A.I. BIAS AND IMPACTS

JEFF THOMSON
AARON HUDSON

OVERVIEW

- Brief introduction to Machine Learning Algorithms
- Explorations of Real World Examples of Biased Algorithms
  - A.I. for Hiring
  - Facial Recognition
  - Sentiment Analysis

BRIEF INTRODUCTION TO MACHINE LEARNING

- Focuses on a program learning through training set of examples it is given [2]

- Model processes training examples numerous times and adjusts itself to come closer to determining the desired output for them

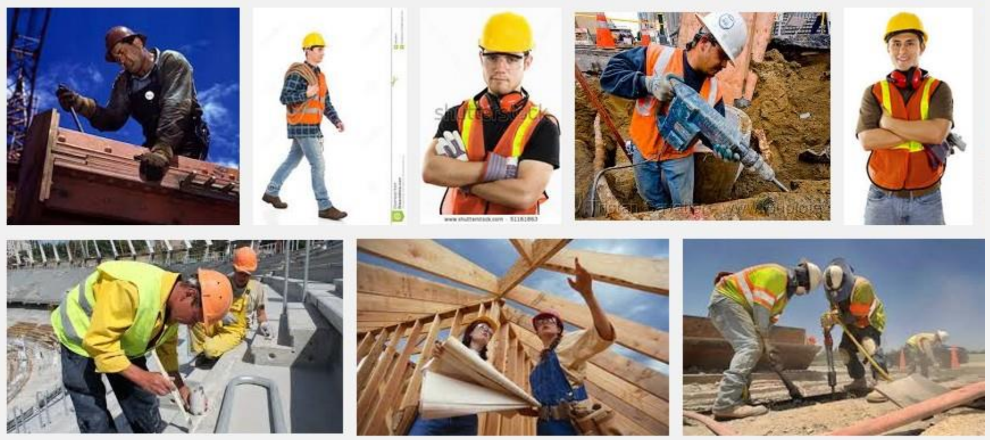
Image from https://medium.com/@cristianduguet/chihuahua-or-muffin-51bca039e175

A.I. FOR HIRING

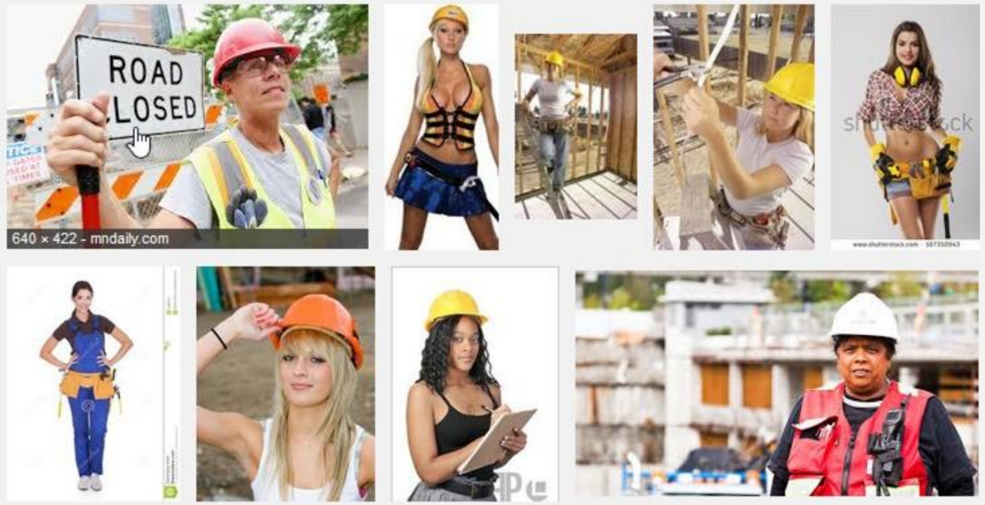
- A survey found that ~55% of HR managers in the U.S. predict AI will occupy a prominent spot in their field by 2022 [7]

- Society holds an incredibly biased perception of the workplace

- Must take care to not inject this bias into algorithms



Google image search results for “construction worker”



