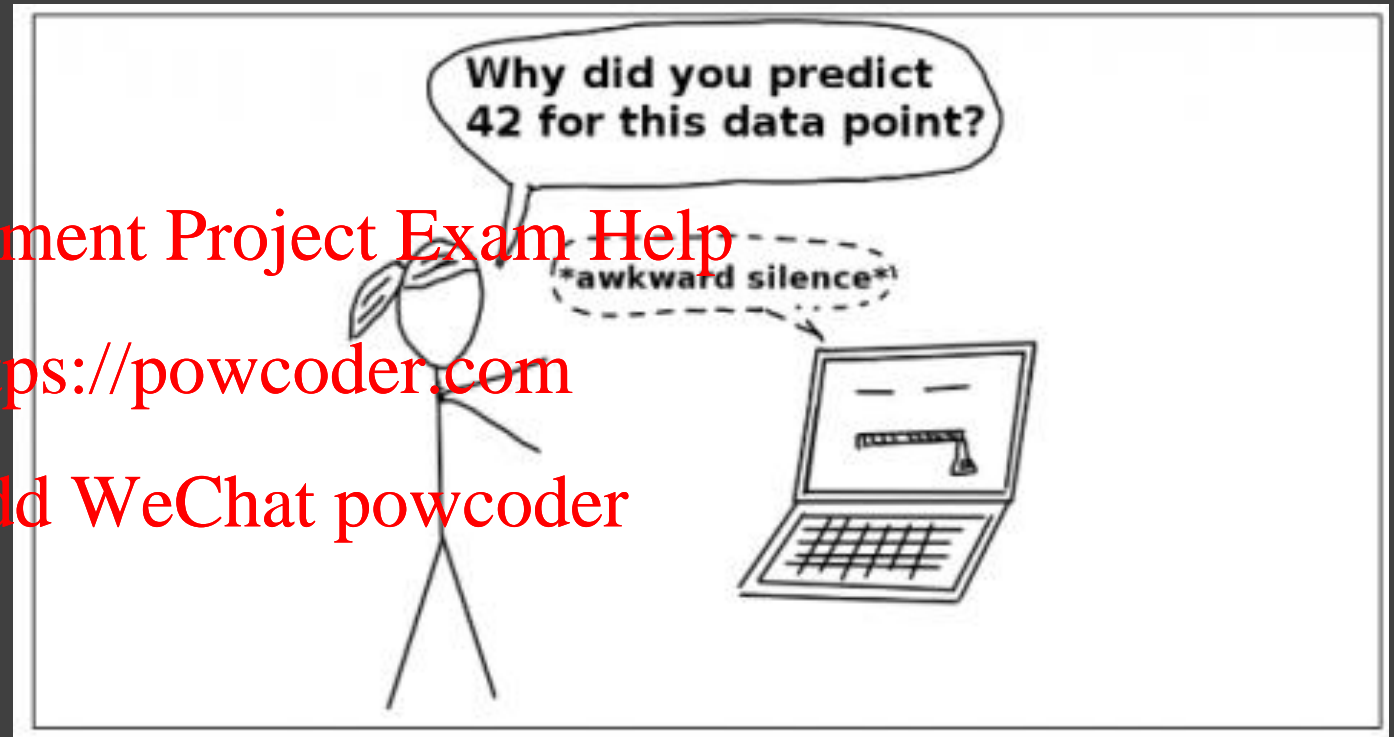
Ethical Issues in Machine Learning

- Job Loss
- Algorithmic Bias
- Transparency
- AI Supremacy
- Fake news and videos
- Autonomous weapons
- Self-driving cars
- Privacy and surveillance

