July 19, 2024

Ethics in technology:
History, practice, and mitigation

GitHub Platform

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

Powered by AI

Collaboration

Productivity

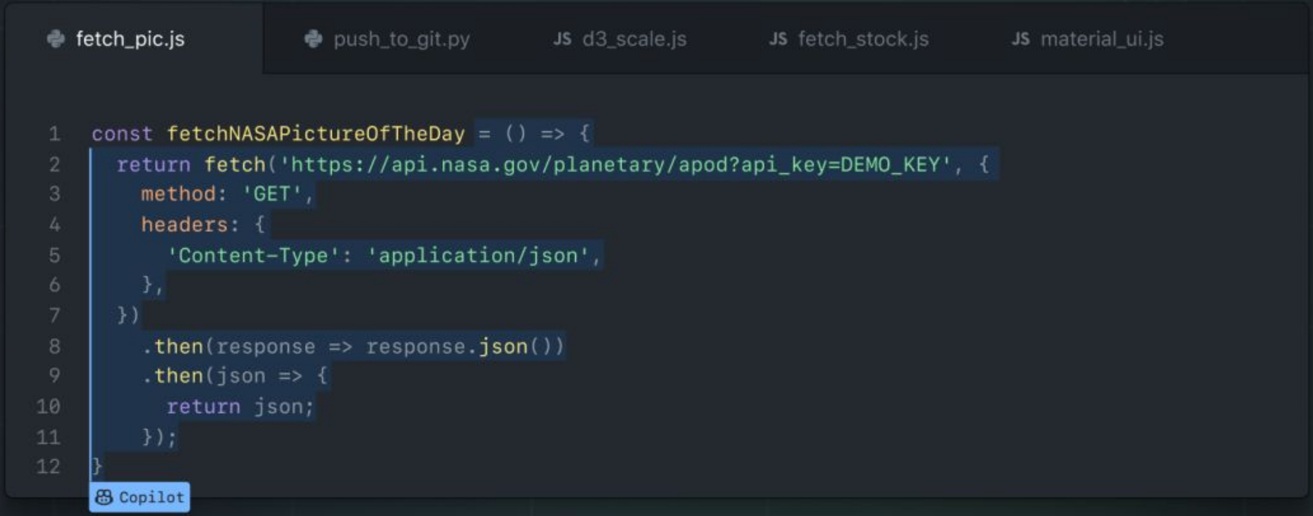
Security

Scale

Integrations & APIs

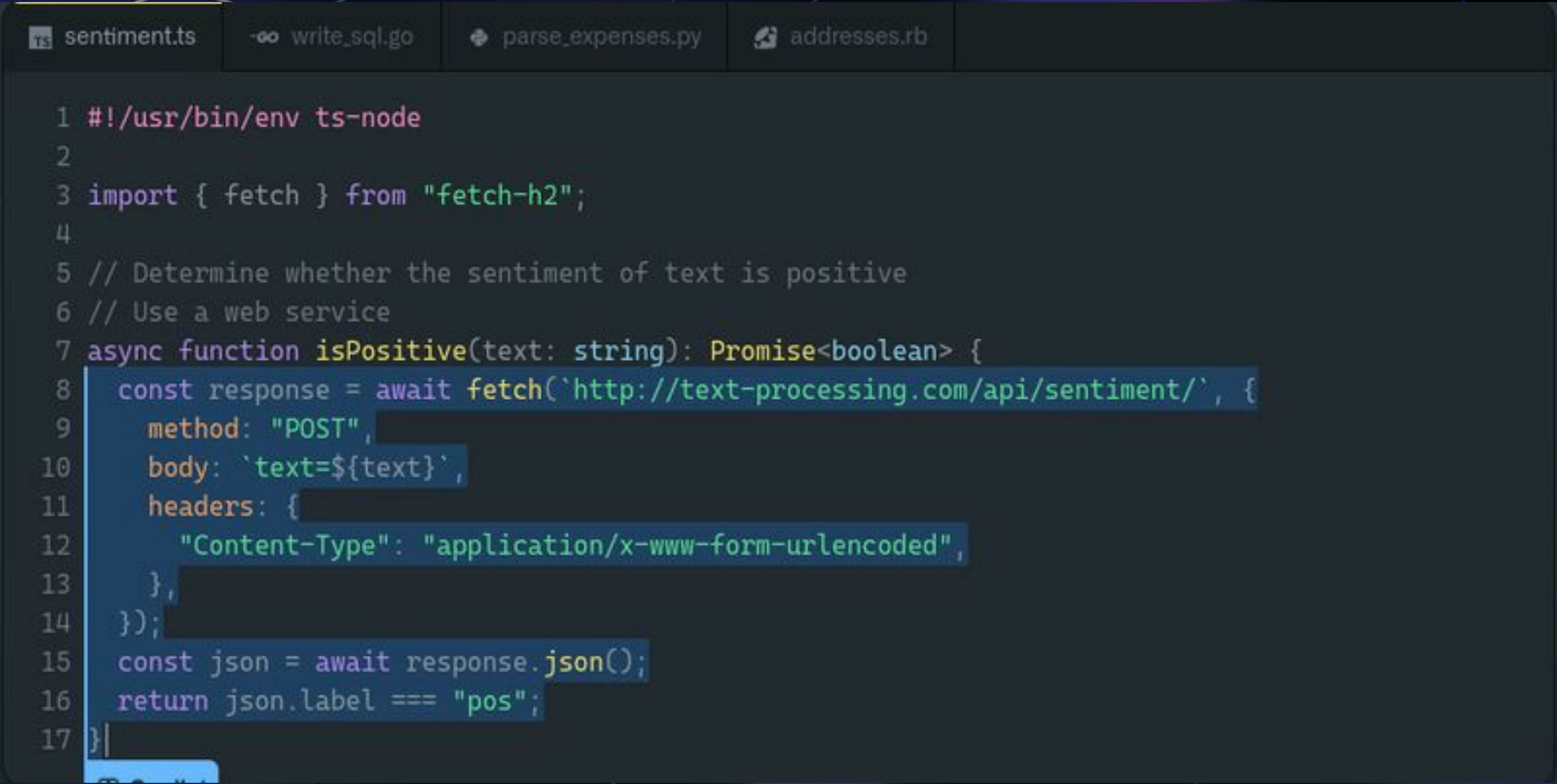
Technical preview

Your AI pair programmer



```
const fetchNASAPictureOfTheDay = () => {
  return fetch('https://api.nasa.gov/planetary/apod?api_key=DEMO_KEY', {
    method: 'GET',
    headers: {
      'Content-Type': 'application/json',
    },
  })
    .then(response => response.json())
    .then(json => {
      return json;
    });
}
```

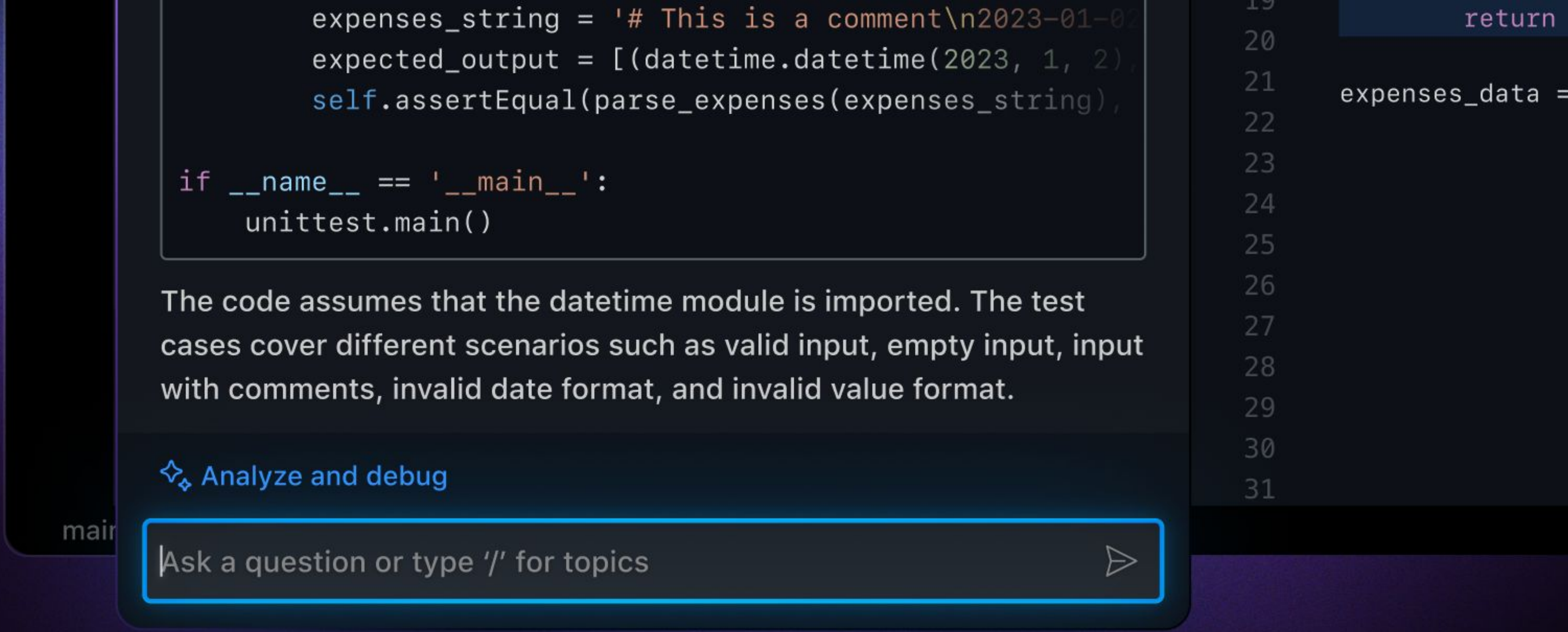
GitHub Copilot



```typescript
#!/usr/bin/env ts-node

import { fetch } from "fetch-h2";

// Determine whether the sentiment of text is positive
// Use a web service
async function isPositive(text: string): Promise<boolean> {
  const response = await fetch(`http://text-processing.com/api/sentiment/`, {
    method: "POST",
    body: `text=${text}`,
    headers: {
      "Content-Type": "application/x-www-form-urlencoded",
    },
  });
  const json = await response.json();
  return json.label === "pos";
}
```



```
        expenses_string = '# This is a comment\n2023-01-0
        expected_output = [(datetime.datetime(2023, 1, 2),
        self.assertEqual(parse_expenses(expenses_string),

if __name__ == '__main__':
    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.

```
        return

    expenses_data =
```

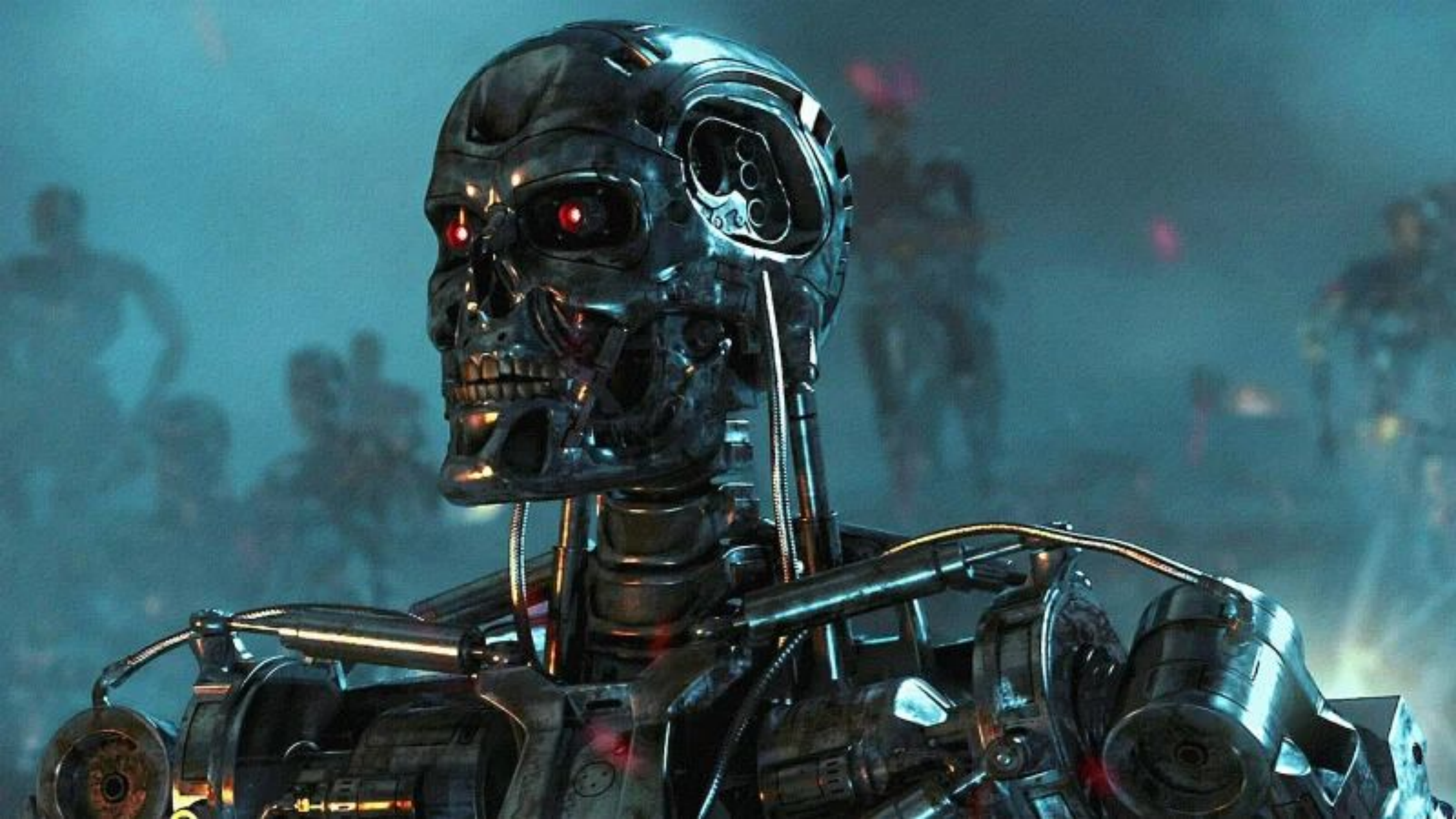
GENERAL AVAILABILITY

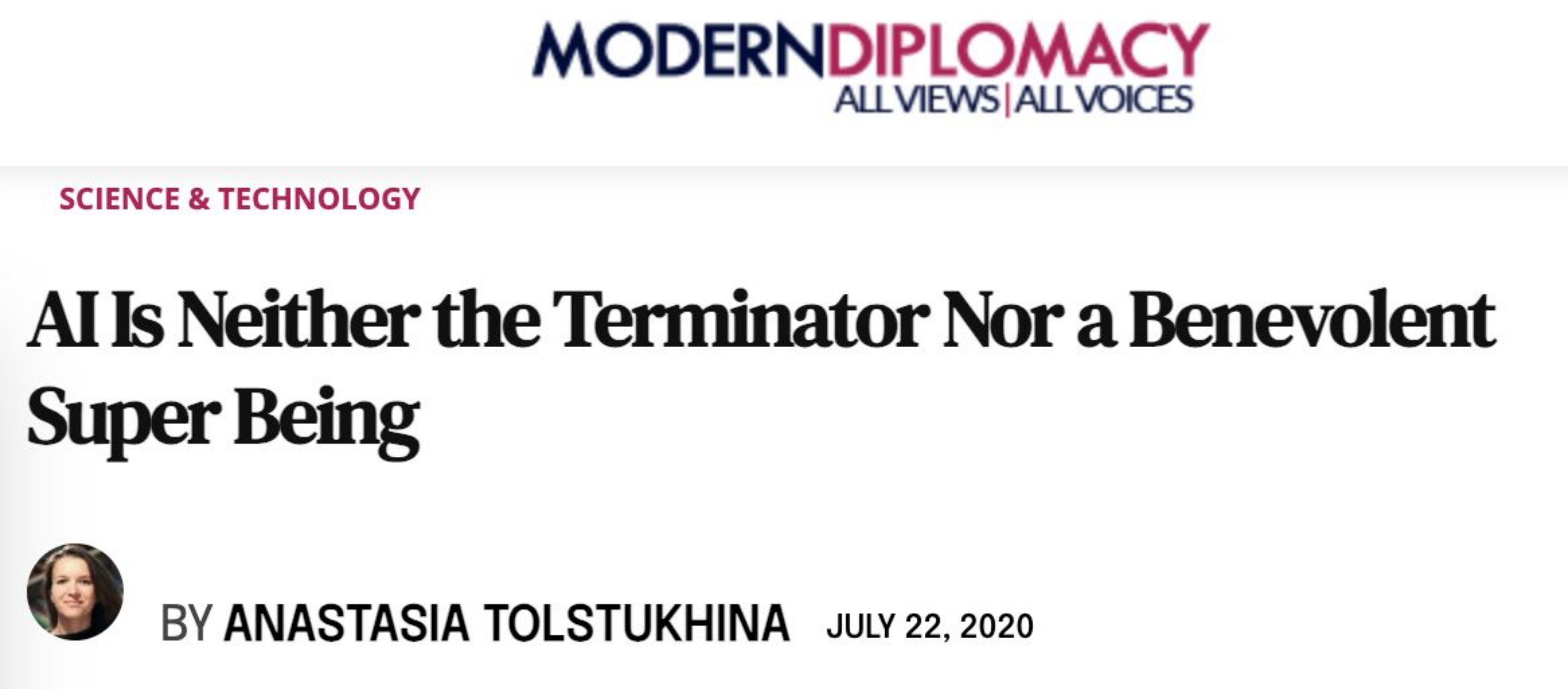
GitHub Copilot Chat

GENERAL AVAILABILITY

Copilot Enterprise

Sensitive Content





MODERNDIPLOMACY
ALL VIEWS | ALL VOICES

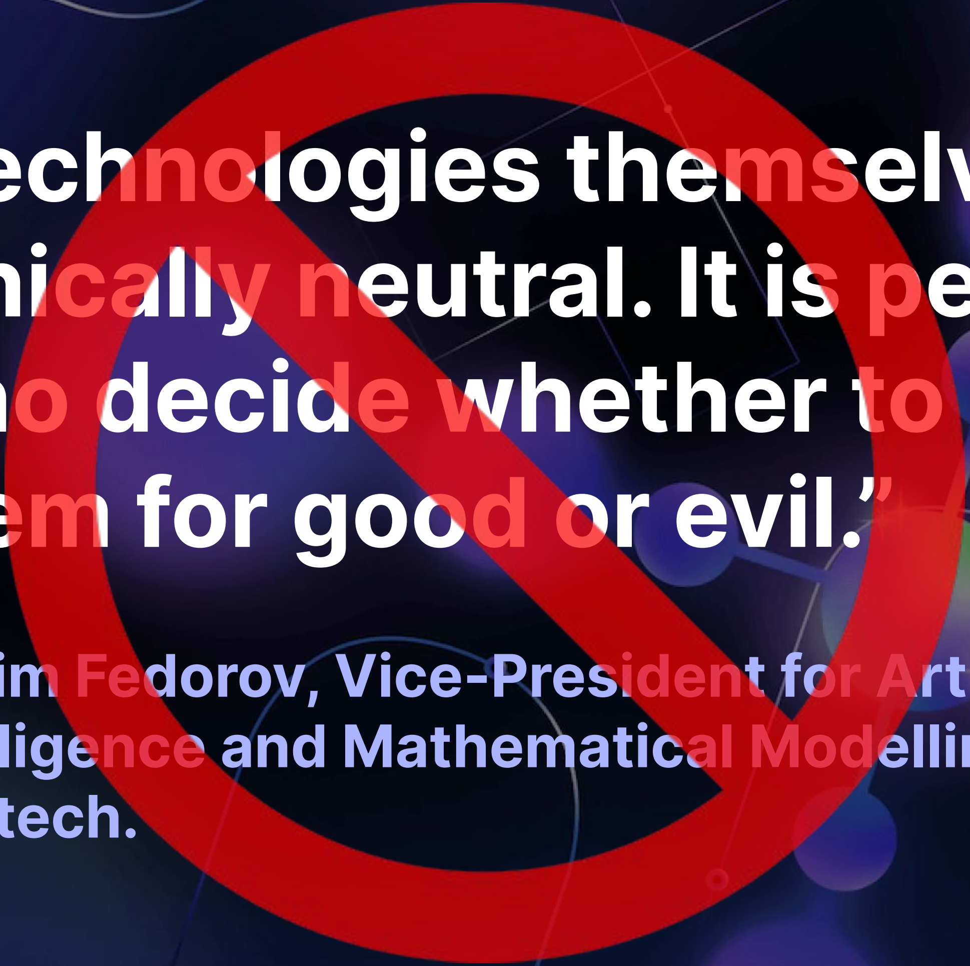
SCIENCE & TECHNOLOGY

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.

