

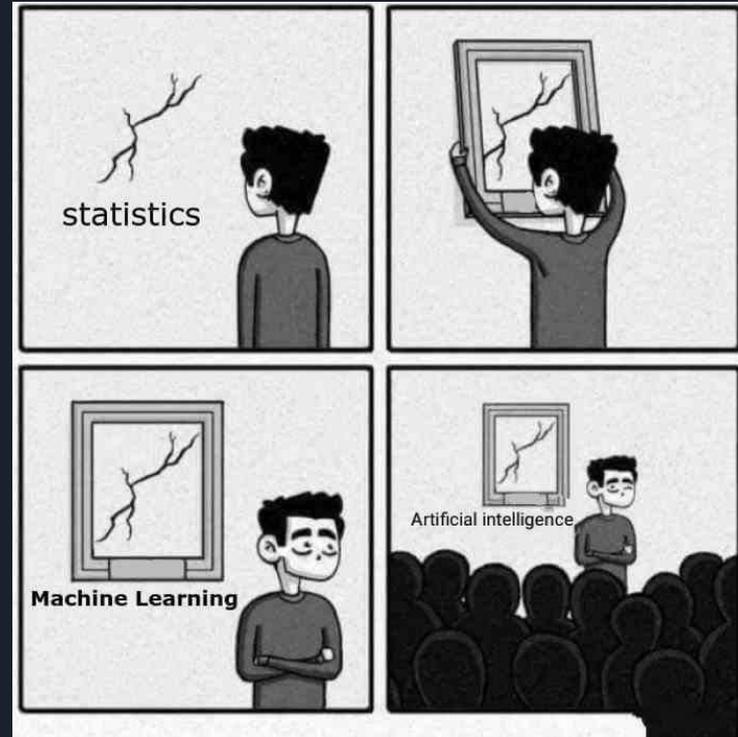


AI (and also other things)

Hien Pham '23 & Aditya Gollapudi '23

What is AI? (And some other things)*

- Artificial Intelligence
 - Mimicking human “behavior”
- Machine Learning
 - Improving at a task with more data
- Deep Learning
 - “Learning” like a human



*These are very loose definitions, everyone uses these terms a bit differently - and sometimes they are just random buzzwords