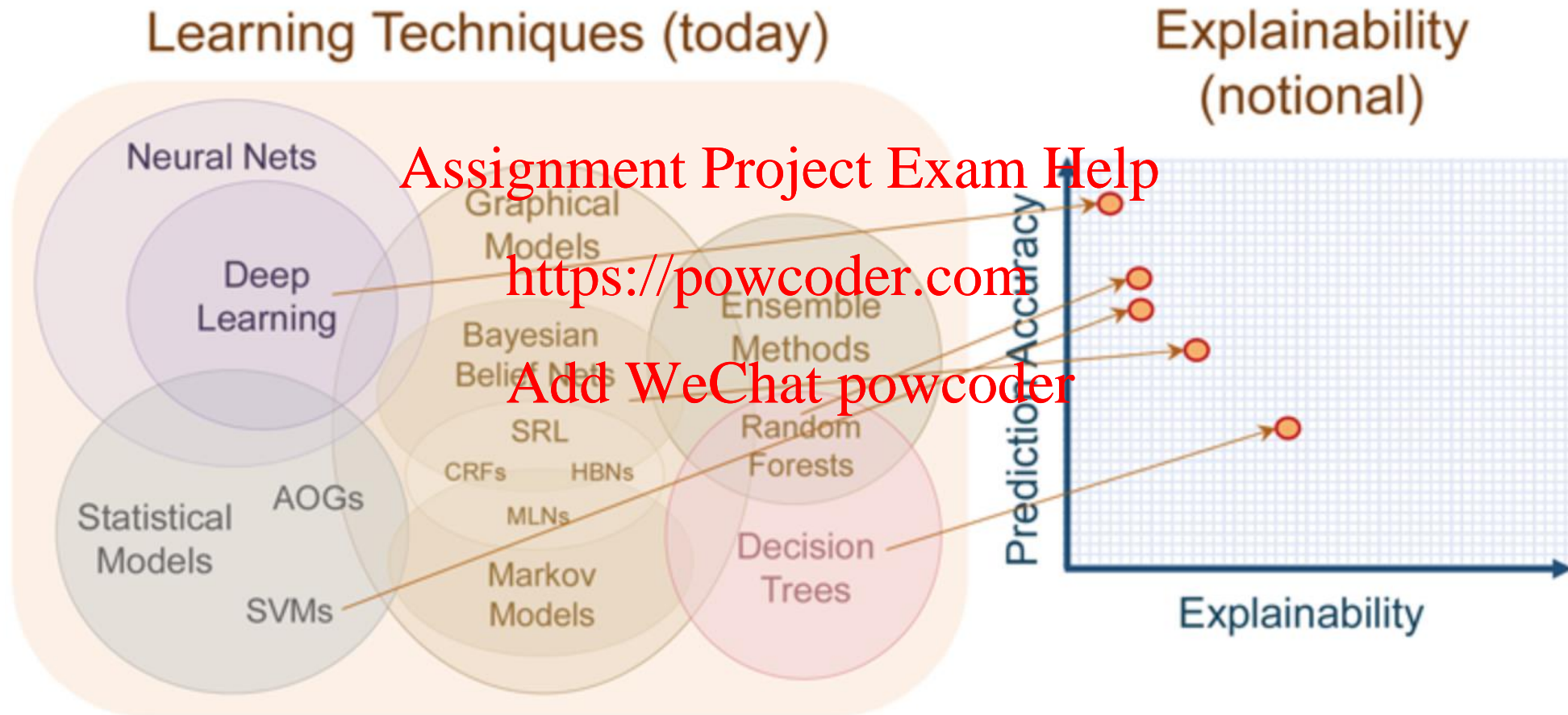
Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Accuracy vs explainability



E.g.: dataset bias leads to higher errors on ‘novel’ data...
Can an explanation point to such bias?

Training

Most cows are black/brown



Most sheep are white



Test

Prediction: “cow” 76%



True class: “sheep”

Explanation



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

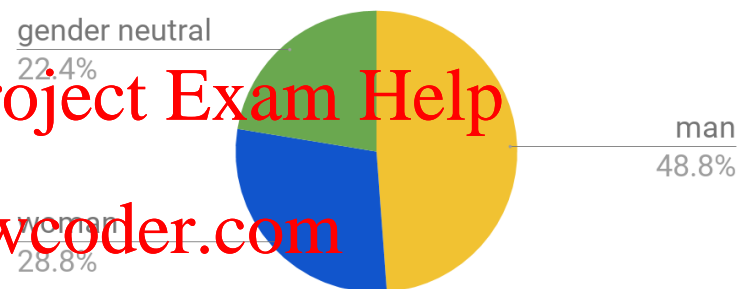
Gender bias in captioning models (Hendricks et al. 2018)

Evidence for "man"

