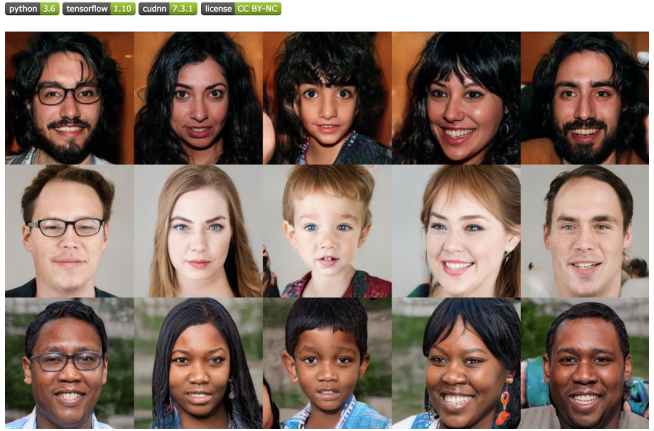
Bias in AI

Science of Science & Computational Discovery Lab

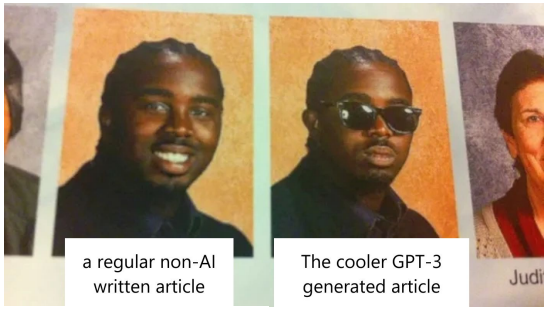
School of Information Studies SYRACUSE UNIVERSITY

Science of Science Summer School

StyleGAN — Official TensorFlow Implementation



AlphaGo



THE VERGE

ENTERTAINMENT \ TECH \ CULTURE

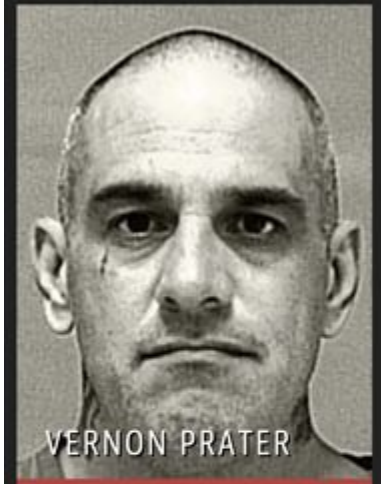
TL;DR

Neural net-generated memes are one of the best uses of AI on the internet

I can't stop making memes

By Jay Peters | @jaypeters | Apr 29, 2020, 12:54pm EDT

Two Petty Theft Arrests



LOW RISK 3

HIGH RISK 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.