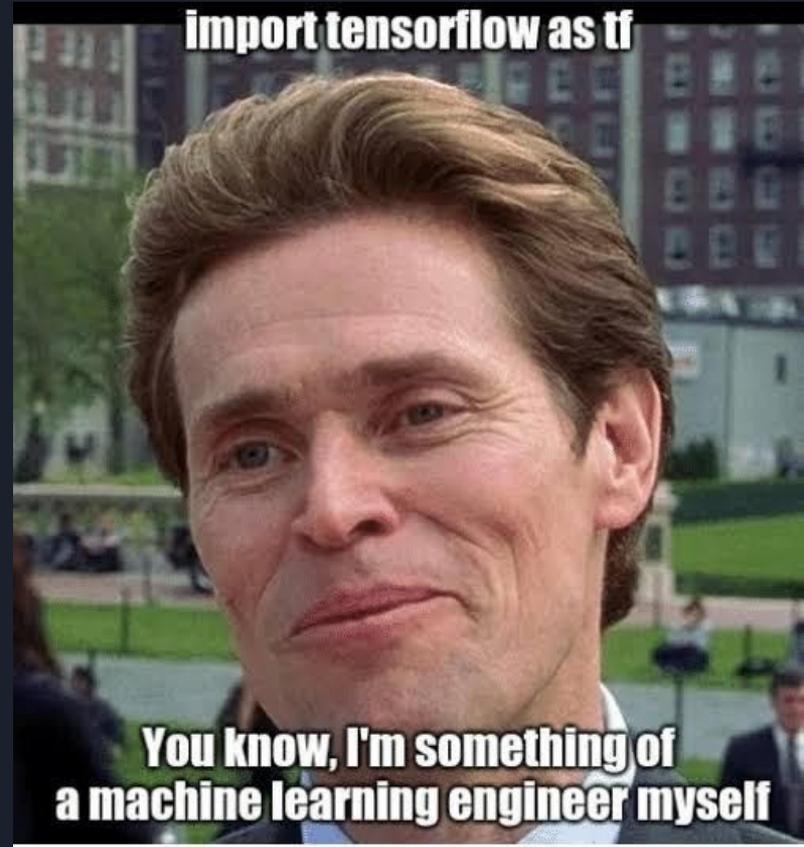


Vocabulary

- Utility Functions
 - Error/Loss
 - Reward/Punishment
- Types of Datasets
 - Training
 - Testing/Validation
- Failure Mode
- Types of Learning
 - Supervised
 - Unsupervised
- Classifier
- Recommender
- Narrow Intelligence
- General Intelligence
- Brittle Intelligence
- Intelligent Agent
- World Model
- NLP, CV

Who is involved in the production of AI?

- ML Researchers
 - Industry Labs (Deepmind, FAIR, Brain, Amazon Science, OpenAI, etc.)
 - Academia
 - Develop new techniques and models
- ML “Practitioners”
 - Apply techniques and models to specific problems
 - SWEs at many firms, data scientists etc.
- ML “Users”
 - Use models built by practitioners in applications
 - E.g. Meta(FB) engineers using the recommender algorithm to decide post order
- Data Collector/Labeller
 - Large apps (eg. FB or Google)
 - Government Agencies
 - MTurk workers
 - etc.



The Blackbox Problem

- It is very hard to determine “why” an AI made a decision
- Can making determining what environments a model will function well in hard
- Makes proving that any individual decision is biased hard



Alignment Problem

- Very hard to tell an AI exactly what you want to do
- Misalignment in narrow AIs
 - Recommender Algorithms
 - Youtube optimizing for clicks, views, watchtime, etc.
 - Want to optimize for total profit
 - Bail algorithms optimizing for matching with judges sentencing
 - Want to optimize for actual flight risk
- Misalignment in general AIs
 - “Paperclip Maximizer”
 - AI doesn’t want to be turned off
 - AI can out think and out perform you
 - AI doesn’t want to have its goal changed



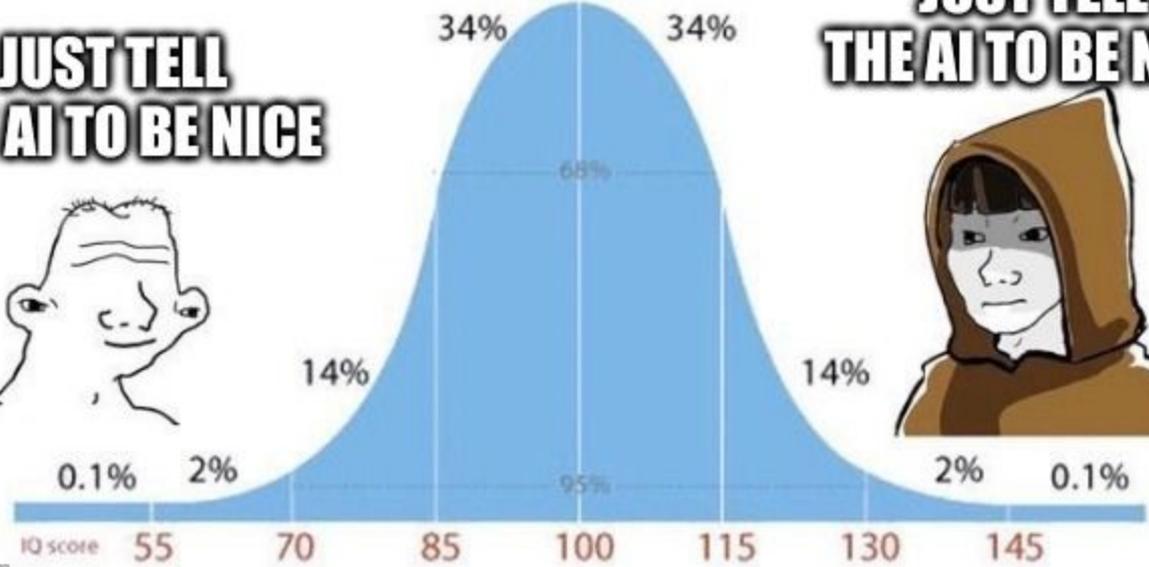
**NOOO, ALIGNMENT REQUIRES
ADVANCED UNDERSTANDING OF AGENT
FOUNDATIONS AND SOLUTIONS TO
OVEROPTIMIZATION OF VALUE PROXIES AND...**