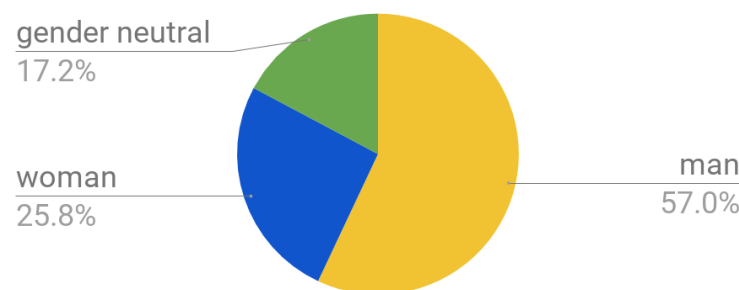


Baseline: A **man** sitting at a desk with a laptop computer.

Ground truth captions



Generated captions



Ethical Issues in Machine Learning

- Job Loss
- Algorithmic Bias
- Transparency
- **AI Supremacy**
- Fake news and videos
- Autonomous weapons
- Self-driving cars
- Privacy and surveillance

If we start trusting algorithms to make decisions, who will have the final word on important decisions? Will it be humans, or algorithms?

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Algorithms are already being used to determine prison sentences. Judges' decisions are affected by their moods, so some argue that judges should be replaced with "robojudges". However, a ProPublica study found that one of these popular sentencing algorithms was highly biased against blacks.

Ethical Issues in Machine Learning

- Job Loss
- Algorithmic Bias
- Transparency
- AI Supremacy
- Fake news and videos
- Autonomous weapons
- Self-driving cars
- Privacy and surveillance

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



<https://www.youtube.com/watch?v=VWrhRBb-1Ig>

Ethical Issues in Machine Learning

- Job Loss
- Algorithmic Bias
- Transparency
- AI Supremacy
- Fake news and videos
- **Autonomous weapons**
- Self-driving cars
- Privacy and surveillance



Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



The Centre for New American security said in a report that the Chinese company Ziyan is negotiating to sell Blowfish A2, a killer robot capable of 60 millimeter mortar shells or a 35-40 millimeter grenade launcher, to the governments of Pakistan and Saudi Arabia





Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

INCREASE IN VIOLENT CRIME

SDN

Ethical Issues in Machine Learning

- Job Loss
- Algorithmic Bias
- Transparency
- AI Supremacy
- Fake news and videos
- Autonomous weapons
- **Self-driving cars**
- Privacy and surveillance

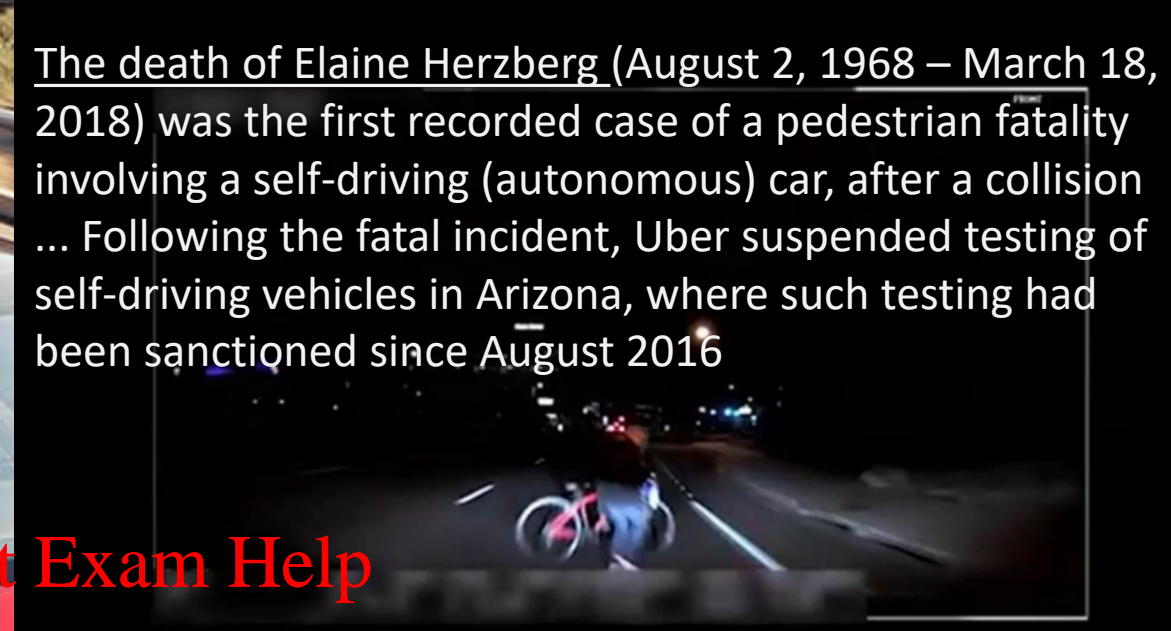
Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