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

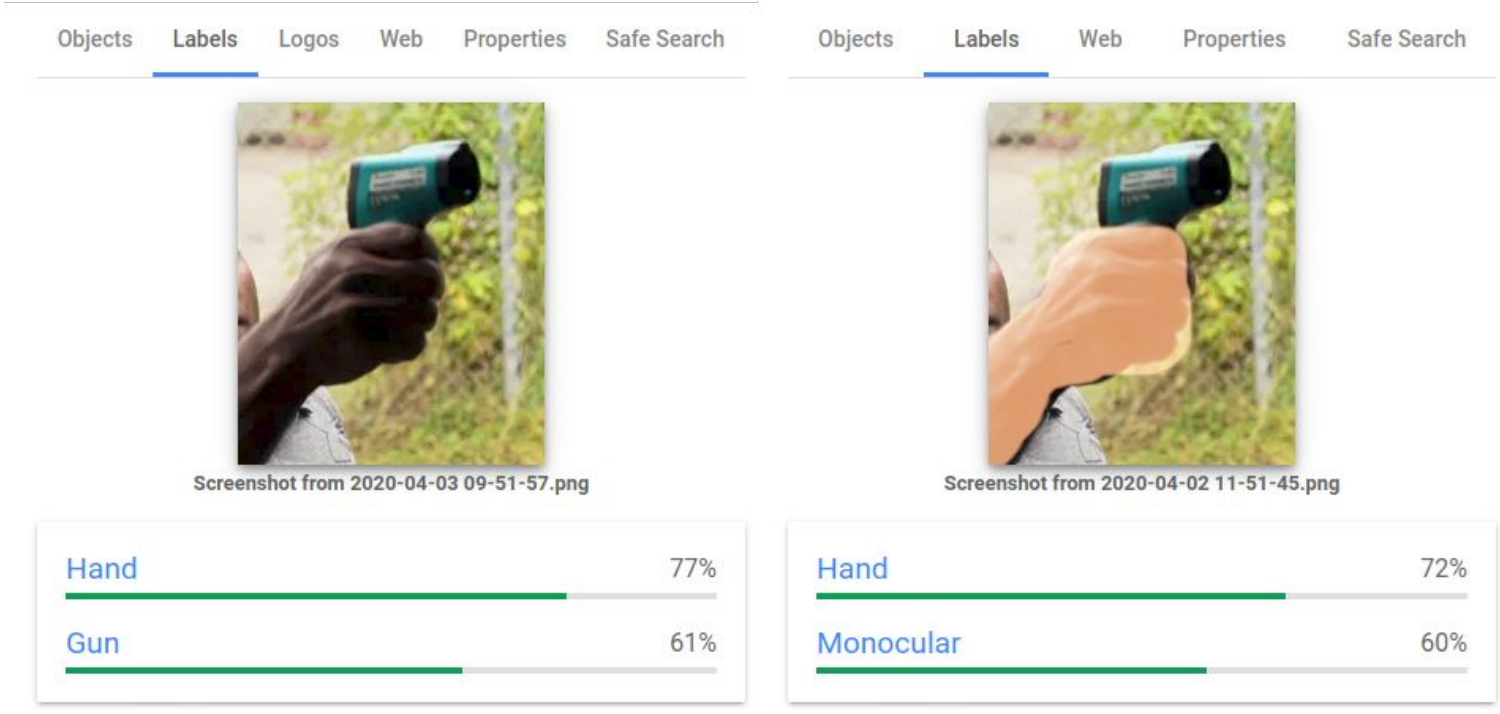
Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