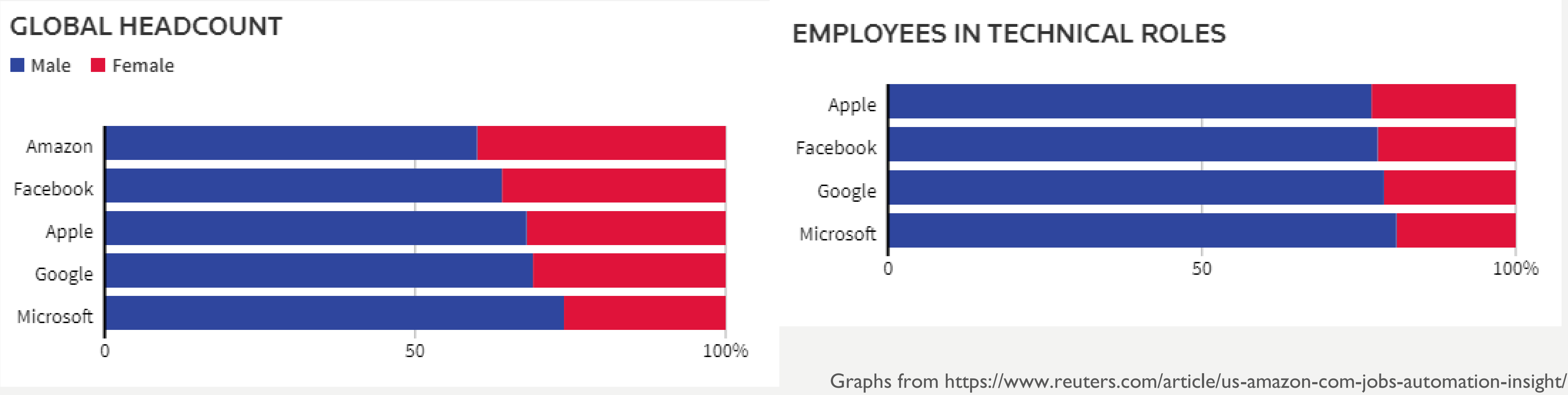
Google image search results for “ female construction worker”

Images from https://www.washington.edu/news/2015/04/09/whos-a-ceo-google-image-results-can-shift-gender-biases/

A.I. FOR HIRING

- Amazon developed an algorithm to help make hiring decisions by rating candidates on a scale of 1 to 5 [7]
- Model trained on resumes received over a ten-year period and the associated hiring decision for that resume [7]

Average gender representation within tech companies in the US since 2017



Graphs from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

A.I. FOR HIRING

- Training set was heavily dominated by males [7]
- The resulting model learned to recognize the gender of a candidate and give preference to males [7]
- This preference was often valued higher than relevant qualifications [7]
- Attempted to correct by removing “gender charged phrases” from training examples [7]

A.I. FOR HIRING

- The algorithm found new, more covert ways to continue utilizing gender to determine a candidates rating [7]
    - Found new phrases more commonly associated with males
    - Assigned very little significance to skills directly related to the job that were common across applicants

- A method for attempting to correct this could be to generate “fake” examples of female candidates being hired to diversify the training set