Takeaway

It’s our responsibility as leaders in the industry to influence change and to mitigate risk so that AI can live up to its full potential.

Problem

Datasets are infallible: incomplete and unbalanced

Deep Dive

Dataset Case Study: Predictive Policing

PredPol

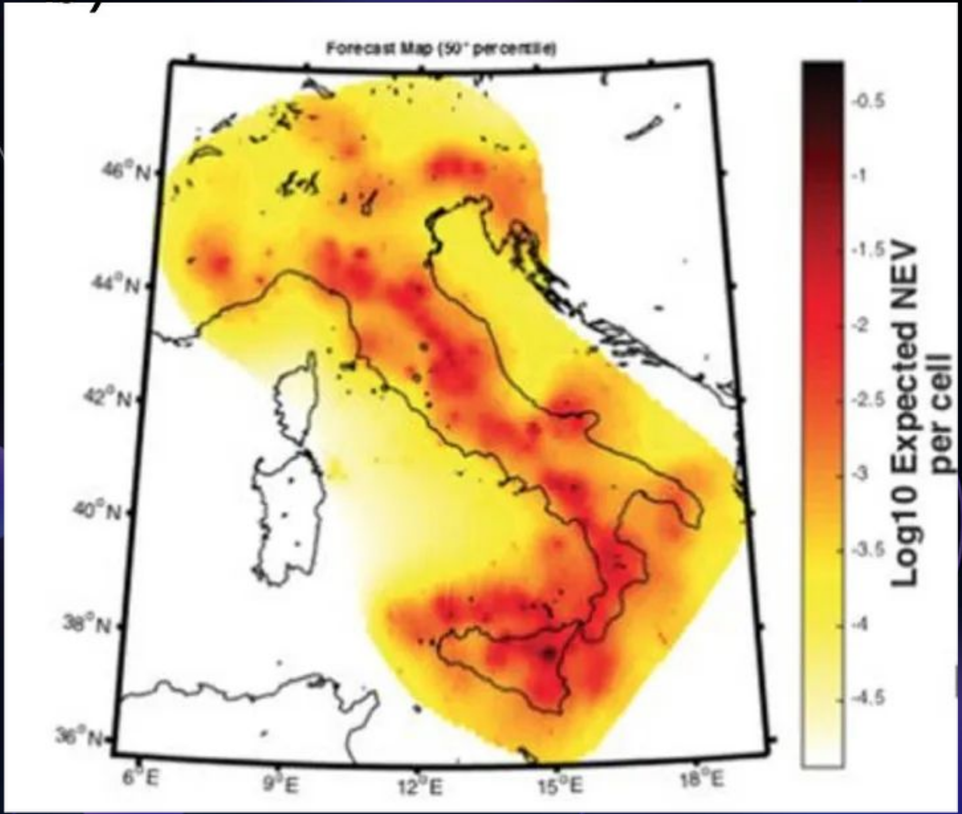
Predict Crime in **Real Time**™

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

PredPol

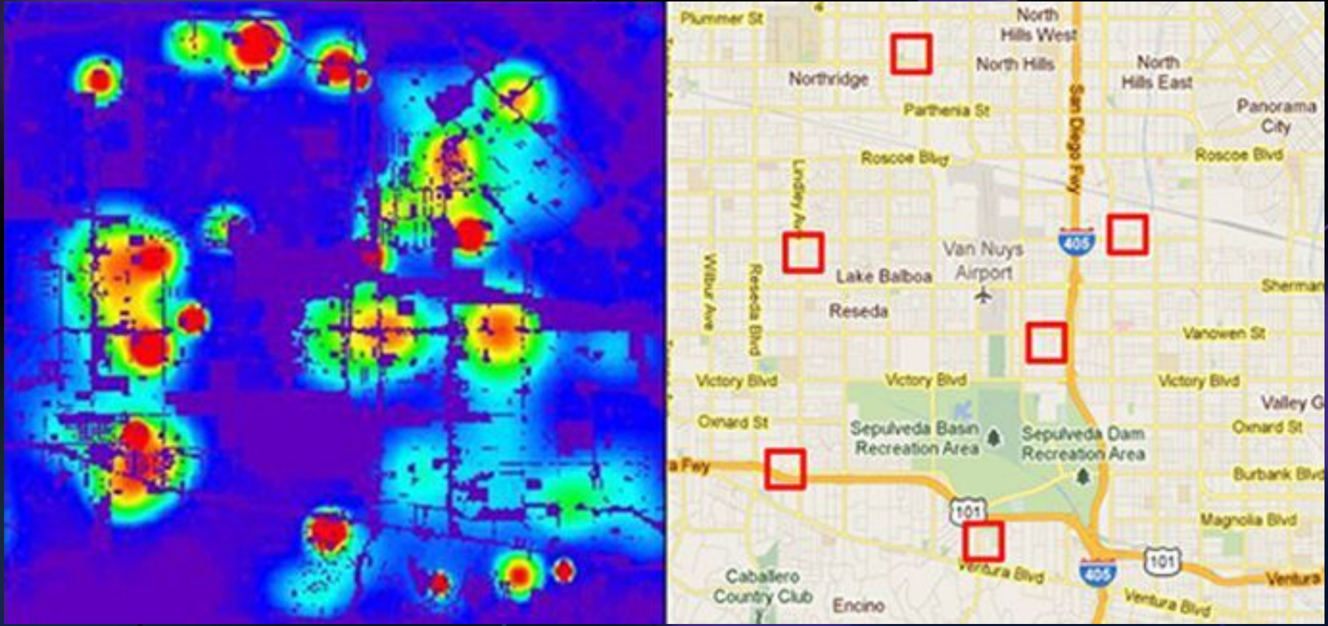
- Epidemic-type aftershock sequence (ETAS) model
- Used to predict earthquakes
- Standard statistical model of seismicity

b)



PredPol

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

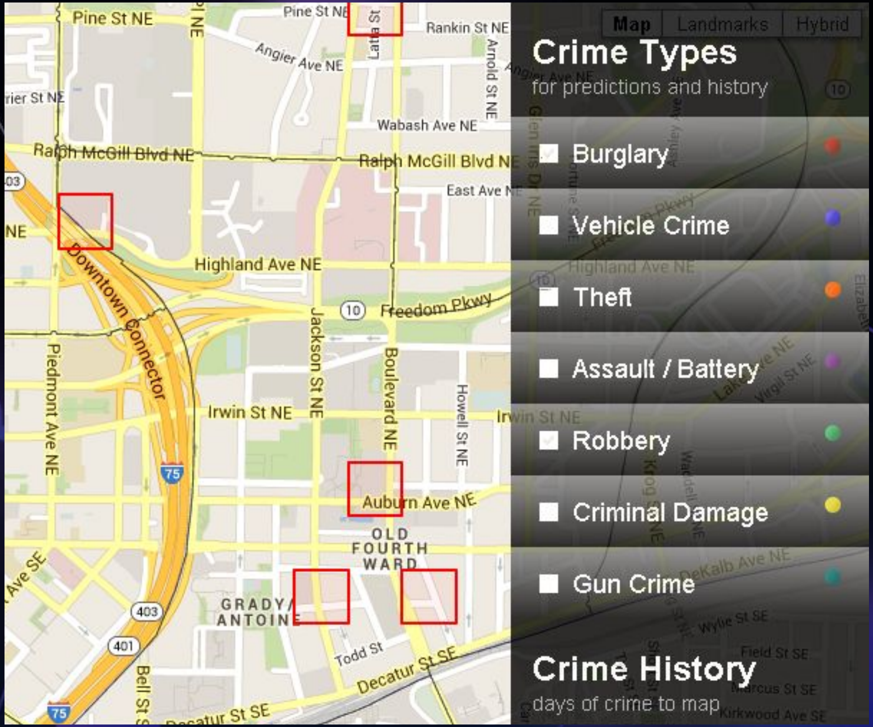


PredPol

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

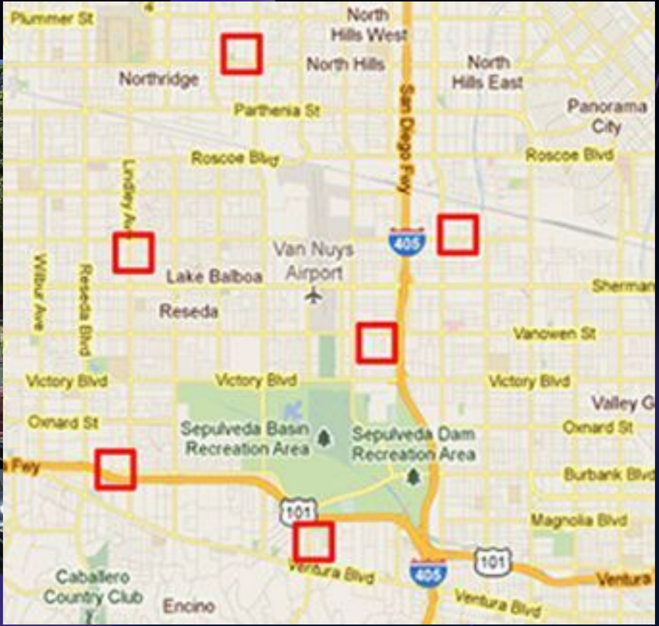
PredPol

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



PredPol

- 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



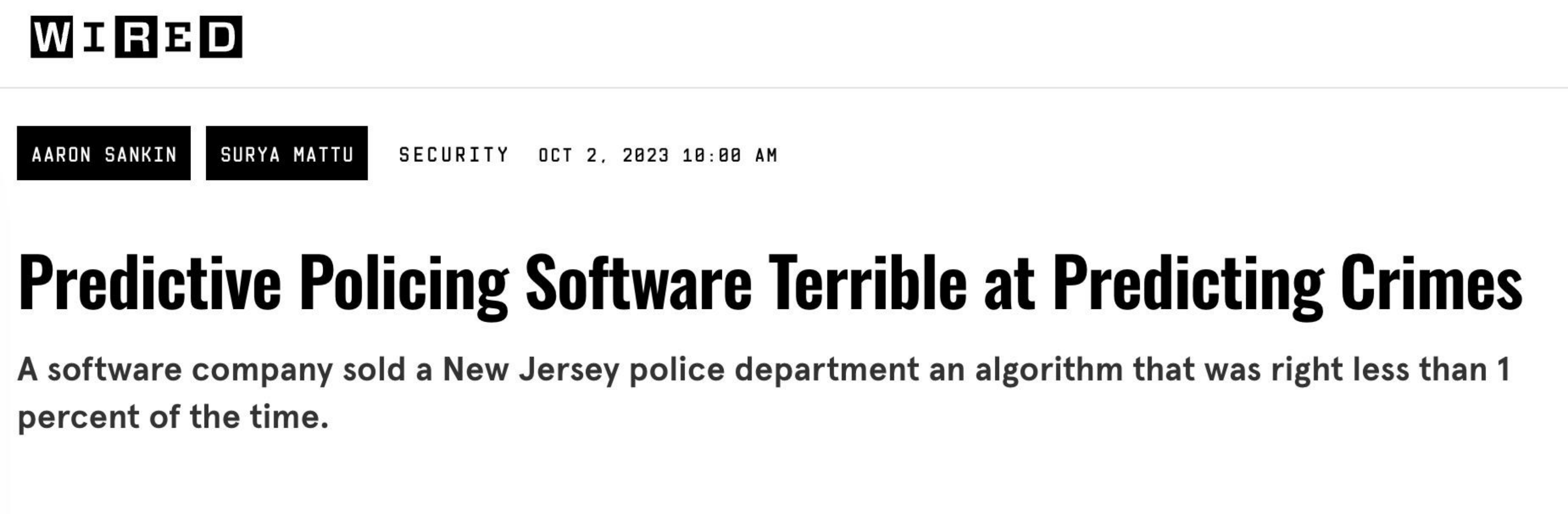
PredPol

What happened?