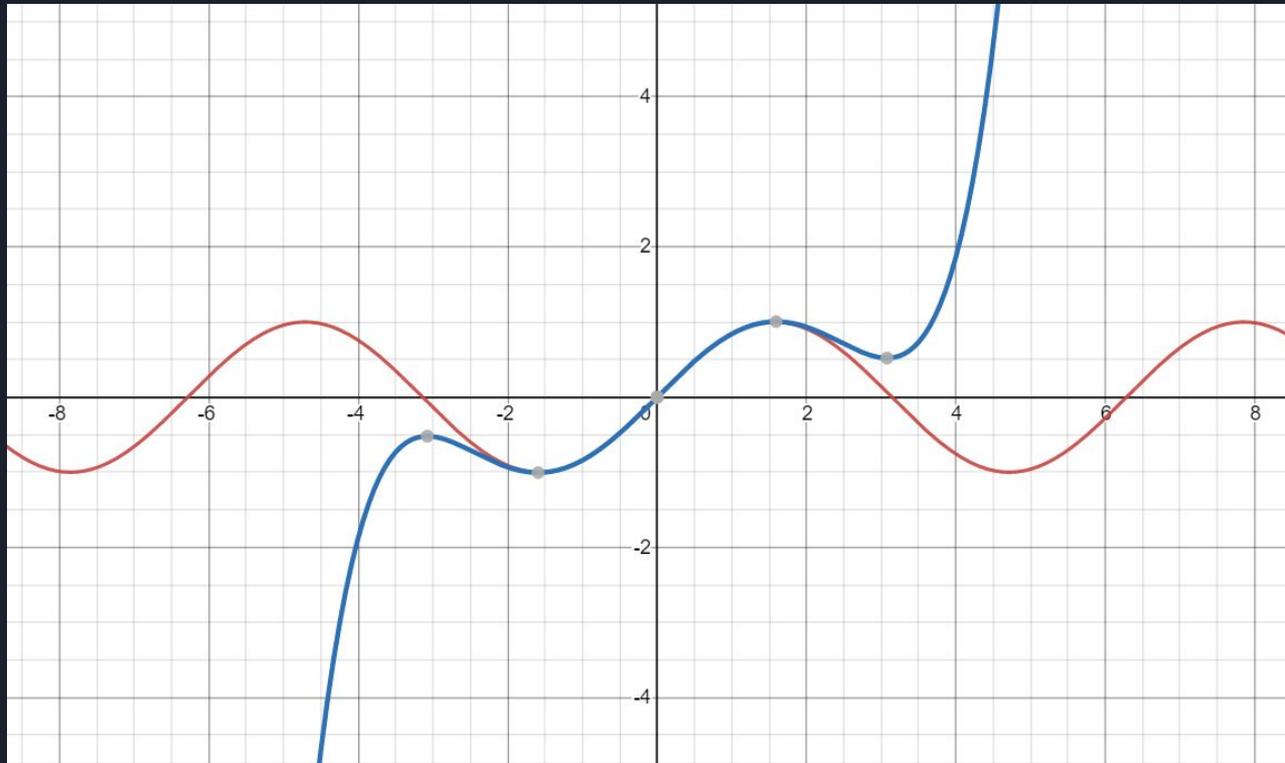
**JUST TELL
THE AI TO BE NICE**



**JUST TELL
THE AI TO BE NICE**

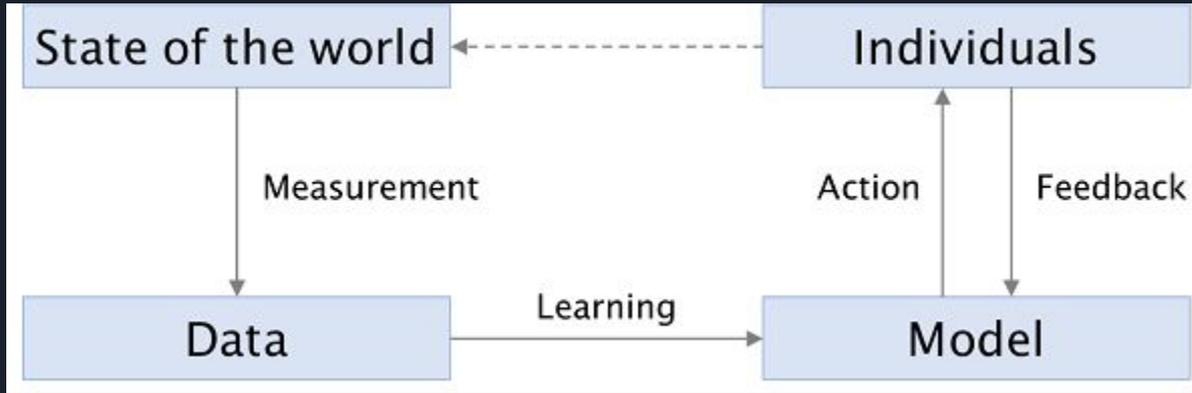


Brittleness



ML Loop

aka. every step where things can go wrong



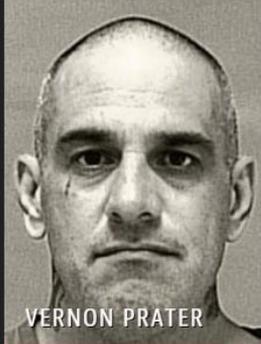
Latent Dimensions aka. Redundant Encoding

- COMPAS recidivism algorithm by Northpointe “includes factors such as education levels, and whether a defendant has a job”
- “Northpointe’s core product is a set of scores derived from 137 questions that are either answered by defendants or pulled from criminal records. Race is not one of the questions.”

Black defendants were twice as likely as white defendants to be misclassified as a higher risk of violent recidivism, and white recidivists were misclassified as low risk 63.2 percent more often than black defendants.

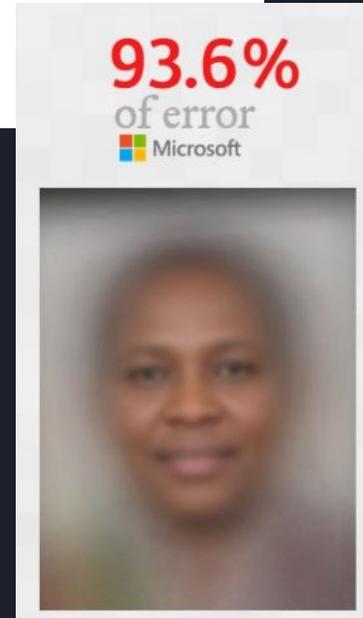
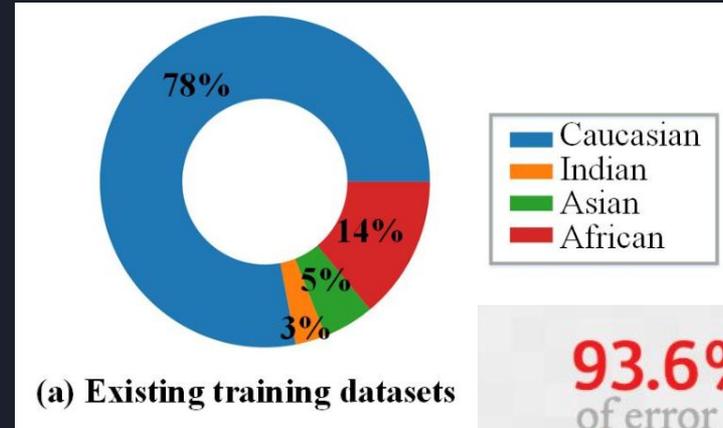
→ Race implicitly included in the algorithm,
racism amplified by ML algorithm

Two Petty Theft Arrests

 VERNON PRATER	 BRISHA BORDEN
LOW RISK 3	HIGH RISK 8
VERNON PRATER Prior Offenses 2 armed robberies, 1 attempted armed robbery Subsequent Offenses 1 grand theft	BRISHA BORDEN Prior Offenses 4 juvenile misdemeanors Subsequent Offenses None
LOW RISK 3	HIGH RISK 8

Population Sampling

- Data → training → prediction
- Less data → worse prediction
- Example: MS-Celeb-1M - 10 million images of 100k celebrities → so white they had to remove the dataset from the internet but people still use it for research + commercial purposes
- Gender Shades study: Among mislabelled people by Microsoft's commercial facial recognition algorithms, 93.6% were of black women



Miscellaneous things

- **Homogeneity of Systems:** Same dataset & 160 systems → same predictions. What if all companies use the same predictive system (i.e. for hiring?)
- **Conflicts of Interest:**

A binary classifier from different perspectives

Decision-maker: of those I've labeled high-risk, how many will recidivate?

Predictive value

Defendant: what's the probability I'll be incorrectly classified high-risk?

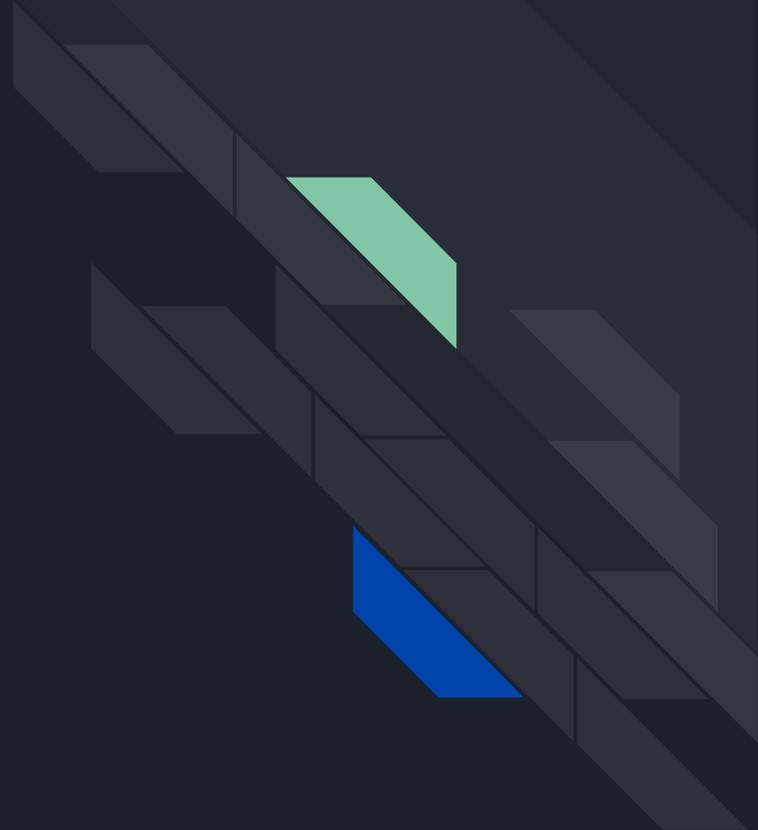
False positive rate

Society [think hiring rather than criminal justice]: is the selected set demographically balanced?

Demography

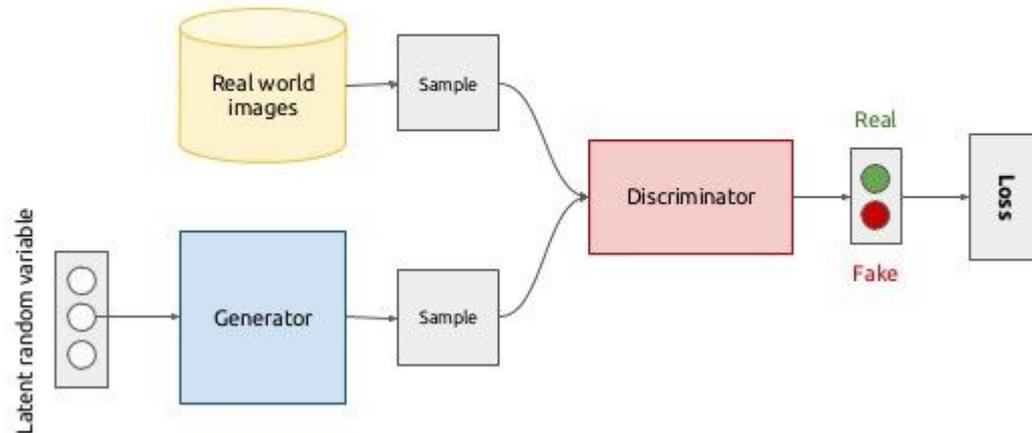
Did not recidivate	TN	<u>FP</u>
Recidivated	<u>FN</u>	TP
	Labeled low-risk	Labeled high-risk

State of the Art Techniques



Generative Adversarial Networks (GANs)

Generative adversarial networks (conceptual)





Notable Examples

- AlphaGo
- AlphaFold
- StyleGAN
- Deepfakes
- ArtGAN

Strengths

- Can significantly exceed human performance
 - E.g. AlphaGo
- Makes it really hard to automatically detect fakes of anything
- Continuously improve in an online setting
- Transfer Learning

Mode Failure



Adversarial Noise



“panda”

57.7% confidence

+ .007 ×



noise

=

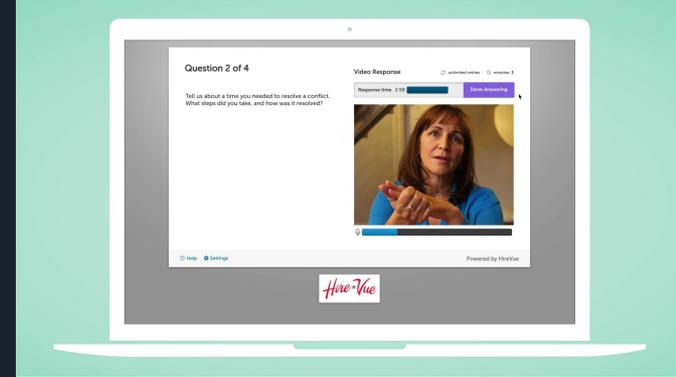


“gibbon”

99.3% confidence

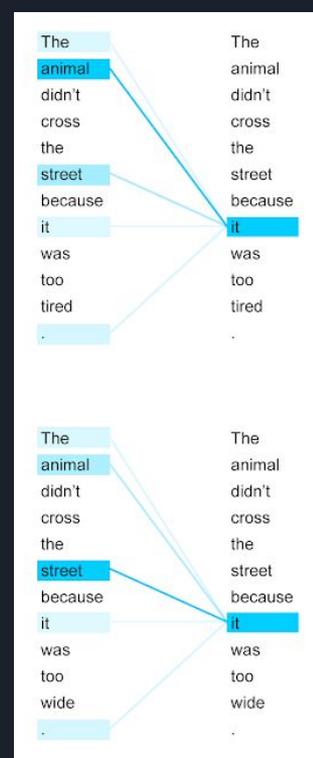
Reinforcement Learning

- Learn without data
- Punished & rewarded based on things that happen in the environment
- Have to balance “exploiting” where you operate in areas where you already know you get reward and “exploring” where you search out for new high reward areas
 - Too little exploration is dangerous e.g. if a hiring algo never hire minorities it will never learn there is reward there
 - Too much exploration is also dangerous e.g. a self driving car probably shouldn't be too experimental



Transformers/Attention

- Given a set of data you choose what you think is an “important subset”
- Either repeat this or run something else on the reduced data
- Compared to other techniques very legible to humans (you can tell why it made its decisions)
- Super powerful for looking at text



Coreference Resolution





Sources

- <https://fairmlbook.org/> and Arvind Narayanan, COS 534 Fairness in Machine Learning
- <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- <http://gendershades.org/>
- <http://www.whdeng.cn/RFW/Trainingdataste.html>
-