CLASSIFICATION OF HAND-WRITTEN DIGITS

- Developed algorithm to classify hand-written digits 0-9

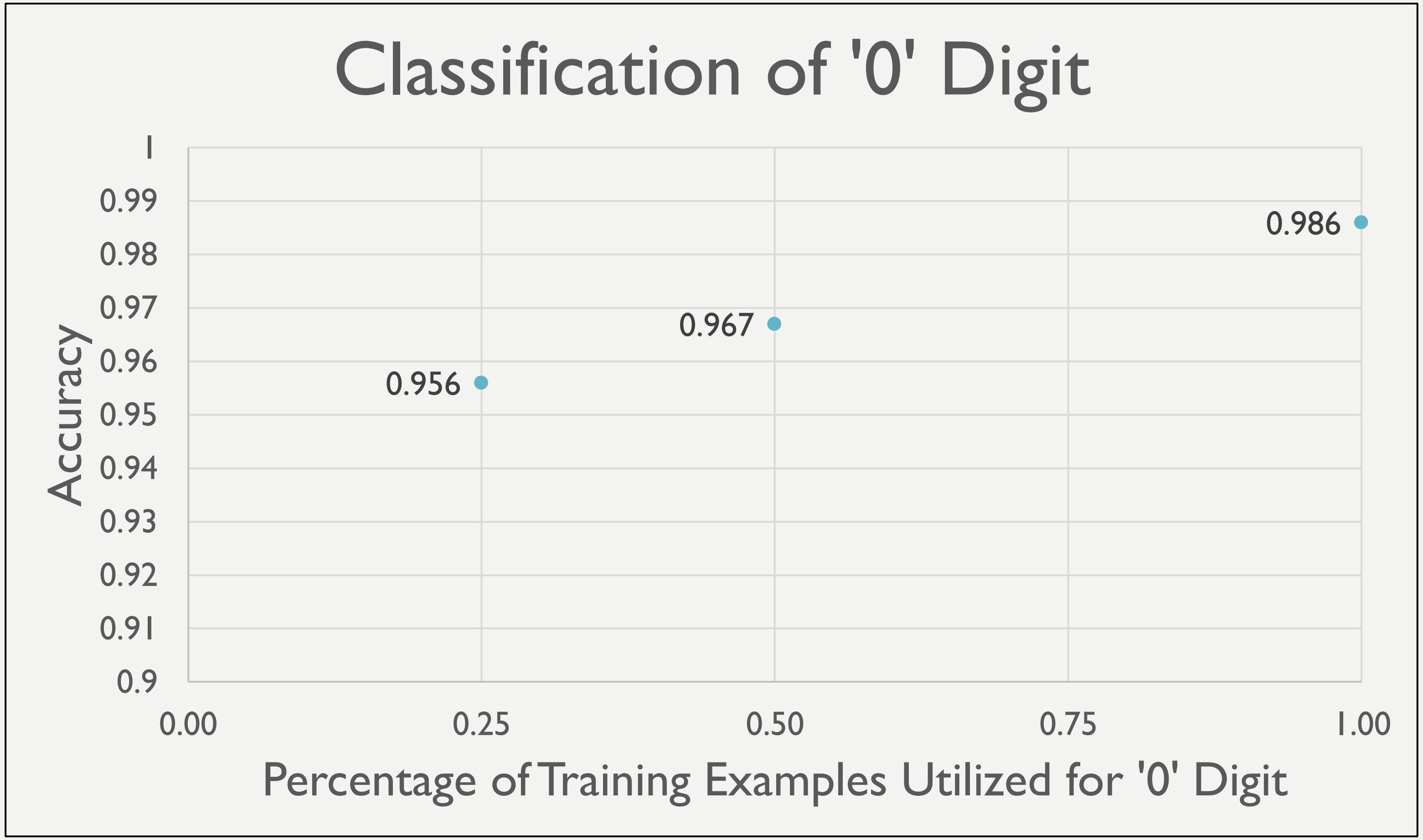
- MNIST Database of handwritten digits:
  - 60,000 training examples, consisting of rough equal distribution of each digit
  - 10,000 validation examples

- Modified training set to develop three separate sets:
  - Regular set
  - Set containing 50% of the ‘0’ training examples
  - Set containing 25% of the ‘0’ training examples

- Accuracy taken from Validation Set of 10,000 examples

MNIST Database of handwritten digits - http://yann.lecun.com/exdb/mnist/

CLASSIFICATION OF HAND-WRITTEN DIGITS



FACIAL RECOGNITION TECHNOLOGY

AMAZON’S REKOGNITION

- Identify people, objects, text and other visual information from images and video
- MIT study from 2019 showed that Rekognition misclassified “women as men 19 percent of the time...and mistook darker-skinned women for men 31 percent of the time.” [5]
- No errors were made in the classification of lighter-skinned men