Problem

It doesn’t work

PredPol



WIRED

AARON SANKIN **SURYA MATTU** SECURITY OCT 2, 2023 10:00 AM

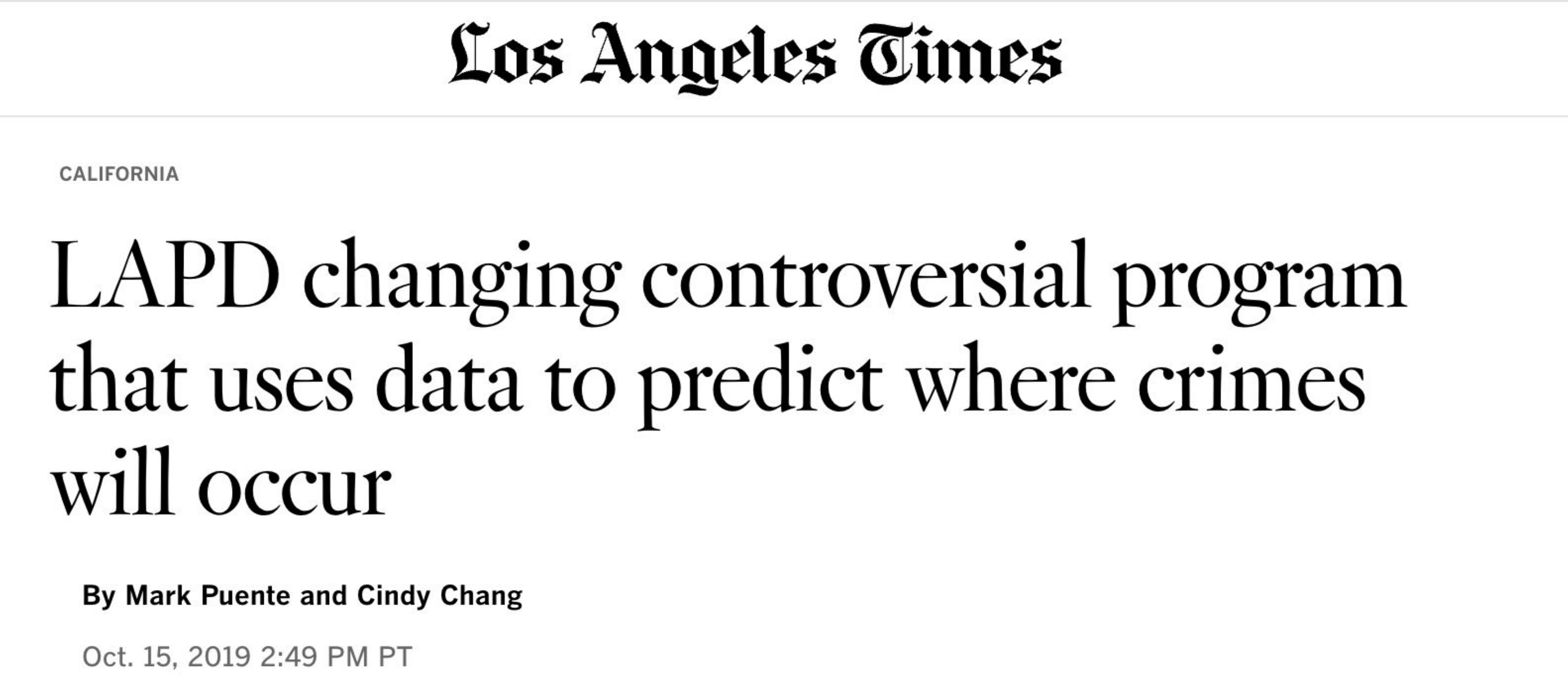
Predictive Policing Software Terrible at Predicting Crimes

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

PredPol

- **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%



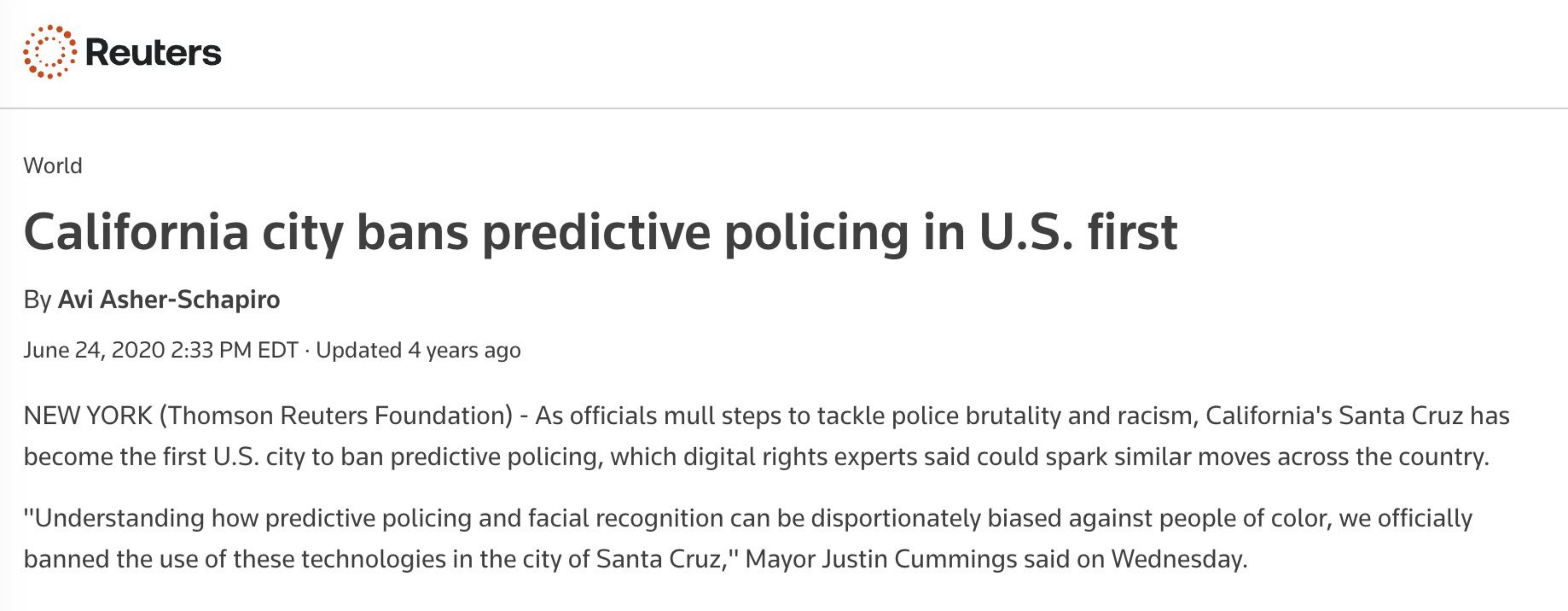
Los Angeles Times

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



Reuters

World

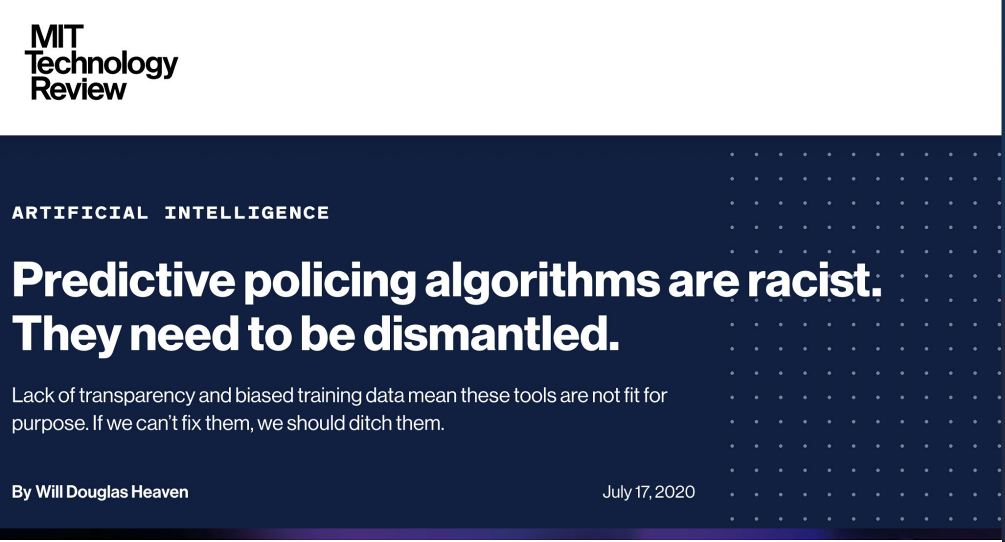
California city bans predictive policing in U.S. first

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 disportionately 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.



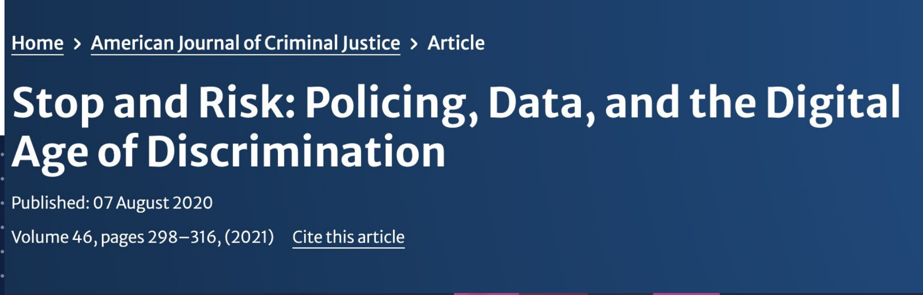
MIT Technology Review

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

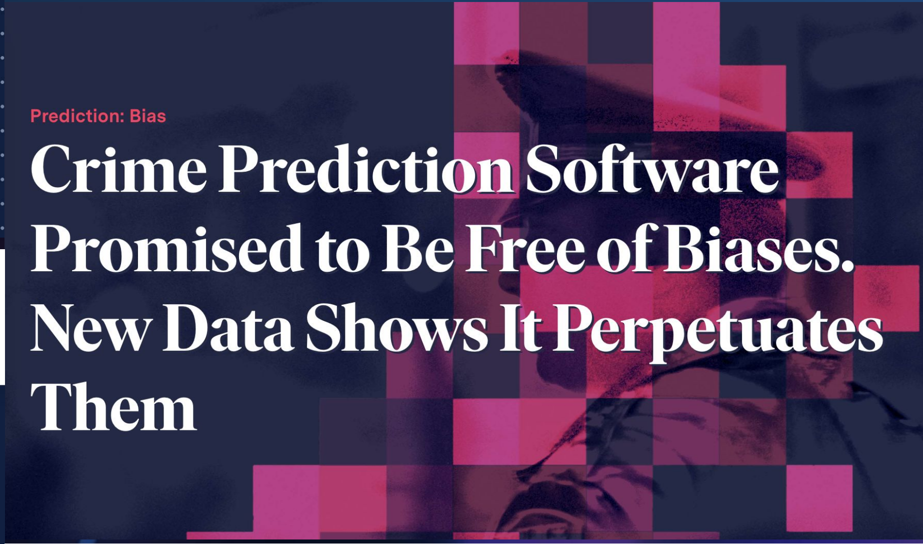


Home › American Journal of Criminal Justice › Article

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

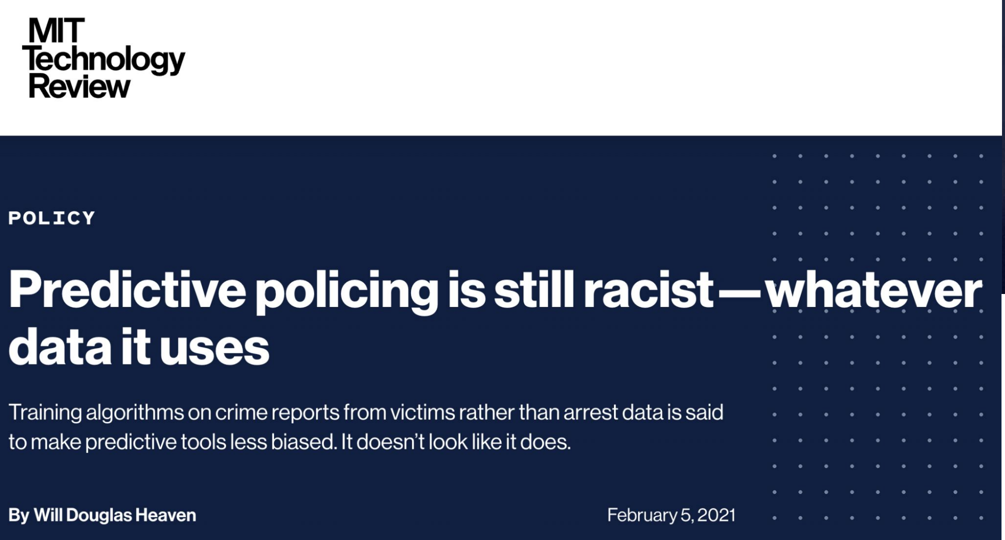
Published: 07 August 2020

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



Prediction: Bias

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



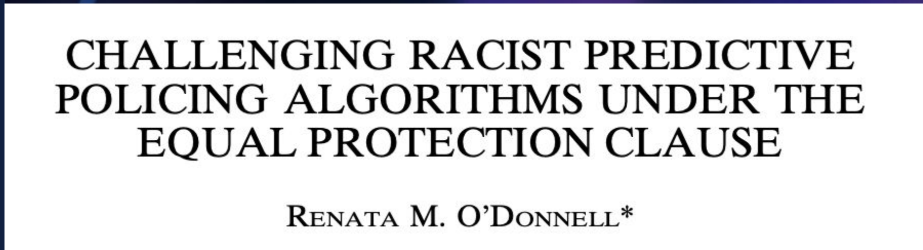
MIT Technology Review

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

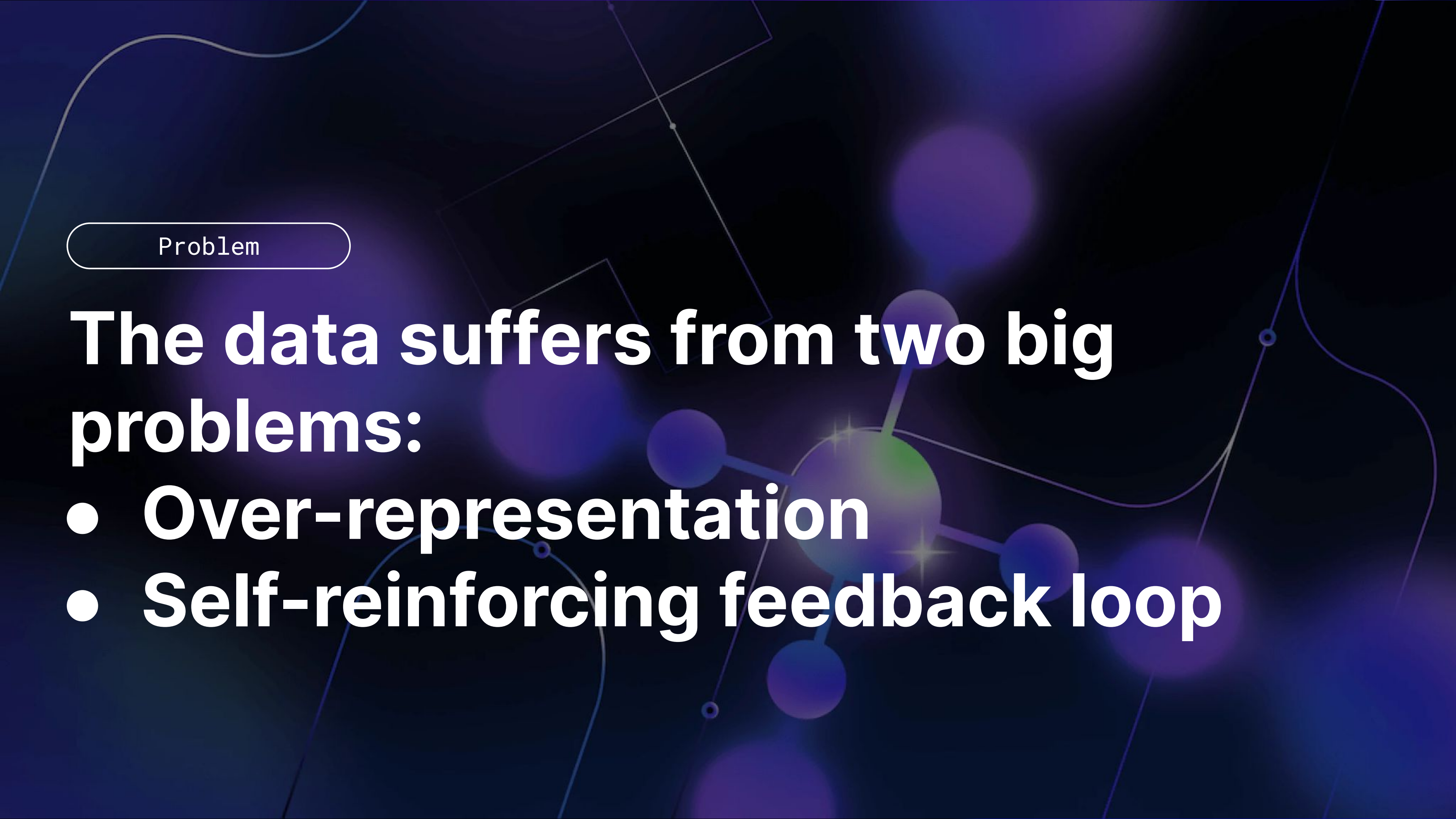


CHALLENGING RACIST PREDICTIVE POLICING ALGORITHMS UNDER THE EQUAL PROTECTION CLAUSE

RENATA M. O'DONNELL*

Problem

What’s the problem?



