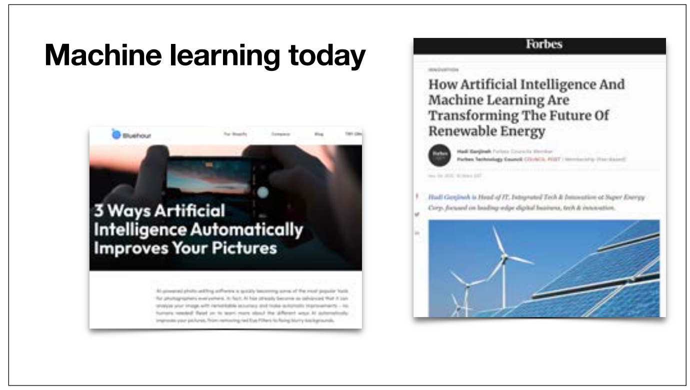
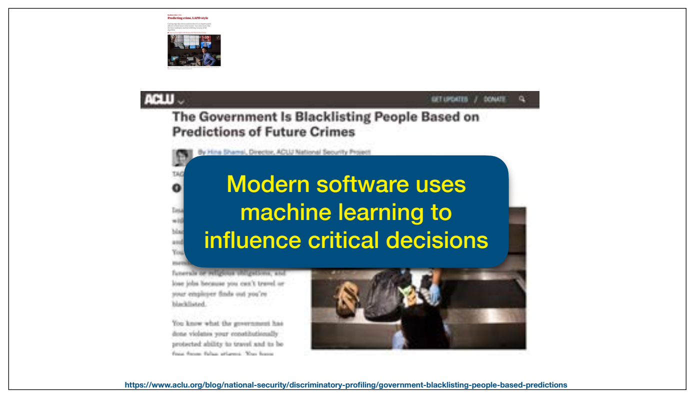
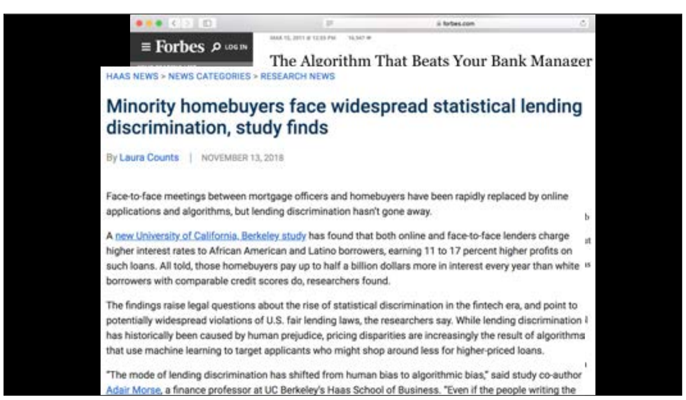

The promise and perils
of using machine learning
when engineering software.
Yuriy Brun
MaLTeSQuE 2022


Machine learning today
3 Ways Artificial Intelligence Automatically Improves Your Pictures
Forbes
How Artificial Intelligence And Machine Learning Are Transforming The Future Of Renewable Energy


Resilient cities Cities
Predicting crime, LAPD-style
Cutting edge data-driven analysis directs Los Angeles patrol officers to likely future crime scenes - but critics worry that decision-making by machine will bring 'tyranny of the algorithm'
Join our live Q&A with Homicide Watch this Friday
https://www.theguardian.com/cities/2014/jun/25/predicting-crime-lapd-los-angeles-police-data-analysis-algorithm-minority-report

