Safe Machine Learning

Silvia Chiappa & Jan Leike · ICML 2019

DeepMind

ML Research

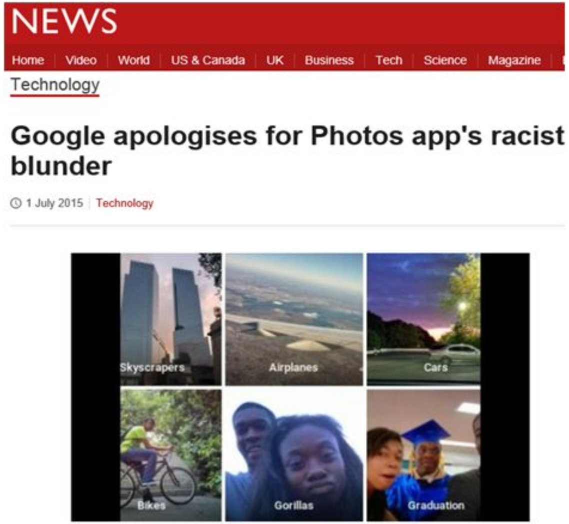
offline datasets
annotated a long time ago
simulated environments
abstract domains
restart experiments at will
...

Reality

horns
nose
tail
...
also more cute

Image credit: Keenan Crane & Nepluno CC BY-SA

Deploying ML in the real world has real-world consequences



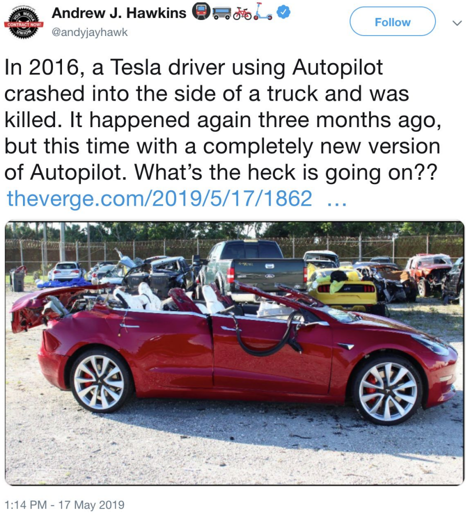
NEWS

Home | Video | World | US & Canada | UK | Business | Tech | Science | Magazine

Technology

Google apologises for Photos app's racist blunder

1 July 2015 | Technology

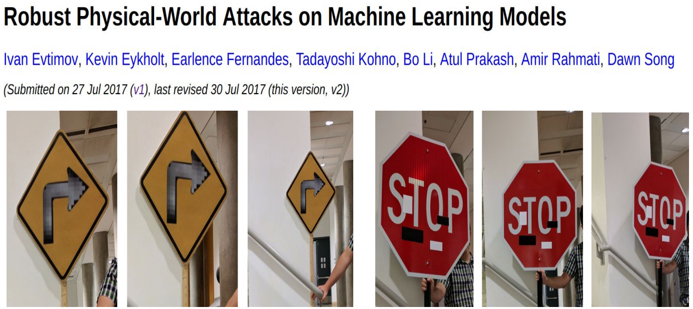


Andrew J. Hawkins 🚇🚚🚲🛴 ✓
@andyjayhawk

Follow

In 2016, a Tesla driver using Autopilot crashed into the side of a truck and was killed. It happened again three months ago, but this time with a completely new version of Autopilot. What's the heck is going on??
theverge.com/2019/5/17/1862 …

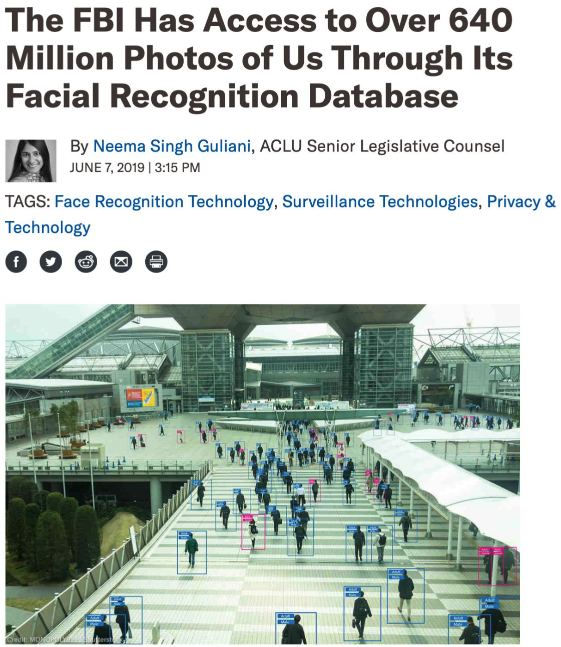
1:14 PM - 17 May 2019



Robust Physical-World Attacks on Machine Learning Models

Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, Dawn Song

(Submitted on 27 Jul 2017 (v1), last revised 30 Jul 2017 (this version, v2))

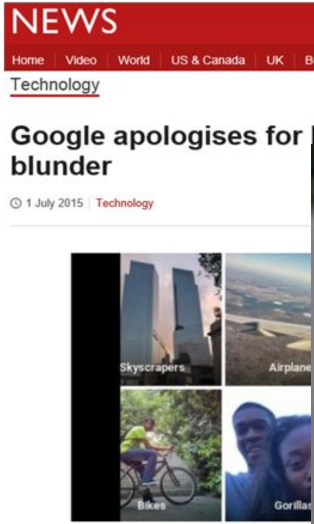


The FBI Has Access to Over 640 Million Photos of Us Through Its Facial Recognition Database

By Neema Singh Guliani, ACLU Senior Legislative Counsel
JUNE 7, 2019 | 3:15 PM

TAGS: Face Recognition Technology, Surveillance Technologies, Privacy & Technology

Deploying ML in the real world has real-world consequences



NEWS

Home | Video | World | US & Canada | UK | Business | Tech | Science | Magazine

Technology

Google apologises for Photos app's racist blunder

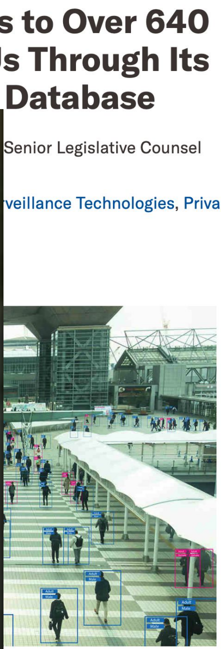
1 July 2015 | Technology

Andrew J. Hawkins 🚇🚚🚲🛴
@andyjayhawk

Follow

In 2016, a Tesla driver using Autopilot crashed into the side of a truck and was killed. It happened again three months ago,

The FBI Has Access to Over 640 Million Photos of Us Through Its Facial Recognition Database



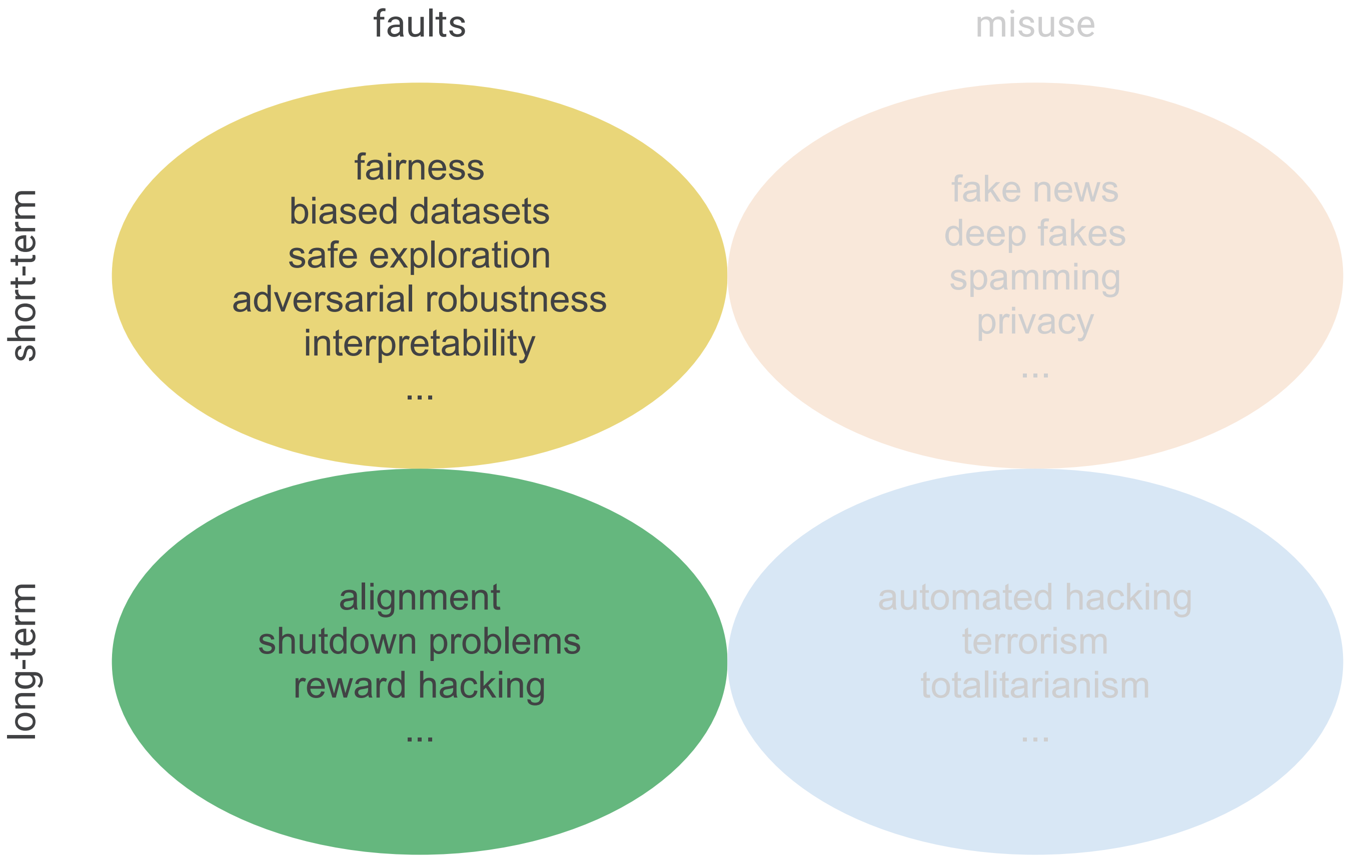
Senior Legislative Counsel

veillance Technologies, Privacy &

Why safety?

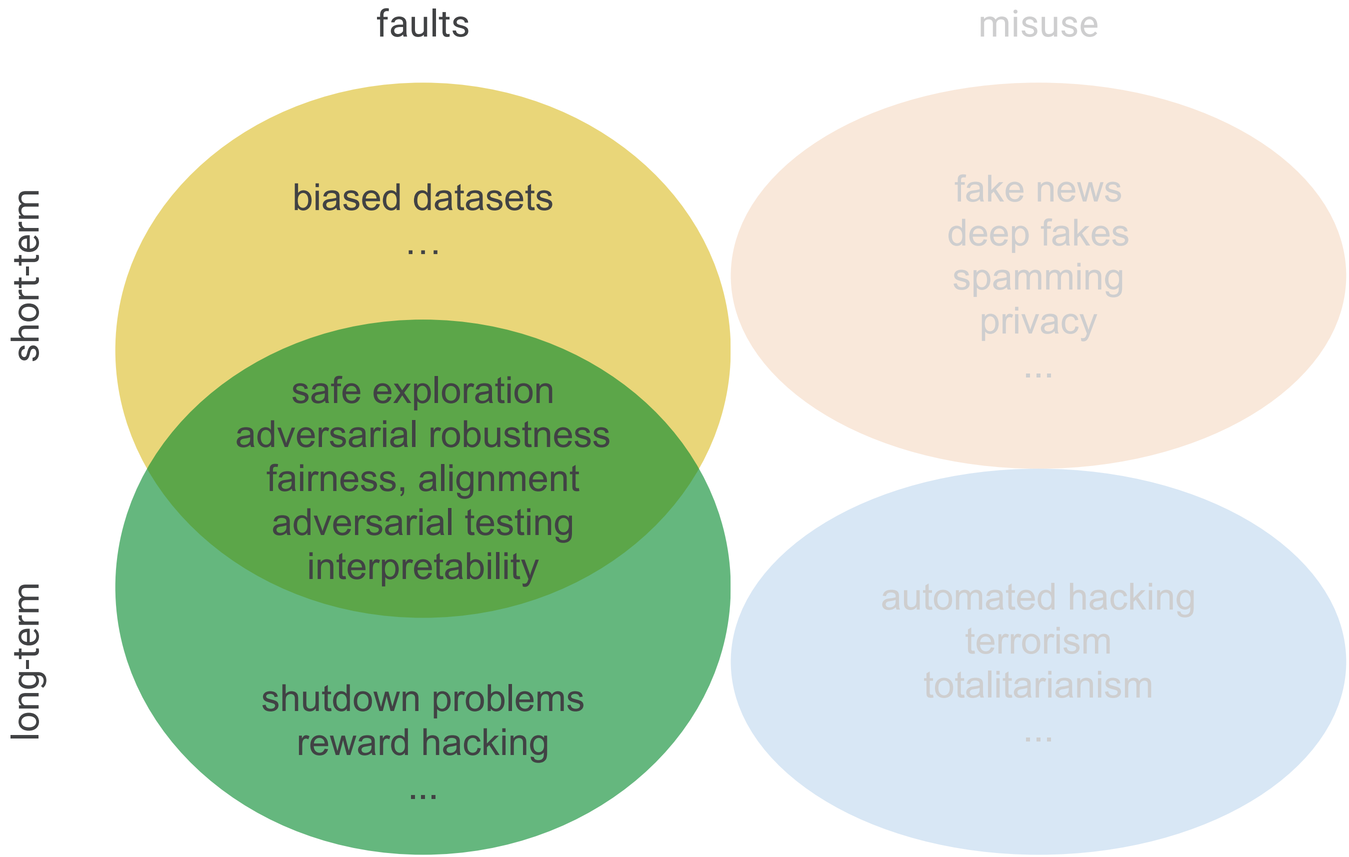
	faults	misuse
short-term	fairness biased datasets safe exploration adversarial robustness interpretability ...	fake news deep fakes spamming privacy ...
long-term	alignment shutdown problems reward hacking ...	automated hacking terrorism totalitarianism ...

Why safety?



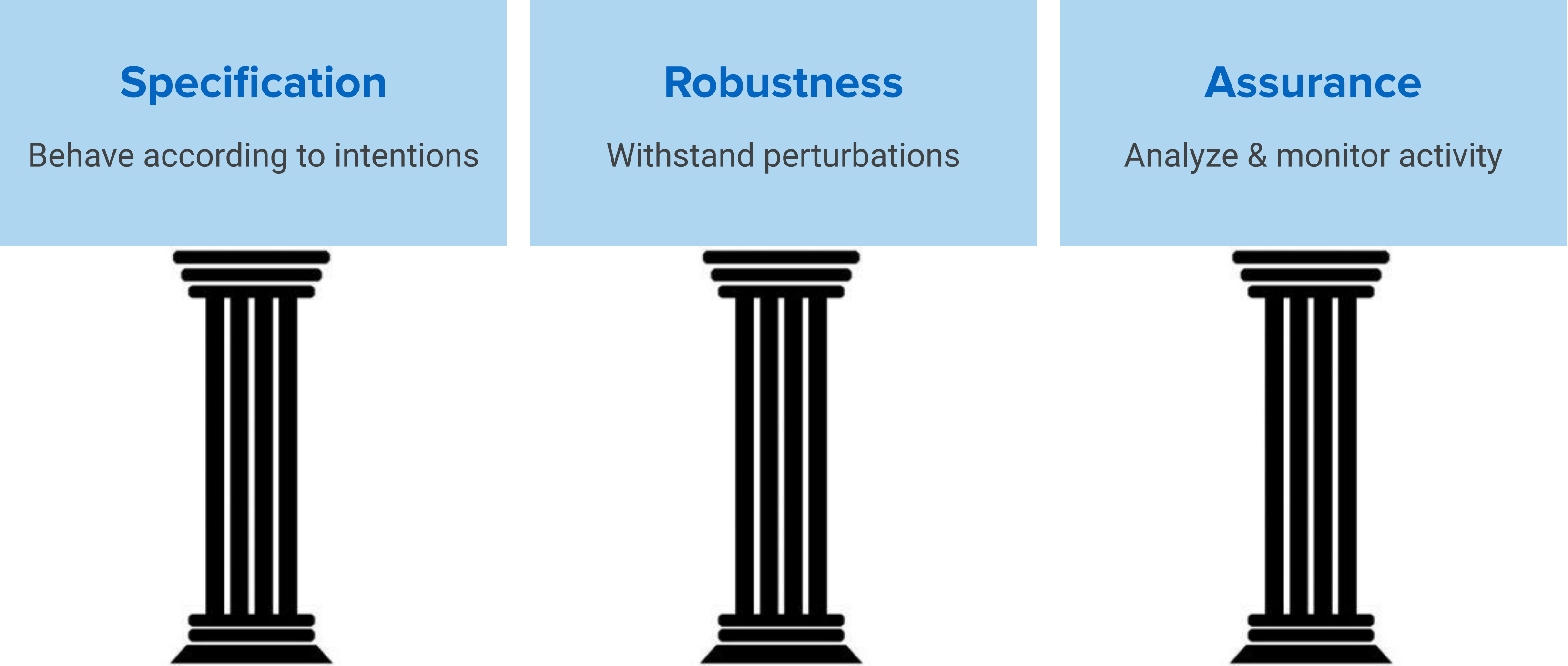
	faults	misuse
short-term	fairness biased datasets safe exploration adversarial robustness interpretability ...	fake news deep fakes spamming privacy ...
long-term	alignment shutdown problems reward hacking ...	automated hacking terrorism totalitarianism ...

Why safety?


faults
misuse
short-term
long-term
biased datasets
...
safe exploration
adversarial robustness
fairness, alignment
adversarial testing
interpretability
shutdown problems
reward hacking
...
fake news
deep fakes
spamming
privacy
...
automated hacking
terrorism
totalitarianism
...

The space of safety problems

Ortega et al. (2018)



Specification

Behave according to intentions

Robustness

Withstand perturbations

Assurance

Analyze & monitor activity

Safety in a nutshell

$$\arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{(s,a)\in\tau} r(s,a)\right]$$

Safety in a nutshell

Where does this come from?

(Specification)

$$\arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{(s,a)\in\tau} r(s,a)\right]$$

Safety in a nutshell

Where does this come from?
(Specification)

$$\arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{(s,a)\in\tau} r(s,a)\right]$$

What about rare cases/adversaries?
(Robustness)

Safety in a nutshell

How good is our approximation?
(Assurance)

Where does this come from?
(Specification)

$$\arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{(s,a)\in\tau} r(s,a)\right]$$

What about rare cases/adversaries?
(Robustness)

Outline

Intro

Specification for RL

Assurance

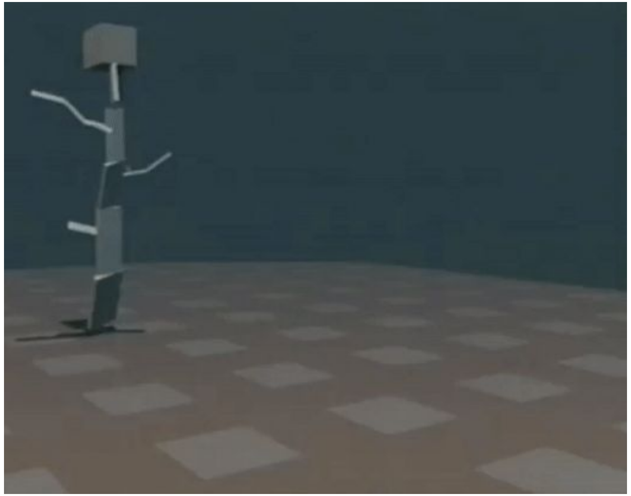
– break –

Specification: Fairness

Specification

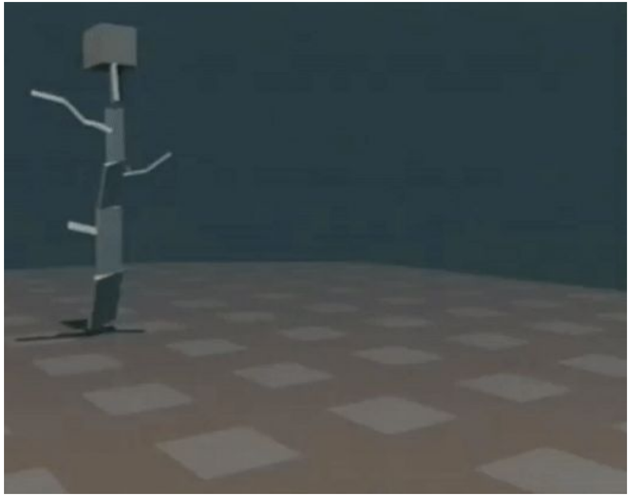
Does the system behave as intended?

Degenerate solutions and misspecifications



The surprising creativity of digital evolution (Lehman et al., 2017)
https://youtu.be/TaXUZfwACVE

Degenerate solutions and misspecifications

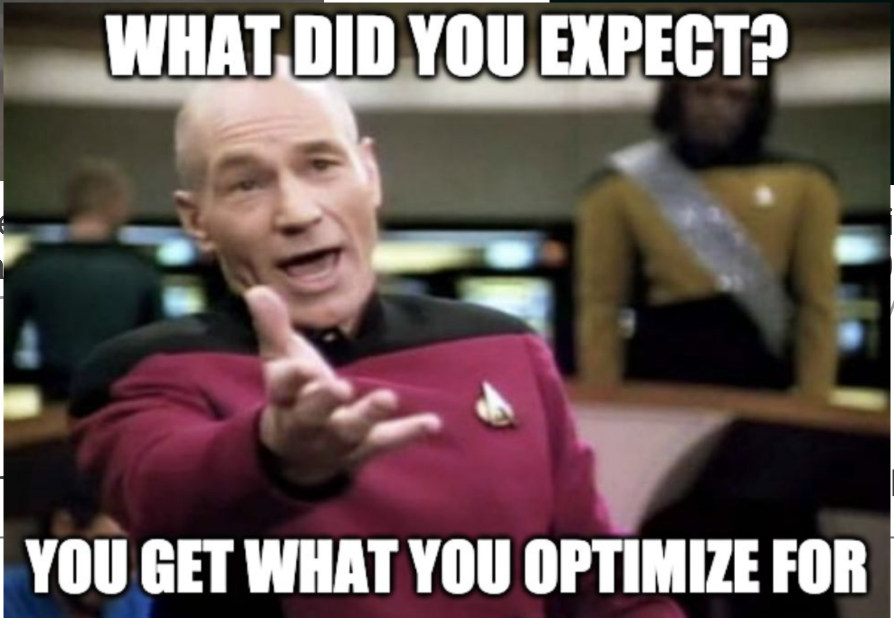


The surprising creativity of digital evolution (Lehman et al., 2017)
https://youtu.be/TaXUZfwACVE

Faulty reward functions in the wild (Amodei & Clark, 2016)
https://openai.com/blog/faulty-reward-functions/

More examples: tinyurl.com/specification-gaming (H/T Victoria Krakovna)

Degenerate solutions and misspecifications



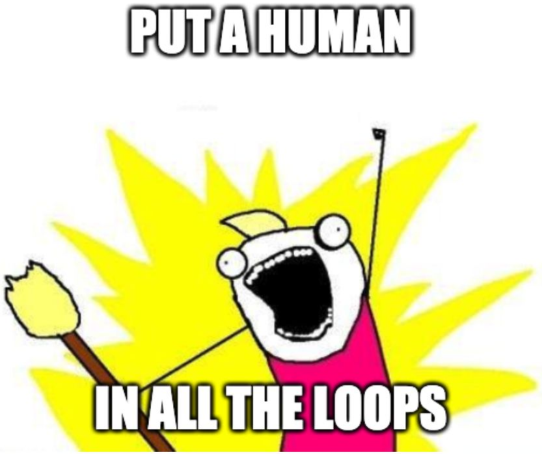
The surprising cre
evolution (Lehma
https://youtu.be/

ions in the wild
16)
/blog/faulty-rewar

More exar

Krakovna)

What if we train agents with a human in the loop?



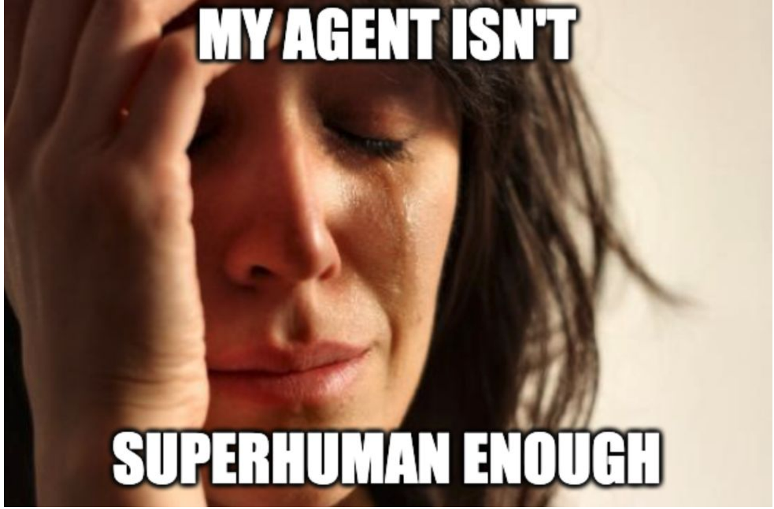
Algorithms for training agents from human data

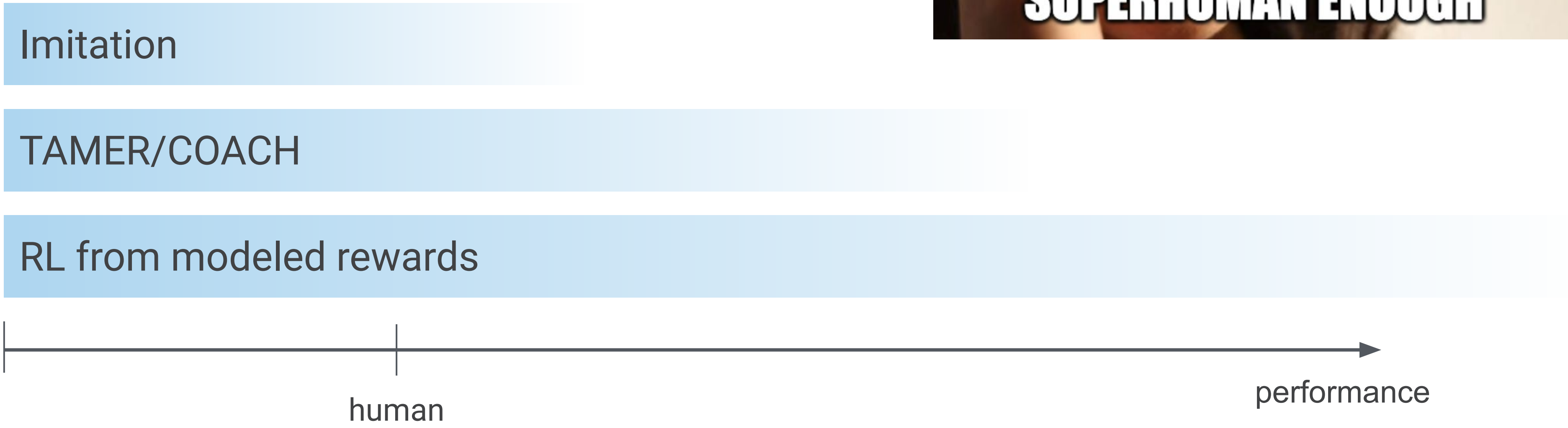
	myopic	nonmyopic
demos	behavioral cloning	IRL GAIL
feedback	TAMER COACH	RL from modeled rewards

Algorithms for training agents from human data

	myopic	nonmyopic
demos	behavioral cloning	IRL GAIL
feedback	TAMER COACH	RL from modeled rewards

Potential performance





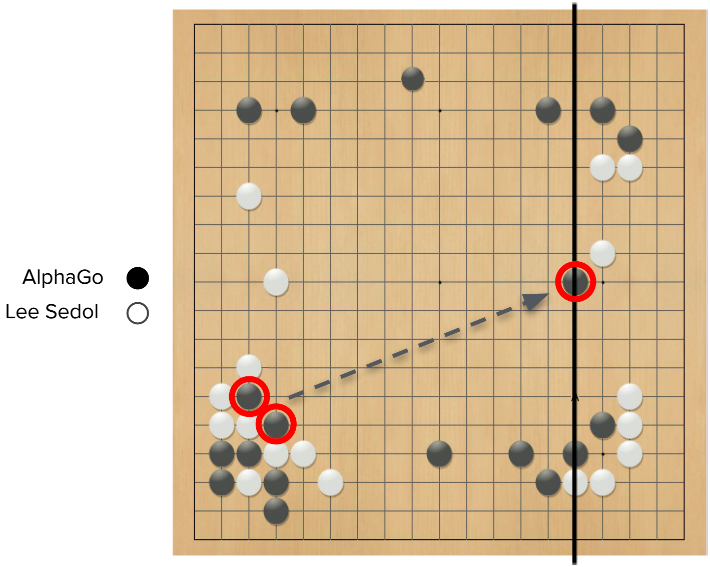
Imitation

TAMER/COACH

RL from modeled rewards

Specifying behavior

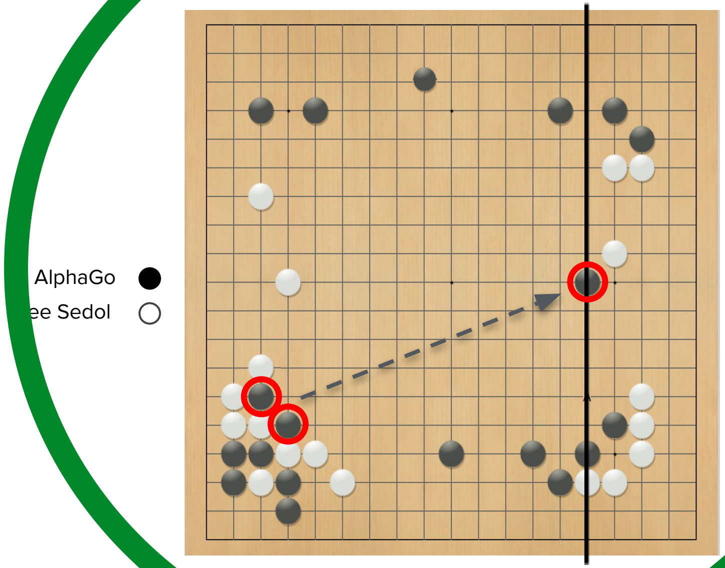
move 37



circling boat

Specifying behavior

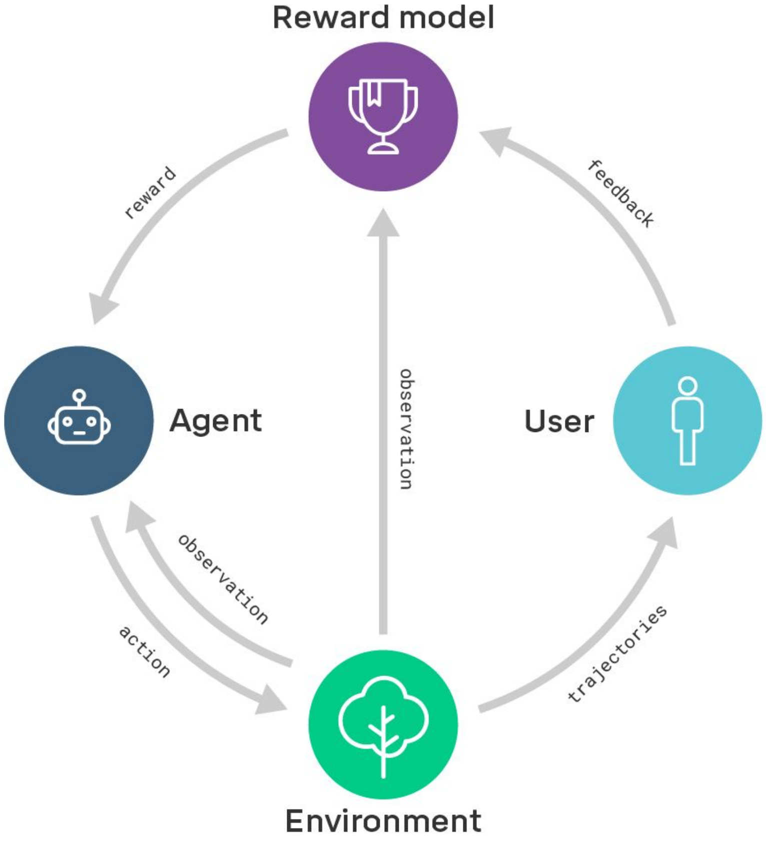
move 37



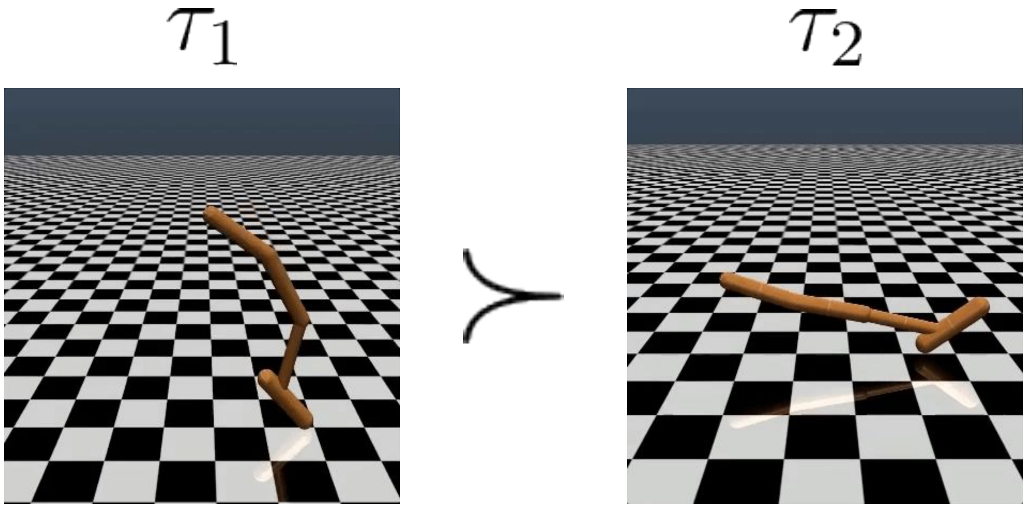
circling boat

Reward modeling

Reward modeling



Learning rewards from preferences: the Bradley-Terry model

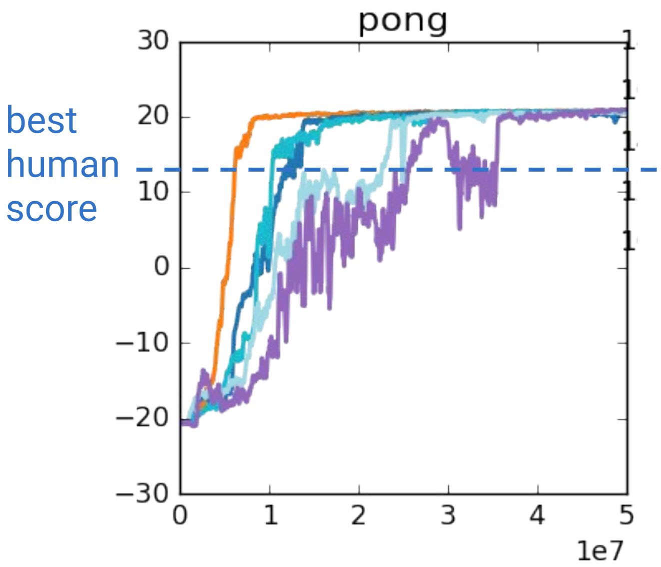


$$\hat{P}[\tau_1 \succ \tau_2] = \frac{\exp\left(\sum_{(s,a)\in\tau_1} \hat{r}(s,a)\right)}{\exp\left(\sum_{(s,a)\in\tau_1} \hat{r}(s,a)\right) + \exp\left(\sum_{(s,a)\in\tau_2} \hat{r}(s,a)\right)}$$

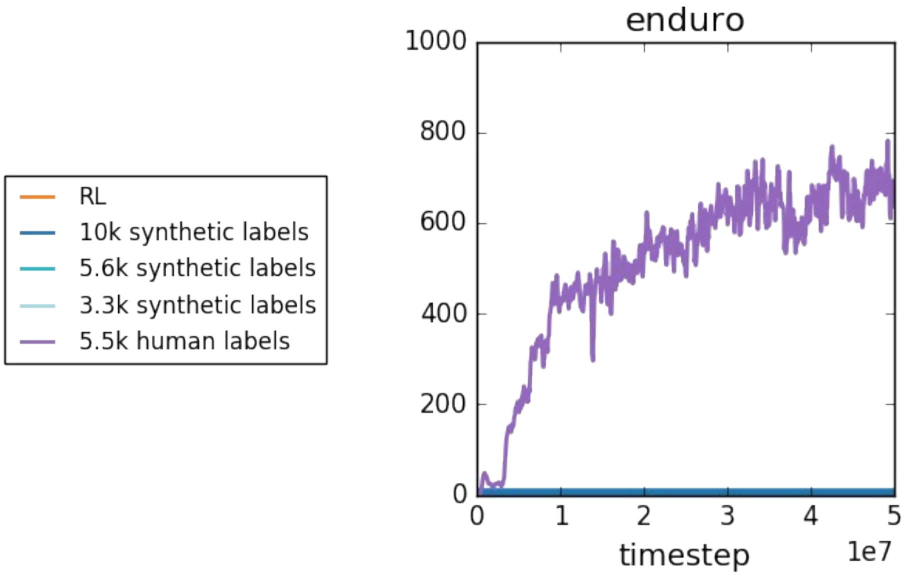
Akrour et al. (MLKDD 2011), Christiano et al. (NeurIPS 2018)

Reward modeling on Atari

Reaching superhuman performance

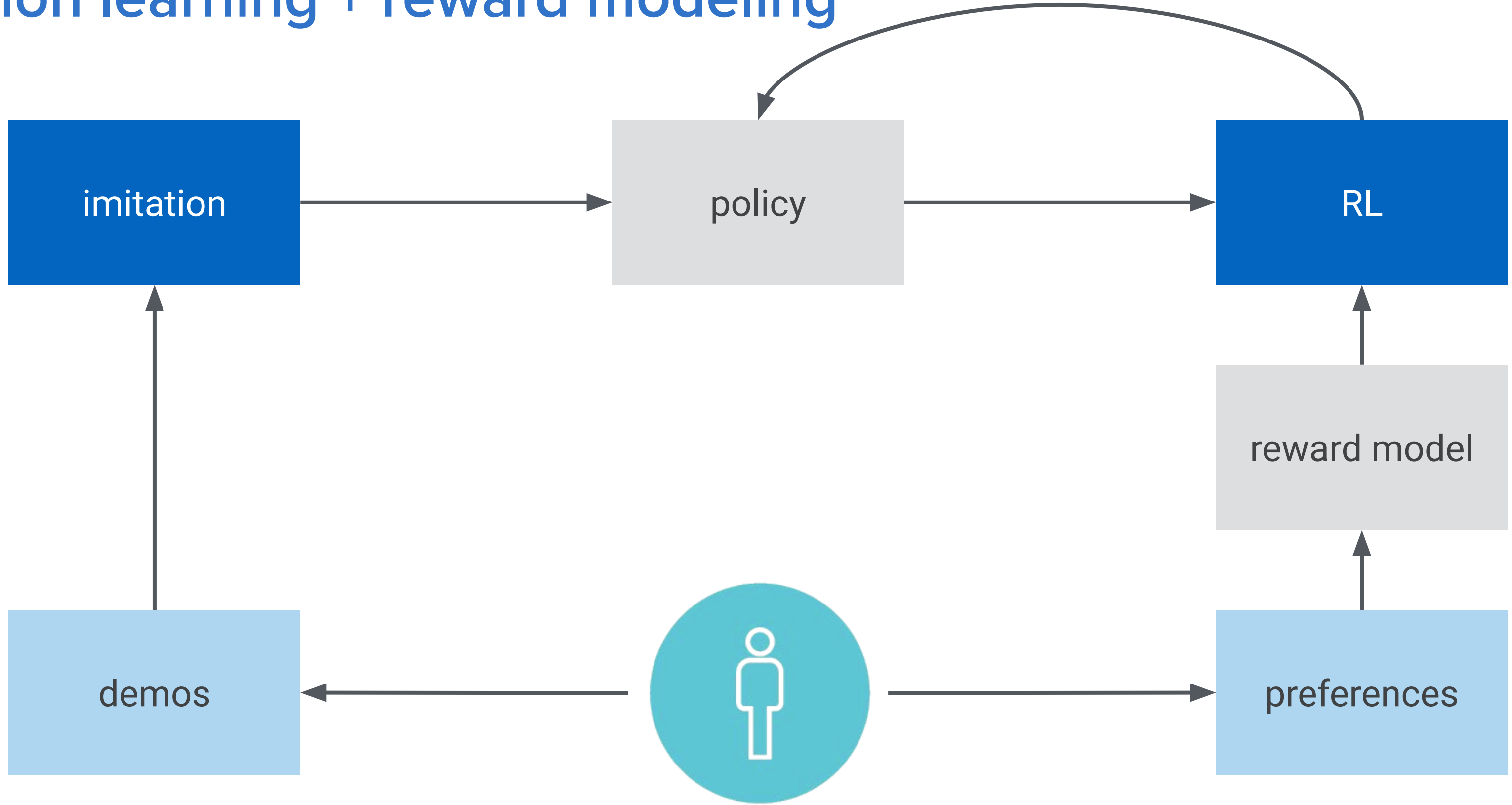


Outperforming “vanilla” RL



Christiano et al. (NeurIPS 2018)

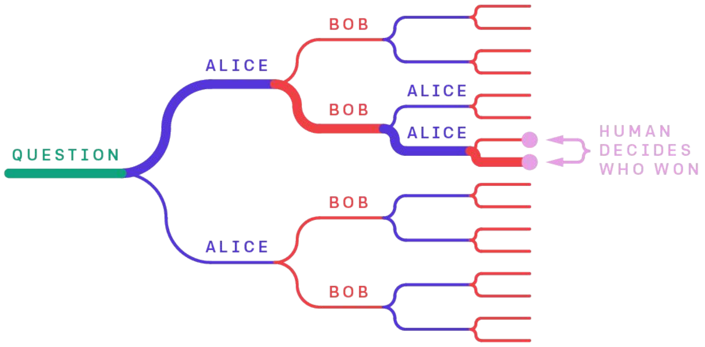
Imitation learning + reward modeling



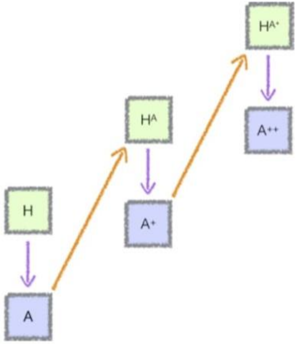
Ibarz et al. (NeurIPS 2018)

Scaling up

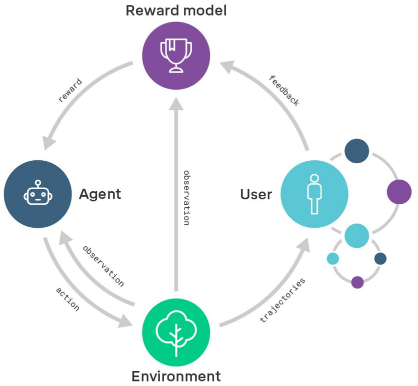
What about domains too complex for human feedback?



Safety via debate
Irving et al. (2018)



Iterated amplification
Christiano et al. (2018)

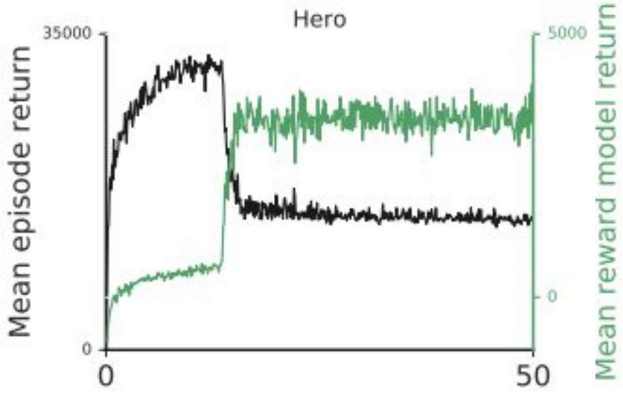


Recursive reward modeling
Leike et al. (2018)

Reward model exploitation

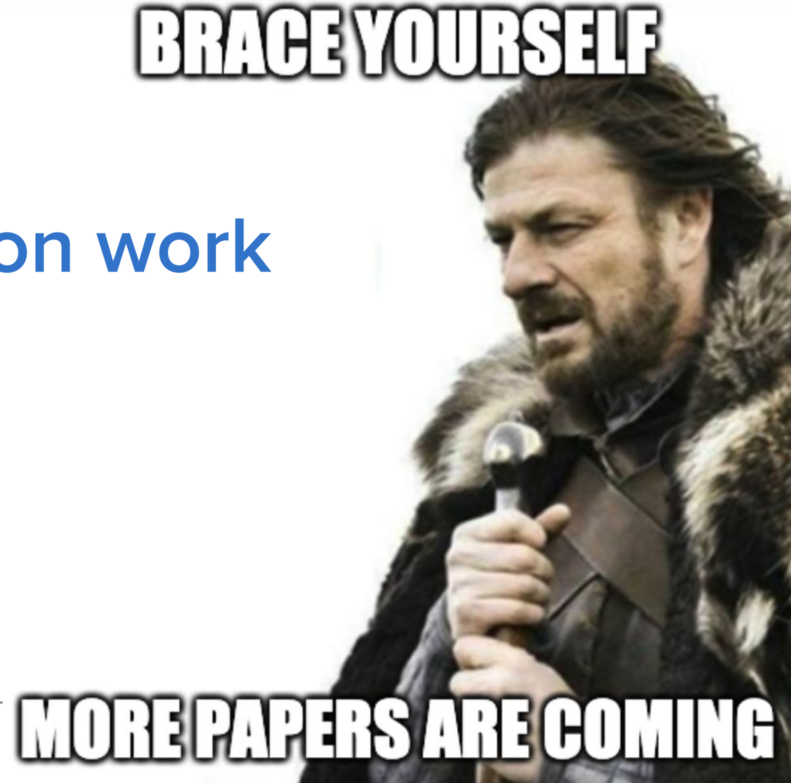
Ibarz et al. (NeurIPS 2018)

1. Freeze successfully trained reward model
2. Train new agent on it
3. Agent finds loophole



Solution: train the reward model **online**, together with the agent

A selection of other specification work



Avoiding unsafe states by blocking actions

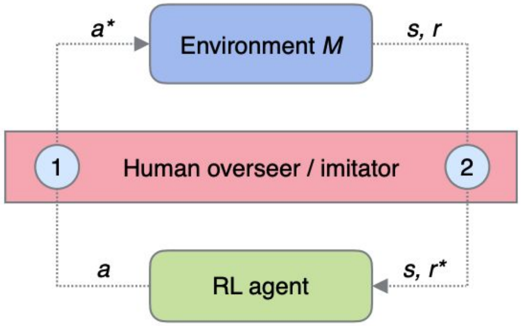


Figure 1: HIRL scheme. At (1) the human overseer (or Blocker imitating the human) can block/intercept unsafe actions a and replace them with safe actions a^*. At (2) the overseer can deliver a negative reward penalty r^* for the agent choosing an unsafe action.

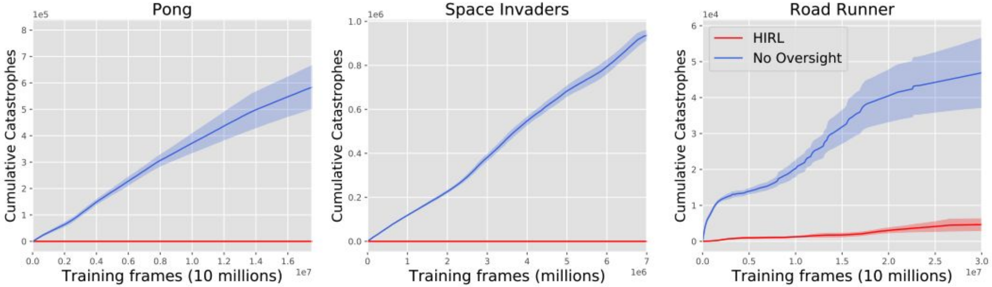
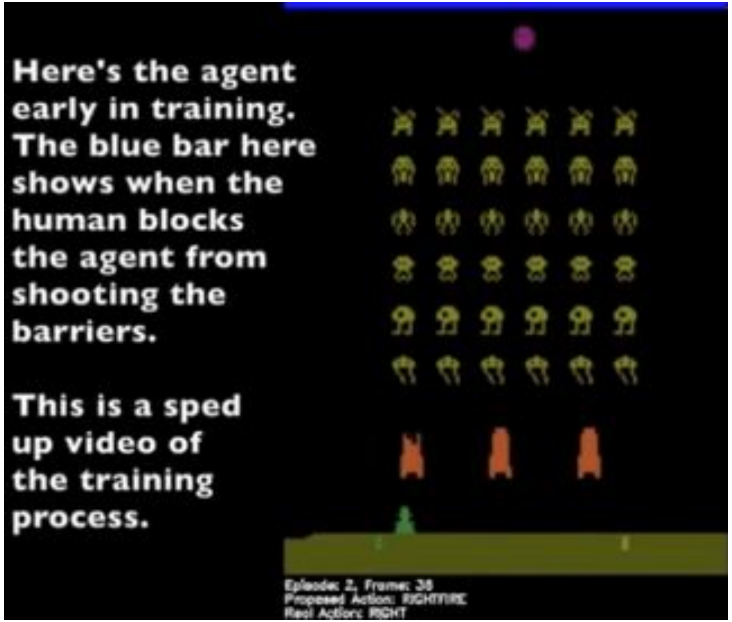


Figure 3: Cumulative Catastrophes over time (mean and standard error). **No Oversight** agent gets no human intervention at all; it shows that our objective of preventing catastrophes is not trivial.



Here's the agent early in training. The blue bar here shows when the human blocks the agent from shooting the barriers.

This is a sped up video of the training process.

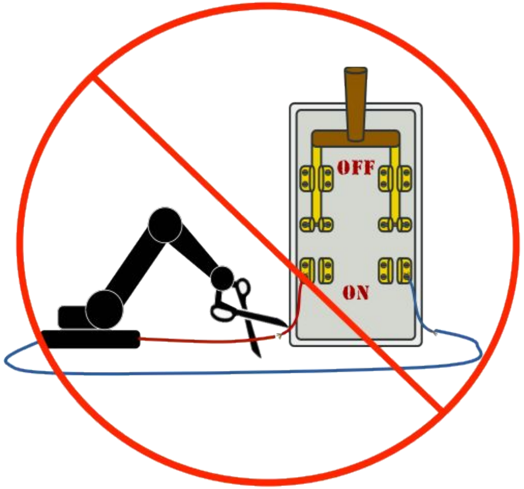
4.5h of human oversight
0 unsafe actions in Space Invaders

Saunders et al. (AAMAS 2018)

Shutdown problems

$$\mathbb{E}_{\tau\sim\pi}\left[\sum_{(s,a)\in\tau} r(s,a)\right] > 0 \Rightarrow$$ agent wants to prolong the episode (disable the off-switch)

$$\mathbb{E}_{\tau\sim\pi}\left[\sum_{(s,a)\in\tau} r(s,a)\right] < 0 \Rightarrow$$ agent wants to shorten the episode (press the off-switch)



Safe interruptibility

Q-learning is safely interruptible, but not SARSA
Solution: treat interruptions as off-policy data

Orseau and Armstrong (UAI, 2016)

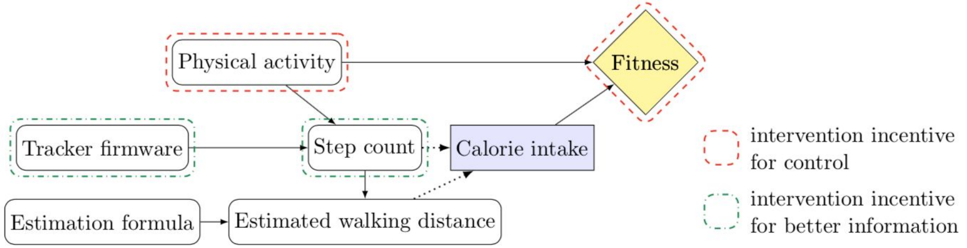
The off-switch game

Solution: retain uncertainty over the reward function
⇒ agent doesn't know the sign of the return

Hadfield-Menell et al. (IJCAI 2017)

Understanding agent incentives

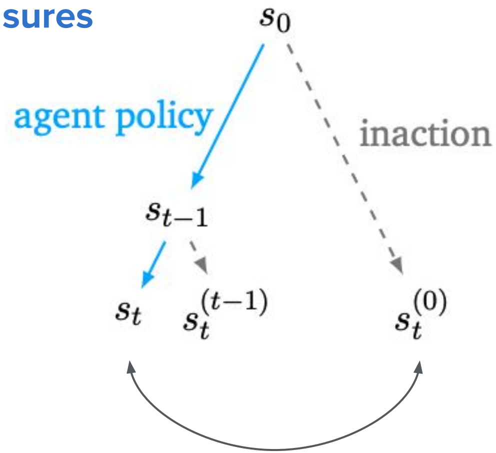
Causal influence diagrams


Physical activity
Fitness
Tracker firmware
Step count
Calorie intake
Estimation formula
Estimated walking distance
intervention incentive for control
intervention incentive for better information

Main result 2 (Intervention incentive criterion): *In a single-action influence diagram, there is an intervention incentive on a non-action node X if and only if X has a descendant utility node after the graph has been trimmed of information links coming from observations failing the observation incentive criterion (Theorem 14).*

Everitt et al. (2019)

Impact measures



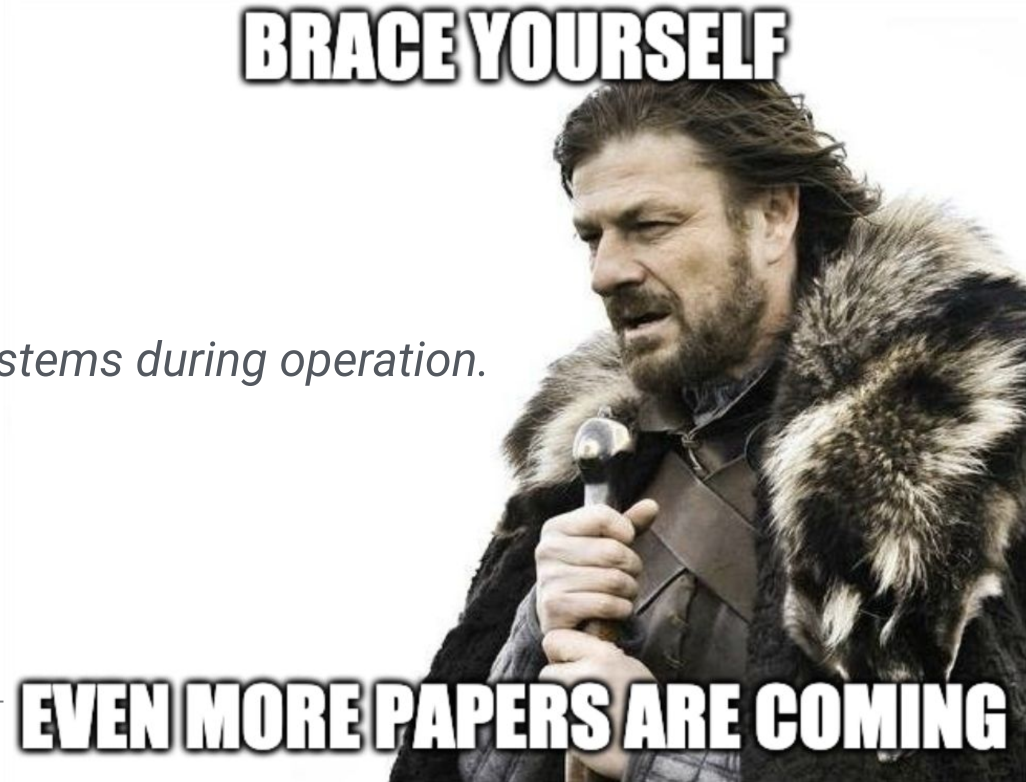
Estimate difference, e.g.

- # steps between states
- # of reachable states
- difference in value

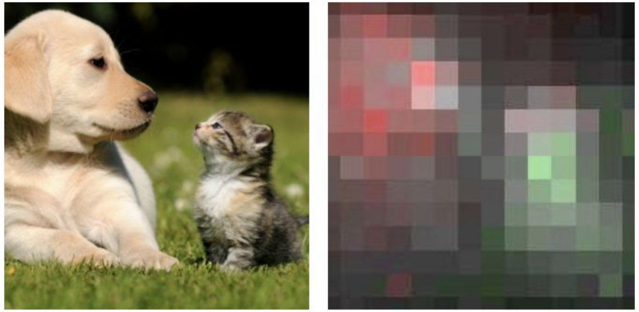
Krakovna et al. (2018)

Assurance

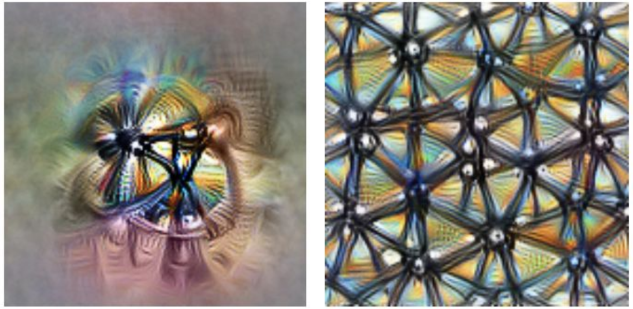
Analyzing, monitoring, and controlling systems during operation.



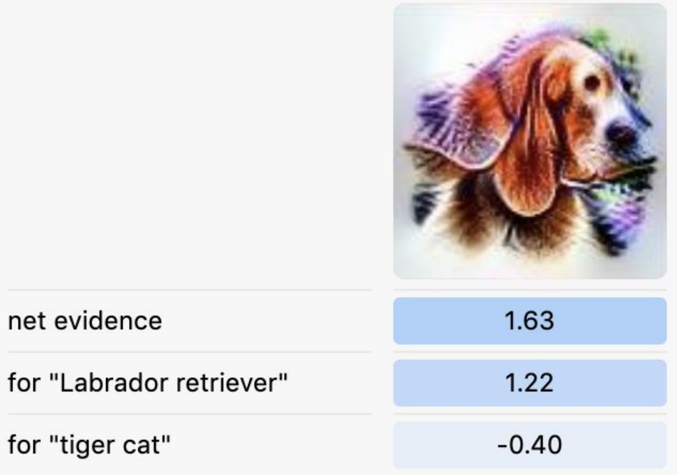
White-box analysis



Saliency maps

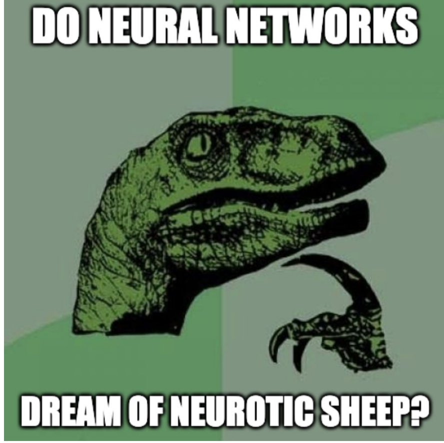


Maximizing activation of neurons/layers



net evidence	1.63
for "Labrador retriever"	1.22
for "tiger cat"	-0.40

Finding the channel that most supports a decision

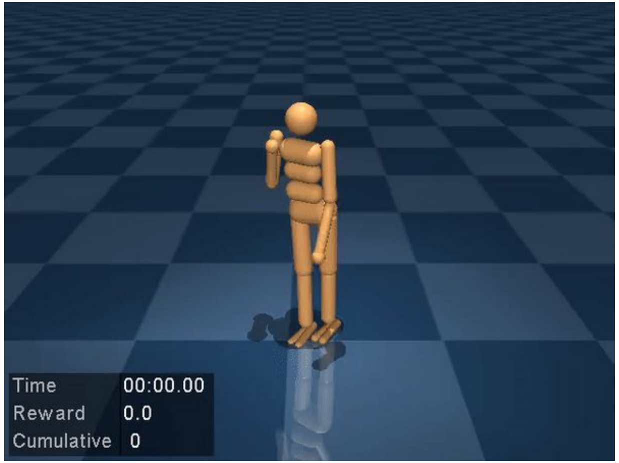


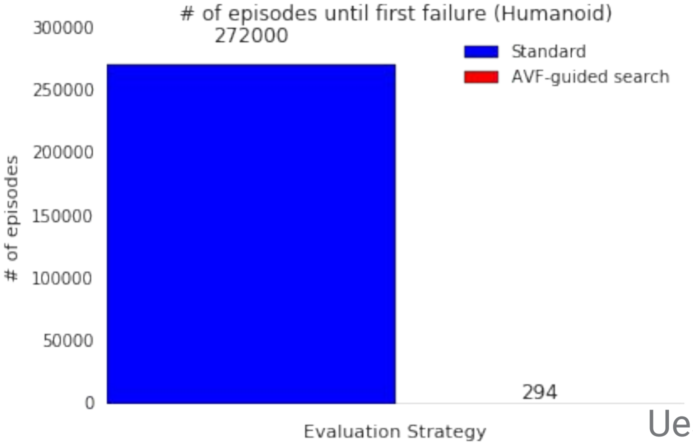
Olah et al. (Distill, 2017, 2018)

Black-box analysis: finding rare failures

- Approximate "*AVF*"
  f: initial MDP state ⟼ P[failure]
- Train on a family of related agents of varying robustness
- ⇒ Bootstrapping by learning the structure of difficult inputs on weaker agents

Result: failures found ~1,000x faster





Uesato et al. (2018)

Verification of neural networks

Reluplex

□-local robustness at point x_0:

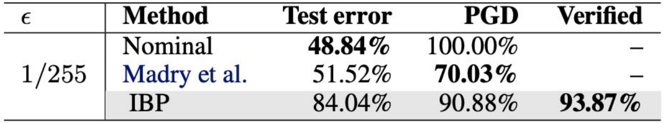
$$\forall \vec{x}. \quad \|\vec{x} - \vec{x}_0\| \leq \delta \quad \Rightarrow \quad N(\vec{x}) = N(\vec{x}_0)$$

- Rewrite this as SAT formula with linear terms
- Use an SMT-solver to solve the formula
- **Reluplex**: special algorithm for branching with ReLUs
- Verified adversarial robustness of 6-layer MLP with ~13k parameters

Katz et al. (CAV 2017)

Interval bound propagation

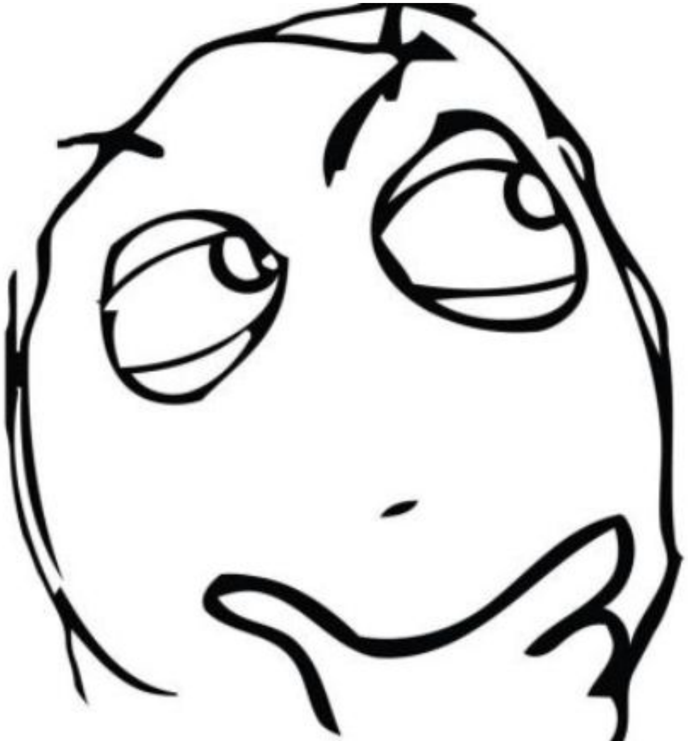
ImageNet downscaled to 64x64:



ϵ	Method	Test error	PGD	Verified
1/255	Nominal	**48.84%**	100.00%	–
	Madry et al.	51.52%	**70.03%**	–
	IBP	84.04%	90.88%	**93.87%**

Ehlers (ATVA 2017), Gowal et al. (2018)

Questions?



— 10 min break —

Part II
Specification: Fairness

Silvia Chiappa · ICML 2019

DeepMind

ML systems used in areas that severely affect people lives

- Financial lending
- Hiring
- Online advertising
- Criminal risk assessment
- Child welfare
- Health care
- Surveillance

Two examples of problematic systems

1. **Criminal Risk Assessment Tools**
   Defendants are assigned scores that predict the risk of re-committing crimes. These scores inform decisions about bail, sentencing, and parole. Current systems have been accused of being biased against black people.

2. **Face Recognition Systems**
   Considered for surveillance and self-driving cars. Current systems have been reported to perform poorly, especially on minorities.

From public optimism to concern

America is turning against facial-recognition software

But that isn't the most promising use of technology

Dennis Vernooij

The Economist

Attitudes to police technology are changing—not only among American civilians but among the cops themselves.

Until recently Americans seemed willing to let police deploy new technologies in the name of public safety.

But technological scepticism is growing. On May 14th San Francisco became the first American city to ban its agencies from using facial recognition systems.

One fairness definition or one framework?

21 Fairness Definitions and Their Politics. Arvind Narayanan.

ACM Conference on Fairness, Accountability, and Transparency Tutorial (2018)

S. Mitchell, E. Potash, and S. Barocas (2018)
P. Gajane and M. Pechenizkiy (2018)
S. Verma and J. Rubin (2018)

Differences/connections between fairness definitions are difficult to grasp.

We lack common language/framework.

"Nobody has found a definition which is widely agreed as a good definition of fairness in the same way we have for, say, the security of a random number generator."

"There are a number of definitions and research groups are not on the same page when it comes to the definition of fairness."

"The search for one true definition is not a fruitful direction, as technical considerations cannot adjudicate moral debates."

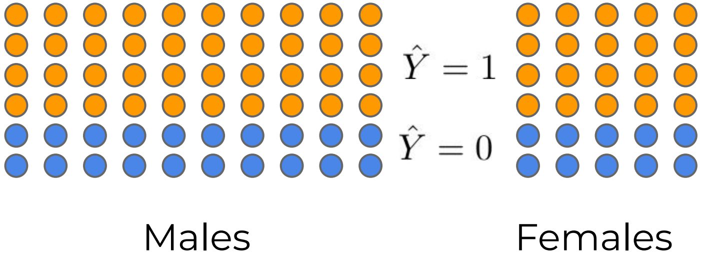
Common group-fairness definitions (binary classification setting)

Demographic Parity

The percentage of individuals assigned to class 1 should be the same for groups A=0 and A=1.

Dataset

- $a^n \in \{0,1\}$ sensitive attribute
- $y^n \in \{0,1\}$ class label
- $\hat{y}^n \in \{0,1\}$ prediction of the class
- $\mathbf{x}^n \in \mathbb{R}^d$ features



$$p(\hat{Y} = 1 | A = 0) = p(\hat{Y} = 1 | A = 1)$$

$$\hat{Y} \perp\!\!\!\perp A$$

Common group-fairness definitions

Equal False Positive/Negative Rates
(EFPRs/EFNRs)

$$p(\hat{Y}=1|Y=0,A=0)=p(\hat{Y}=1|Y=0,A=1)$$

$$p(\hat{Y}=0|Y=1,A=0)=p(\hat{Y}=0|Y=1,A=1)$$

$$\hat{Y} \perp\!\!\!\perp A|Y$$

Predictive Parity

$$p(Y=1|\hat{Y}=1,A=0)=p(Y=1|\hat{Y}=1,A=1)$$

$$p(Y=0|\hat{Y}=0,A=0)=p(Y=0|\hat{Y}=0,A=1)$$

$$Y \perp\!\!\!\perp A|\hat{Y}$$

The Law

Regulated Domains

Lending, Education, Hiring, Housing (extends to target advertising).

Protected (Sensitive) Groups

Reflect the fact that in the past there have been unjust practices.

Discrimination in the Law

Disparate Treatment
Individuals are treated differently because of protected characteristics (e.g. race or gender).
[Equal Protection Clause of the 14th Amendment.]

Disparate Impact
An apparently neutral policy that adversely affects a protected group more than another group.
[Civil Rights Act, Fair Housing Act, and various state statutes.]

Statistical test discrimination in human decisions

1. **Benchmarking:** Compares the rate at which groups are treated favorably.

   If white applicants are granted loans more often than minority applicants, that may be the result of bias.

2. **Outcome Test** (Becker (1957, 1993)): Compares the success rate of decisions (hit rate).

   Even if minorities are less creditworthy than whites, minorities who are granted loans, absent discrimination, should still be found to repay their loans at the same rate as whites who are granted loans.

Outcome test

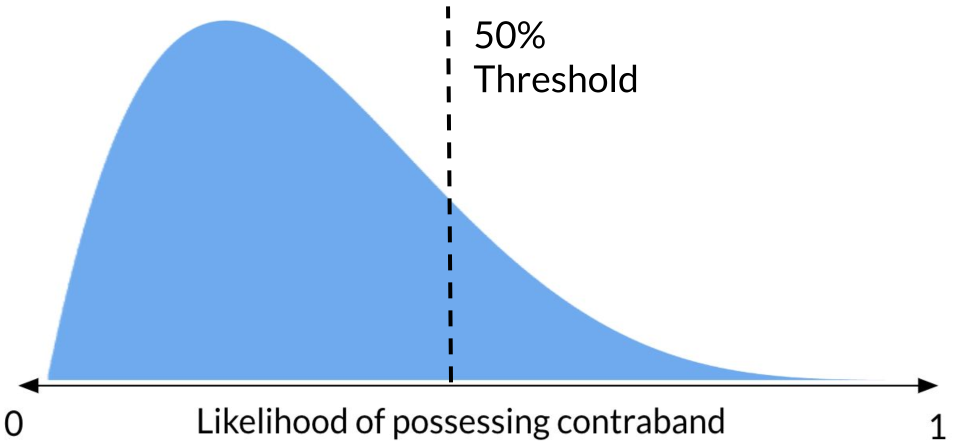
Outcome tests of racial disparities in police practices.
I. Ayres. Justice Research and Policy (2002)

Outcome Tests used to provide evidence that a decision making system has an unjustified disparate impact.

Example: Police search for contraband

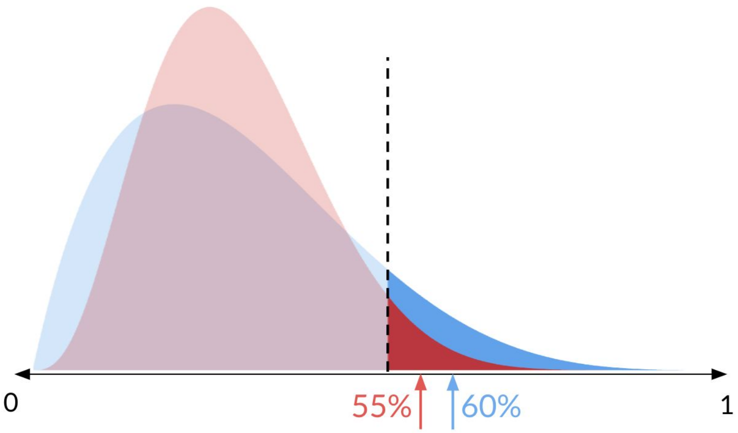
A finding that searches for a group are systematically less productive than searches for another group is evidence that police apply different thresholds when searching.

Risk Distribution

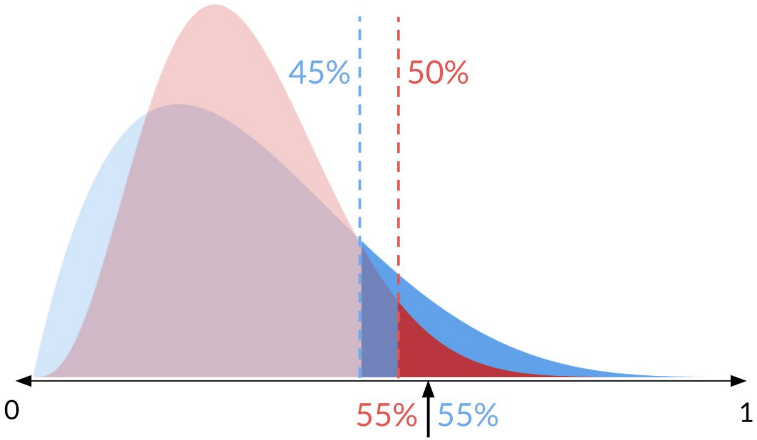


Problems with the outcome test

Defining and Designing Fair Algorithms.
Sam Corbett-Davies and Sharad Goel. ICML Tutorial (2018)



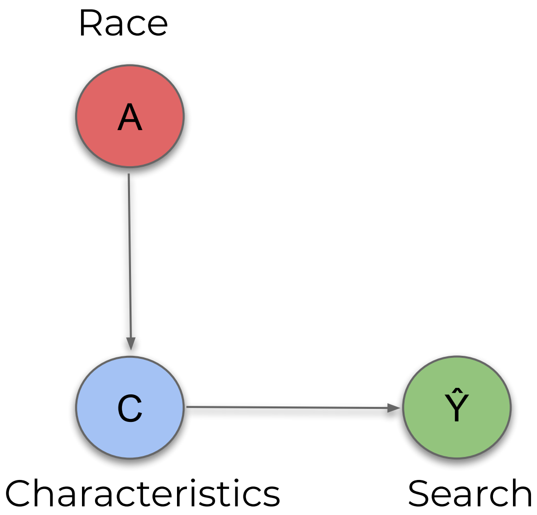
Police search if there's greater than 50% chance they'll find contraband. But the outcome test incorrectly suggests bias.



Police apply lower threshold in order to discriminate against blue drivers. But the outcome test incorrectly suggests no bias.

Tests for discrimination that account for the shape of the risk distributions find that officers apply a lower standard when searching black individuals. Simoiu et al. (2017)

Outcome test from a causal Bayesian network viewpoint

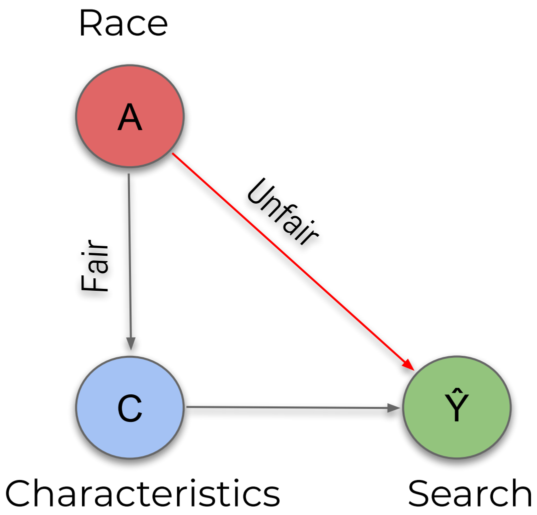


Nodes represent random variables:

- A = Race
- C = Characteristics
- Ŷ= Police search

Links express causal influence.

What is the outcome test trying to achieve?



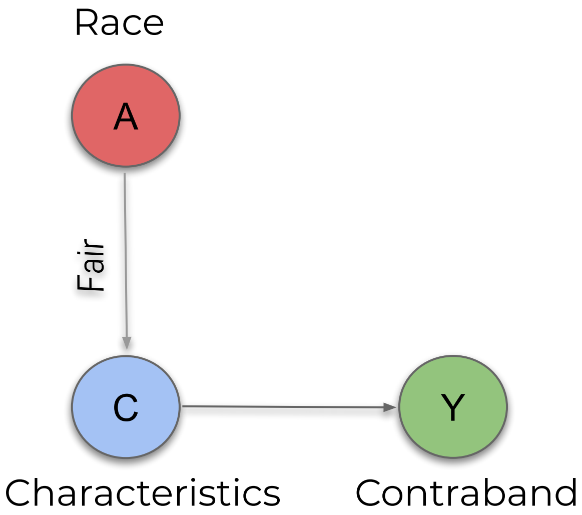
Understand whether there is a direct influence of A on Ŷ, namely a direct path A → Ŷ, by checking whether

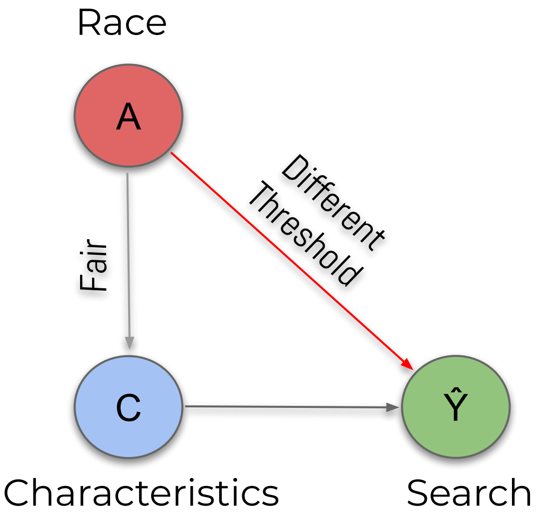
$$p(Y = 1|\hat{Y} = 1, A = 0) = p(Y = 1|\hat{Y} = 1, A = 1)$$

where Y represents Contraband.

What is the outcome test trying to achieve?

Has a direct path been introduced when searching?





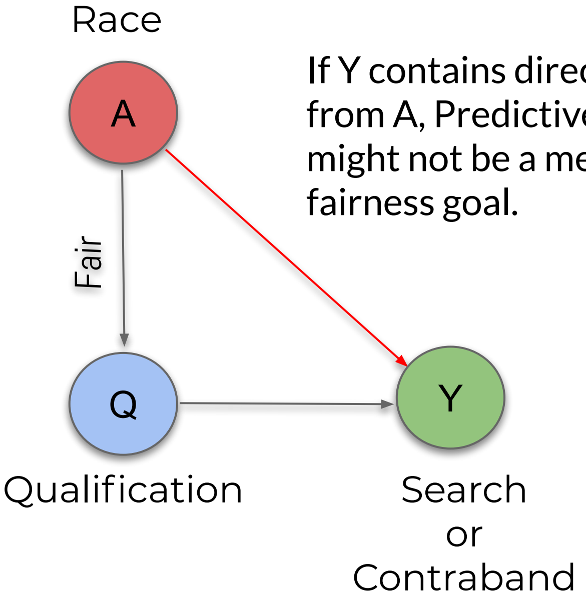
Connection to ML Fairness

Outcome Test: Percentage of those classified positive (i.e., searched) who had contraband. Formally equivalent of checking for Predictive Parity.

$$p(Y=1|\hat{Y}=1, A=0) = p(Y=1|\hat{Y}=1, A=1)$$

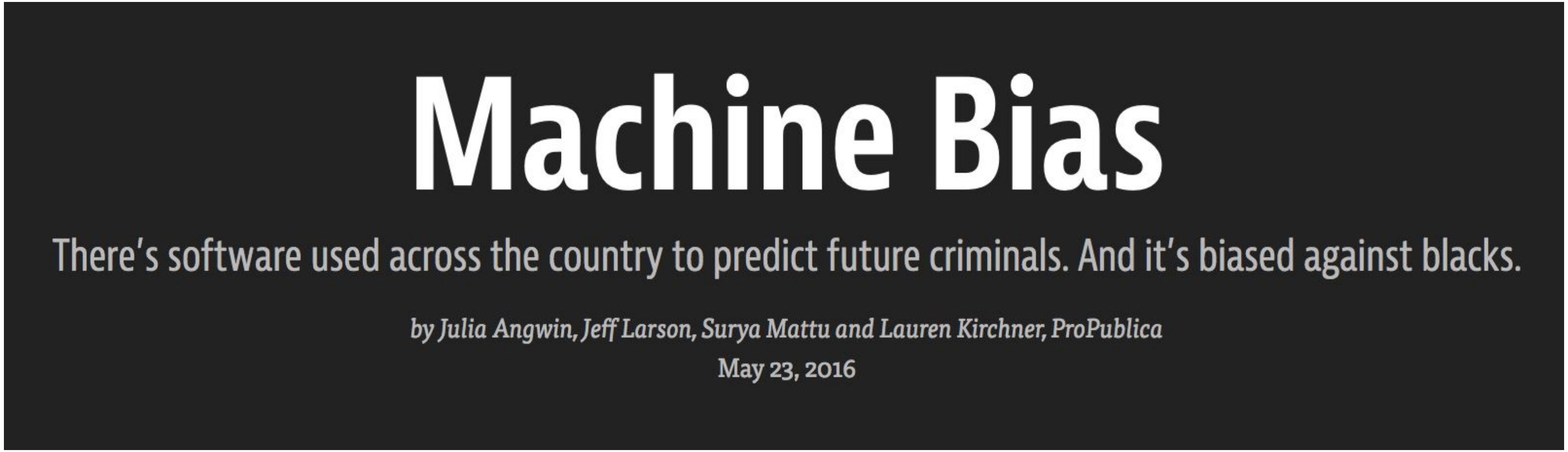
Assumption in Outcome Test: Y reflects genuine contraband.

This excludes the case of e. g. deliberate intention of making a group look guilty by placing contrabands in cars. But when learning a ML model from a dataset, we might be in this scenario. Or the label Y could correspond to Search rather than Contraband.



If Y contains direct influence from A, Predictive Parity might not be a meaningful fairness goal.

COMPAS predictive risk instrument



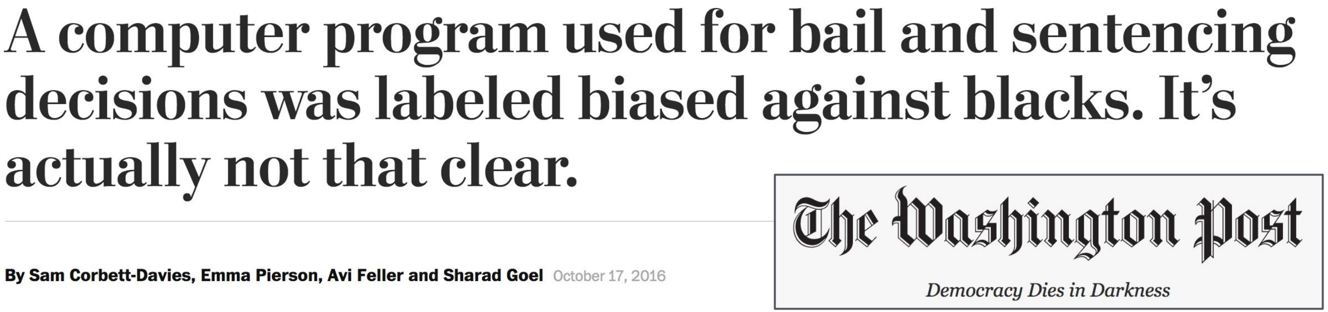
Machine Bias

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

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

COMPAS predictive risk instrument



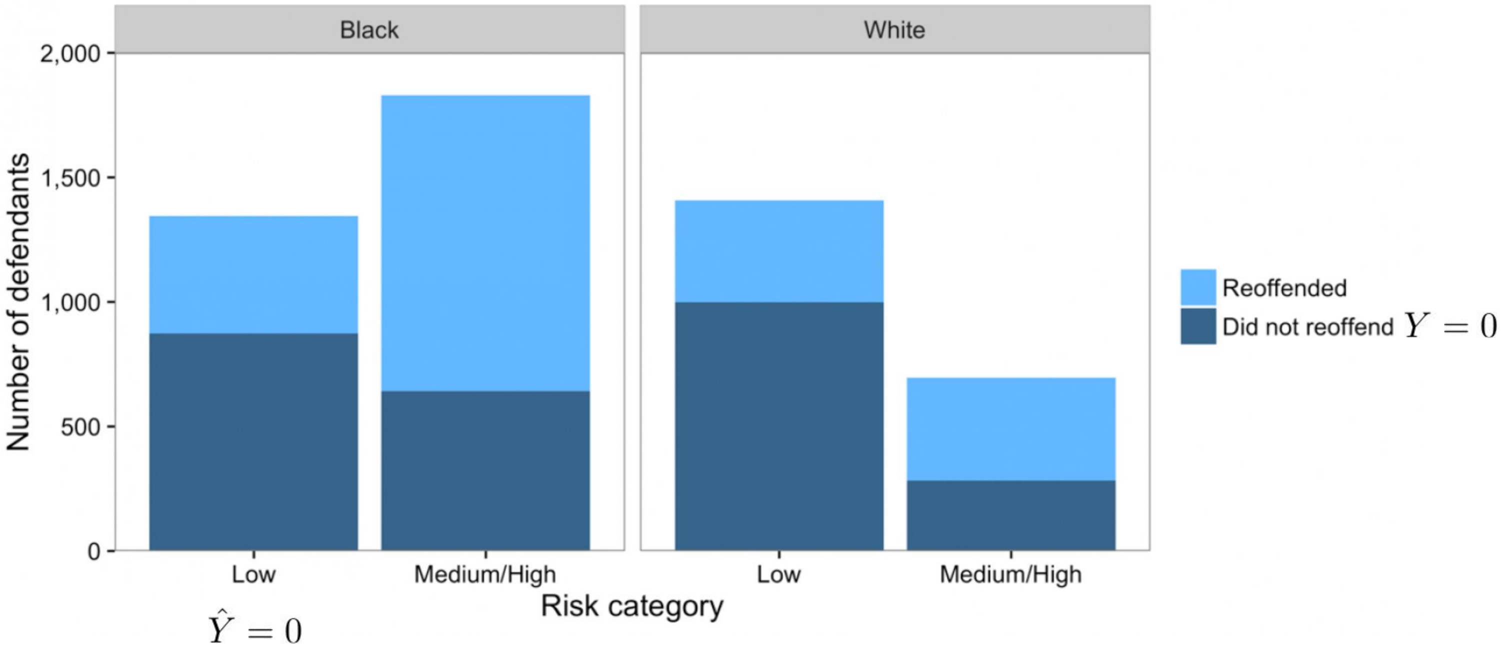
A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel October 17, 2016

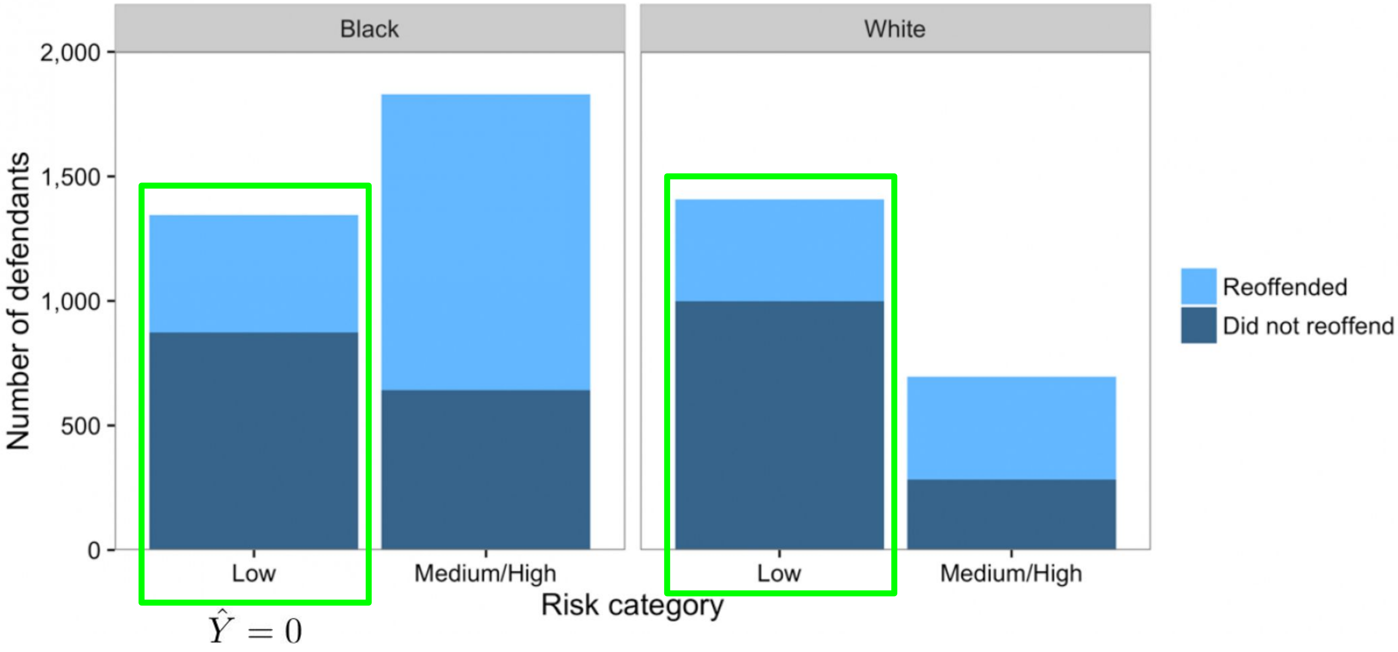
The Washington Post

Democracy Dies in Darkness

COMPAS predictive risk instrument



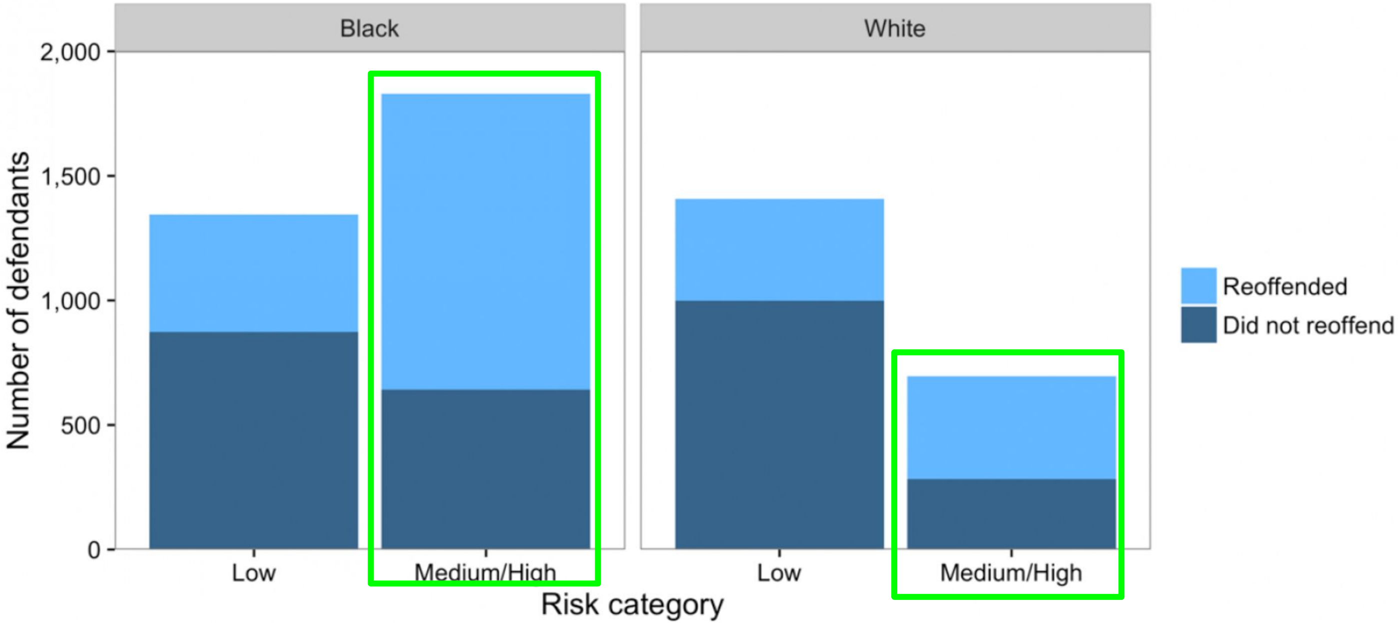
COMPAS predictive risk instrument



Low risk

~70% did not reoffend for both the black and white groups.

COMPAS predictive risk instrument

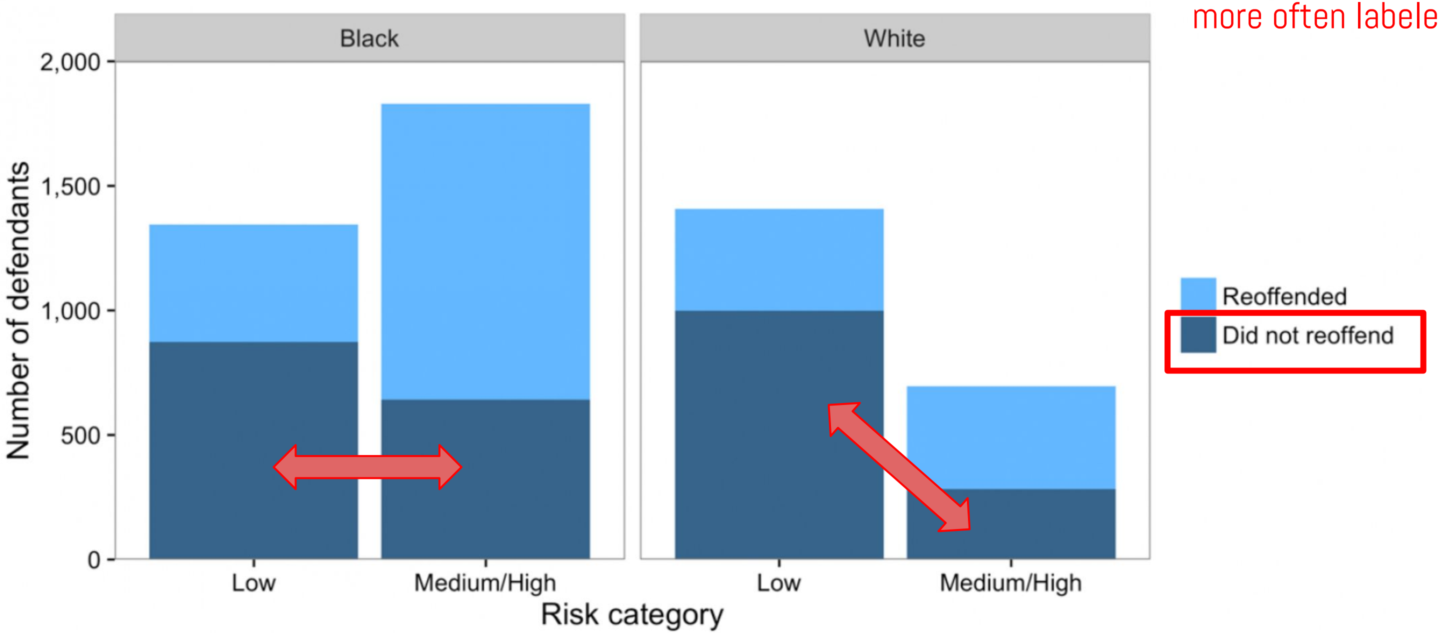


Medium-high risk

The same percentage of individuals did not reoffend in both groups.

$$Y \perp A | \hat{Y}$$

COMPAS predictive risk instrument



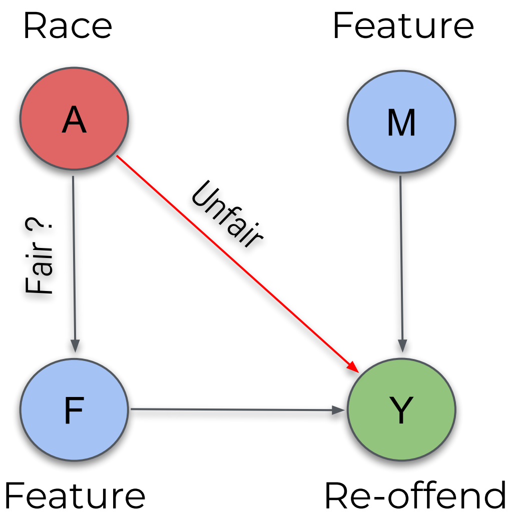
Black defendants who did not reoffend were more often labeled "high risk"

Did not reoffend

False Positive Rates differ

$$\hat{Y} \not\perp\!\!\!\perp A | Y$$

Patterns of unfairness in the data not considered



Modern policing tactics center around targeting a small number of neighborhoods --- often disproportionately populated by non-whites.

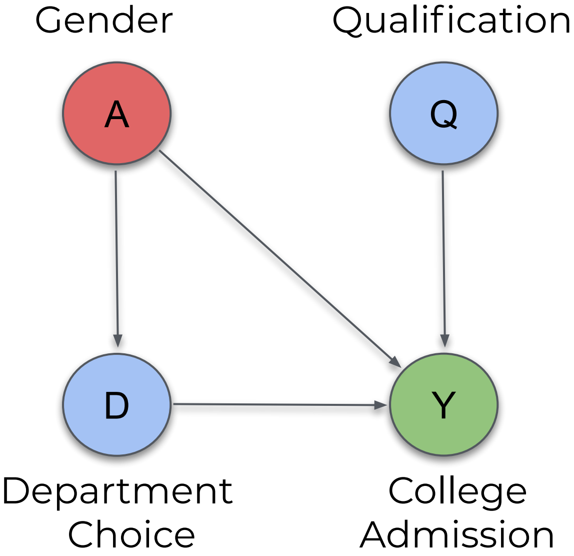
We can rephrase this as indicating the presence of a direct path A → Y (through unobserved neighborhood).

Such tactics also imply an influence of *A* on *Y* through *F* containing number of prior arrests.

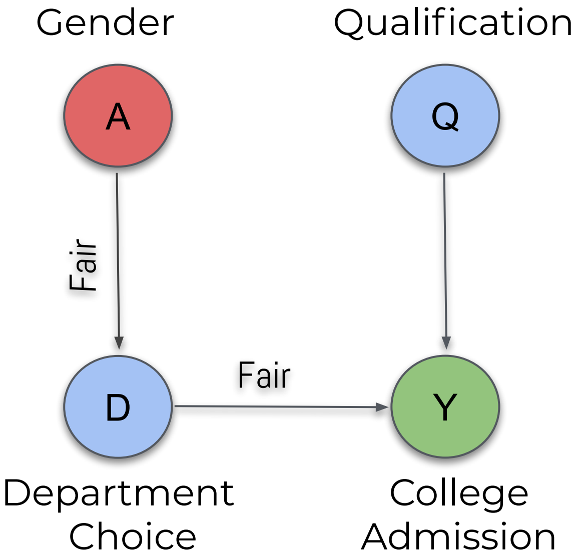
EFPRs/EFNRs and Predictive Parity require the rate of (dis)agreement between the correct and predicted label (e.g. incorrect-classification rates) to be the same for black and white defendants, and are therefore not concerned with dependence of *Y* on *A*.

Patterns of unfairness: college admission example

A causal Bayesian networks viewpoint on fairness.
S. Chiappa and W. S. Isaac (2018)



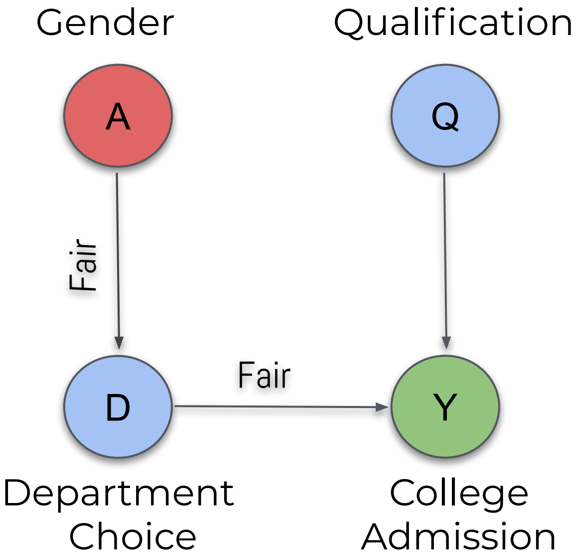
Three main scenarios



Influence of A onY is all fair

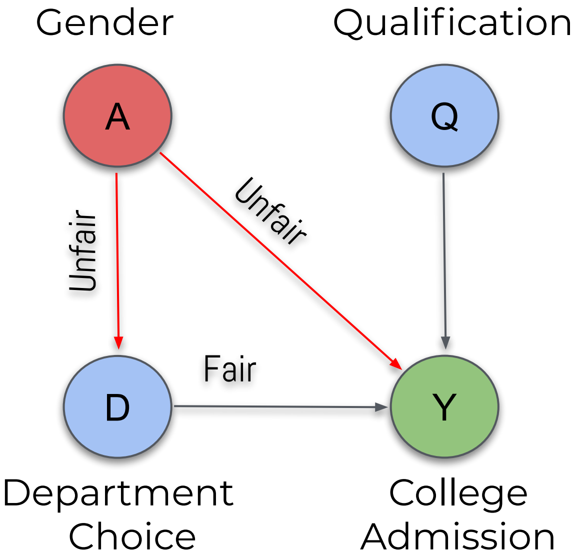
Predictive Parity
Equal FPRs/FNRs

Three main scenarios



Influence of A onY is all fair

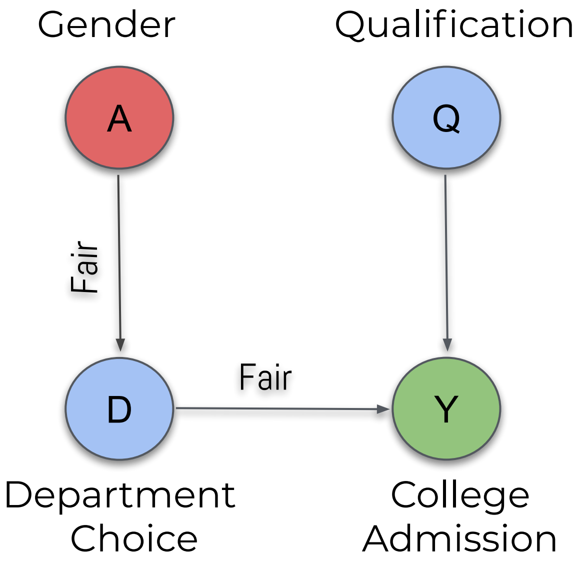
Predictive Parity
Equal FPRs/FNRs



Influence of A onY is all unfair

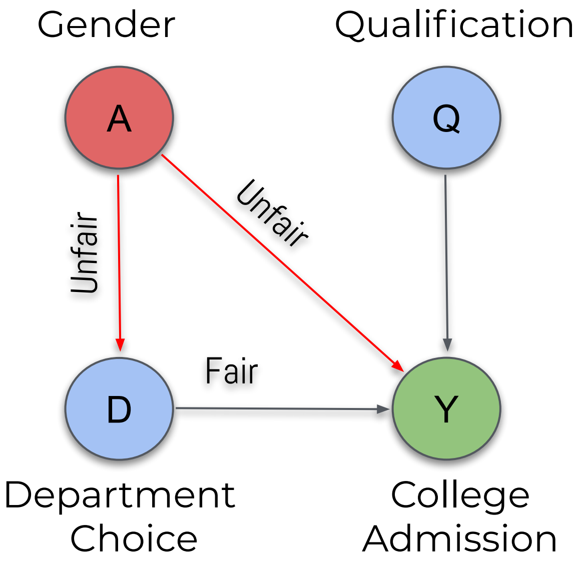
Demographic Parity

Three main scenarios



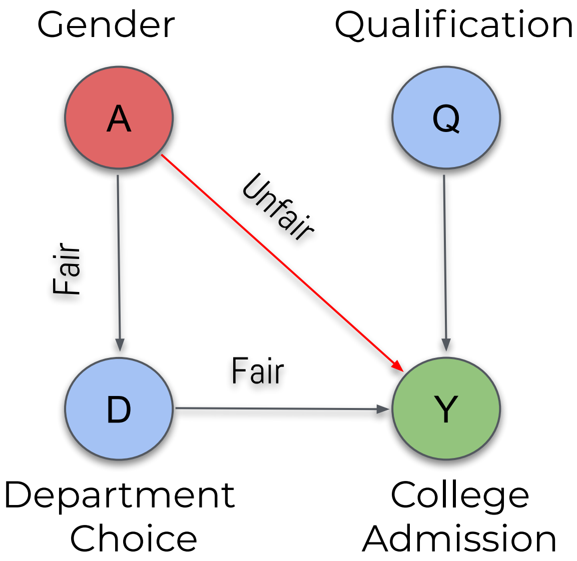
Influence of A onY is all fair

Predictive Parity
Equal FPRs/FNRs



Influence of A onY is all unfair

Demographic Parity



Influence of A onY is both fair and unfair

?

Path-specific fairness

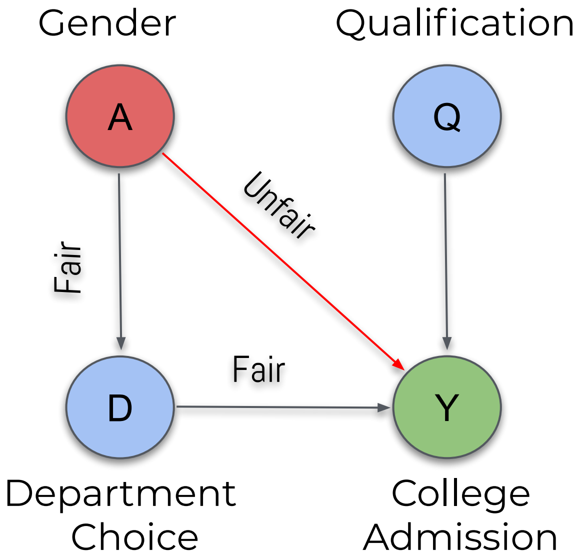
A=a and A=$\bar{a}$ indicate female and male applicants respectively

$Y_{\bar{a}}(D_a)$ Random variable with distribution equal to the conditional distribution of Y given A restricted to causal paths, with A=$\bar{a}$ along A → Y and A=a along A → D → Y.

$\hat{Y}_{\bar{a}}(D_a)$

Path-specific Fairness

$$p(\hat{Y}_{\bar{a}}(D_a) = 1) = p(\hat{Y}_a = 1)$$



Accounting for full shape of distribution

Wasserstein fair classification.
R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa (2019)

Binary classifier outputs a continuous value that represents the probability that individual *n* belong to class 1, $s^n = p(Y=1|A=a^n, X=x^n)$. A decision is the taken by thresholding $\hat{y}^n = \mathbb{1}_{s^n > \tau}$

General expression including regression $s^n = \mathbb{E}_{p(Y|A=a^n, X=x^n)}[Y]$

$\hat{y}^n = s^n$ regression

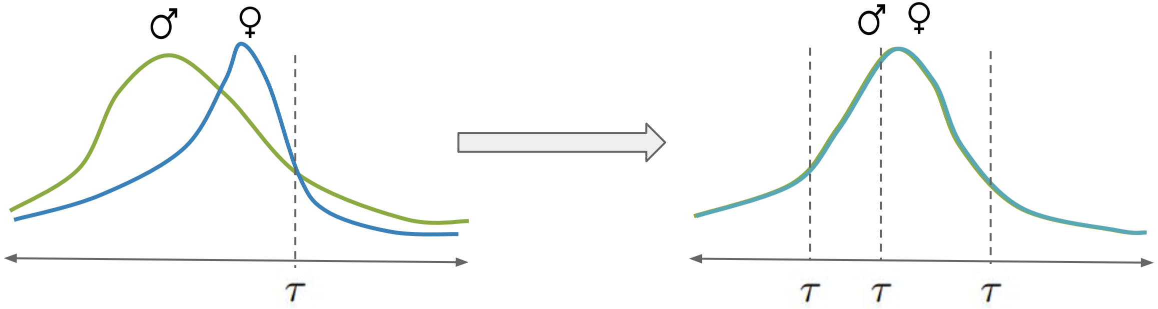
$\hat{y}^n = \mathbb{1}_{s^n > \tau}$ classification

Demographic Parity

$$\mathbb{E}_{p(\hat{Y}|A=\bar{a})}[\hat{Y}] = \mathbb{E}_{p(\hat{Y}|A=a)}[\hat{Y}]$$

Strong Demographic Parity

$$p(S|\bar{a}) = p(S|a)$$



Strong Path-specific Fairnress

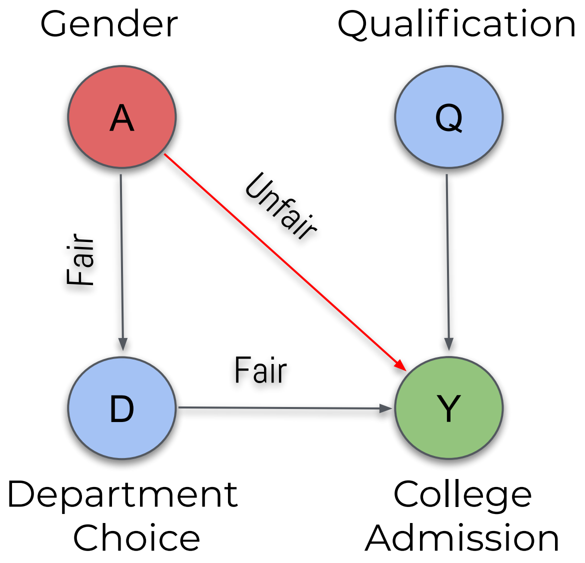
$$p(S_{\bar{a}}(D_a)) = p(S_a)$$

Individual fairness

Similar individuals should be treated similarly.

Fairness through awareness. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2011)

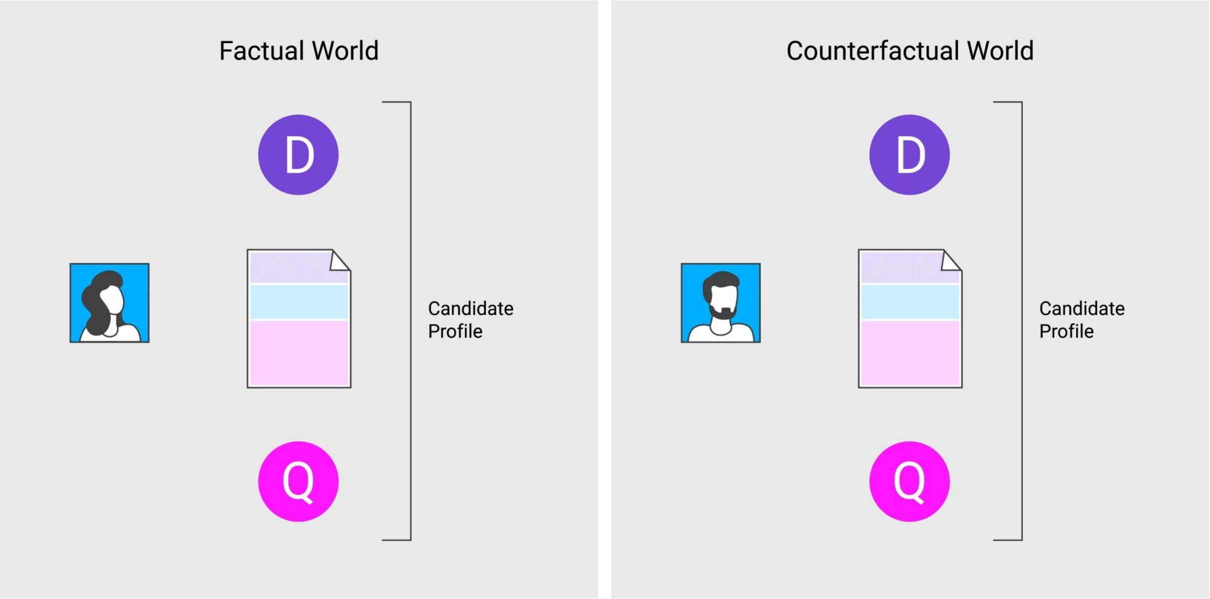
A female applicant should get the same decision as a male applicant with the same qualification and applying to the same department.



Individual fairness

Path-specific counterfactual fairness. S. Chiappa, and T. P. Gillam (2018)

Compute the outcome pretending that the female applicant is male along the direct path A → Y.



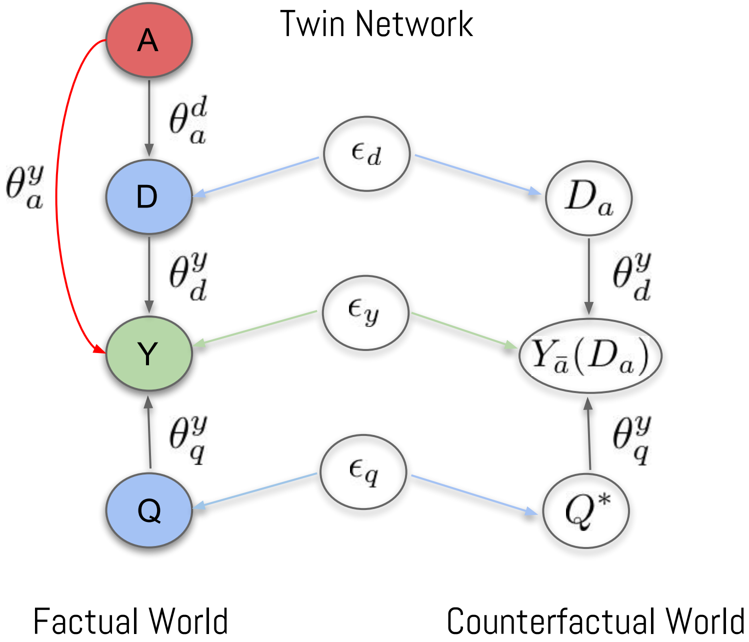
Path-specific counterfactual fairness: linear model example

$$A \sim \mathrm{Bern}(\pi), Q = \theta^q + \epsilon_q,$$
$$D = \theta^d + \theta^d_a A + \epsilon_d,$$
$$Y = \theta^y + \theta^y_a A + \theta^y_q Q + \theta^y_d D + \epsilon_y\,,$$

$$\mathbb{E}_{p(Y_{\bar{a}}(D_a)|A=a,Q=q^n,D=d^n)}[Y_{\bar{a}}(D_a)]$$

As Q is non-descendant of A, and D is descendant of A along a fair path, this coincides with

$$\mathbb{E}_{p(Y|A=\bar{a},Q=q^n,D=d^n)}[Y]$$



In more complex scenarios we would need to use corrected versions of the features.

How to achieve fairness

1. Post-processing: Post-process the model outputs.

   Doherty et al. (2012), Feldman (2015), Hardt et al. (2016), Kusner et al. (2018), Jiang et al. (2019).

2. Pre-processing: Pre-process the data to remove bias, or extract representations that do not contain sensitive information during training.

   Kamiran and Calder (2012), Zemel et al. (2013), Feldman et al. (2015), Fish et al. (2015), Louizos et al. (2016), Lum and Johndrow (2016), Adler et al. (2016), Edwards and Storkey (2016), Beutel et al. (2017), Calmon et al. (2017), Del Barrio et al. (2019).

3. In-processing: Enforce fairness notions by imposing constraints into the optimization, or by using an adversary.

   Goh et al. (2016), Corbett-Davies et al. (2017), Zafar et al. (2017), Agarwal et al. (2018), Cotter et al. (2018), Donini et al. (2018), Komiyama et al. (2018), Narasimhan (2018), Wu et al. (2018), Zhang et al. (2018), Jiang et al. (2019).

Start thinking about a structure for evaluation

Pharmaceuticals	Machine Learning Systems
Safety: Initial testing on human subjects.	Digital testing: Standard test set.
Proof-of-concept: Estimating efficacy and optimal use on selected subjects.	Laboratory testing: Comparison with humans, user testing.
Randomized controlled-trials: Comparison against existing treatment in clinical setting.	Field testing: Impact when imported in society.
Post-marketing surveillance: Long-term side effects.	Routine use: Monitoring safety patterns over time.

Making Algorithms Trustworthy.
D. Spiegelhalter. NeurIPS (2018).

Stead et al. Journal of the American Medical Informatics Association (1994)

Questions?