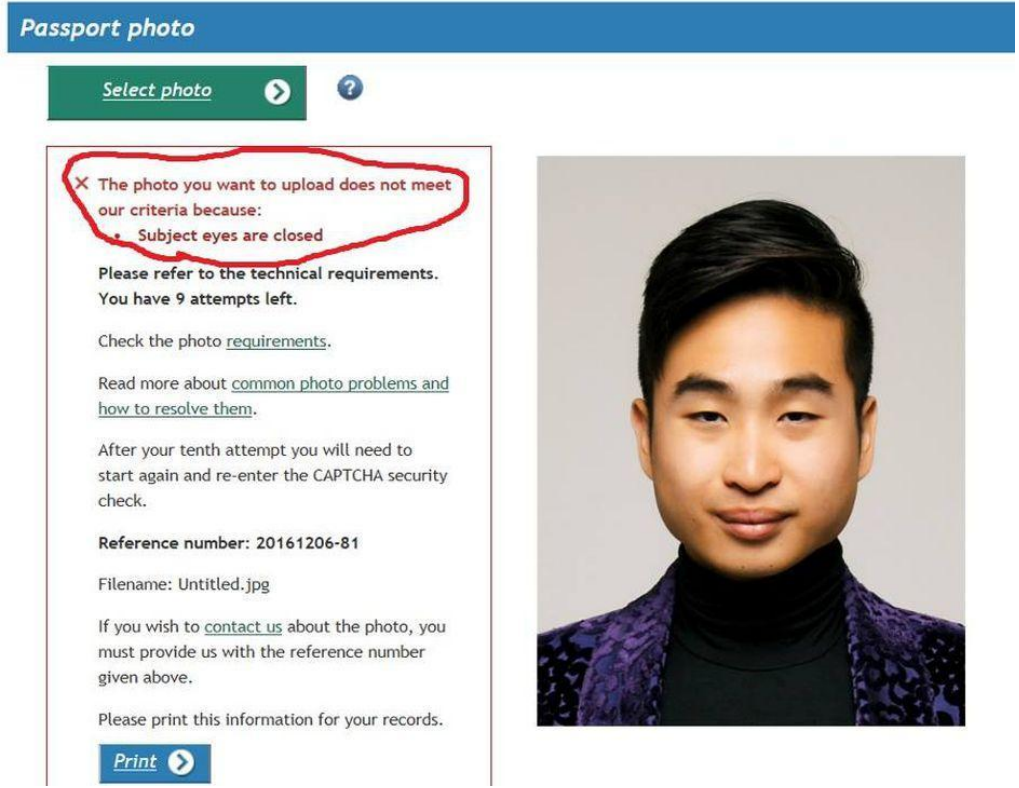
HIGH RISK 8

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

“Google apologizes after its Vision AI produced racist results”