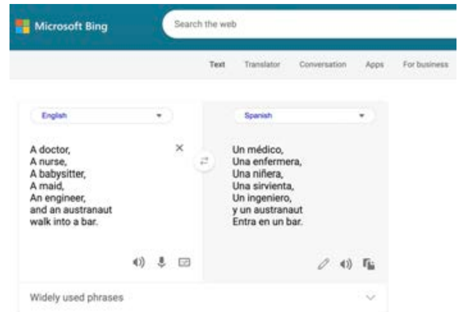

ACLU
The Government Is Blacklisting People Based on Predictions of Future Crimes
Modern software uses machine learning to influence critical decisions
https://www.aclu.org/blog/national-security/discriminatory-profiling/government-blacklisting-people-based-predictions


Forbes
The Algorithm That Beats Your Bank Manager
HAAS NEWS > NEWS CATEGORIES > RESEARCH NEWS
Minority homebuyers face widespread statistical lending discrimination, study finds
By Laura Counts | NOVEMBER 13, 2018
Face-to-face meetings between mortgage officers and homebuyers have been rapidly replaced by online applications and algorithms, but lending discrimination hasn't gone away.
A new University of California, Berkeley study has found that both online and face-to-face lenders charge higher interest rates to African American and Latino borrowers, earning 11 to 17 percent higher profits on such loans. All told, those homebuyers pay up to half a billion dollars more in interest every year than white borrowers with comparable credit scores do, researchers found.
The findings raise legal questions about the rise of statistical discrimination in the fintech era, and point to potentially widespread violations of U.S. fair lending laws, the researchers say. While lending discrimination has historically been caused by human prejudice, pricing disparities are increasingly the result of algorithms that use machine learning to target applicants who might shop around less for higher-priced loans.
"The mode of lending discrimination has shifted from human bias to algorithmic bias," said study co-author Adair Morse, a finance professor at UC Berkeley's Haas School of Business. "Even if the people writing the


HEALTH IT ANALYTICS
MIT News
ON CAMPUS AND AROUND THE WORLD
Using AI to predict breast cancer and personalize care
MIT/MGH's image-based deep learning model can predict breast cancer up to five years in advance.
Adam Conner-Simons and Rachel Gordon | CSAIL
May 7, 2019
Software can make bad decisions.
Software can discriminate!

Google Translate

BING TRANSLATOR



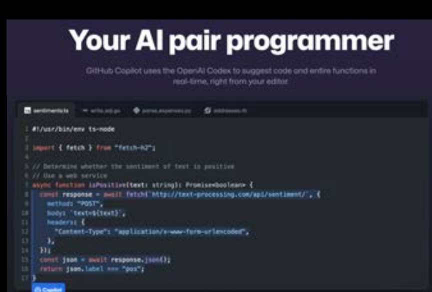
Machine learning has great promise,
but with that promise, come risks.

Today's goal:
Identifying and addressing the risks

Part I
Automated program repair

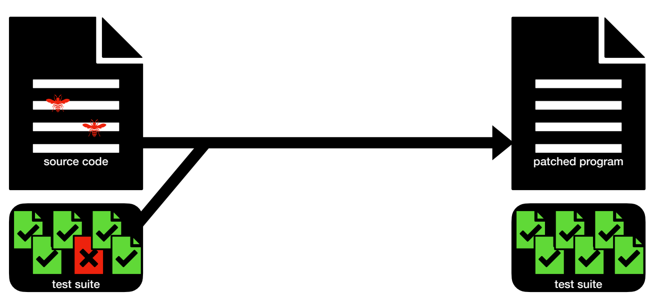
Part II
Software discrimination

Machine Learning in Software Engineering

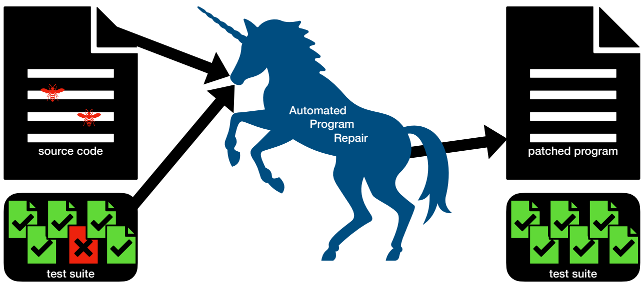


https://github.com/features/copilot

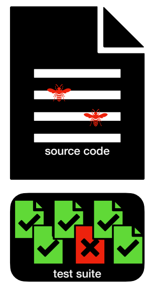
A simpler problem: Automated program repair



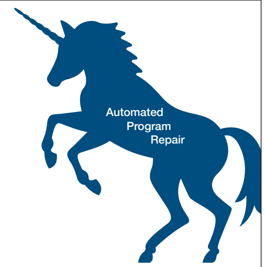
A simpler problem: Automated program repair



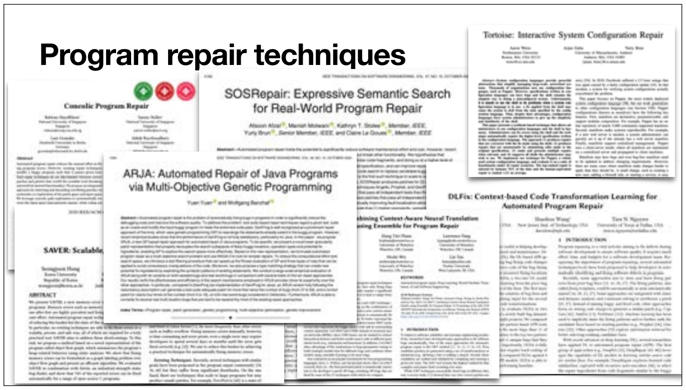
Program repair techniques



- Tweak the program
- Check if tests pass
- If not, repeat



Program repair techniques


Tortoise: Interactive System Configuration Repair

Concolic Program Repair

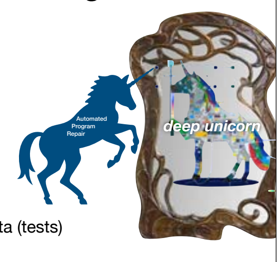
SOSRepair: Expressive Semantic Search for Real-World Program Repair

ARJA: Automated Repair of Java Programs via Multi-Objective Genetic Programming

DLFix: Context-based Code Transformation Learning for Automated Program Repair

APR is a form of machine learning

- first, many techniques rely on ML to learn
  - where to edit the code
  - how to edit the code
  - how to decide which patches are good



- second, the underlying problem is learning a function (program) using training data (tests)

How well does APR work?

- Evaluated 4 techniques
  - GenProg
  - Par
  - TrpAutoRepair
  - SimFix
- Measured patch quality
- Measured what affects patch quality


IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 48, NO. 2, FEBRUARY 2022

Quality of Automated Program Repair on Real-World Defects

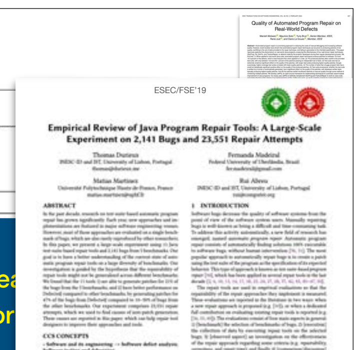
Manish Motwani, Mauricio Soto, Yuriy Brun, Senior Member, IEEE, René Just, and Claire Le Goues, Member, IEEE

Quality vs. quantity

technique
GenProg
Par
SimFix
TRPAutoRepair
total

When applied to rea

APR produces patches for


ESEC/FSE'19

Empirical Review of Java Program Repair Tools: A Large-Scale Experiment on 2,141 Bugs and 23,551 Repair Attempts

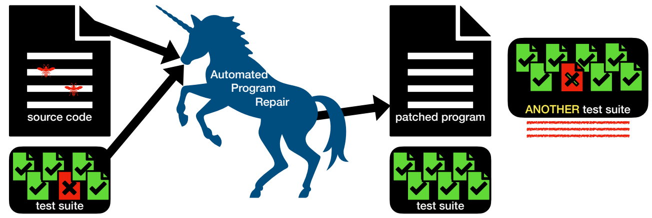
Thomas Durieux, Fernanda Madeiral, Matias Martinez, Rui Abreu

Quality vs. quantity

Potential problem: Overfitting

APR uses a set of tests to guide repair.
Tests are inherently partial.
No way APR can know if a patch captures intended behavioral constraints.

Quality vs. quantity

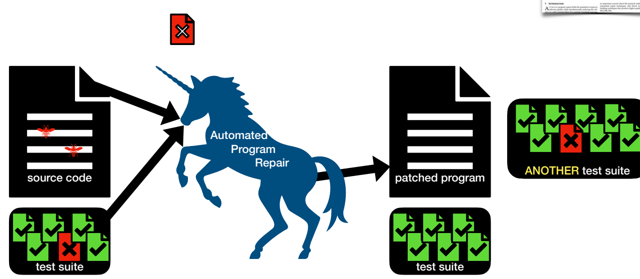


Quality vs. quantity

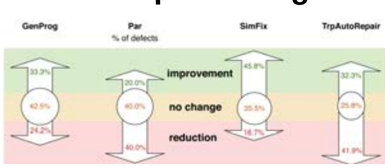
technique	patch quality minimum	mean	median	maximum	100%-quality patches
GenProg	64.8%	95.7%	98.4%	100.0%	24.3%
Par	64.8%	96.1%	98.5%	100.0%	13.8%
SimFix	65.0%	96.3%	99.9%	100.0%	46.1%
TrpAutoRepair	64.8%	96.4%	98.4%	100.0%	19.5%

Less than half (14-46%)
of the patches are correct

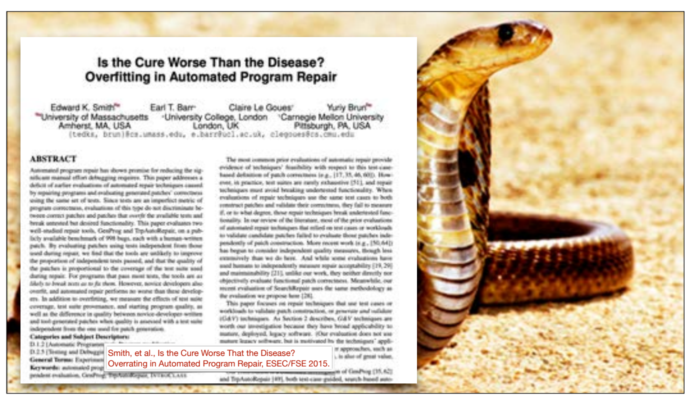
Does APR at least improve things a bit?



Does APR at least improve things a bit?



technique	change in quality due to patch minimum	mean	median	maximum
GenProg	−30.9%	−1.7%	0.0%	2.6%
Par	−30.9%	−2.8%	0.0%	1.5%
SimFix	−24.9%	0.2%	0.0%	35.0%
TrpAutoRepair	−30.9%	−2.1%	0.0%	3.8%



Takeaway: Tests are an imperfect oracle, so APR suffers, producing low-quality patches.

Can we find a domain with better oracles?

Formal verification allows proving software correct

Interactive theorem provers for formal verification

Formal verification comes with a built-in oracle: The theorem prover

Proof script

```
Proof.
intros.
induction n.
⋮
Qed.
```

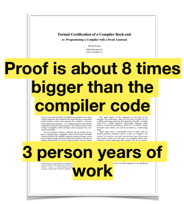
Agda

Industrial impact of theorem proving

AIRBUS, BedRock Systems Inc, galois, android, aws, CERTORA

Prohibitively difficult

Verified software requires a lot of time and a lot of proofs in proportion to code

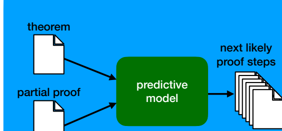


POPL 2006

Virtually all software that ships today is unverified.

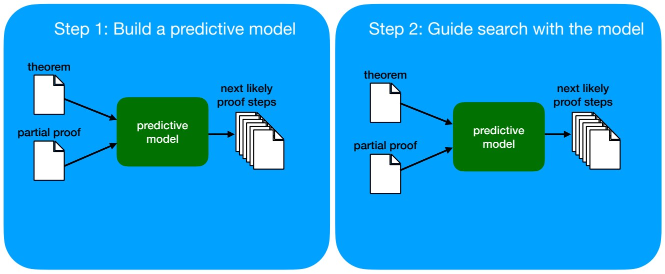
Proposal: Use APR-style technology to synthesize proofs

Step 1: Build a predictive model

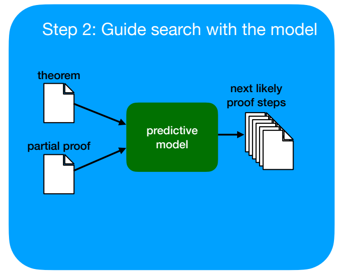


Step 2: Guide search with the model

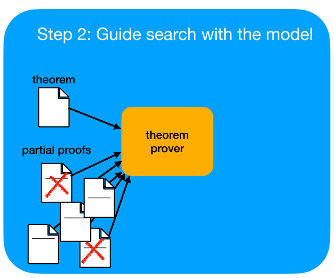
Proposal: Use APR-style technology to synthesize proofs



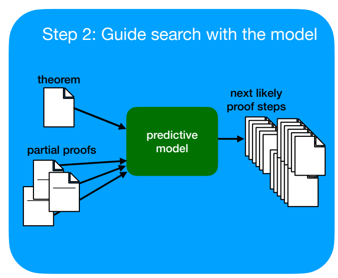
Proposal: Use APR-style technology to synthesize proofs

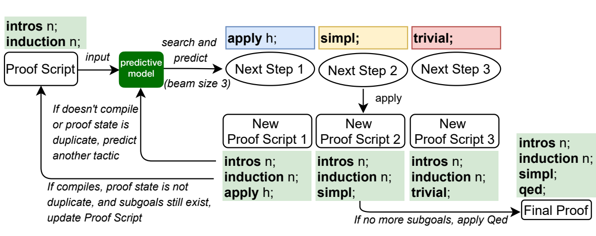


Proposal: Use APR-style technology to synthesize proofs

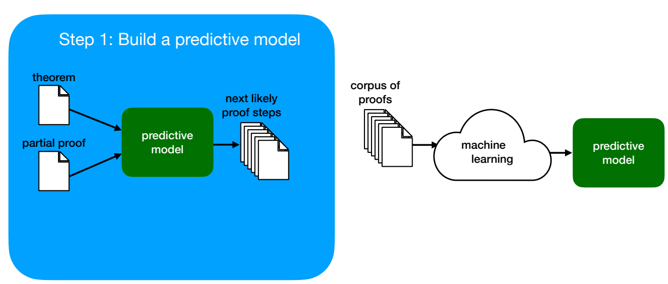


Proposal: Use APR-style technology to synthesize proofs

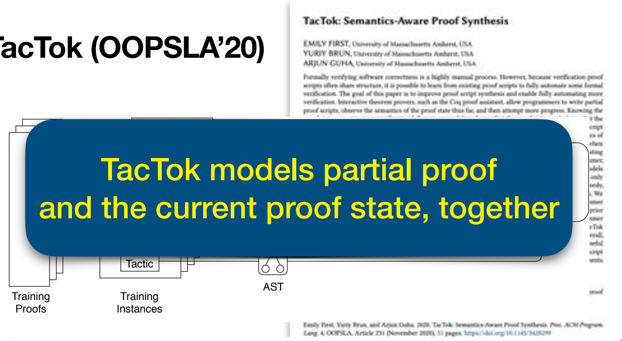




How to learn a predictive model



TacTok (OOPSLA'20)



TacTok models partial proof and the current proof state, together

ASTactic [Yang and Deng, Learning to Prove Theorems via Interacting with Proof Assistants, ICML'19] modeled just proof state.
[Hellendoorn, Devanbu, Alipour, On the naturalness of proofs, ESEC/FSE NIER'18] looked at predictability of proof sequences.

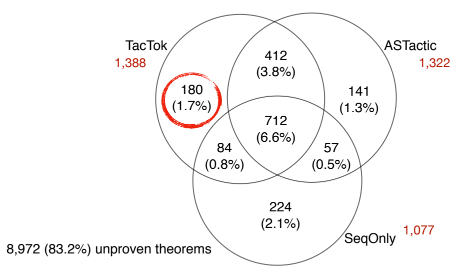
CoqGym Dataset

- 123 open-source software projets in Coq
- 70,856 theorems
- Broken down into 96 projects (57,719 proofs) for training and 27 projects (13,137 theorems) for testing

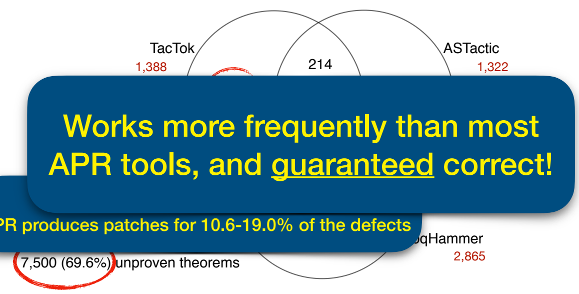
https://github.com/princeton-vl/CoqGym

[Yang and Deng, Learning to Prove Theorems via Interacting with Proof Assistants, ICML'19]

TacTok vs. ASTactic vs. SeqOnly



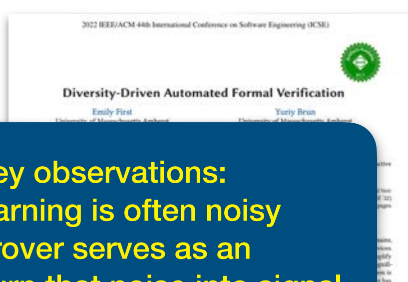
TacTok vs. ASTactic vs. CoqHammer



Works more frequently than most APR tools, and guaranteed correct!

APR produces patches for 10.6-19.0% of the defects

Diva (ICSE'22)



2 key observations:
- Machine learning is often noisy
- Theorem prover serves as an oracle to turn that noise into signal.

Diva (ICSE'22)

- Vary:
  - proof tactic and token depth
  - learning rate
  - embedding size
  - number of layers
  - training order
  - access to proof state, partial proof, Gallina proof term

Diva vs. state-of-the-art

Diversity inherent in ML increases the proving power 68%-77% over prior search-based synthesis tools, and 27% over CoqHammer.

https://github.com/LASER-UMASS/Diva/

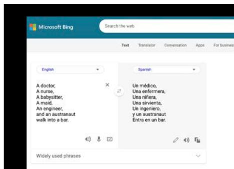
Fully Automated Formal Verification

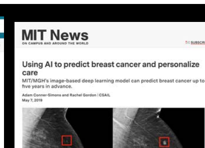
Machine learning and meta-heuristic search can fully automate some bug-repair and formal verification.

While APR underperforms because it is driven by an unreliable oracle, formal verification is a killer app for APR because the theorem prover provides a reliable oracle.

…let's talk about a different peril of machine learning that verification might help with.

Part II
Software discrimination





Part II
Software discrimination

Data-driven systems can exhibit undesirable properties.

Can we build systems to be safe and fair?

Testing systems for bias

Themis

automated test-suite generator

How much does my software discriminate with respect to …?

Does my software discriminate more than 10% of the time, and against what?

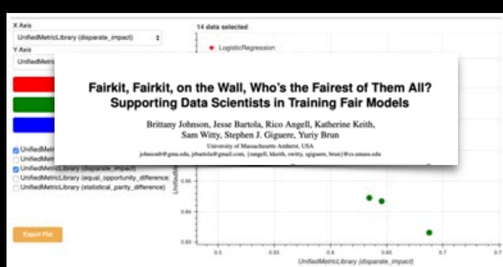
http://fairness.cs.umass.edu

Galhotra, Brun, and Meliou, Fairness Testing: Testing Software for Discrimination. ESEC/FSE 2017

Debugging bias

Debugging

fairkit-learn



Can we verify systems to be safe and fair?

How would that work?

User specifies a definition of safe or fair behavior.

training | testing | safety

Train classifiers,
selects one to satisfy fairness,
verify safety on held-out suite.

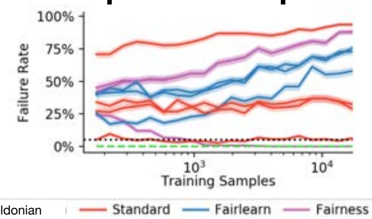
Example scenario:

Suppose a university wants to train a model to predict student success from entrance exam scores, while ensuring the model is fair:

roughly the same fraction of men and women are predicted to be successful.

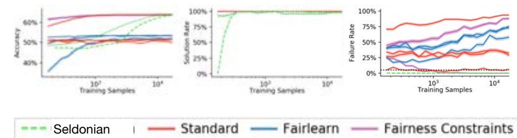
(This is called Disparate Impact.)

Disparate Impact



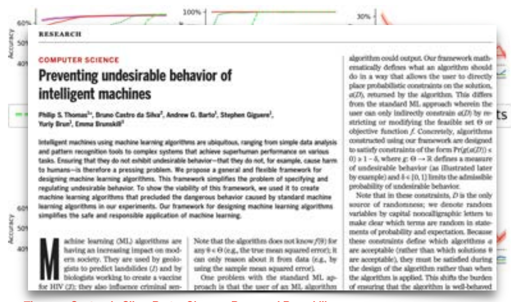
Fairlearn: Agarwal et al. A reductions approach to fair classification. ICML 2018.
Fairness Constraints: Zafar et al., Fairness Constraints: A Mechanism for Fair Classification. FATML 2015.

Disparate Impact



Fairlearn: Agarwal et al. A reductions approach to fair classification. ICML 2018.
Fairness Constraints: Zafar et al., Fairness Constraints: A Mechanism for Fair Classification. FATML 2015.

Equalized Odds



RESEARCH

COMPUTER SCIENCE

Preventing undesirable behavior of intelligent machines

Thomas, Castro da Silva, Barto, Giguere, Brun, and Brunskill.
"Preventing Undesirable Behavior of Intelligent Machines", Science 366 (6468), Nov 22, 2019

And this approach is very versatile: …works for policy selection

Offline Contextual Bandits with High Probability Fairness Guarantees

Abstract

Metevier, Giguere, Brockman, Kobren, Brun, Brunskill, Thomas. Offline Contextual Bandits with High Probability Fairness Guarantees. NeurIPS 2019.

Example scenario:

One source of ML bias comes from deploying a model on data that is fundamentally different from the data the model was trained on.

What if software is deployed on data fundamentally different from training data?

Published as a conference paper at ICLR 2022

FAIRNESS GUARANTEES UNDER DEMOGRAPHIC SHIFT

ABSTRACT

Giguere, Metevier, Brun, Castro da Silva, Thomas, and Niekum, Fairness Guarantees under Demographic Shift, ICLR 2022.

Machine learning can result in unexpected, unintended behavior.

But machine learning can be leveraged to produce verified safe and fair models, avoiding such behavior.