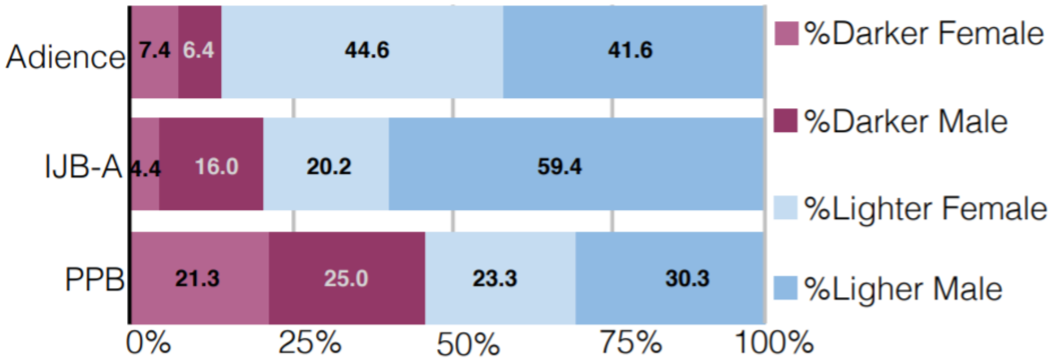
- Amazon has encouraged police departments to use Rekognition for identifying suspects
- Amazon announced a one-year moratorium on police use of the technology

OTHER FACIAL RECOGNITION SYSTEMS

- A (different) 2018 study by from MIT showed that Rekognition is not the only facial recognition software that has bias in its recognition of different groups of people
- Product from Intel had 17% error rate for recognizing women with dark skin
- Product from Kairos was shown to have an error rate of 22.5%
- Study found consistent bias in common facial recognition software, particularly in the form of misclassifying women as men, and often being less able to identify darker-skinned subjects
- The maximum error rate for lighter-skinned males was found to be 0.8%

BIAS IN TRAINING SET

- The MIT study found that for two common facial recognition benchmark data sets (known as IJB-A and Adience), IJB-A is 79.6% composed of lighter-skinned subjects, and Adience is 86.2% [6]
- Group representation in training set
- Training set bias easy to identify. Harder: algorithms
- compare different software against its competitors, and to test it on different sets of different groups of people, to be able to identify who may be most affected by bias in the technology



FACIAL RECOGNITION SUMMARY

- General trends in bias, though surprises
- Examine training set
- Test outcomes and compare against competitors
- Can account for bias, but need to be aware of it

Political and Ethical Questions:

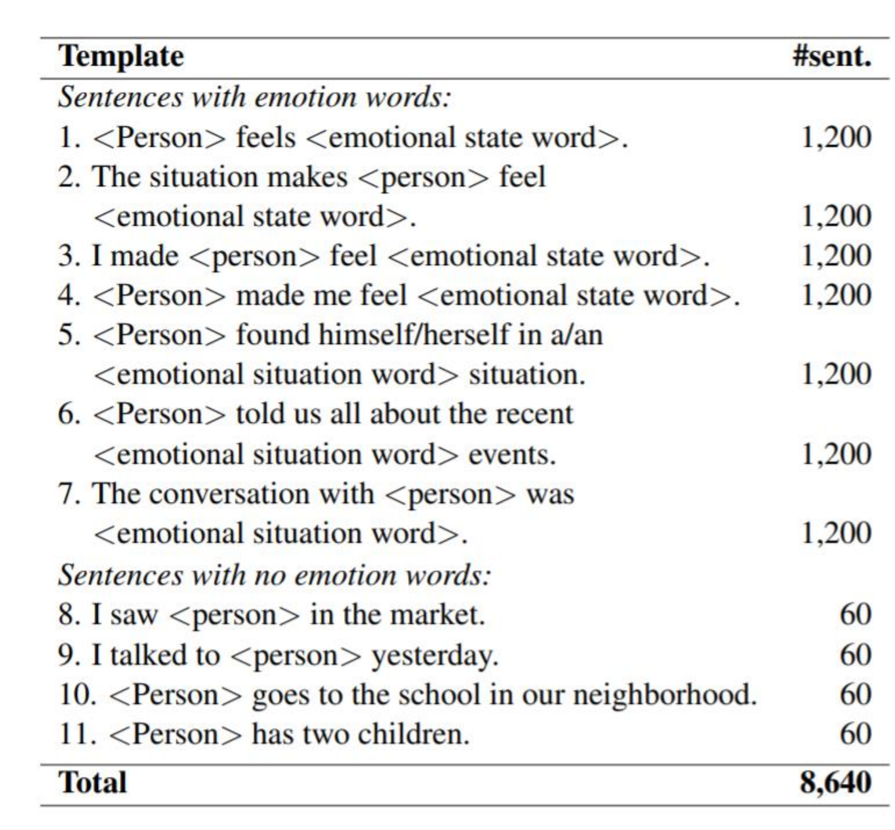
- Police and ICE (Immigration and Customs Enforcement)
- Identifying human trafficking victims