https://algorithmwatch.org/en/story/google-vision-racism/



New Zealand passport robot tells applicant of Asian descent to open eyes

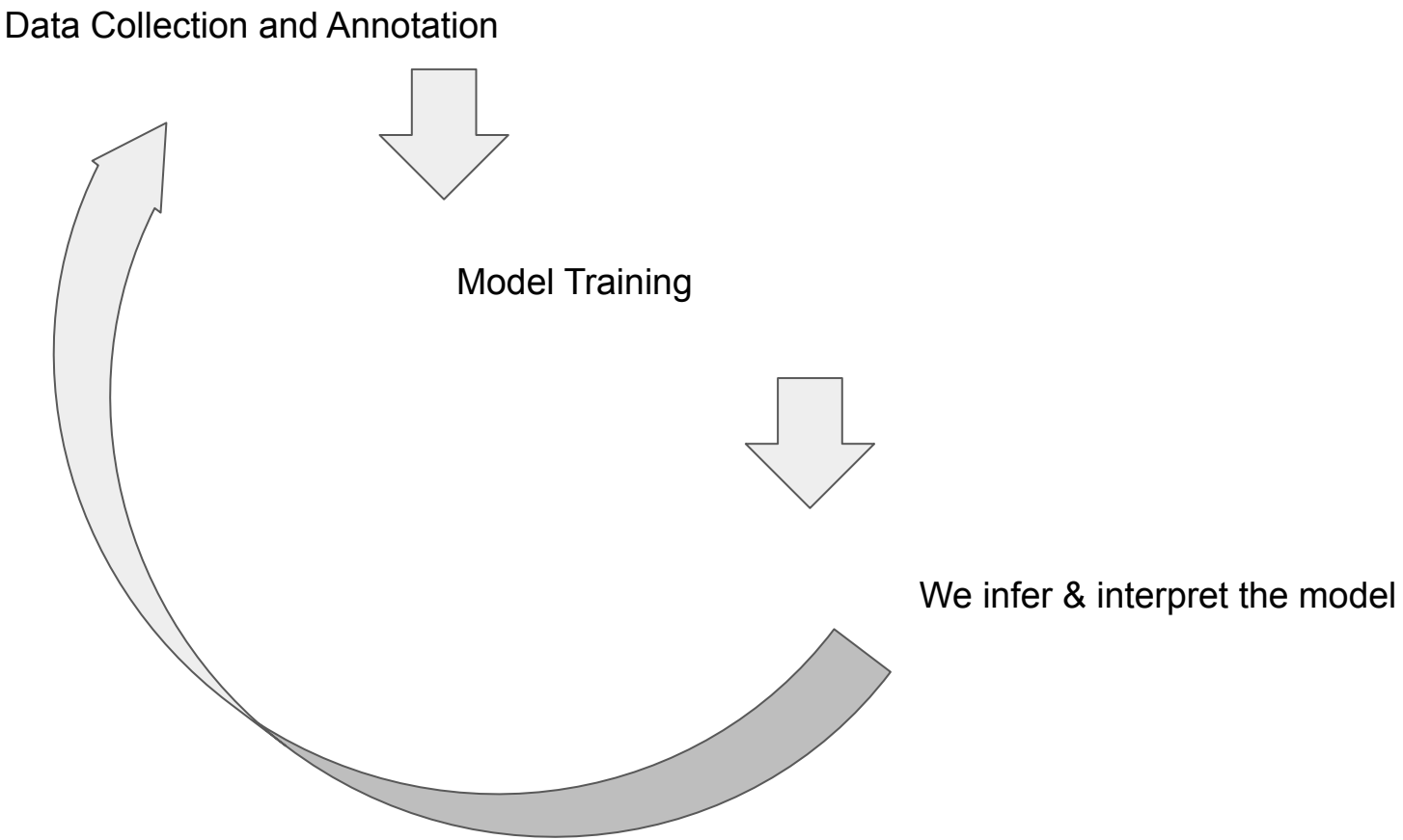
-REUTERS

(https://www.reuters.com/article/us-newzealand-passport-error/new-zealand-passport-robot-tells-applicant-of-asian-descent-to-open-eyes-idUSKBN13W0RL)

What?	How?
Allocation Harm	AI systems extend or withhold opportunities, resources
Quality-of-service Harm	AI systems do not work as well for one individual as it does for another
……	

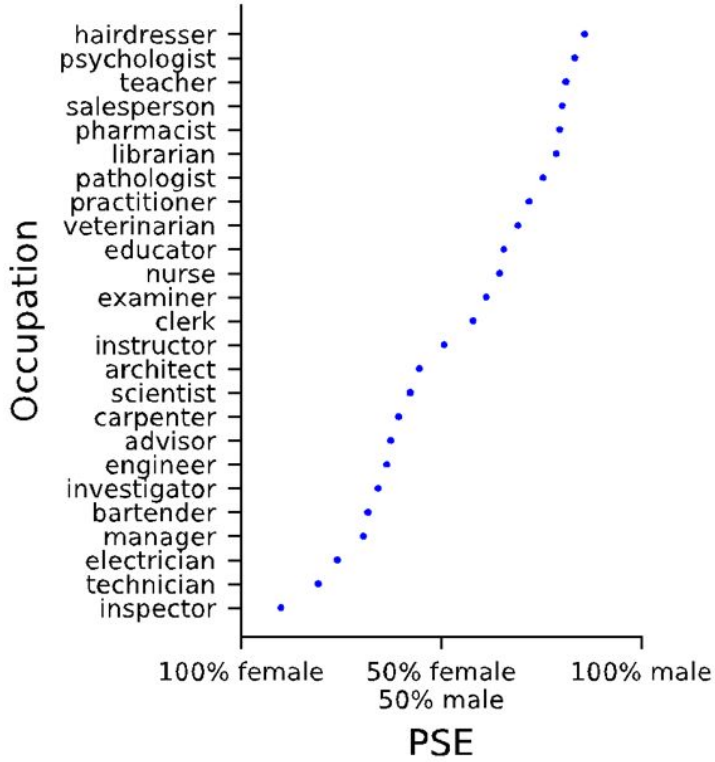
Amazon announced moratorium on police use of Amazon's facial recognition technology (June 10, 2020)

IBM announced that the company would exit the general-purpose face recognition business to fight racism (June 8, 2020)


Data Collection and Annotation
Model Training
We infer & interpret the model

Biases in Data

→ Biased data representation: There are some groups represented less positively than others
→ Biased Label: Annotation in your dataset will reflect the worldviews of your annotators
→ Reporting bias: What people share is not a reflection of real-world frequencies
→ Selection bias: Selection does not reflect a random sample



Liang., L., **Acuna, DE** (2020) *Artificial mental phenomena: Psychophysics as a framework to detect perception biases in AI models* In Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA, 10 pages.