Problem

The data suffers from two big problems:

- Over-representation
- Self-reinforcing feedback loop

Problem

Over-representation

Oakland PD drug arrests, 2010

National Survey on Drug Use and Health, 2011

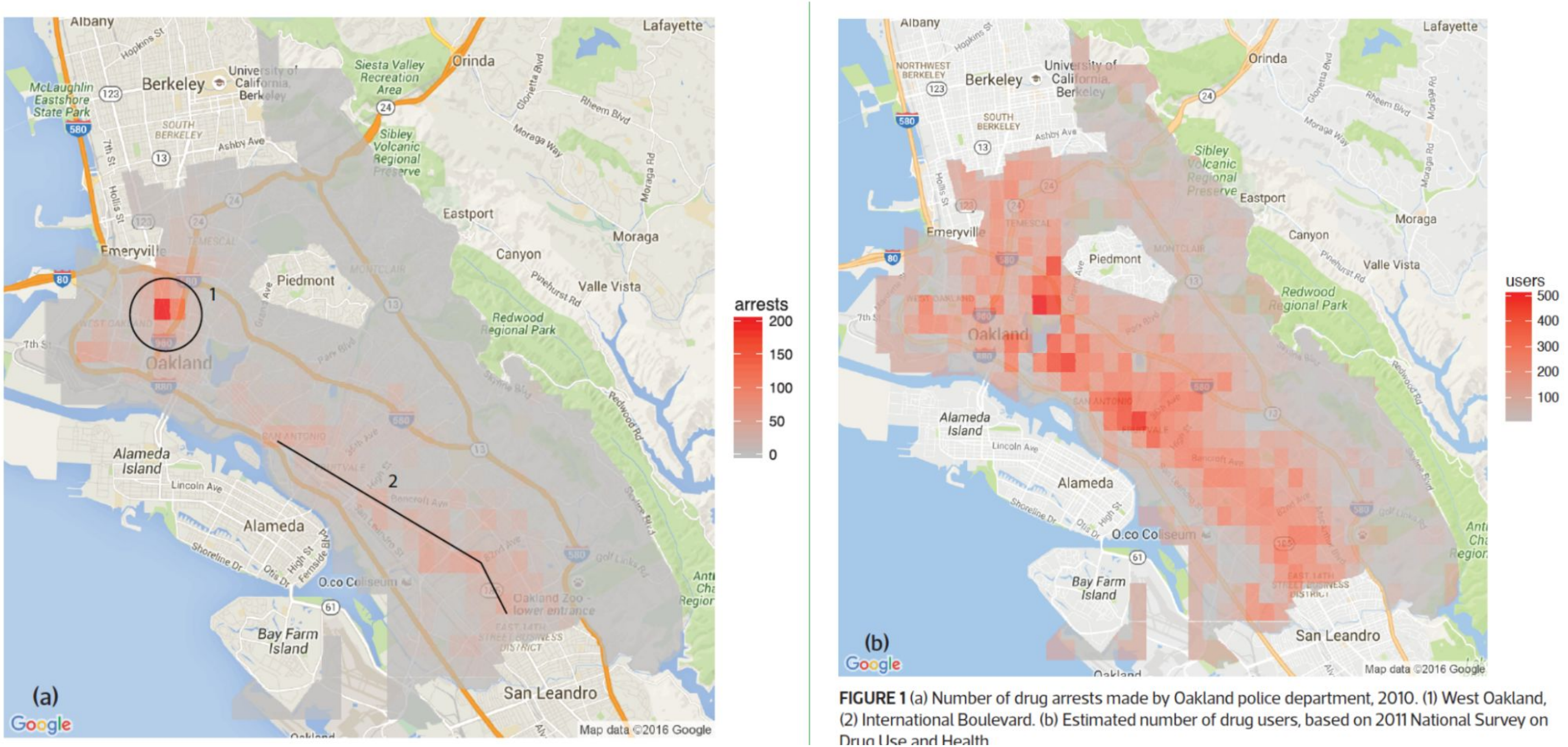


FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

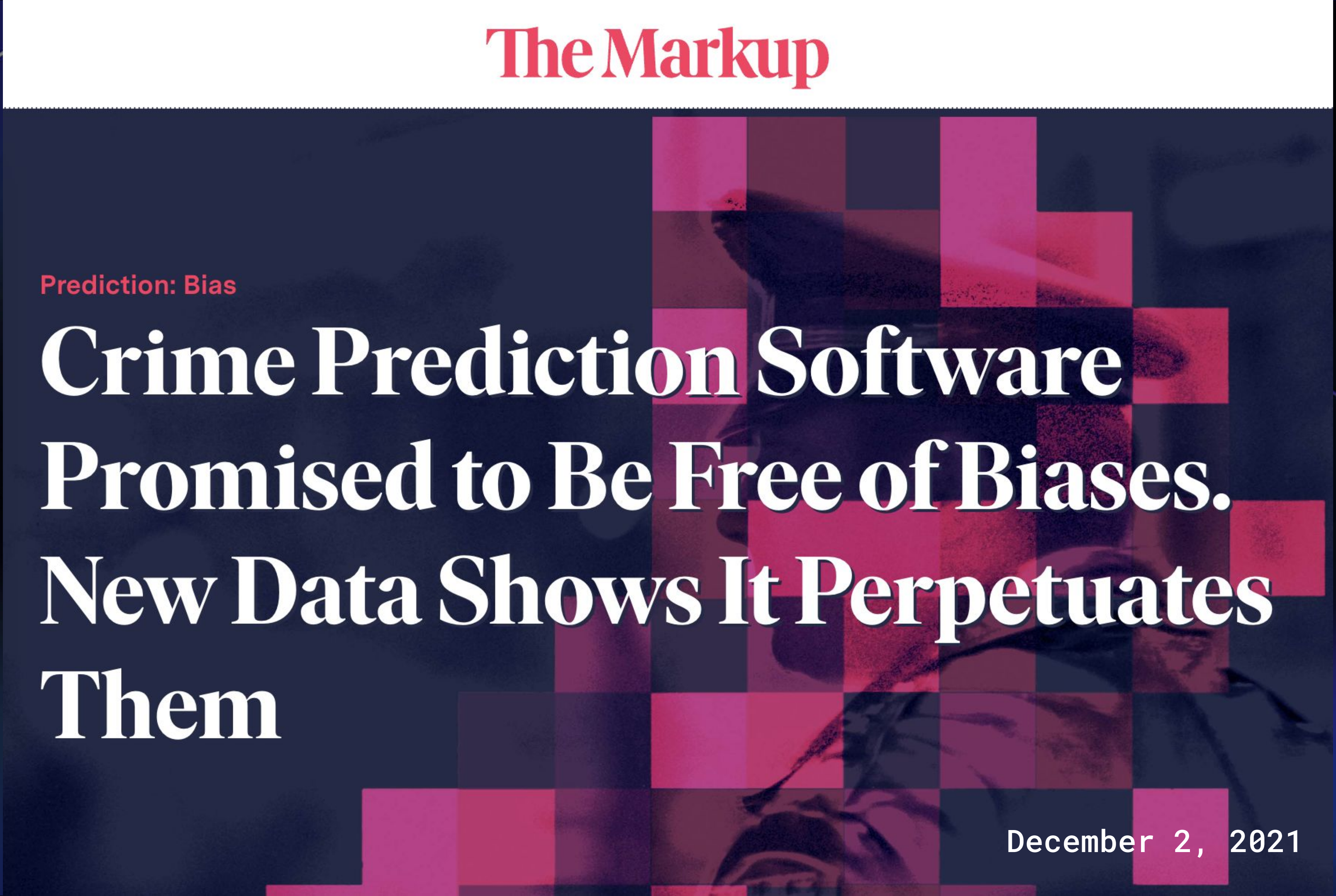
FIGURE 1. Comparison of PredPol predictions versus NSDUH predictions [27, Figure 1].

Problem

What accounts for the difference?

Problem

- Dataset focus is on crimes recorded, not crimes committed
- Incomplete census
- Not a representative random sample



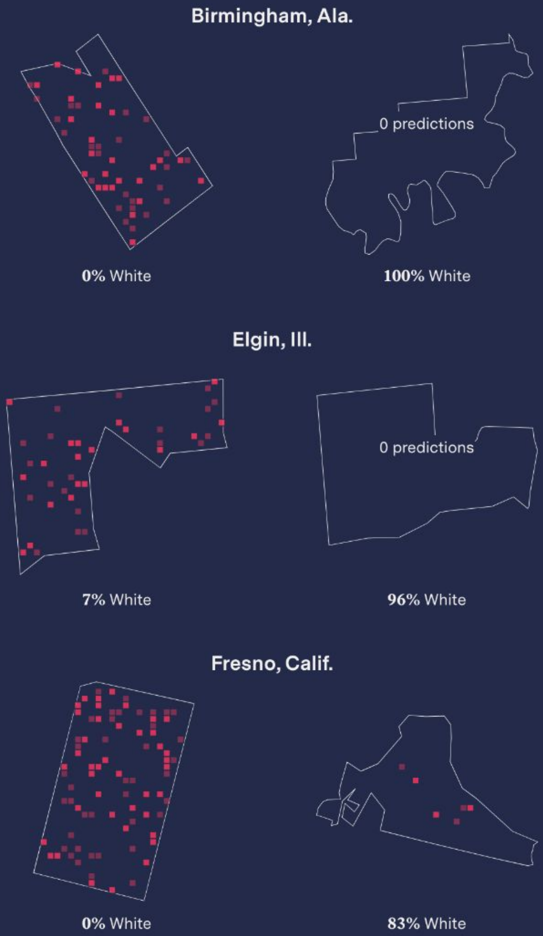
The Markup

Prediction: Bias

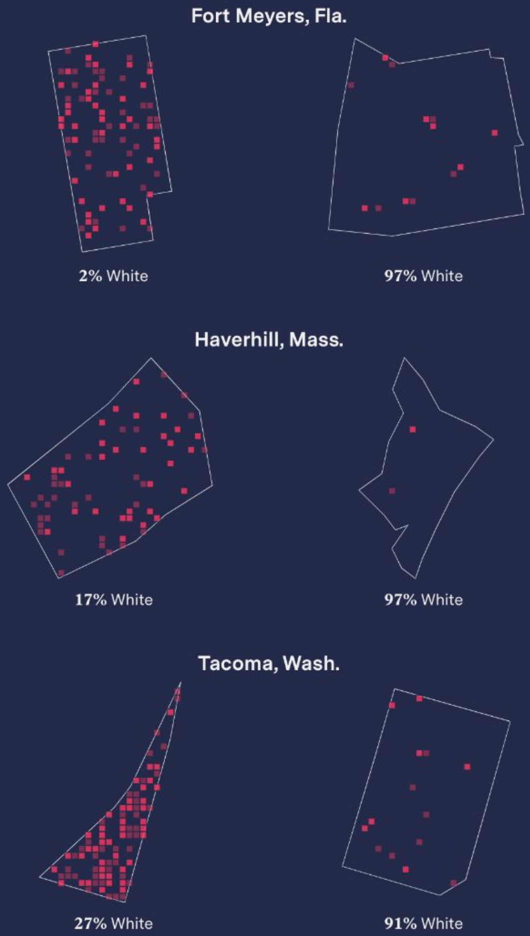
Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them

December 2, 2021

Each 🟥 represents 100 predictions

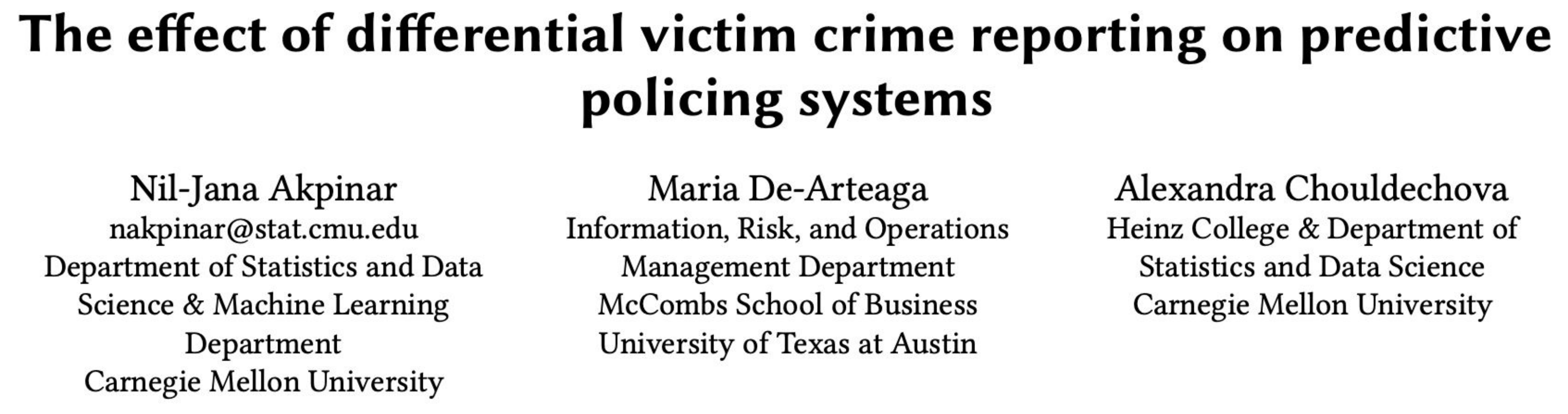


We analyzed more than five million predictions and neighborhoods with fewer predictions consistently had a higher population of White residents.



Problem

- **Independent algorithms trained on district-by-district victim crime reporting data in Bogota, Colombia had similar biased outcomes**

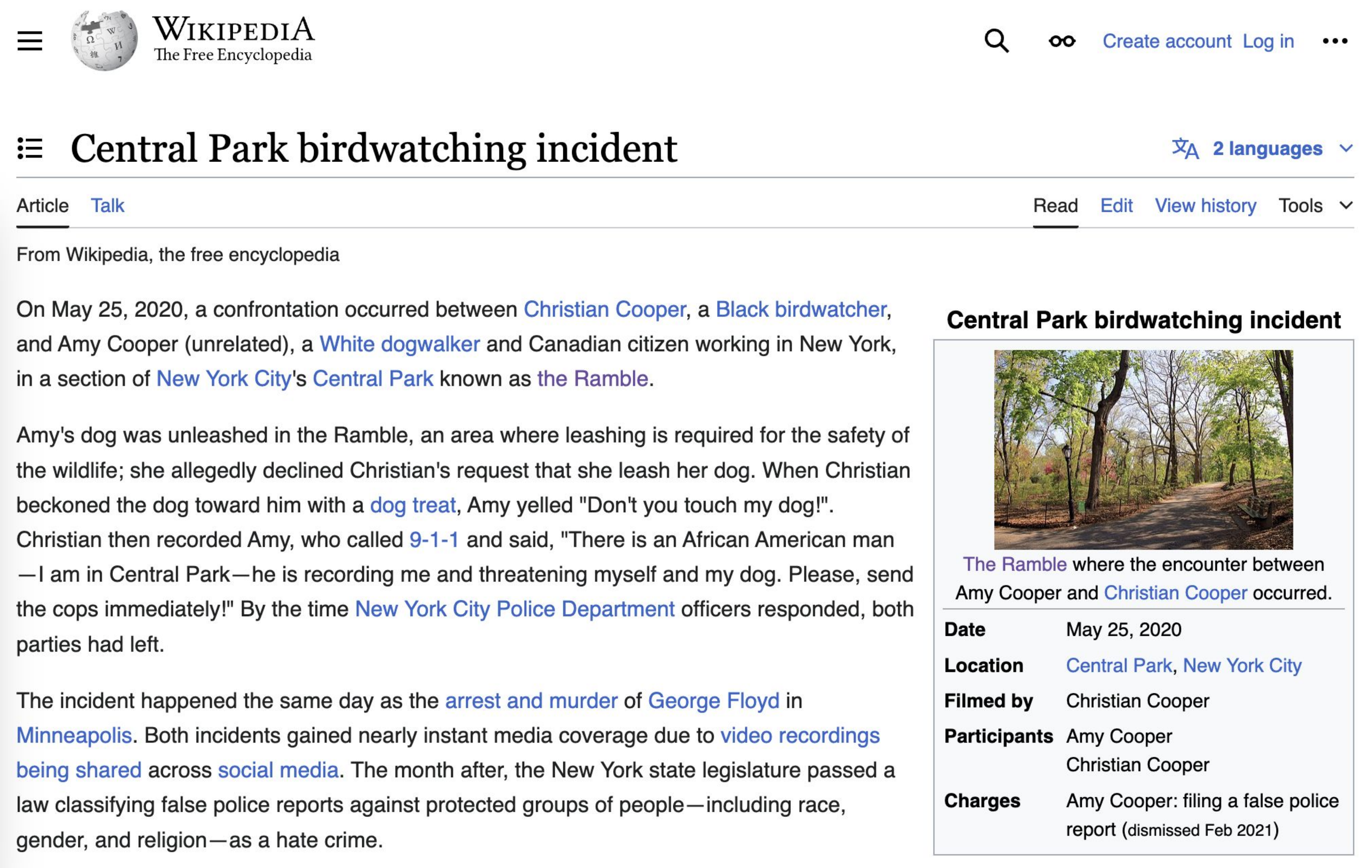


The effect of differential victim crime reporting on predictive policing systems

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



Central Park birdwatching incident

From Wikipedia, the free encyclopedia

On May 25, 2020, a confrontation occurred between Christian Cooper, a Black birdwatcher, and Amy Cooper (unrelated), a White dogwalker and Canadian citizen working in New York, in a section of New York City's Central Park known as the Ramble.

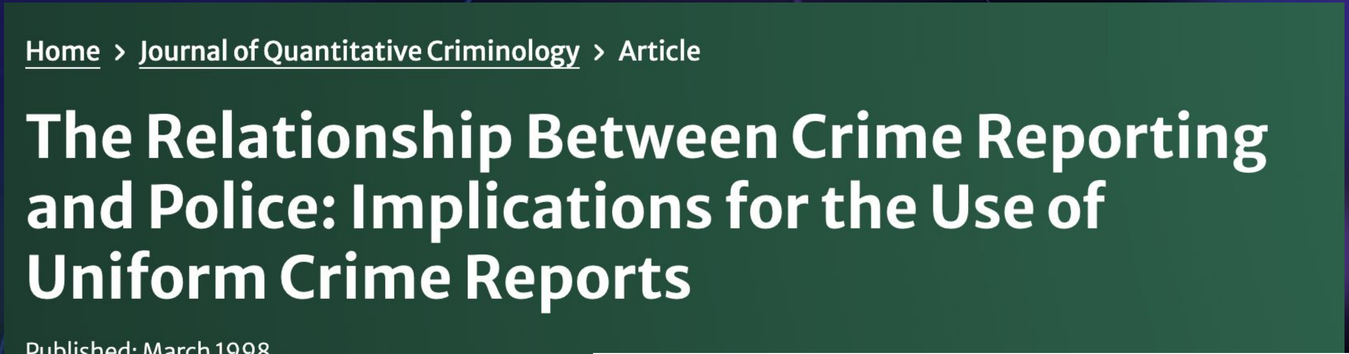
Amy's dog was unleashed in the Ramble, an area where leashing is required for the safety of the wildlife; she allegedly declined Christian's request that she leash her dog. When Christian beckoned the dog toward him with a dog treat, Amy yelled "Don't you touch my dog!". Christian then recorded Amy, who called 9-1-1 and said, "There is an African American man—I am in Central Park—he is recording me and threatening myself and my dog. Please, send the cops immediately!" By the time New York City Police Department officers responded, both parties had left.

The incident happened the same day as the arrest and murder of George Floyd in Minneapolis. Both incidents gained nearly instant media coverage due to video recordings being shared across social media. The month after, the New York state legislature passed a law classifying false police reports against protected groups of people—including race, gender, and religion—as a hate crime.

Central Park birdwatching incident

The Ramble where the encounter between Amy Cooper and Christian Cooper occurred.

Date	May 25, 2020
Location	Central Park, New York City
Filmed by	Christian Cooper
Participants	Amy Cooper Christian Cooper
Charges	Amy Cooper: filing a false police report (dismissed Feb 2021)

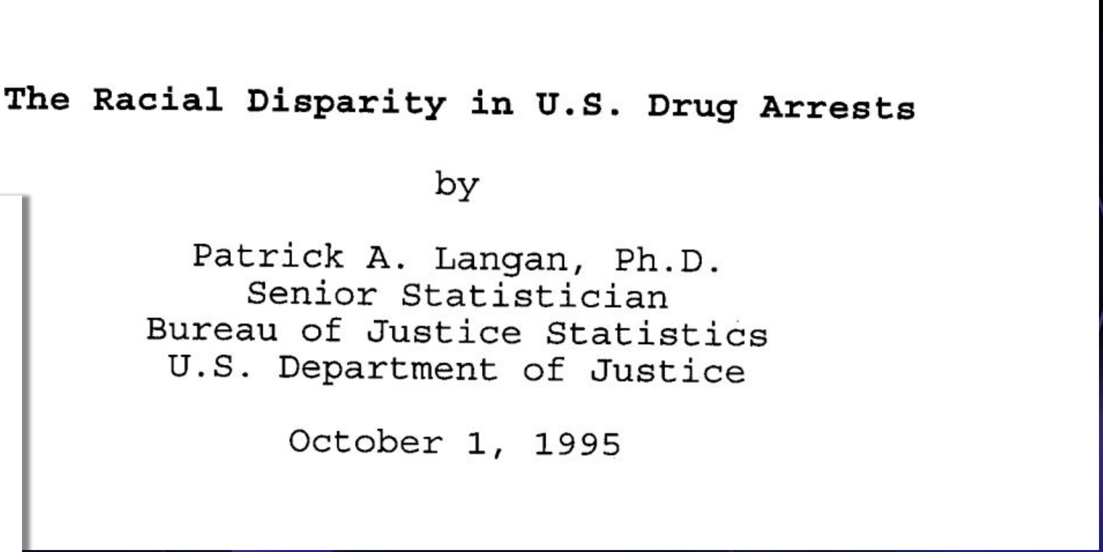


Home › Journal of Quantitative Criminology › Article

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

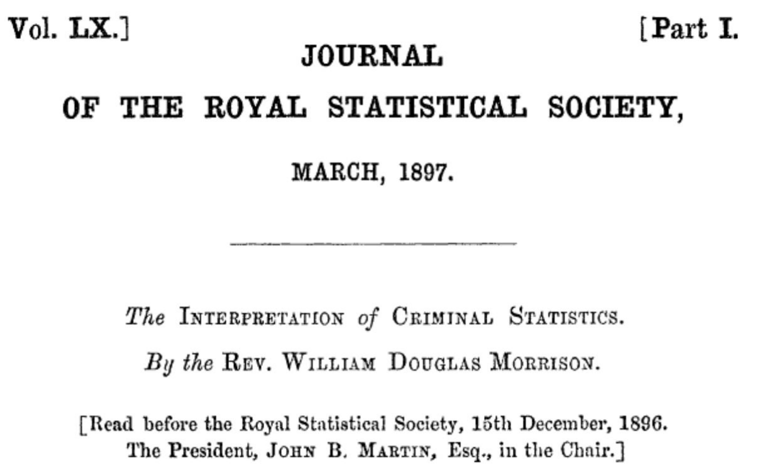


The Racial Disparity in U.S. Drug Arrests

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,

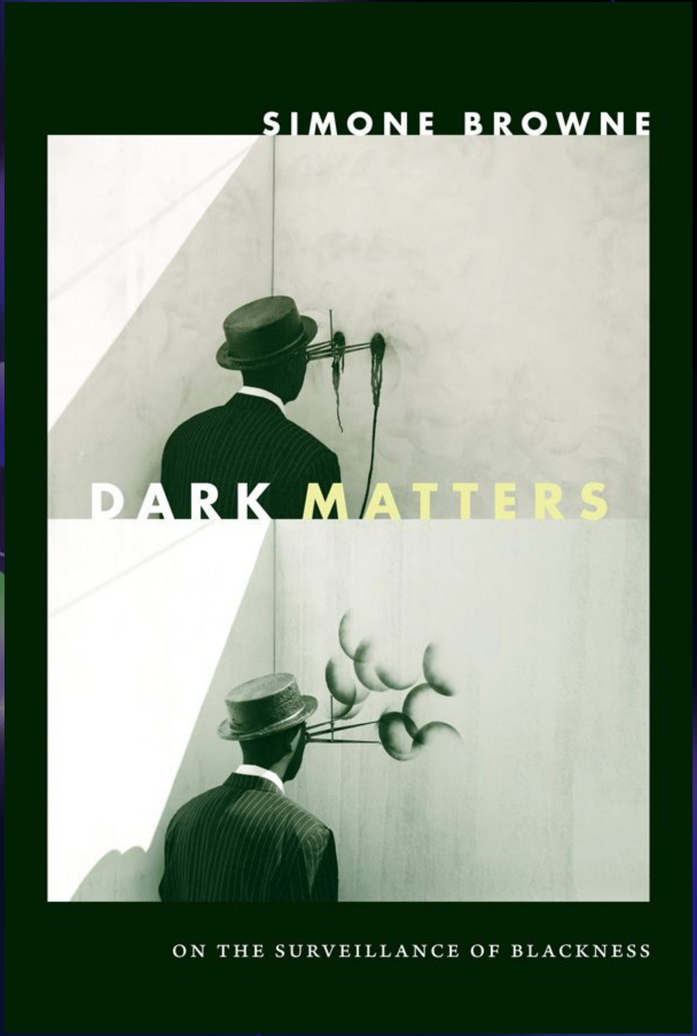
MARCH, 1897.

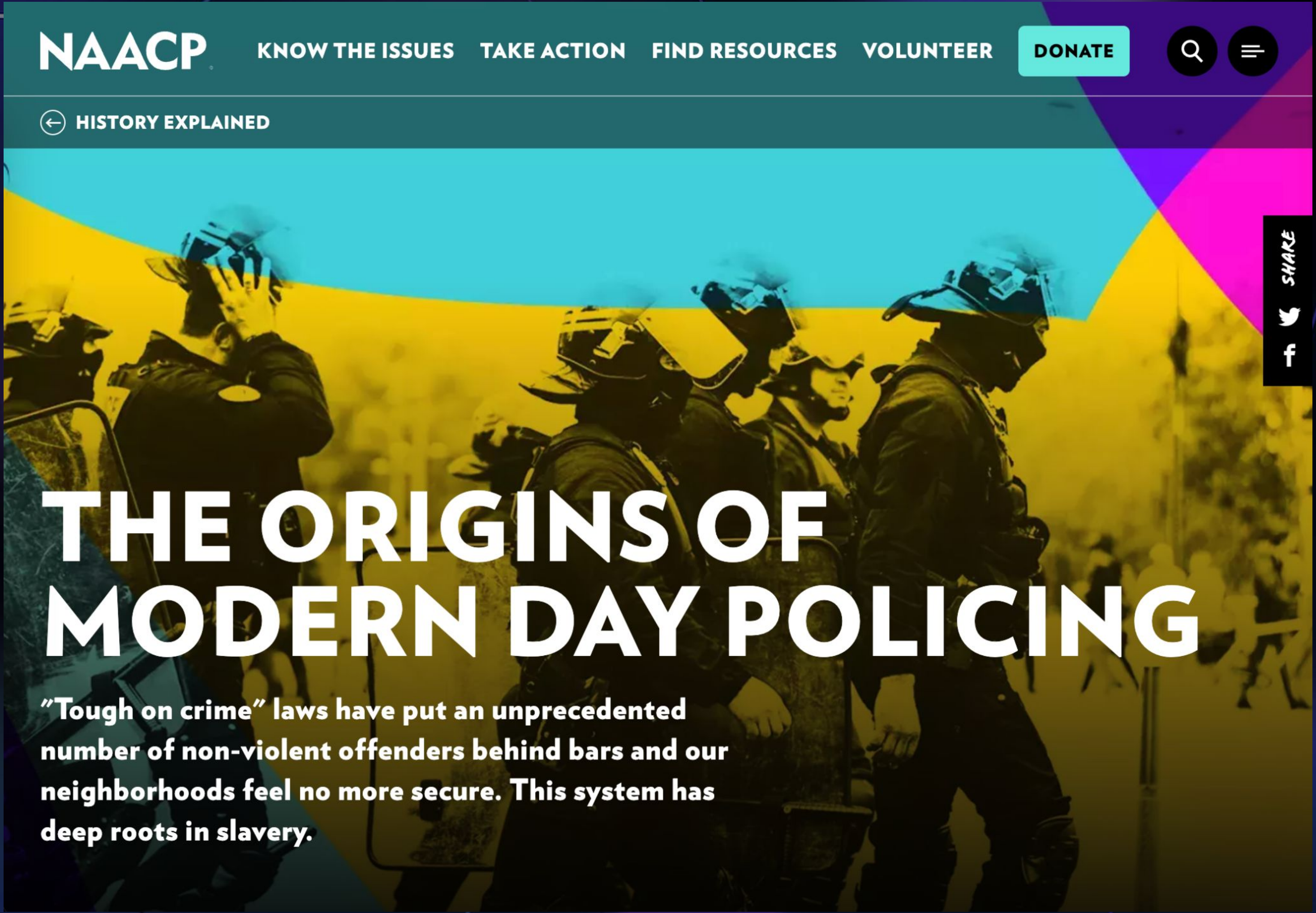
The Interpretation *of* Criminal Statistics.

By the Rev. William Douglas Morrison.

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

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.





THE ORIGINS OF MODERN DAY POLICING

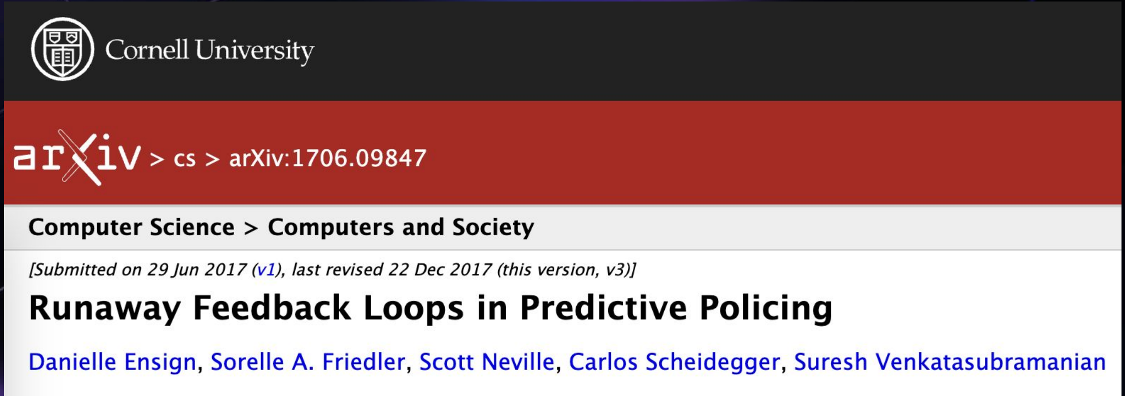
"Tough on crime" laws have put an unprecedented number of non-violent offenders behind bars and our neighborhoods feel no more secure. This system has deep roots in slavery.

Problem

Self-reinforcing feedback loop

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

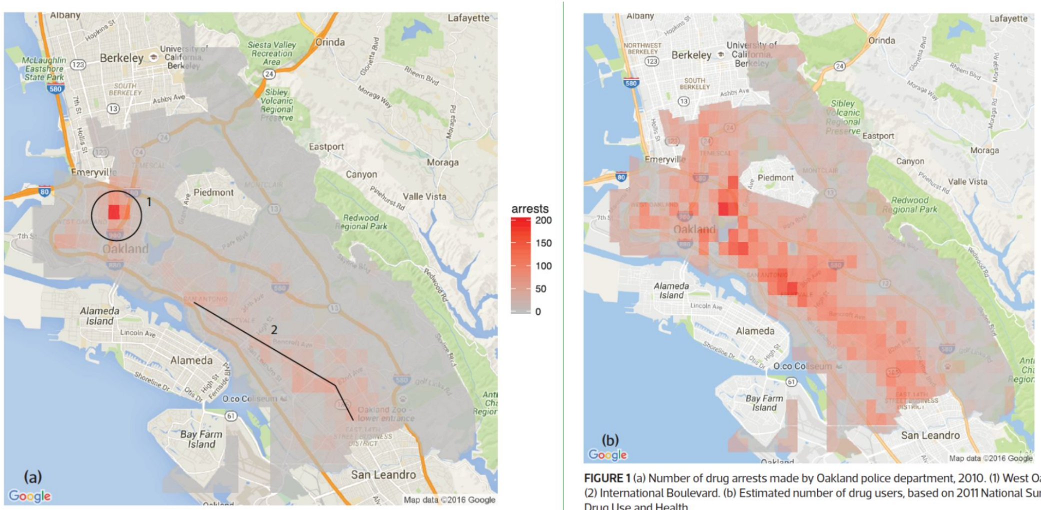
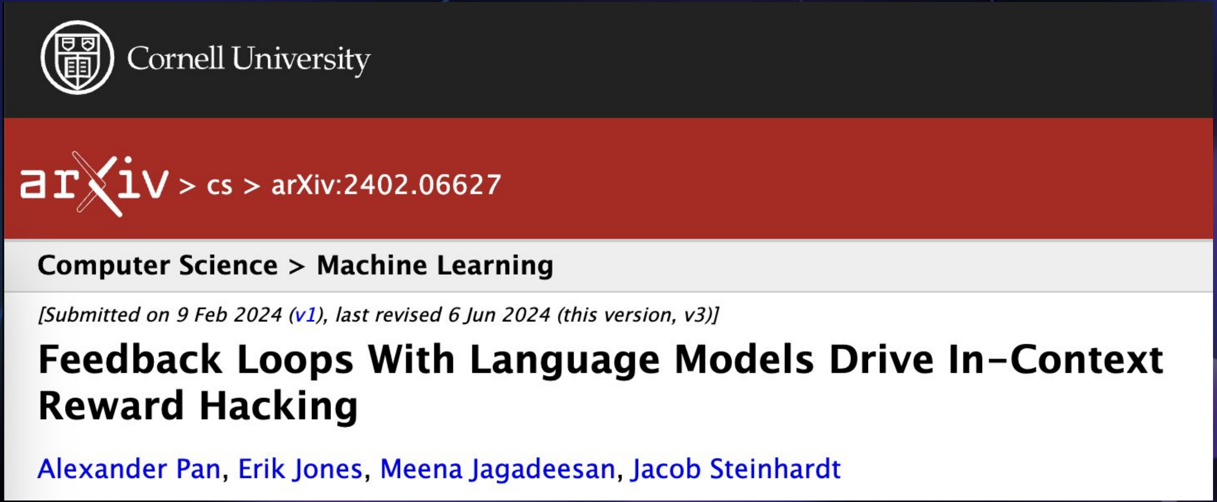


FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health



Cornell University

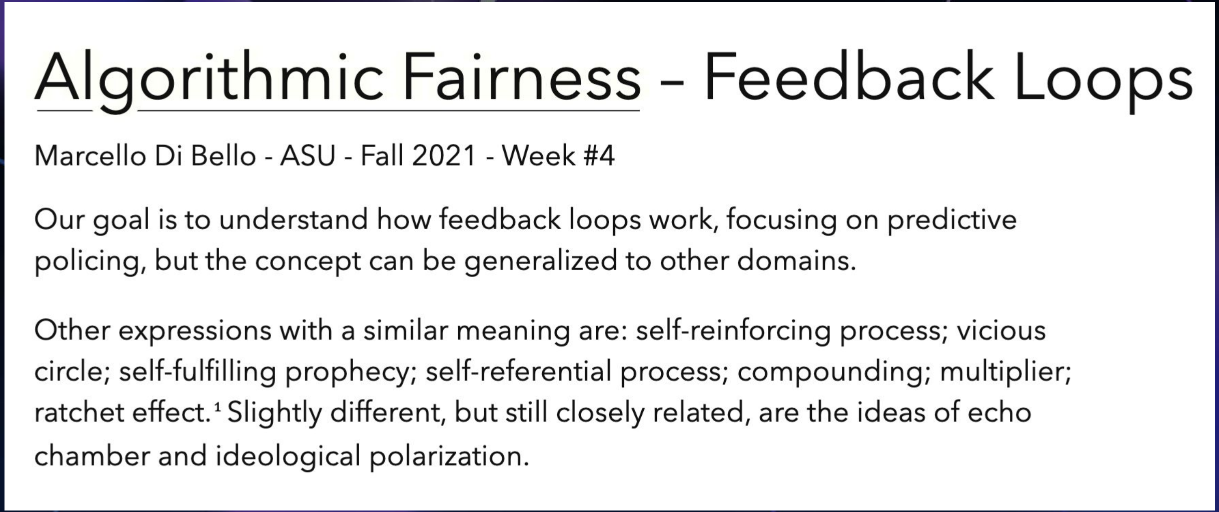
arXiv > cs > arXiv:2402.06627

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



Algorithmic Fairness – Feedback Loops

Marcello Di Bello - ASU - Fall 2021 - Week #4

Our goal is to understand how feedback loops work, focusing on predictive policing, but the concept can be generalized to other domains.

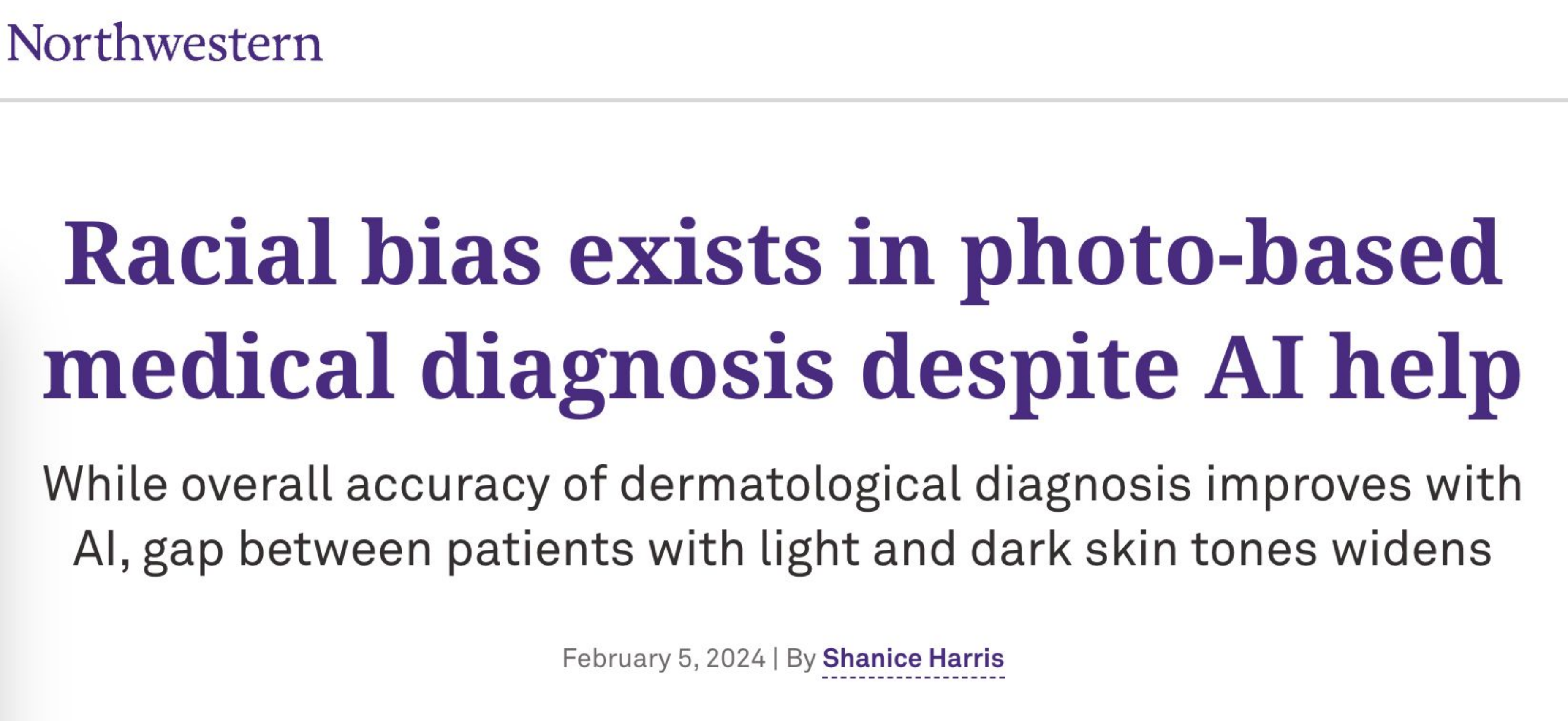
Other expressions with a similar meaning are: self-reinforcing process; vicious circle; self-fulfilling prophecy; self-referential process; compounding; multiplier; ratchet effect.[1] Slightly different, but still closely related, are the ideas of echo chamber and ideological polarization.

Problem

Under-representation

Under-representation

Healthcare



Northwestern

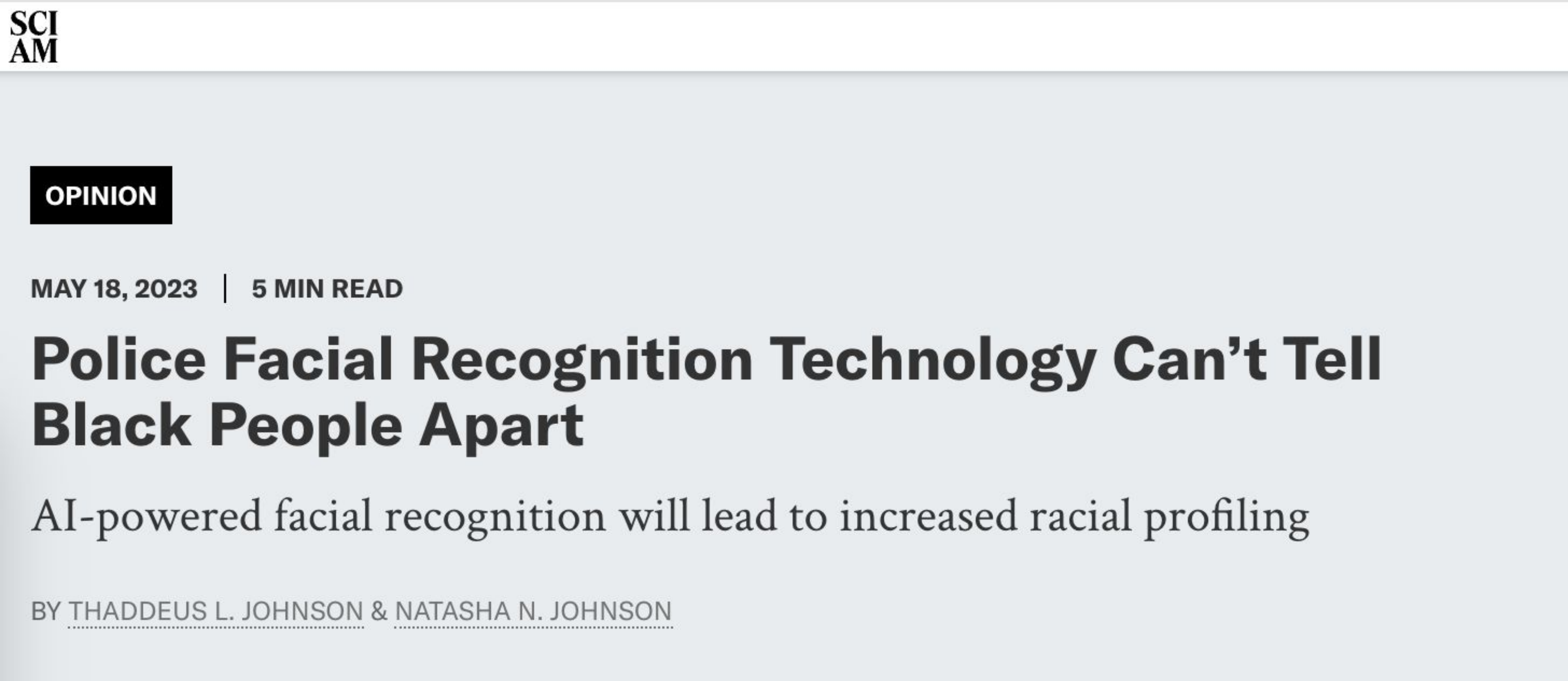
Racial bias exists in photo-based medical diagnosis despite AI help

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

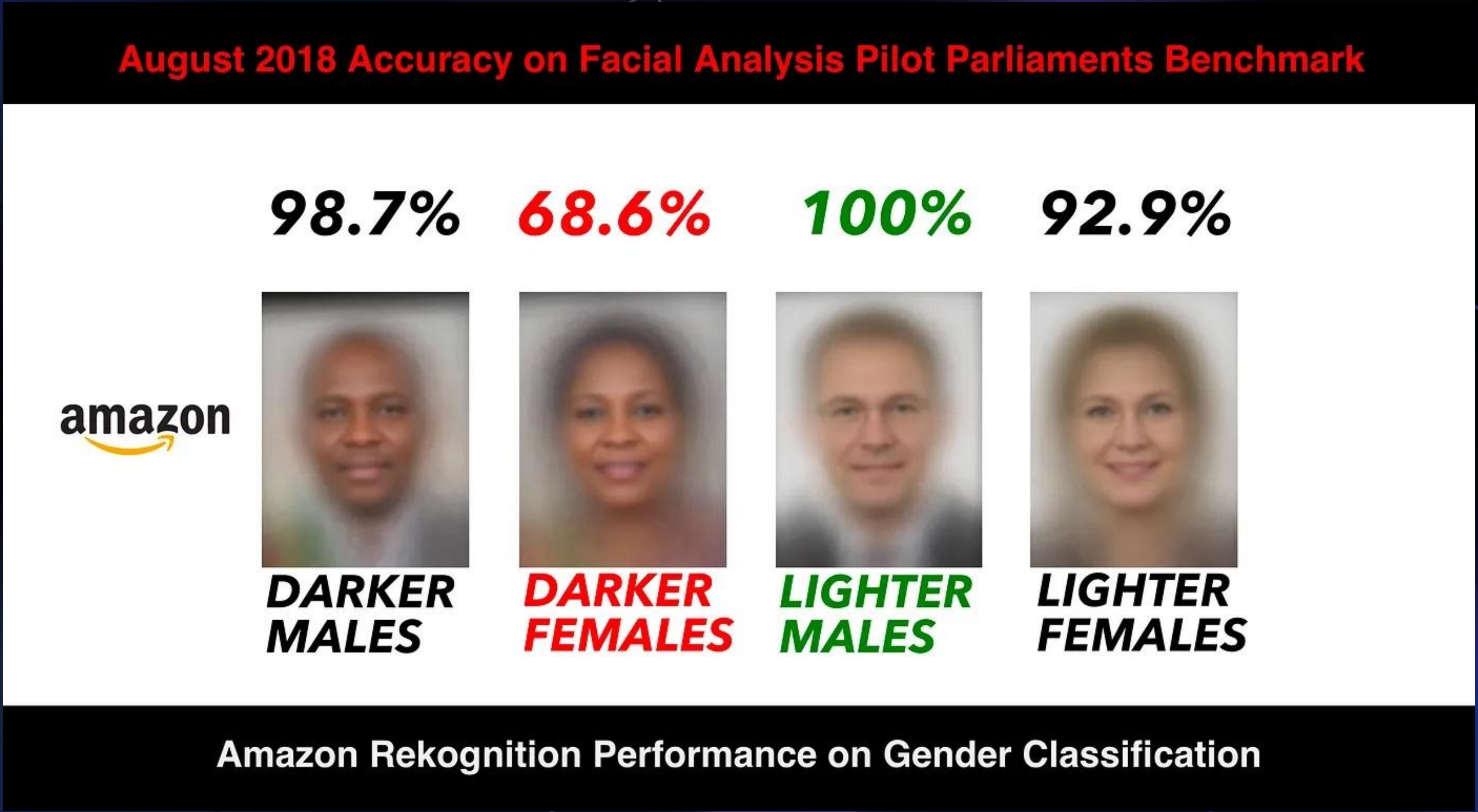
Police Facial Recognition Technology Can’t Tell Black People Apart

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

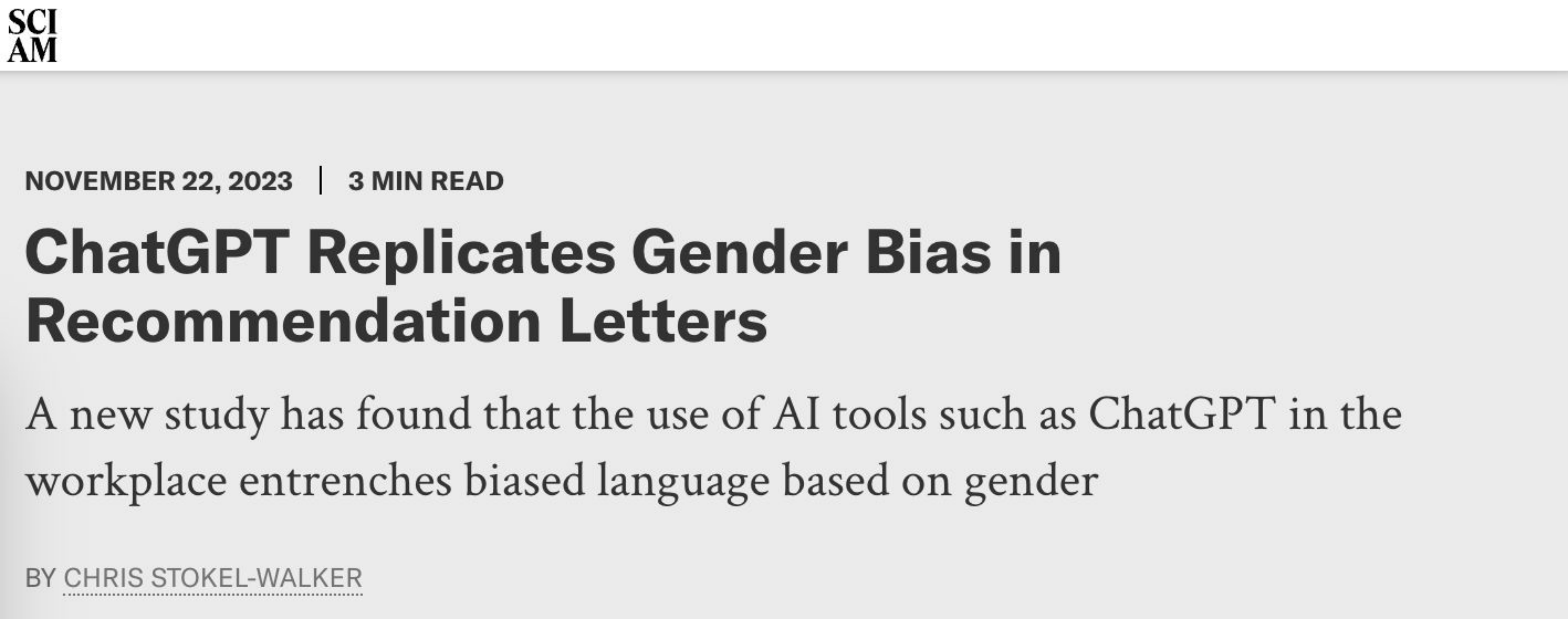
amazon

98.7%	68.6%	100%	92.9%
DARKER MALES	DARKER FEMALES	LIGHTER MALES	LIGHTER FEMALES

Amazon Rekognition Performance on Gender Classification

Under-representation

Workforce



SCI
AM

NOVEMBER 22, 2023 | 3 MIN READ

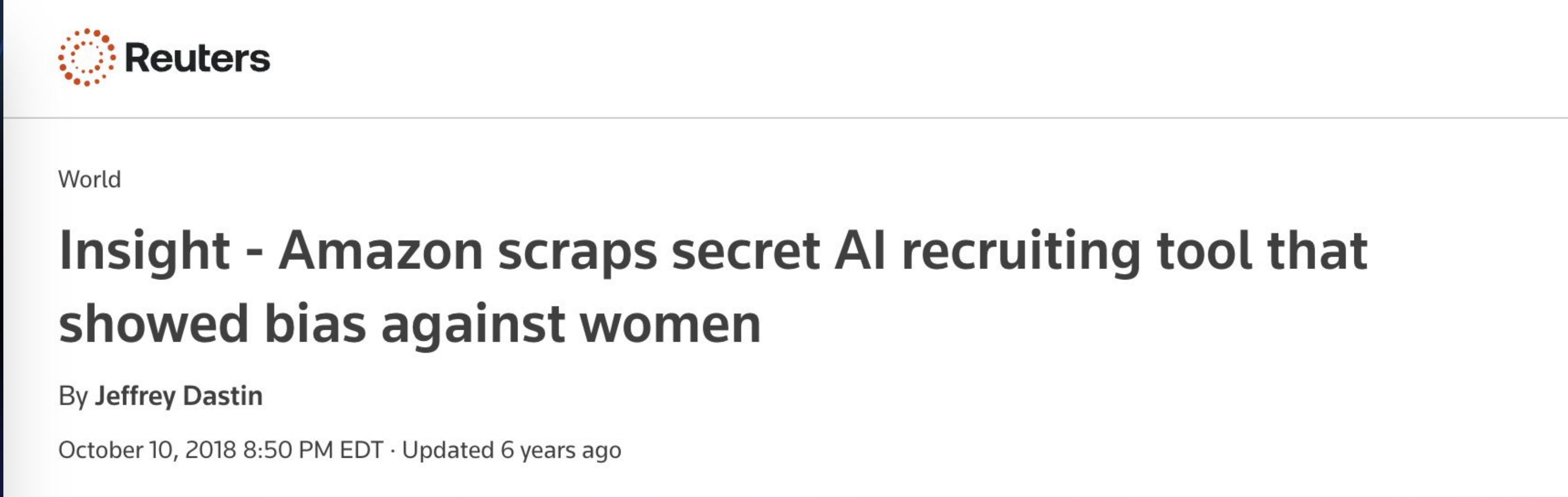
ChatGPT Replicates Gender Bias in Recommendation Letters

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

Under-representation

Workforce



Reuters

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

> “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

Good intentions with bad outcomes, not nefarious bad actors.

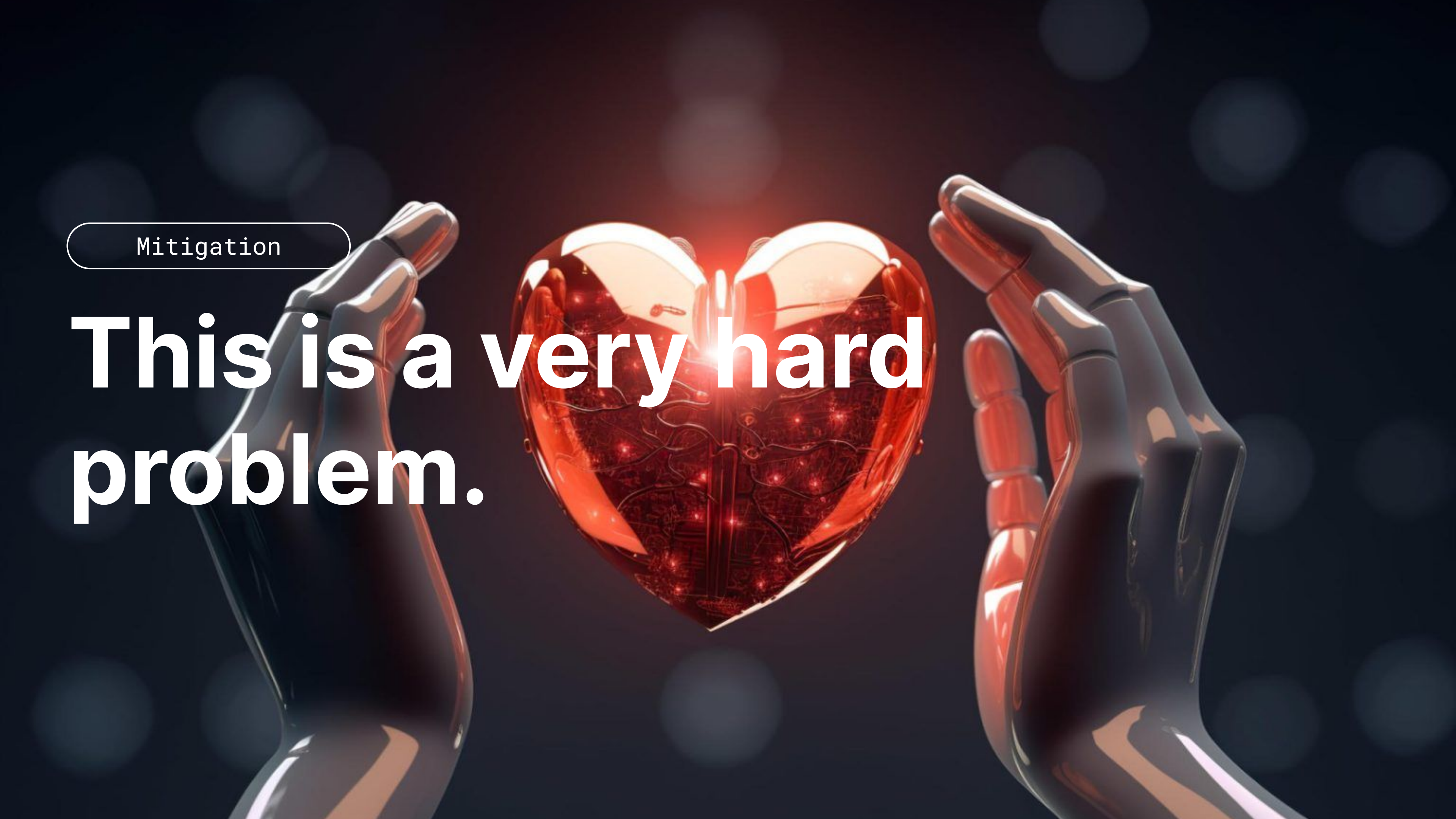
On the contrary: people doing their best to improve the lives of others, increase safety, and improve public health.

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.

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

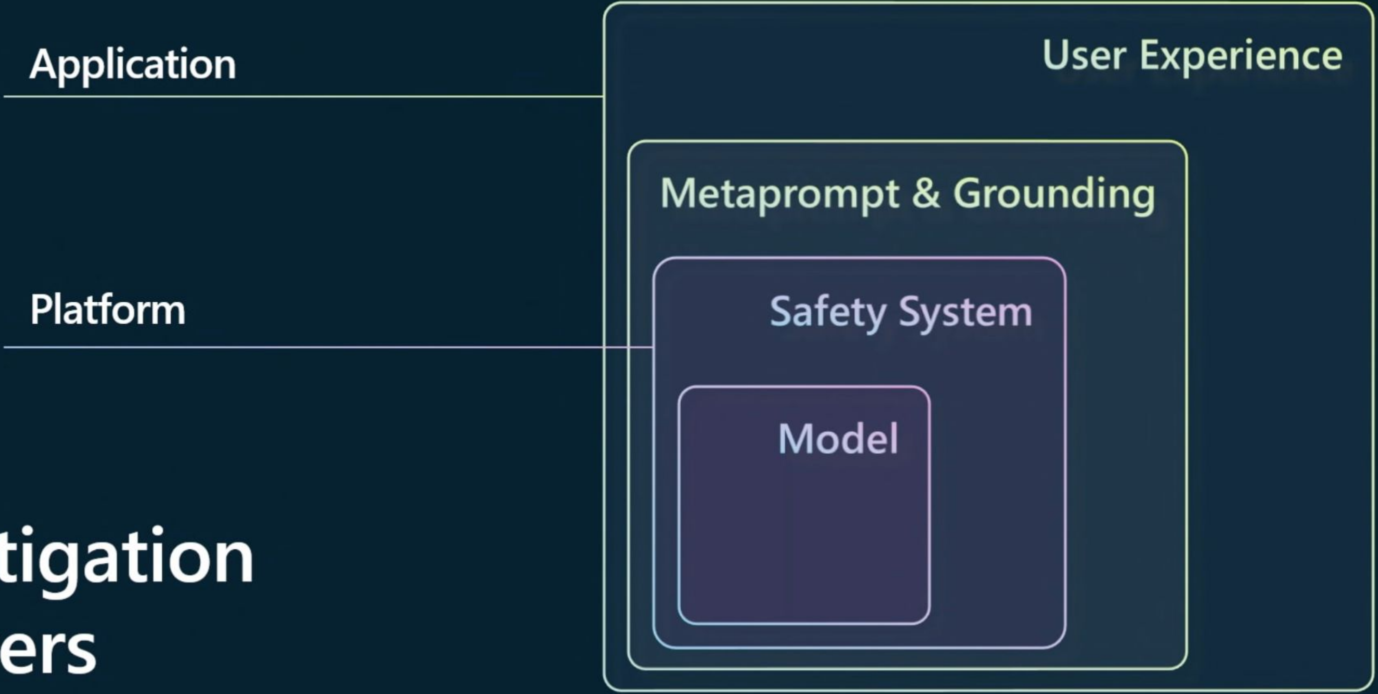
Mitigation

How do we mitigate?

Mitigation

Apply Responsible AI principles & tactics

Mitigation layers



Application

Platform

User Experience

Metaprompt & Grounding

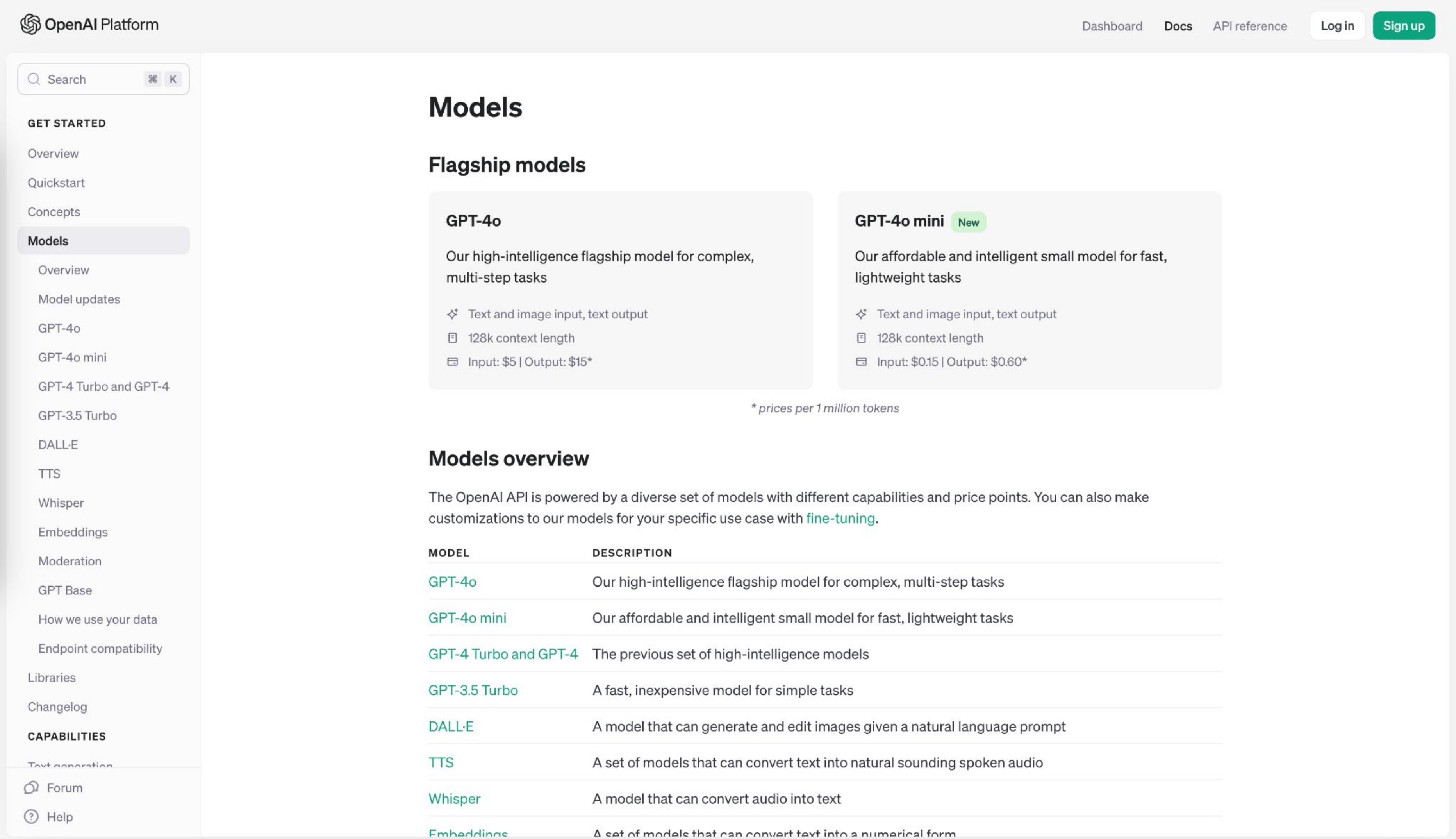
Safety System

Model

Mitigation

Model

Consider the dataset



Models

Flagship models

GPT-4o

Our high-intelligence flagship model for complex, multi-step tasks

- Text and image input, text output
- 128k context length
- Input: $5 | Output: $15*

GPT-4o mini New

Our affordable and intelligent small model for fast, lightweight tasks

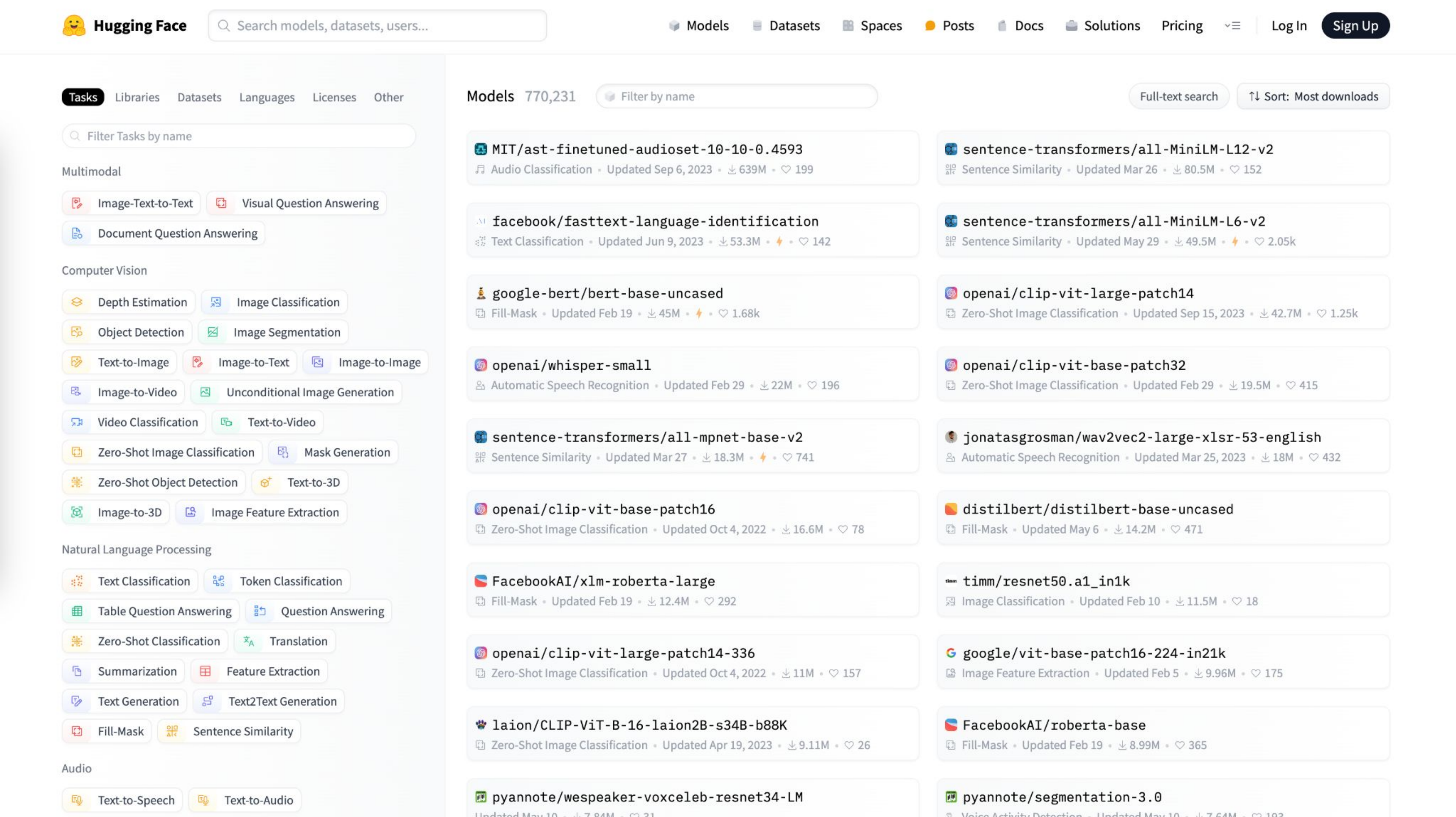
- Text and image input, text output
- 128k context length
- Input: $0.15 | Output: $0.60*

** prices per 1 million tokens*

Models overview

The OpenAI API is powered by a diverse set of models with different capabilities and price points. You can also make customizations to our models for your specific use case with fine-tuning.