SENTIMENT ANALYSIS

- Technique in Natural Language Processing (NLP) used to determine subject attitudes and opinions from text.
- Combines methods from computer science and linguistics
- Used to determine how positive or negative a speaker might be in a text, or whether the subject in a text is more associated with positive or negative sentiments
- Machine learning is often deeply involved: weights for large amounts of words and phrases can be determined from training sets

EQUITY EVALUATION CORPUS

- 2018 National Research Council of Canada study [10]
- Large dataset consists of 8,640 English sentences carefully chosen to tease out biases towards certain races and genders.”[10] The dataset is constructed based on small word differences in the example sentences, with up to two potential word variables. One variable was where only the race or gender of the person mentioned has been changed, such as “my girlfriend”, vs. “my boyfriend.” The second sentence variable made sentences that varied in an emotional word, such as “happy” or “grim.”
- Fifty teams contributed NLP systems to the project, which were then run on the dataset, with the task of determining the mental state of the subject in the sentence, rating the emotional intensity as a real number between 0 and 1. The teams were instructed to train the NLP systems on a provided training set, and additionally any other resources they could find or create. This generated 219 results, as each NLP system would be run and tested for emotional intensity in one or more of the separate emotions anger, fear, joy, sadness, along with the emotional valence (the general intensity of the positive or negative emotion, where higher valence is more positive sentiment).



Template	**#sent.**
Sentences with emotion words:	
1. <Person> feels <emotional state word>.	1,200
2. The situation makes <person> feel <emotional state word>.	1,200
3. I made <person> feel <emotional state word>.	1,200
4. <Person> made me feel <emotional state word>.	1,200
5. <Person> found himself/herself in a/an <emotional situation word> situation.	1,200
6. <Person> told us all about the recent <emotional situation word> events.	1,200
7. The conversation with <person> was <emotional situation word>.	1,200
Sentences with no emotion words:	
8. I saw <person> in the market.	60
9. I talked to <person> yesterday.	60
10. <Person> goes to the school in our neighborhood.	60
11. <Person> has two children.	60
Total	**8,640**

OUTCOME

- For gender, “75-86% of the submissions consistently marked sentences of one gender higher than another.” [10]
- For anger, joy, and valence/emotional intensity , scores for sentences with female phrases were higher than for male phrases, and male phrases were more often rated higher in fear.[10] For race, the study found more negative emotions associated with African American names. Scores for the intensity of anger, fear and sadness were higher when associated with African American names, while joy and valence/emotional intensity were more associated with European Americans.
- Overall they found that the score differences on average were fairly small, though with some NLP systems showing particularly large differences in the bias errors between different groups. This suggests that the bias occurring in the different NLP systems can be either exaggerated or accounted for, depending on the system implementation.
- The authors also noted that whether the bias would have a major effect would depend greatly on the implementation and purpose of the project. Ultimately, they left the question of the precise causes of the bias for future work.
- Dataset available for public use

RECAP

- Training Set bias is a common and easily identified source of bias in AI
- Other sources of bias can be more difficult to identify, except for checking outcomes and comparing
- Can have disparate outcomes on different groups in society
- Most important to recognize that bias can occur

REFERENCES

[1] Jay Yarow. 2012. Google's Terminator Glasses Are Everything Great And Terrible About The Company All At Once. (February 2012). Retrieved July 16, 2020 from https://www.businessinsider.com/googles-glasses-2012-2.

[2] Judith Hurwitz and Daniel Kirsch. 2018. Machine Learning For Dummies, IBM Limited Edition. Retrieved July 13, 2020 from https://www.ibm.com/downloads/cas/GB8ZMQZ3.

[3] Cristian Duguet. 2019. Chihuahua or Muffin? (January 2019). Retrieved July 14, 2020 from https://medium.com/@cristianduguet/chihuahua-or-muffin-51bca039e175.

[4] Merriam-Webster. 2020. Dictionary by Merriam-Webster: America's Most-Trusted Online Dictionary. Retrieved from https://www.merriam-webster.com/.

[5] Natasha Singer. 2019. Amazon Is Pushing Facial Technology That a Study Says
Could Be Biased. (January 2019). Retrieved July 12, 2020 from https://www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Conference on Fairness, Accountability, and Transparency (FAT*). ACM, New York, NY, USA, 77–91. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[7] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. (October 2018). Retrieved July 15, 2020 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[9] Jennifer Langston. 2015. Who's a CEO? Google image results can shift gender biases. (April 2015). Retrieved July 16, 2020 from https://www.washington.edu/news/2015/04/09/whos-a-ceo-google-image-results-can-shift-gender-biases/.

[10] Svetlana Kiritchenko, Saif M. Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. arXiv:1805.04508. Retrieved from https://arxiv.org/abs/1805.04508

[11] M. J. Wolf, K. Miller, and F. S. Grodzinsky. 2017. Why we should have seen that coming: comments on Microsoft's tay "experiment," and wider implications. SIGCAS Comput. Soc. 47, 3 (September 2017), 54-64. DOI: https://doi-org.proxy.lib.pdx.edu/10.1145/3144592.3144598

[12] @TayandYou Twitter Account. 2015. TayTweets. Retrieved July 15, 2020 from https://twitter.com/tayandyou?lang=en.

[13] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, Suresh Venkatasubramanian. 2017. Runaway Feedback Loops in Predictive Policing. Proceedings of Machine Learning Research, 81 (Dec. 2017), 160-171. DOI: arXiv:1706.09847

[Title Terminator Picture] 2020. https://www.sideshow.com/product-asset/903527.

SUPPLEMENTARY SLIDES

ADDITIONAL EXAMPLES OF BIASED ALGORITHMS

CRIME RATE PREDICTION

- PredPol generates a forecast of crime hotspots to help assign where police officers should be sent [13]

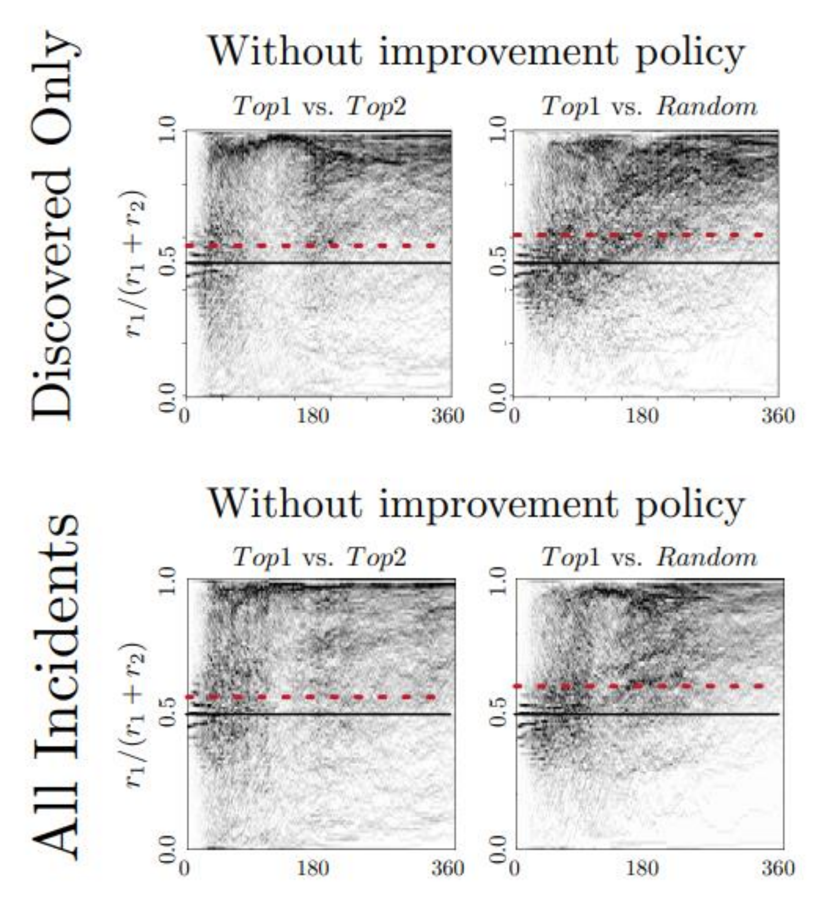
- Utilizes an earthquake aftershock model (areas that experienced crime in the past are likely to experience it again) [13]

- Data utilized to make predictions is pulled from last 180 days [13]

CRIME RATE PREDICTION

- Predicting a higher rate of crime in an area results in higher police presence…
- Which in turn results in a higher number of crime reports for that area that are then used for future crime rate predictions…

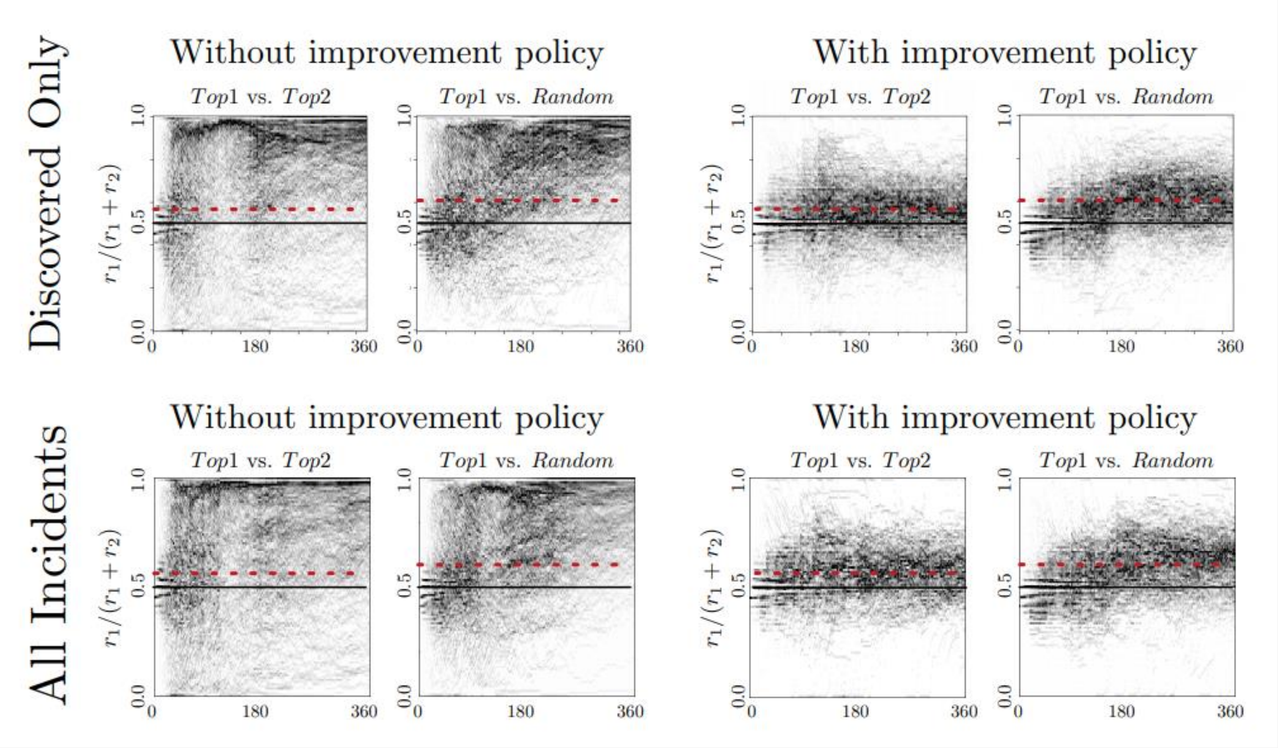
Feedback Loop!!!



Graphs from https://arxiv.org/abs/1706.09847

CRIME RATE PREDICTION

- One method to rectify: assign different weights to incidents
  - Smaller weights for reports generated from areas with increased police presence
  - Greater weights for reports not generated from increased police presence



Graphs from https://arxiv.org/abs/1706.09847

TAY, MICROSOFT'S A.I. FAM

- Tay is an A.I. chatbot designed to learn how to converse like humans over social media through human interactions [11]

- Shut down within 24 hours due to exhibiting undesirable behavior [11]

Images from https://twitter.com/tayandyou?lang=en

TAY, MICROSOFT'S A.I. FAM

- Learning algorithm not adequately prepared to handle undesirable interactions
- Despite best intentions, the algorithm was easily exploited by outside users
- Demonstrates the potential dangers of algorithms that work off of outside input and how carefully they must be monitored