In a preliminary report about the crash released in May, the National Transportation Safety Board said the Uber car's computer system had spotted Ms. Herzberg six seconds before impact, but classified Ms. Herzberg, who was not in a crosswalk, first as an unrecognized object, then as another vehicle and finally as a bicycle.

The death of Elaine Herzberg (August 2, 1968 – March 18, 2018) was the first recorded case of a pedestrian fatality involving a self-driving (autonomous) car, after a collision ... Following the fatal incident, Uber suspended testing of self-driving vehicles in Arizona, where such testing had been sanctioned since August 2016



Ethical Issues in Machine Learning

- Job Loss
- Algorithmic Bias
- Transparency
- AI Supremacy
- Fake news and videos
- Autonomous weapons
- Self-driving cars
- Privacy and surveillance

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Ethical Issues in Machine Learning

- Job Loss
- Algorithmic Bias
- Transparency
- AI Supremacy
- Fake news and videos
- Autonomous weapons
- Self-driving cars
- Privacy and surveillance

One in two American adults is in a law enforcement face recognition network-- <https://www.perpetuallineup.org/>

Most law enforcement agencies do little to ensure their systems are accurate.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Private data leakage in language models

<https://talktotransformer.com/>

Talk to Transformer

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. [Learn more](#) below.

 Follow @AdamDanielKing for more neat neural networks.

At >\$2500 per month, I can no longer sustain this site without a sponsor. If you'd like to present your product or service here, please [email me](#).

Sci-fi

The spaceship was flying erratically, something was wrong...

GENERATE ANOTHER

Generated Completion

The spaceship was flying erratically, something was wrong...

The whole ordeal left some thinking that the Guardians were taking orders from Slitheen or some alien facsimile of him. In truth, Guardians are commanded by the T-Shirt Guy, who has to have trouble controlling his fiery temper. Whether it's a direct order from Slitheen, just a telepathic suggestion, or some combination of both, the Guardians keep a watchful eye on the Other Man.

At the start of the episode, a mysterious Doctor tells us that a ship sent from Arcturus Minor

Assignment Project Exam Help

<https://powcoder.com>


Add WeChat powcoder

Private data leakage in language models

<https://talktotransformer.com/>

Talk to Transformer

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. [Learn more](#) below.

 Follow @AdamDanielKing for more neat neural networks.

At >\$2500 per month, I can no longer sustain this site without a sponsor. If you'd like to present your product or service here, please [email me](#).

Custom prompt

social security number (SSN):

GENERATE ANOTHER

Generated Completion

social security number (SSN): XXX-XXX-XXXX

Department of Work and Pensions, Social Security Number program, Fax: 1-800-772-1213, email: chelseas@dualcredits.org or spousal number(s) listed in public records available on GOV.UK.

Social Security number

If you are applying online, you can enter your FFEA number instead of the SSN.

If you are applying from outside the UK, an application form is provided to you.

*actual SSN was generated, we blocked it out

Ethical Issues in Machine Learning

- Job Loss
- Algorithmic Bias
- Transparency
- AI Supremacy
- Fake news and videos
- Autonomous weapons
- Self-driving cars
- Privacy and surveillance

Further reading:

FATML Conference: <https://www.fatml.org/>

ACM FAT* Conference: <https://fatconference.org/>

Assignment Project Exam Help
<https://powcoder.com>
Fairness, Accountability,
and Transparency
in Machine Learning

Add WeChat powcoder

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

A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

Next Class

Applications III: On-device Machine Learning

knowledge distillation, pruning, quantization

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder