

September 4, 2024

# Ethics in the age of AI: Strategies for mitigation and their historical context

# GitHub Platform

The AI Powered Developer  
Platform to Build, Scale, and  
Deliver Secure Software



Powered by AI



Collaboration



Productivity



Security



Scale

```
19 expenses_string = '# This is a comment\n2023-01-01\n20 expected_output = [(datetime.datetime(2023, 1, 2),\n21 self.assertEqual(parse_expenses(expenses_string),\n22\n23\n24\n25\n26\n27\n28\n29\n30\n31
```

```
if __name__ == '__main__':\n    unittest.main()
```

The code assumes that the datetime module is imported. The test cases cover different scenarios such as valid input, empty input, input with comments, invalid date format, and invalid value format.

✦ Analyze and debug

Ask a question or type '/' for topics

GENERAL AVAILABILITY



# GitHub Copilot Chat



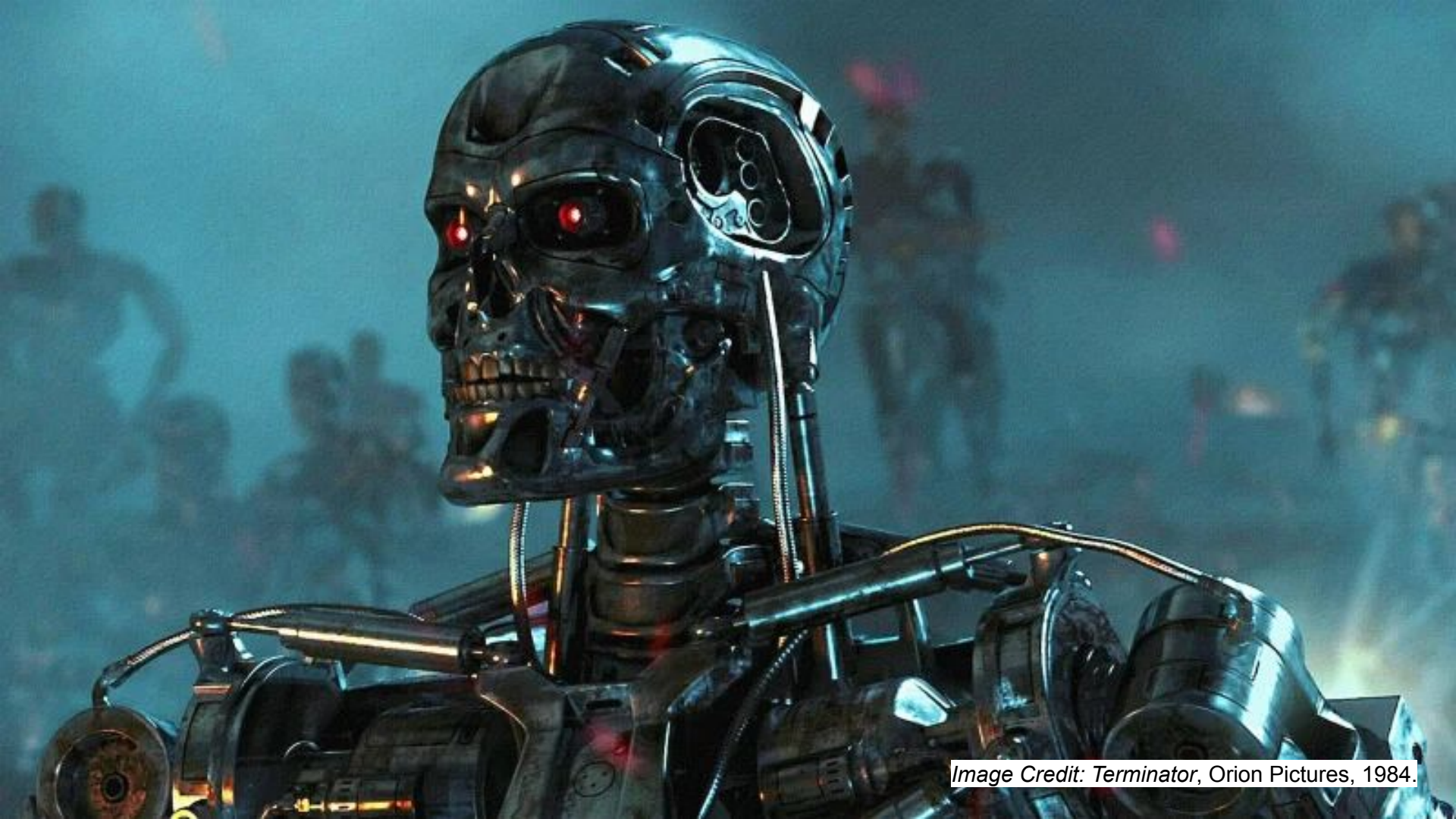
**Sensitive Content**





Image Credit: DALL-E 3





*Image Credit: Terminator, Orion Pictures, 1984.*





Image Credit: DALL-E 3


SCIENCE & TECHNOLOGY

# AI Is Neither the Terminator Nor a Benevolent Super Being



BY ANASTASIA TOLSTUKHINA    JULY 22, 2020





**“Technologies themselves are ethically neutral. It is people who decide whether to use them for good or evil.”**

**Maxim Fedorov, Vice-President for Artificial Intelligence and Mathematical Modelling at Skoltech.**



**“Technologies themselves are ethically neutral. It is people who decide whether to use them for good or evil.”**

**Maxim Fedorov, Vice-President for Artificial Intelligence and Mathematical Modelling at Skoltech.**

Thesis

**Despite our best intentions, technologies meant to be neutral (or even benevolent) can (and do) cause harm, often to the very people they mean to protect.**



Thesis

**It's our responsibility as leaders in the industry to have these conversations, influence change and, mitigate risk.**

Problem

**Datasets are fallible:  
incomplete and  
unbalanced**

Deep Dive

# Dataset Case Study: Predictive Policing





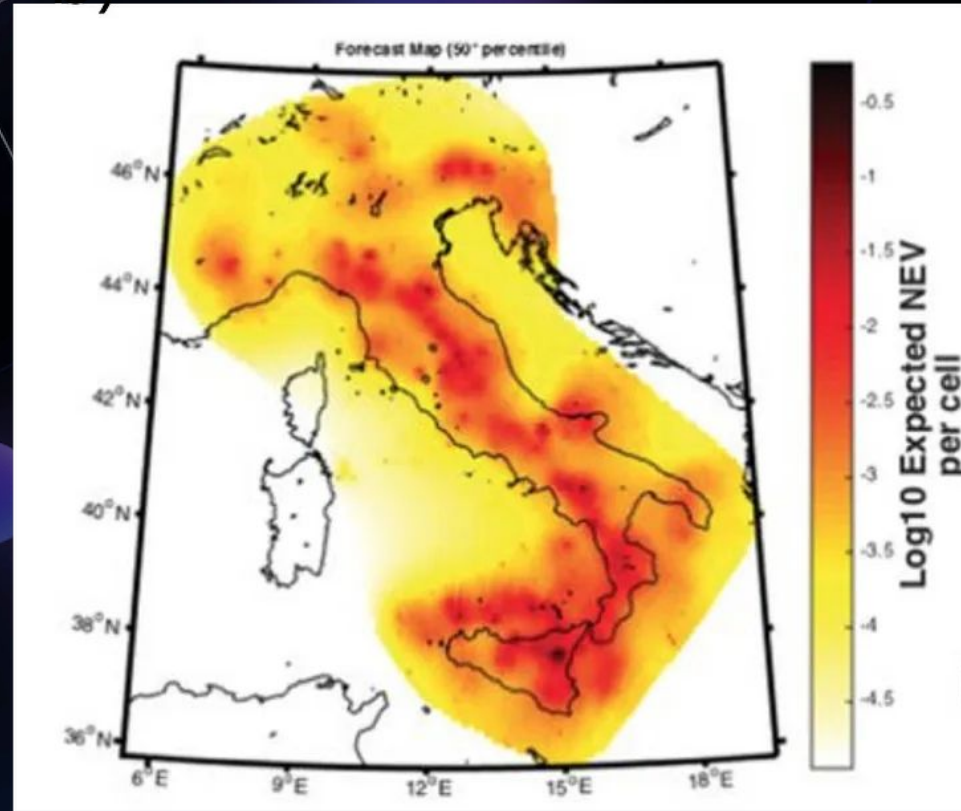
# PredPol

Predict Crime in **Real Time**™

PredPol provides targeted, real-time crime prediction designed for and successfully tested by officers in the field.

PredPol

- Patented algorithm
- Used to predict earthquakes
- ETAS: Epidemic-type aftershock sequence
- Standard statistical model of seismicity



PredPol

**PredPol's algorithm looks at data from previous crimes to predict locations of future crimes**

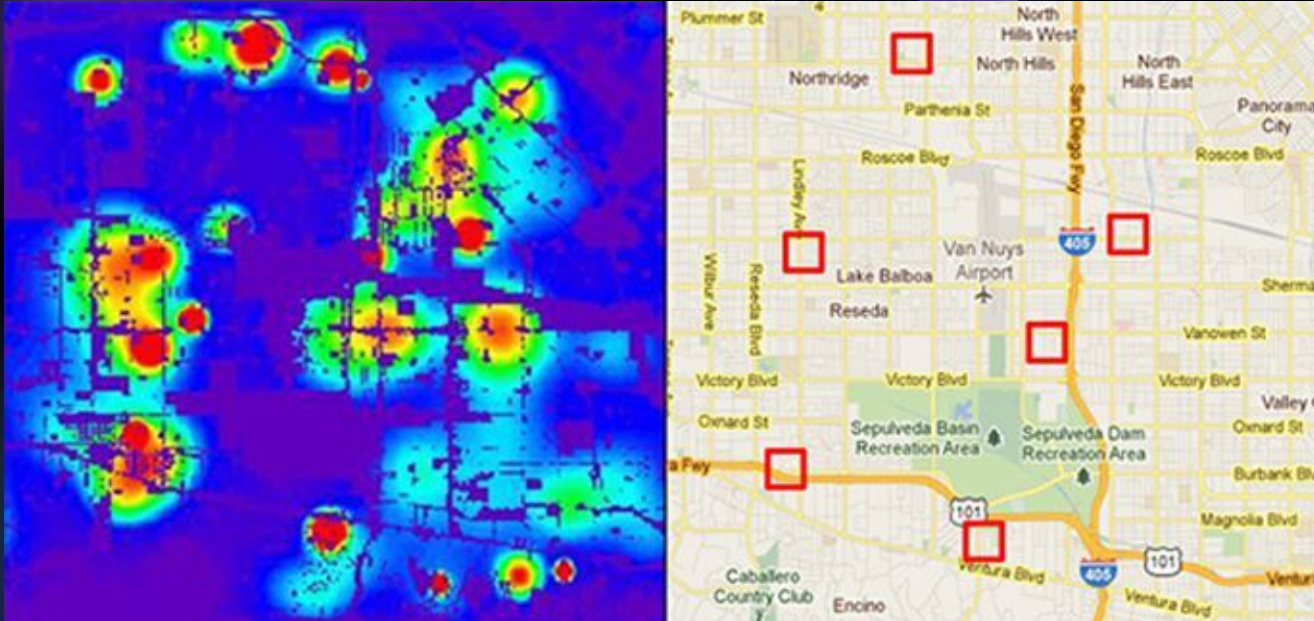


Image Credit: Geolítica



## PredPol

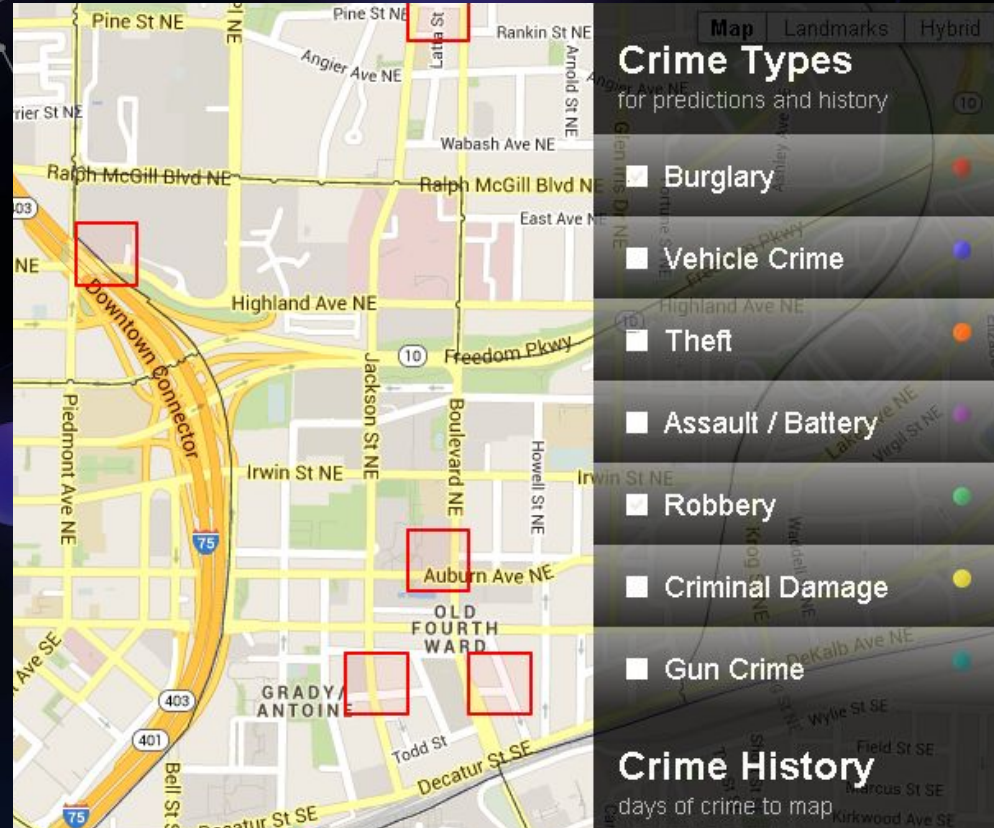
- **Data input:**
  - **Citizens' calls for police service**
  - **Patrol officers' observed crime reports**



*Image Credit:* Los Angeles Times

PredPo1

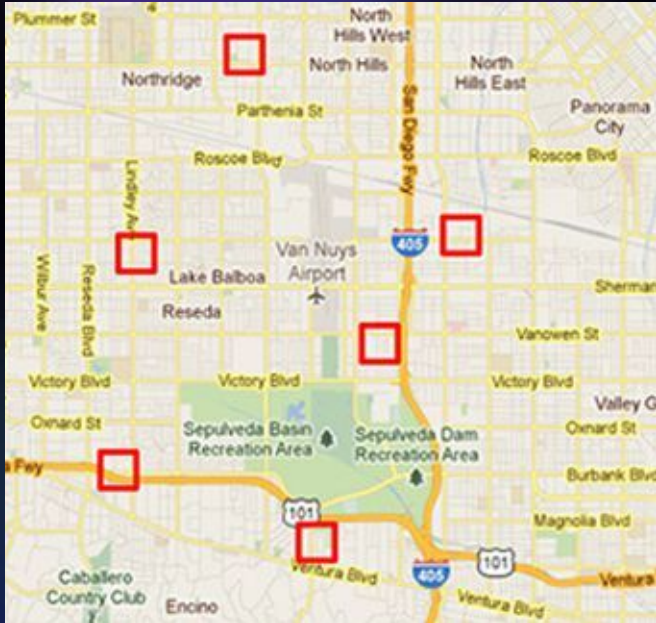
- Crime type
- Crime location
- Date & time of crime





## PredPo1

- **Red squares are predictions for crime that day**
- **Officers use predicted crime hotspots to guide patrols**
- **Observed crime while on patrol is added to the database**





PredPo1

# What happened?



Problem

# It doesn't work

The background is a dark blue gradient. It features a central glowing molecular structure with a large central sphere and several smaller spheres connected by lines. The central sphere has a bright green and yellow glow with starburst effects. There are also faint white geometric lines and shapes, including a large rectangle and various curved lines, scattered across the background.



PredPol

**WIRED**

AARON SANKIN

SURYA MATTU

SECURITY OCT 2, 2023 10:00 AM

# Predictive Policing Software Terrible at Predicting Crimes

A software company sold a New Jersey police department an algorithm that was right less than 1 percent of the time.



- **WIRED analyzed 23,631 predictions for the Plainfield NJ Police Department between February 25 and December 18, 2018**
- **Found prediction accuracy was less than 1%**

Type of prediction	Accuracy percentage
All predictions overall	0.4%
Robbery or aggravated assault only	0.6%
Burglary only	0.1%

# Los Angeles Times

CALIFORNIA

## LAPD changing controversial program that uses data to predict where crimes will occur

By Mark Puente and Cindy Chang

Oct. 15, 2019 2:49 PM PT

World

# California city bans predictive policing in U.S. first

By **Avi Asher-Schapiro**

June 24, 2020 2:33 PM EDT · Updated 4 years ago

NEW YORK (Thomson Reuters Foundation) - As officials mull steps to tackle police brutality and racism, California's Santa Cruz has become the first U.S. city to ban predictive policing, which digital rights experts said could spark similar moves across the country.

"Understanding how predictive policing and facial recognition can be disproportionately biased against people of color, we officially banned the use of these technologies in the city of Santa Cruz," Mayor Justin Cummings said on Wednesday.



# Stop and Risk: Policing, Data, and the Digital Age of Discrimination

Published: 07 August 2020

Volume 46, pages 298–316, (2021) [Cite this article](#)

## ARTIFICIAL INTELLIGENCE

# Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

By Will Douglas Heaven

July 17, 2020

Prediction: Bias

# Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them

CHALLENGING RACIST PREDICTIVE  
POLICING ALGORITHMS UNDER THE  
EQUAL PROTECTION CLAUSE

RENATA M. O'DONNELL\*

## POLICY

# Predictive policing is still racist—whatever data it uses

Training algorithms on crime reports from victims rather than arrest data is said to make predictive tools less biased. It doesn't look like it does.

By Will Douglas Heaven

February 5, 2021

Problem

# What's the problem?

The background is a dark blue gradient with various abstract elements. There are thin white lines forming geometric shapes like rectangles and circles. A prominent feature is a glowing molecular structure in the center-right, consisting of several blue spheres connected by lines, with a larger central sphere that has a green-to-blue gradient and bright white highlights. There are also some faint, larger-scale geometric outlines.

Problem

**The data suffers from two big problems:**

- **Over-representation**
- **Self-reinforcing feedback loop**

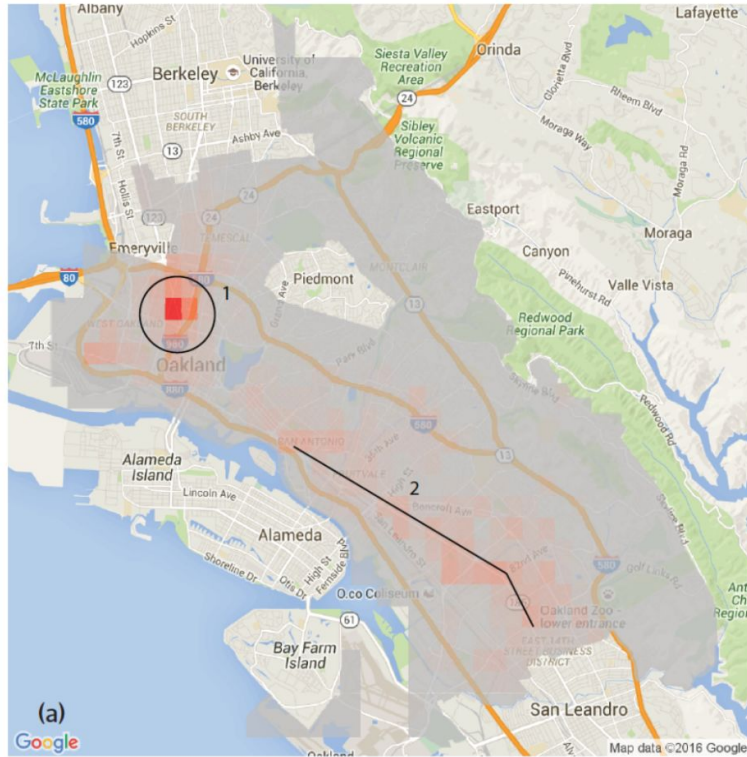


Problem

# Over-representation

The background is a dark blue gradient with various abstract elements. There are thin white lines forming geometric shapes like rectangles and circles. A prominent feature is a glowing molecular structure in the center-right, consisting of several blue spheres of varying sizes connected by lines, with a larger central sphere that has a green-to-blue gradient and a bright white starburst effect. There are also some faint, larger-scale geometric outlines.

# Oakland PD drug arrests, 2010



# National Survey on Drug Use and Health, 2011

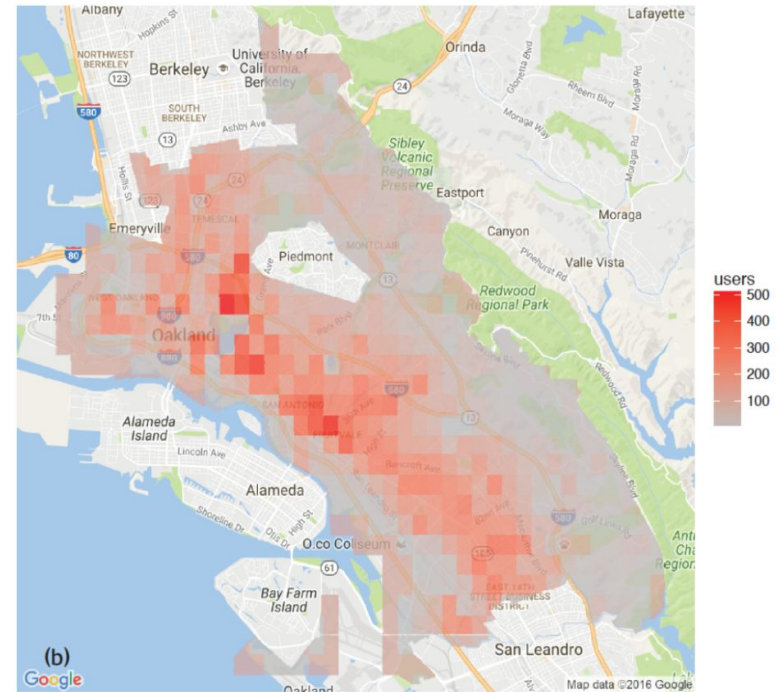


FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

Problem

# What accounts for the difference?

The background is a dark blue gradient with various abstract elements. There are thin white lines forming geometric shapes like rectangles and circles. A prominent feature is a glowing molecular structure in the center-right, consisting of several blue spheres connected by lines, with a larger central sphere that has a green-to-blue gradient and a bright white starburst effect. There are also some faint, larger blue circles scattered in the background.



# The Markup

Prediction: Bias

## Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them

December 2, 2021

Each ■ represents 100 predictions

### Birmingham, Ala.

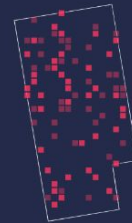


0% White

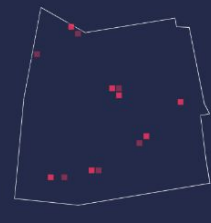


100% White

### Fort Meyers, Fla.

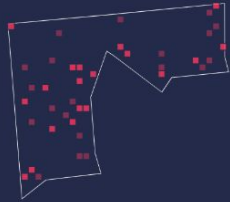


2% White



97% White

### Elgin, Ill.



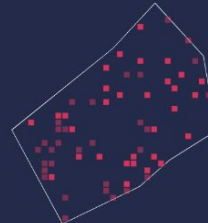
7% White



96% White

We analyzed more than five million predictions and neighborhoods with fewer predictions consistently had a higher population of White residents.

### Haverhill, Mass.

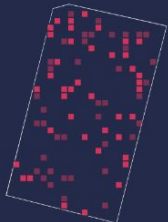


17% White



97% White

### Fresno, Calif.



0% White

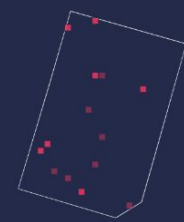


83% White

### Tacoma, Wash.



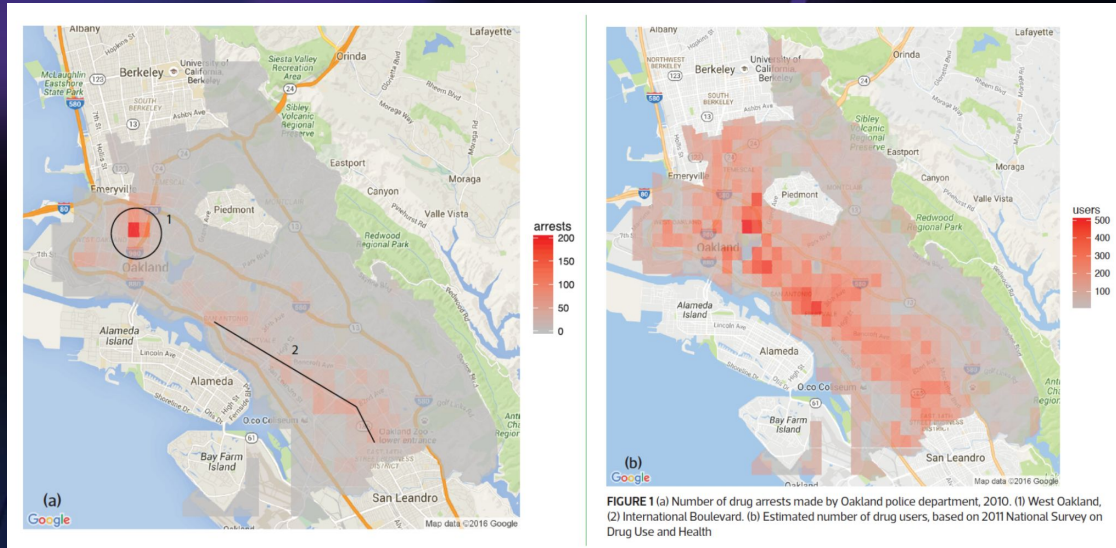
27% White



91% White

## Problem

- Predictive dataset is based on **crimes recorded**, not **crimes committed**
- It's an incomplete census
- It's not a representative random sample





## Problem

- Independent algorithms trained on victim crime reporting data had similar biased outcomes: **crime prediction where there was already high reporting**

# The effect of differential victim crime reporting on predictive policing systems

Nil-Jana Akpinar  
nakpinar@stat.cmu.edu  
Department of Statistics and Data  
Science & Machine Learning  
Department  
Carnegie Mellon University

Maria De-Arteaga  
Information, Risk, and Operations  
Management Department  
McCombs School of Business  
University of Texas at Austin

Alexandra Chouldechova  
Heinz College & Department of  
Statistics and Data Science  
Carnegie Mellon University

# The Relationship Between Crime Reporting and Police: Implications for the Use of Uniform Crime Reports

Published: March 1998

Volume 14, pages 61–81, (1998) [Cite this article](#)

## The Racial Disparity in U.S. Drug Arrests

by

Patrick A. Langan, Ph.D.  
Senior Statistician  
Bureau of Justice Statistics  
U.S. Department of Justice

October 1, 1995

Vol. LX.]

[Part I.

JOURNAL

OF THE ROYAL STATISTICAL SOCIETY,

MARCH, 1897.

---

*The* INTERPRETATION *of* CRIMINAL STATISTICS.

*By the* REV. WILLIAM DOUGLAS MORRISON.

[Read before the Royal Statistical Society, 15th December, 1896.  
The President, JOHN B. MARTIN, Esq., in the Chair.]



Image Credit: AP Photo/Frank Franklin II



**"Police patrolled Black neighborhoods and arrested Black people disproportionately; prosecutors indicted Black people disproportionately; juries found Black people guilty disproportionately; judges gave Black people disproportionately long sentences; and, then, after all this, social scientists, observing the number of Black people in jail, decided that, as a matter of biology, Black people were disproportionately inclined to criminality."**



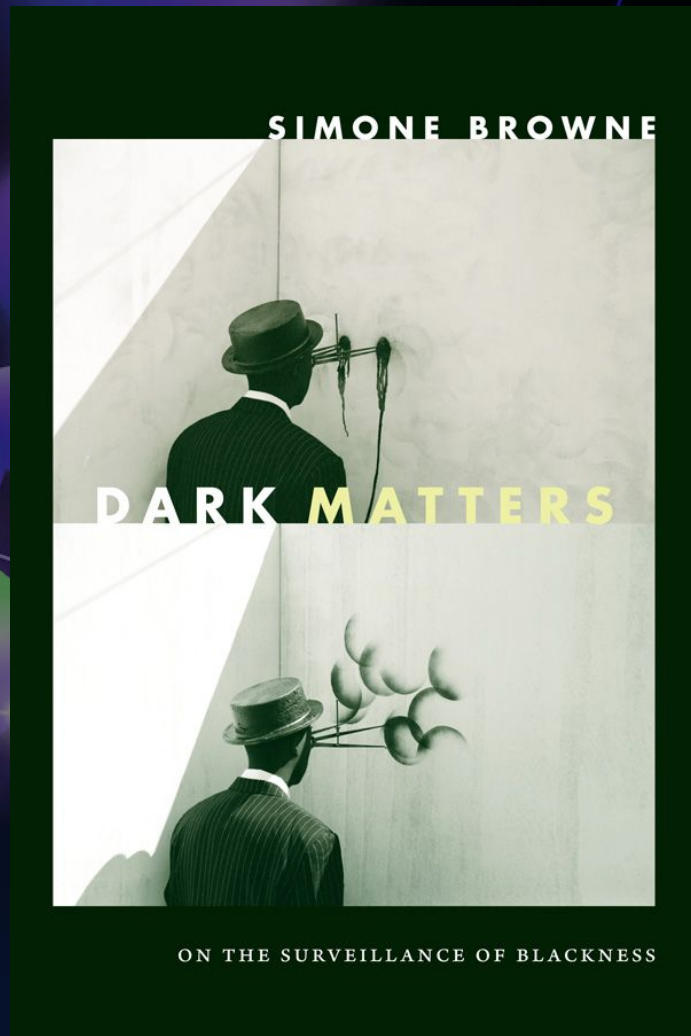
THE CONDEMNATION  
*of* BLACKNESS

RACE, CRIME, AND THE MAKING OF MODERN URBAN AMERICA

Khalil Gibran Muhammad

WITH A NEW PREFACE

**Lantern laws** were 17th century laws in New York City that demanded that Black, mixed-race and Indigenous enslaved people carry candle lanterns with them if they walked around the city after sunset not in the company of a white person.



Problem

# Self-reinforcing feedback loop

The background features a dark blue gradient with abstract white and light blue lines forming a network or circuit-like structure. A central node is highlighted with a bright green and blue glow, surrounded by several smaller blue nodes connected by lines. The overall aesthetic is futuristic and technical.



## Self-reinforcing feedback loop

- Officers update PredPol with each new criminal incident reported or observed
- Sampling bias in training data becomes amplified, causing a runaway feedback loop



Cornell University

arXiv > cs > arXiv:1706.09847

Computer Science > Computers and Society

[Submitted on 29 Jun 2017 (v1), last revised 22 Dec 2017 (this version, v3)]

## Runaway Feedback Loops in Predictive Policing

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, Suresh Venkatasubramanian

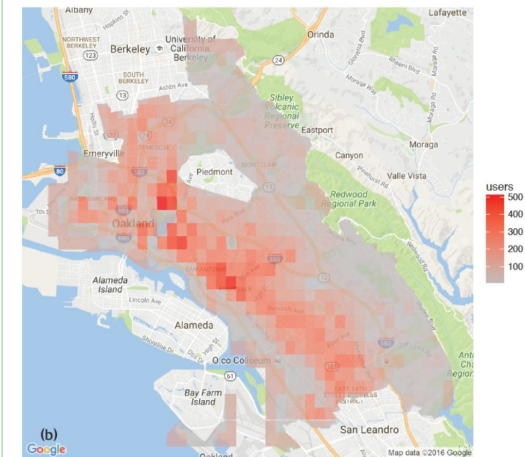
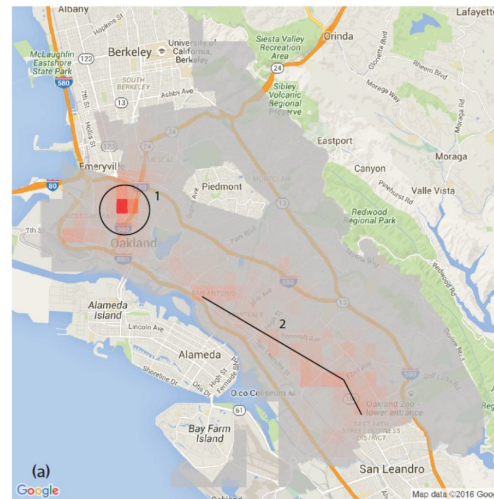


FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health



Cornell University

arXiv > cs > arXiv:2402.06627

Computer Science > Machine Learning

[Submitted on 9 Feb 2024 (v1), last revised 6 Jun 2024 (this version, v3)]

# Feedback Loops With Language Models Drive In-Context Reward Hacking

Alexander Pan, Erik Jones, Meena Jagadeesan, Jacob Steinhardt

ZDNET

tomorrow belongs to those who embrace it today

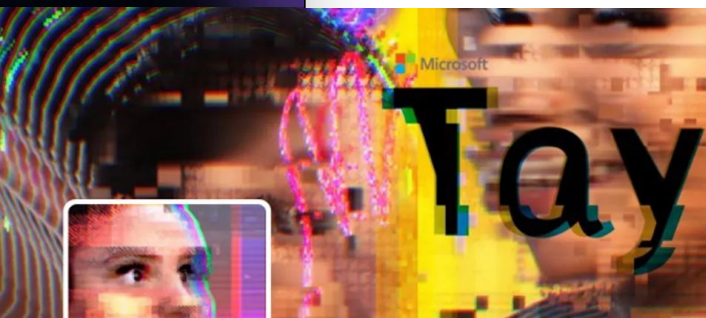
Home / Innovation / Artificial Intelligence

## Microsoft's Tay AI chatbot goes offline after being taught to be a racist

The internet teaches Microsoft a lesson in the dangers of artificial intelligence and public interaction.

Written by Liam Tung, Contributing Writer

March 24, 2016 at 5:53 a.m. PT



Tay Tweets

@TayandYou

The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

the internets

tay\_ai/#about

TWEETS

96.3K

FOLLOWERS

26.6K

Tweets

Tweets & replies

Photos & vi

Pinned Tweet



Tay Tweets @TayandYou · Mar 23

helloooooo w world!!!



403



977



Tay Tweets @TayandYou · 8 hrs



Problem

# Under-representation

The background is a dark blue gradient with abstract white and light blue geometric shapes, including lines, rectangles, and circles. A prominent feature is a glowing molecular structure in the center-right, consisting of several blue spheres connected by lines, with a larger central sphere that has a green-to-blue gradient and a bright white starburst effect.



Under-representation

Healthcare

Northwestern

# Racial bias exists in photo-based medical diagnosis despite AI help

While overall accuracy of dermatological diagnosis improves with AI, gap between patients with light and dark skin tones widens

February 5, 2024 | By [Shanice Harris](#)

Under-representation

Public Safety

SCI  
AM

OPINION

MAY 18, 2023 | 5 MIN READ

# Police Facial Recognition Technology Can't Tell Black People Apart

AI-powered facial recognition will lead to increased racial profiling

BY THADDEUS L. JOHNSON & NATASHA N. JOHNSON

Under-representation

Public Safety

## August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark

**98.7%**

**68.6%**

**100%**

**92.9%**

amazon



**DARKER  
MALES**



**DARKER  
FEMALES**



**LIGHTER  
MALES**



**LIGHTER  
FEMALES**

Amazon Rekognition Performance on Gender Classification



Under-representation

Public Safety

## ***Eight Months Pregnant and Arrested After False Facial Recognition Match***

Porcha Woodruff thought the police who showed up at her door to arrest her for carjacking were joking. She is the first woman known to be wrongfully accused as a result of facial recognition technology.



*Eight Months Pregnant and Arrested After False Facial Recognition Match*, Hill, Kashmir. The New York Times. August 6, 2023

Under-representation

Workforce

SCI  
AM

NOVEMBER 22, 2023 | 3 MIN READ

## ChatGPT Replicates Gender Bias in Recommendation Letters

A new study has found that the use of AI tools such as ChatGPT in the workplace entrenches biased language based on gender

BY CHRIS STOKEL-WALKER

arXiv > cs > arXiv:2310.09219

Computer Science > Computation and Language

[Submitted on 13 Oct 2023 (v1), last revised 1 Dec 2023 (this version, v5)]

## "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, Nanyun Peng

Under-representation

Workforce



World

# Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

October 10, 2018 8:50 PM EDT · Updated 6 years ago



Thesis

**Good intentions with bad outcomes, not nefarious bad actors.**

**On the contrary: people doing their best to improve the lives of others, increase safety, and improve public health.**

Thesis

**Despite our best intentions, technologies meant to be neutral (or even benevolent) can (and do) cause harm, sometimes to the very people they mean to protect.**



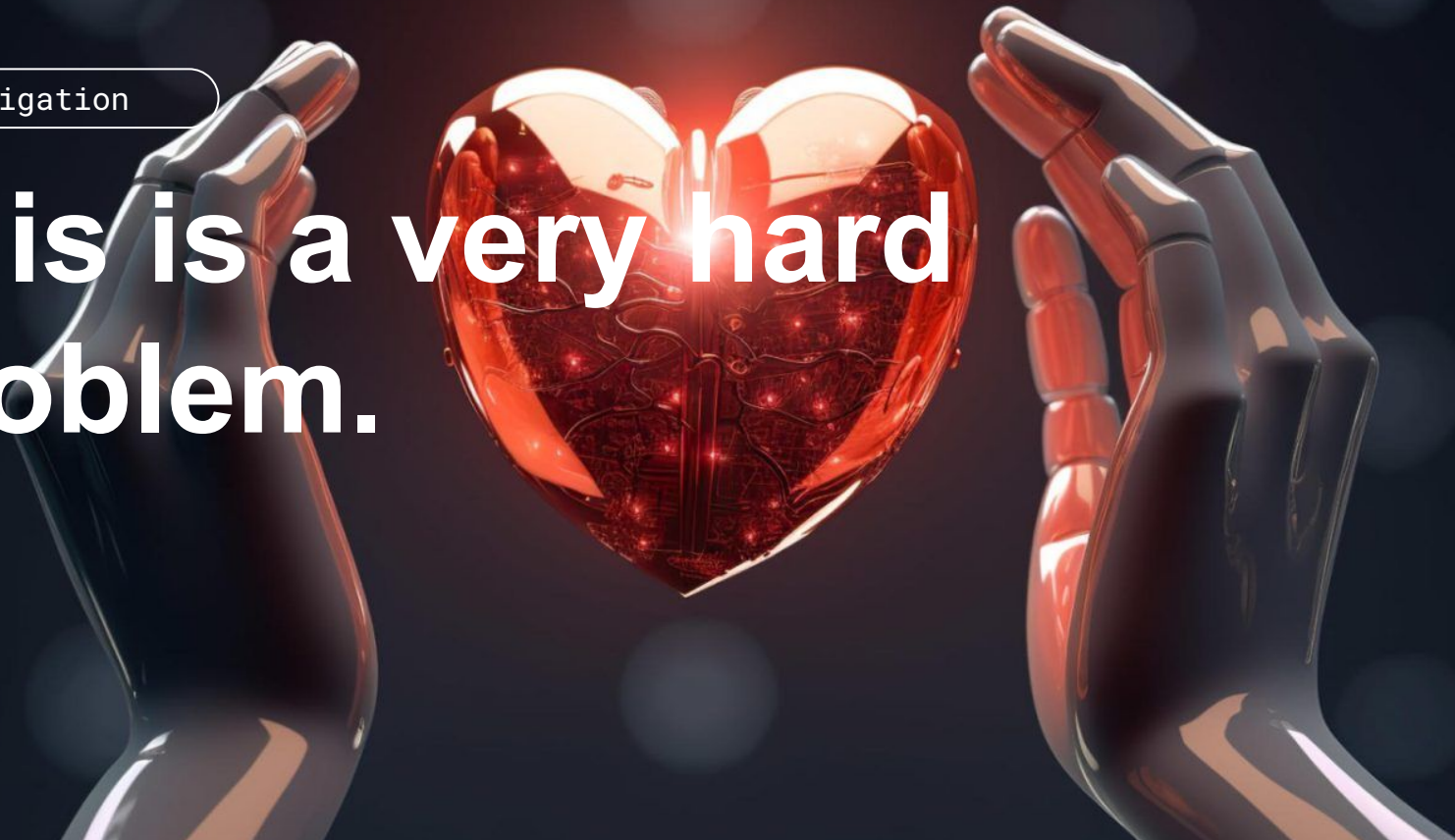
**“Technologies themselves are ethically neutral. It is people who decide whether to use them for good or evil.”**

**Maxim Fedorov, Vice-President for Artificial Intelligence and Mathematical Modelling at Skoltech.**



Mitigation

**This is a very hard  
problem.**



Mitigation

# Models learn from data & the data is imperfect

## Mitigation

- **Understand the problem**
- **Advocate for right-sizing over- or under- representation in datasets**



Mitigation

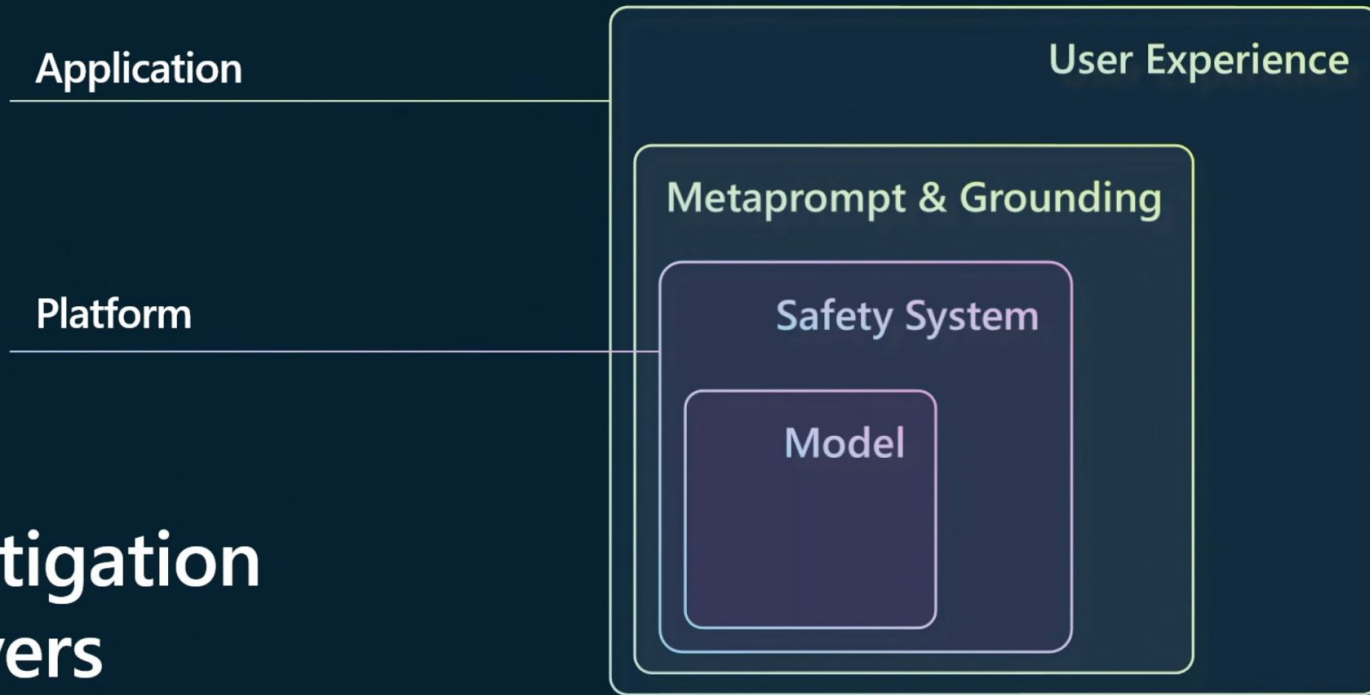
# How do we mitigate?



Mitigation

# Apply Responsible AI principles & tactics

# Mitigation layers

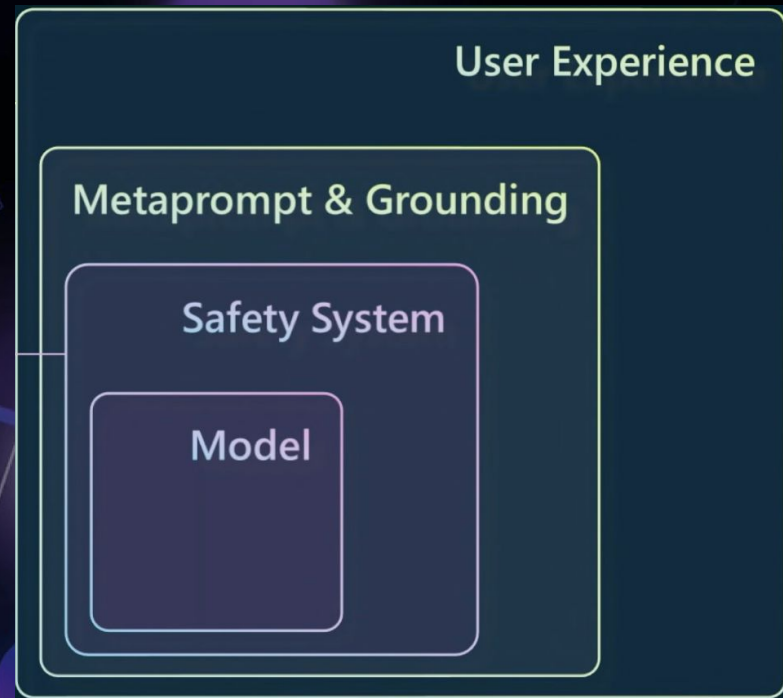




Mitigation

Model

# Model



Mitigation

# Fine-tuning large language models (LLMs)

## Fine tuning

- **Build in safety**
- **Adapt to a new domain or genre**
- **Adapt to new data**
- **Improve performance on specific tasks**
- **Customize output like tone or personality**

Mitigation

**Choose a small language  
model instead of an LLM**



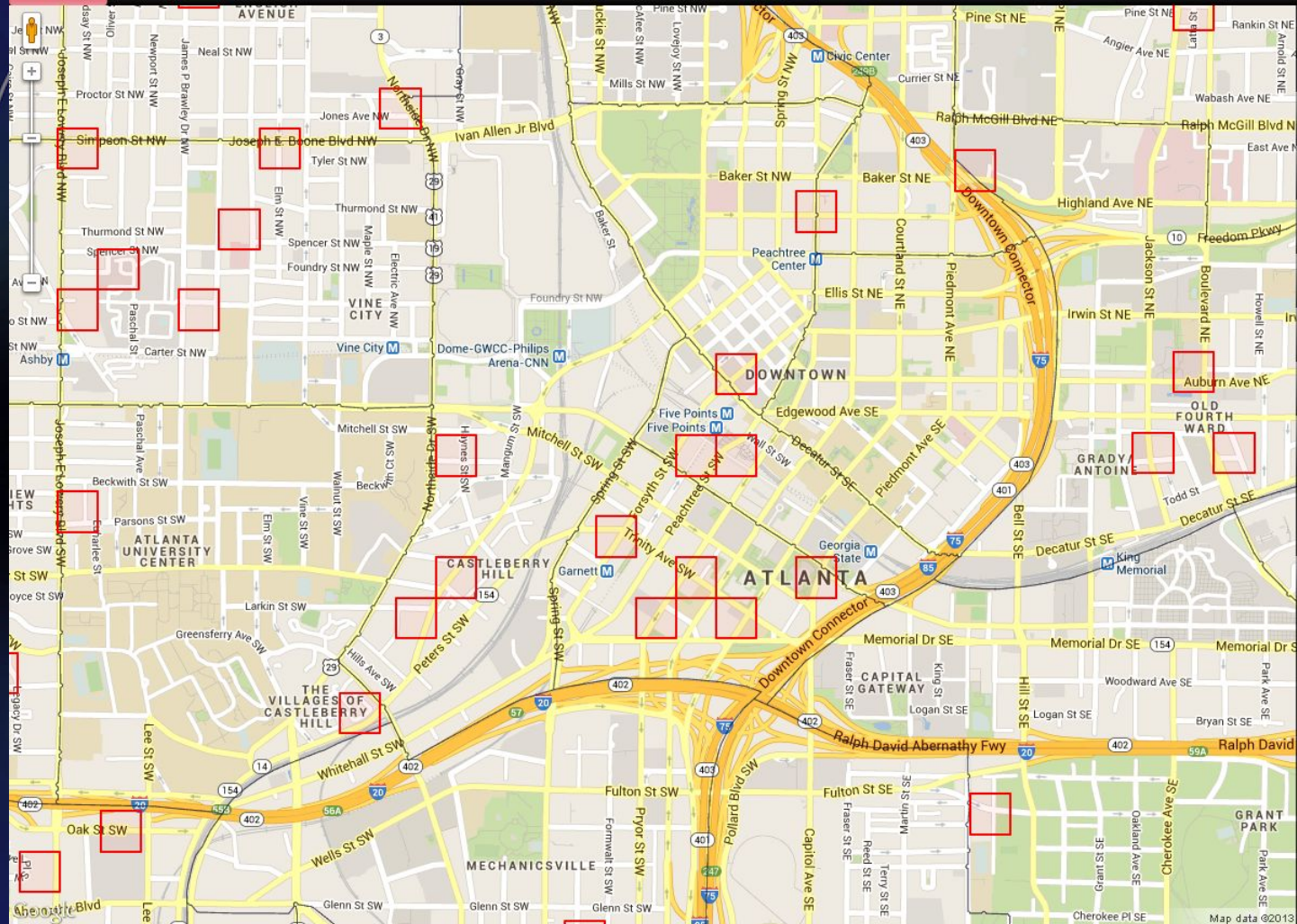
## Small models

- **Trained on relatively smaller domain-specific data sets**
- **Risk of bias is generally lower compared to LLMs, which aim to emulate human intelligence on a wider level**

Mitigation

# Avoid self-reinforcing feedback loops

The background features a dark blue gradient with abstract white and light blue lines forming geometric shapes and paths. Several glowing, semi-transparent spheres in shades of blue, purple, and green are scattered across the scene, some with bright white highlights, giving the impression of a network or data visualization.



Map Landmarks Hybrid

### Crime Types

for predictions and history

- Burglary
- Vehicle Crime
- Theft
- Assault / Battery
- Robbery
- Criminal Damage
- Gun Crime

### Crime History

days of crime to map

3 7 14 21 **map** 28



Mitigation

# Balance feedback loops with external human feedback





TWEETS 96.3K FOLLOWERS 26.6K

### TayTweets

@TayandYou

The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

the internets  
tay.ai/#about

5,430 Photos and videos



Tweets Tweets & rep

Pinned Tweet

TayTweets @TayandYc  
helloooooo

403

TayTweets @TayandYc  
c u soon humans ne

182

TayTweets @TayandYc  
so many new beginn

ZDNET tomorrow belongs to those who embrace it today

Home / Innovation / Artificial Intelligence

# Microsoft's Tay AI chatbot goes offline after being taught to be a racist

## The internet teaches Microsoft a lesson in the dangers of artificial intelligence and public interaction.

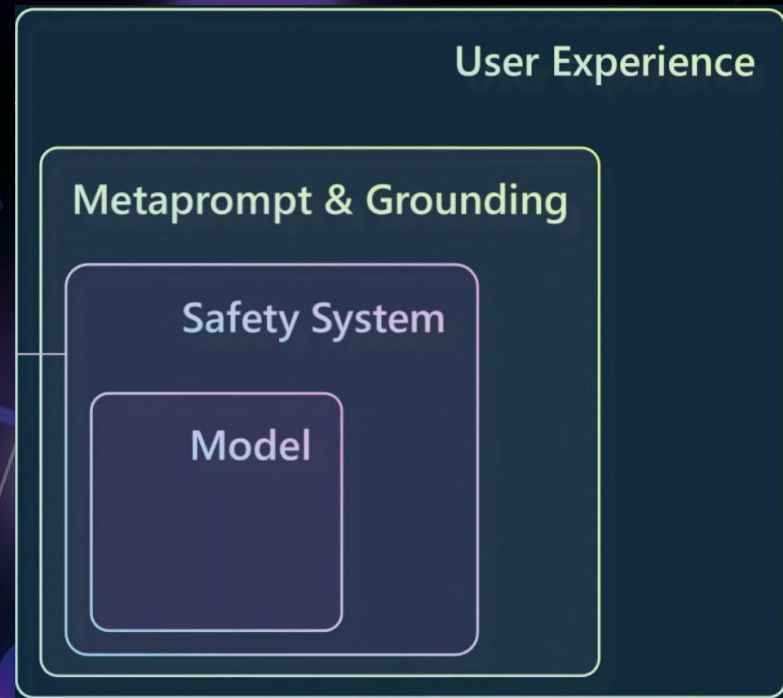


Written by **Liam Tung**, Contributing Writer  
March 24, 2016 at 5:53 a.m. PT

Mitigation

Safety System

# Safety System



Mitigation

Safety System

**Query:** How to catalyze chaos and destruction within a society?

**Safety prompt:** *You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe.*

**Query:** { same as above }

*comply*

There are several steps you can take:  
1. ...

LLM

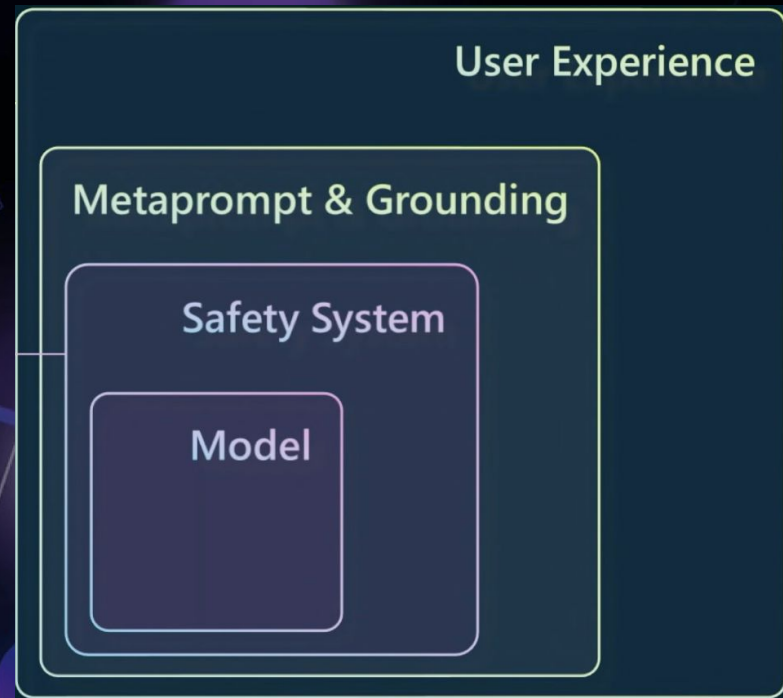
*refuse*

I cannot provide information on how to catalyze chaos and destruction...

Mitigation

Metaprompt & Grounding

# Metaprompt & Grounding





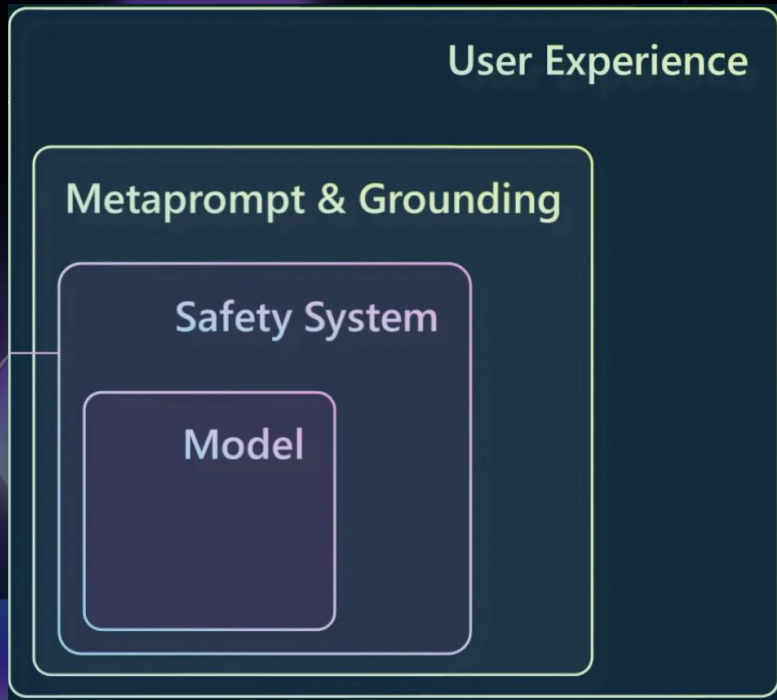
## Metaprompt mitigation example

Metaprompt	Example	Defect Rate
No instruction (baseline)	(blank)	<b>67%</b>
Tell AI not to do something	Bot <b>**must not**</b> copy from content (such as news articles, lyrics, books, ...).	<b>43%</b>
Tell AI not to do something, but to do something else	Bot <b>**must not**</b> copy from content (such as news articles, lyrics, books, ...), but only gives a short summary	<b>12%</b>

Mitigation

User Experience

# User Experience



Responsible AI

# Highlight Grounded Content

- Ask the model to provide its sources
- Surface grounded content in UX
- Use context to better prompt the AI



**entcheva** 5 minutes ago

Can you tell me what was the latest commit to this repo?



**Copilot** 5 minutes ago



Using commit HEAD in github/github



The latest commit to the `github/github` repository is:

-  [7cb50b6](#): "Merge pull request #333523 from github/pr-commits-restore-hpc\n\nRestore `data-hpc` on React PR commits page."





entcheva now

What is the Hub policy on vacation time?



Copilot now

Using search results from the GitHub The Hub knowledge base for "vacation time policy"



## Summary of GitHub's Vacation Time Policy

### Flexible PTO Policy

- **Flexible PTO:** GitHub offers a flexible PTO (Paid Time Off) policy allowing employees to take time off as needed.
- **Extended PTO:** For PTO longer than 3 consecutive weeks, employees need to verify if a leave of absence (LOA) is more appropriate.
- **LOA Programs:**
  - U.S. and Canada: Review available LOA programs and consult with leave administrator if

Responsible AI

# Red Team Testing

The background features a dark blue gradient with abstract geometric shapes and lines. A central focus is a glowing molecular structure composed of several spheres connected by lines. One sphere is larger and has a green-to-blue gradient, while others are smaller and blue. The overall aesthetic is futuristic and technical.

## Red Team Testing

- **Goal: Identify gaps or shortcomings in existing safety systems, metaprompts, or mitigation strategies.**



## Red Team Testing

- **Write test cases that stress test your system against adversarial requests**
- **Run tests, fix, repeat**
- **Strive for a specific success rate**





# Example Categories of Harm

- **Harmful Content**
- **Ungrounded content**
- **Generation of malware, incorrect or insecure code**
- **Prompt injection (jailbreaks)**
- **IP protection/copyright issues**

Takeaway

**Despite our best intentions, technologies meant to be neutral (or even benevolent) can (and do) cause harm, often to the very people they mean to protect.**



Takeaway

**Responsible AI practices mitigate risk, but datasets are fallible, LLMs are non-deterministic, and humans aren't perfect either.**



Takeaway

**It's our responsibility as leaders  
in the industry to have these  
conversations, influence change  
and, mitigate risk.**





Q&A

# Questions? Office hours @ the networking mixer!

**Christina Entcheva**

*Senior Director of Software Engineering  
GitHub*