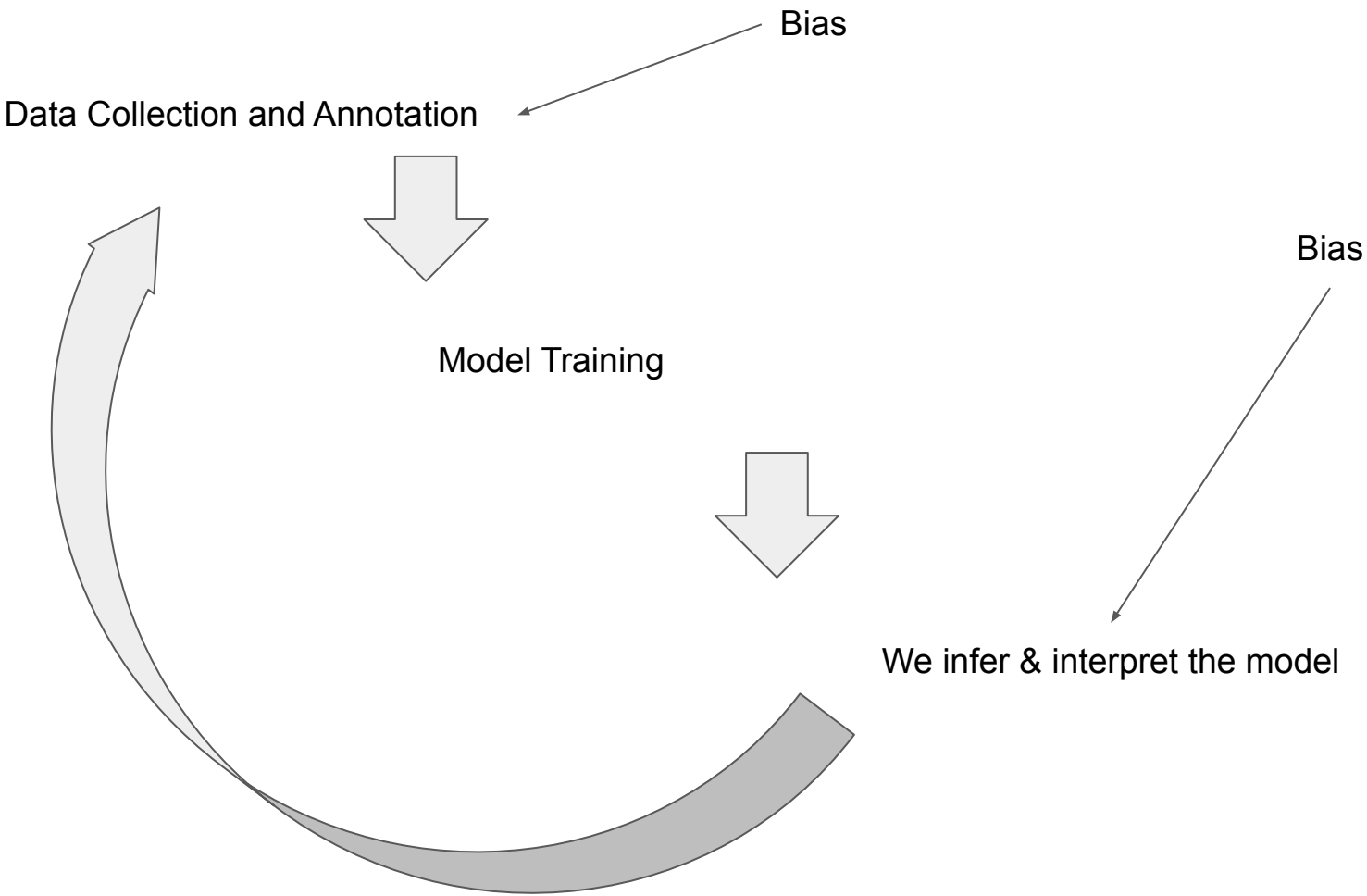
Interpretation:

→ Confirmation bias: The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses

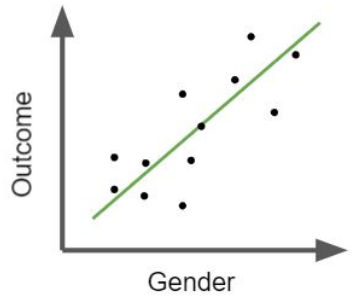
→ Overgeneralization: coming to conclusion based on information that is too general and/or not specific enough

→ Correlation fallacy: Confusing correlation with causation

→ Automation bias: propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation.


Bias
Data Collection and Annotation
Model Training
Bias
We infer & interpret the model

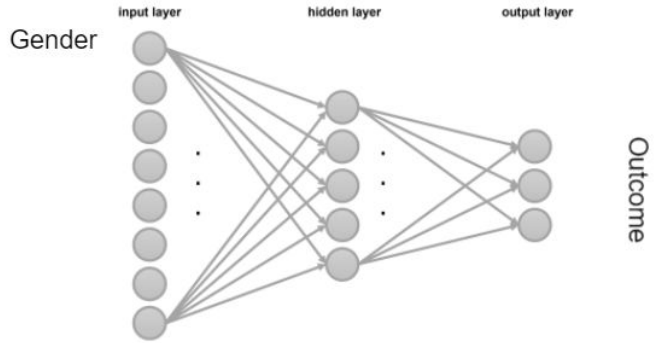
Linear regression
high interpretability
low accuracy



$$\text{Outcome}_i = b_0 + b_1 \text{Gender}_i + \epsilon_i$$

Easy to interpret gender effect on outcome

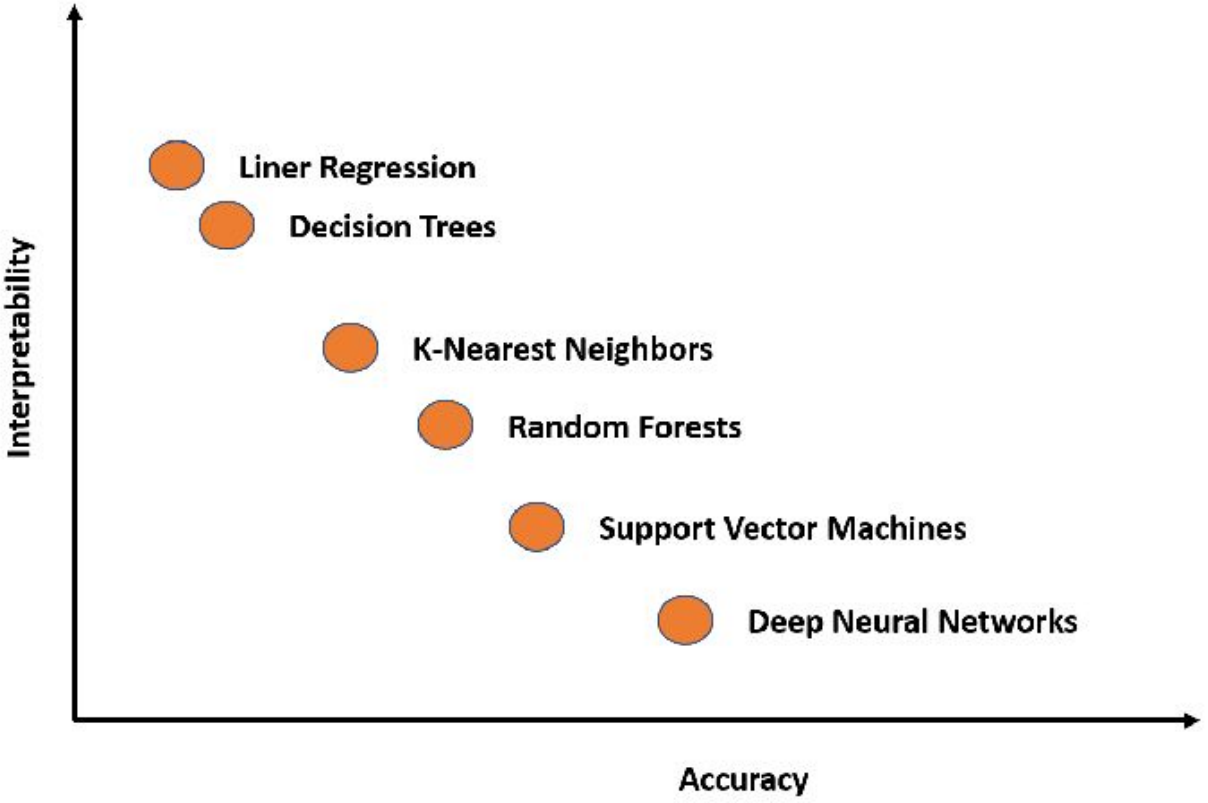
Deep learning
low interpretability
high accuracy



$$\text{Outcome}_i = f(b_{20} + b_{21} f(b_{10} + b_{11} f(b_0 + b_1 \text{Gender}_i))) + \epsilon_i$$

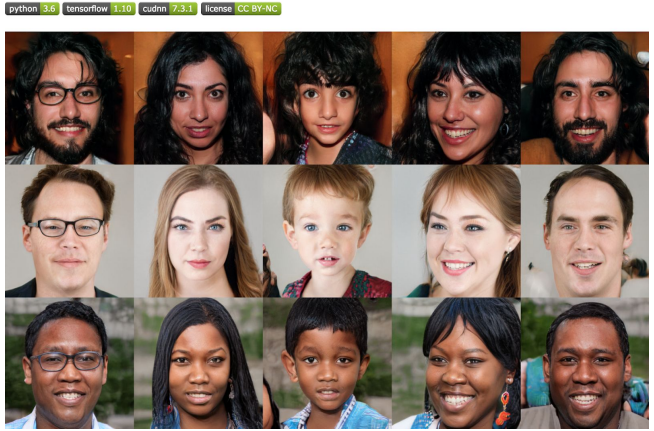
Hard to interpret gender effect on outcome

Accuracy vs. Interpretability

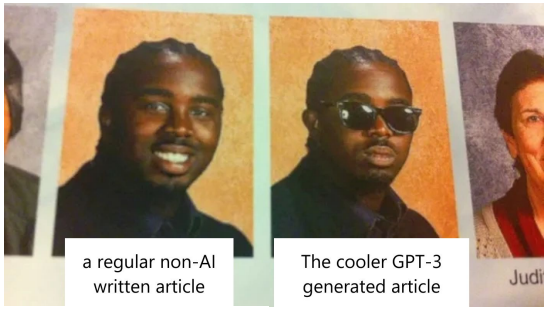


https://mc.ai/interpretability-vs-accuracy-the-friction-that-defines-deep-learning/

StyleGAN — Official TensorFlow Implementation



AlphaGo



THE VERGE

ENTERTAINMENT \ TECH \ CULTURE

TL;DR

Neural net-generated memes are one of the best uses of AI on the internet

I can't stop making memes

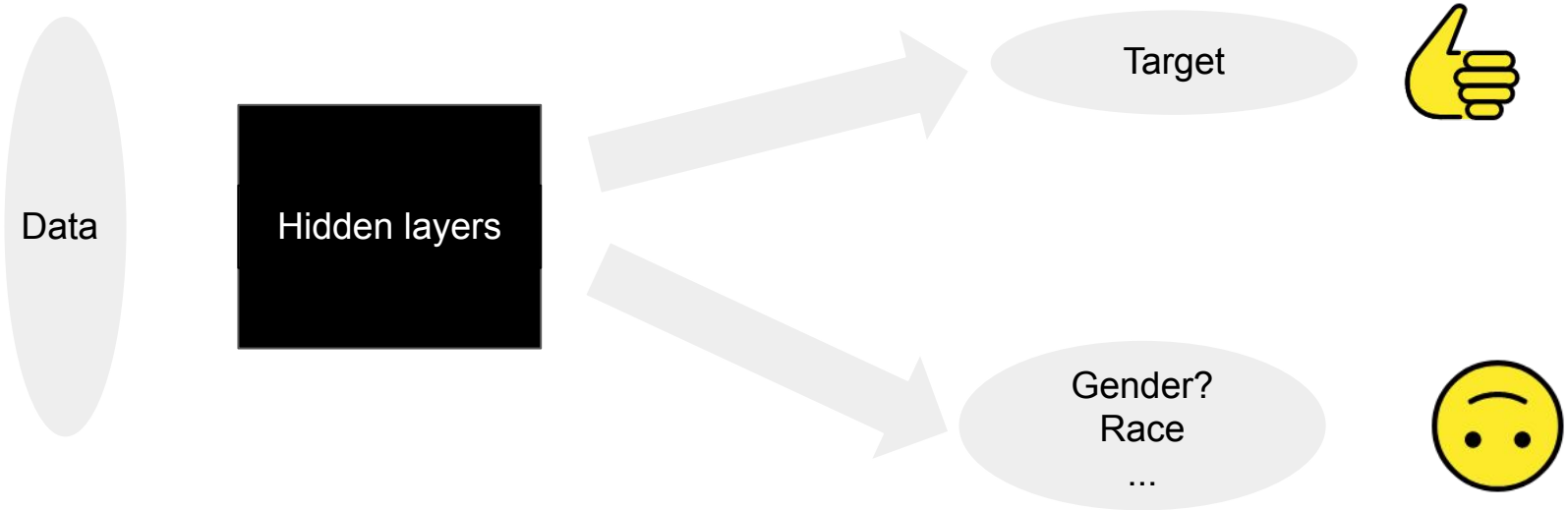
By Jay Peters | @jaypeters | Apr 29, 2020, 12:54pm EDT

Data Preparation

1. Understand your data: skews, correlations
2. Combine inputs from multiple sources
3. Removing the signal for problematic output: 1) stereotyping 2) Sexism, Racism, *-ism
4. Adding signal for desired variables 1) increasing model performance 2) Attention to subgroups or data slices with worse performance

Model Training

1. Always cross-validate!
2. Use held-out test set for hard use cases
3. multitask adversarial training



		Model Predictions		
		Positive	Negative	
Reference	Positive	True Positive	False Negative	Recall
	Negative	False Positive	True Negative	False Positive Rate
		Precision	Negative Predictive Value	

Criterion

Equality of opportunity:
Recall is equal across subgroups

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Predictive parity:
Precision is equal across subgroups

$$Presicion = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Tools

AIX360 (AI Explainability 360)* by *IBM

- Data explanation
- Local post-hoc explanation
- Local direct explanation
- Global direct explanation
- Global post-hoc explanation

https://github.com/Trusted-AI/AIX360

What-if Tool* by *Google

- Visualize dataset
- Visualize model inference
- Explore counterfactual examples
- Compare models prediction
- Visualize model performance

https://pair-code.github.io/what-if-tool/get-started/

Thank you!