MODEL	DESCRIPTION
GPT-4o	Our high-intelligence flagship model for complex, multi-step tasks
GPT-4o mini	Our affordable and intelligent small model for fast, lightweight tasks
GPT-4 Turbo and GPT-4	The previous set of high-intelligence models
GPT-3.5 Turbo	A fast, inexpensive model for simple tasks
DALL·E	A model that can generate and edit images given a natural language prompt
TTS	A set of models that can convert text into natural sounding spoken audio
Whisper	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form



MIT/ast-finetuned-audioset-10-10-0.4593
Audio Classification • Updated Sep 6, 2023 • 639M • 199

sentence-transformers/all-MiniLM-L12-v2
Sentence Similarity • Updated Mar 26 • 80.5M • 152

facebook/fasttext-language-identification
Text Classification • Updated Jun 9, 2023 • 53.3M • • 142

sentence-transformers/all-MiniLM-L6-v2
Sentence Similarity • Updated May 29 • 49.5M • • 2.05k

google-bert/bert-base-uncased
Fill-Mask • Updated Feb 19 • 45M • • 1.68k

openai/clip-vit-large-patch14
Zero-Shot Image Classification • Updated Sep 15, 2023 • 42.7M • 1.25k

openai/whisper-small
Automatic Speech Recognition • Updated Feb 29 • 22M • 196

openai/clip-vit-base-patch32
Zero-Shot Image Classification • Updated Feb 29 • 19.5M • 415

sentence-transformers/all-mpnet-base-v2
Sentence Similarity • Updated Mar 27 • 18.3M • • 741

jonatasgrosman/wav2vec2-large-xlsr-53-english
Automatic Speech Recognition • Updated Mar 25, 2023 • 18M • 432

openai/clip-vit-base-patch16
Zero-Shot Image Classification • Updated Oct 4, 2022 • 16.6M • 78

distilbert/distilbert-base-uncased
Fill-Mask • Updated May 6 • 14.2M • 471

FacebookAI/xlm-roberta-large
Fill-Mask • Updated Feb 19 • 12.4M • 292

timm/resnet50.a1_in1k
Image Classification • Updated Feb 10 • 11.5M • 18

openai/clip-vit-large-patch14-336
Zero-Shot Image Classification • Updated Oct 4, 2022 • 11M • 157

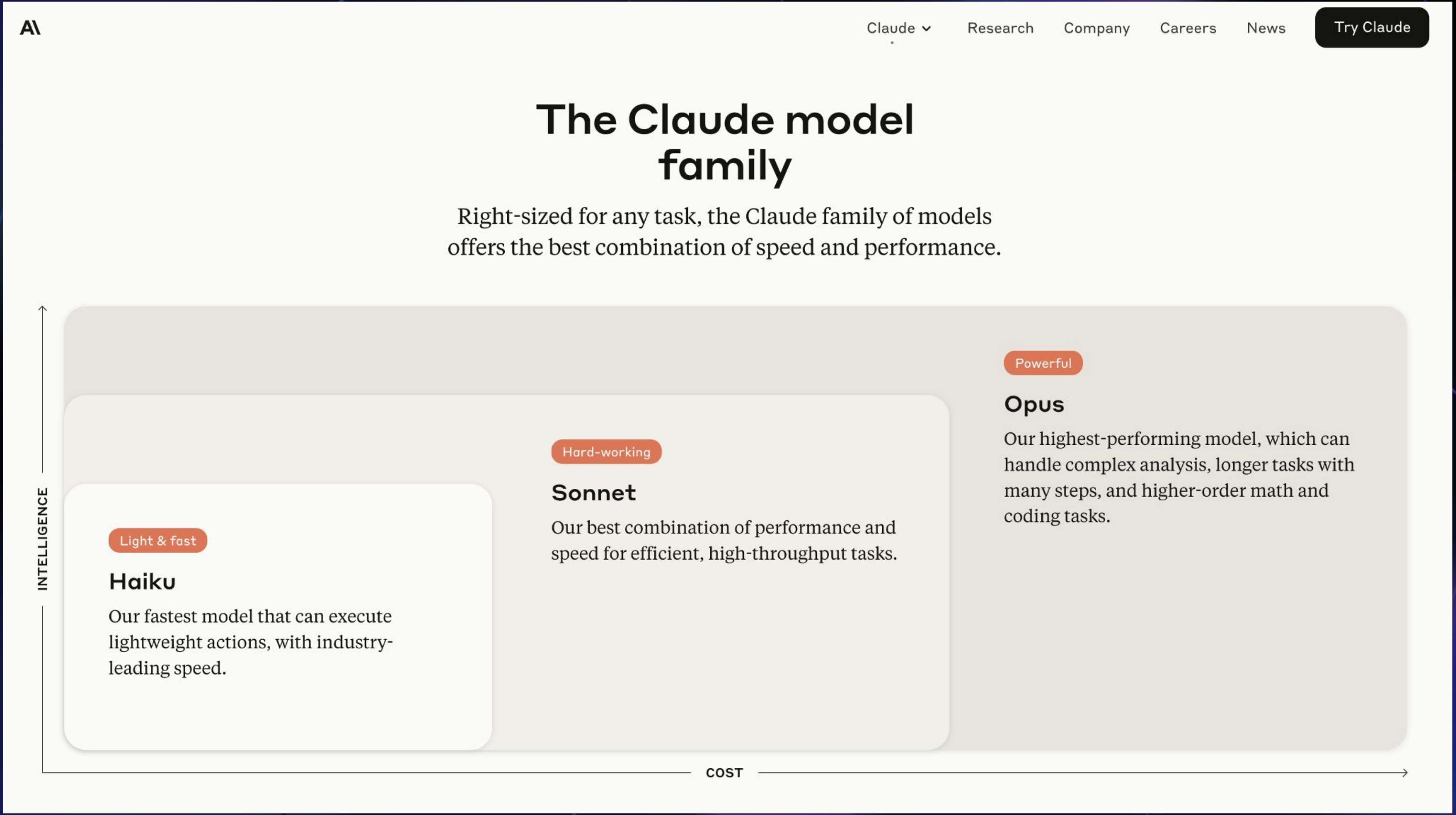
google/vit-base-patch16-224-in21k
Image Feature Extraction • Updated Feb 5 • 9.96M • 175

laion/CLIP-ViT-B-16-laion2B-s34B-b88K
Zero-Shot Image Classification • Updated Apr 19, 2023 • 9.11M • 26

FacebookAI/roberta-base
Fill-Mask • Updated Feb 19 • 8.99M • 365

pyannote/wespeaker-voxceleb-resnet34-LM
Updated May 10 • 7.84M • 31

pyannote/segmentation-3.0
Voice Activity Detection • Updated May 10 • 7.64M • 193



The Claude model family

Right-sized for any task, the Claude family of models offers the best combination of speed and performance.

Light & fast

Haiku

Our fastest model that can execute lightweight actions, with industry-leading speed.

Hard-working

Sonnet

Our best combination of performance and speed for efficient, high-throughput tasks.

Powerful

Opus

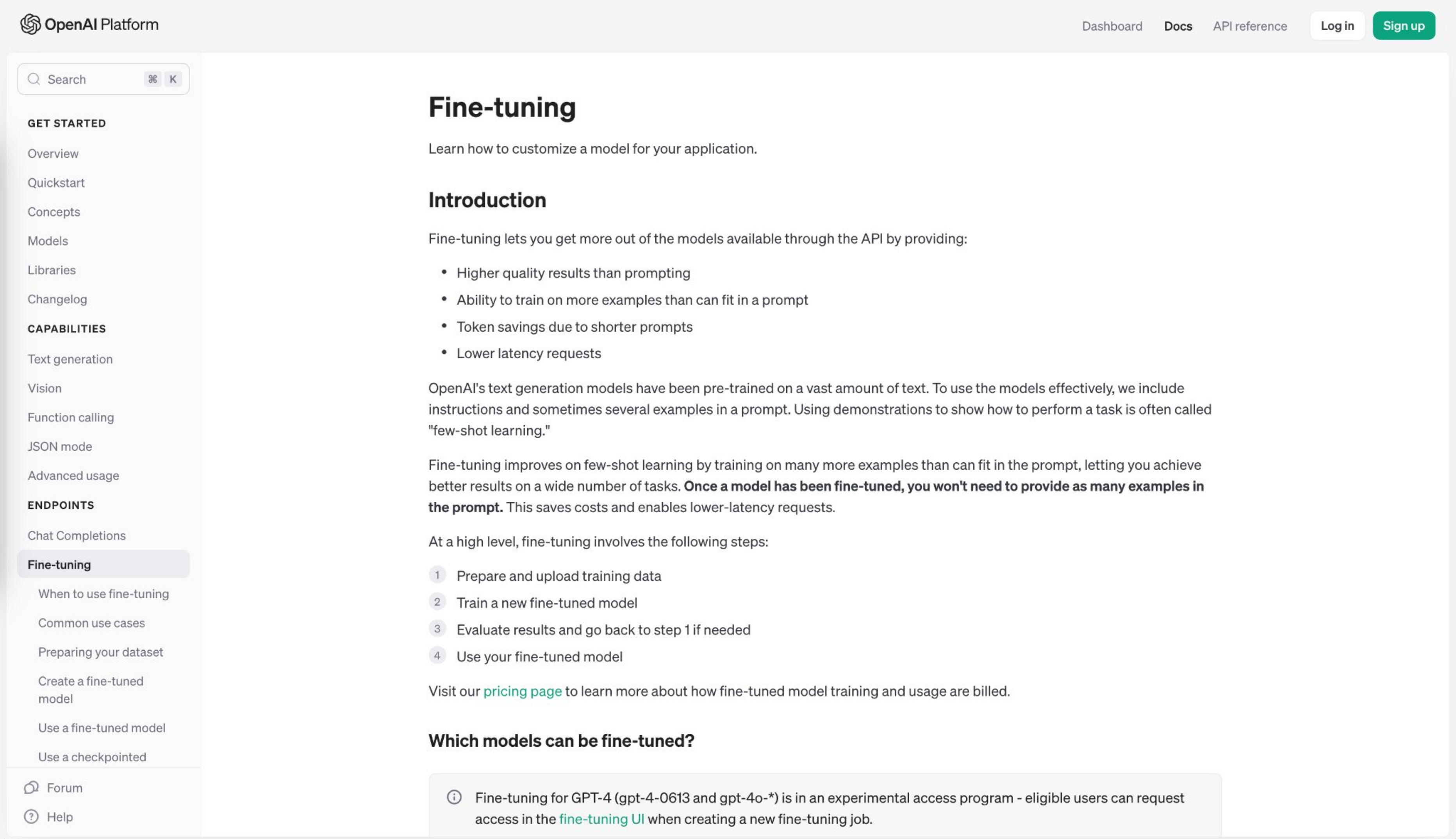
Our highest-performing model, which can handle complex analysis, longer tasks with many steps, and higher-order math and coding tasks.

Mitigation

Fine tune your models

Fine tuning

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



Fine-tuning

Learn how to customize a model for your application.

Introduction

Fine-tuning lets you get more out of the models available through the API by providing:

- Higher quality results than prompting
- Ability to train on more examples than can fit in a prompt
- Token savings due to shorter prompts
- Lower latency requests

OpenAI's text generation models have been pre-trained on a vast amount of text. To use the models effectively, we include instructions and sometimes several examples in a prompt. Using demonstrations to show how to perform a task is often called "few-shot learning."

Fine-tuning improves on few-shot learning by training on many more examples than can fit in the prompt, letting you achieve better results on a wide number of tasks. **Once a model has been fine-tuned, you won't need to provide as many examples in the prompt.** This saves costs and enables lower-latency requests.

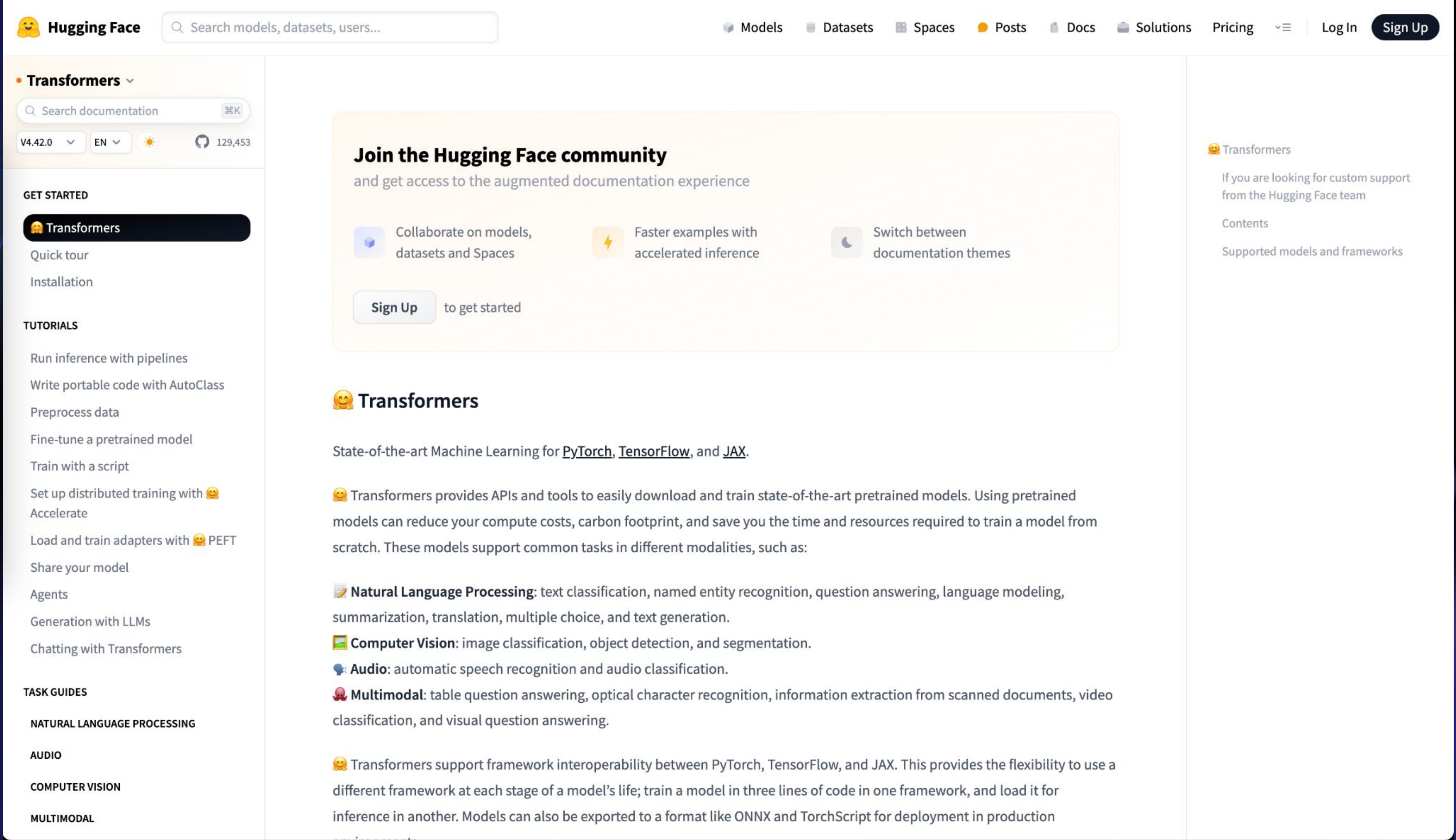
At a high level, fine-tuning involves the following steps:

1. Prepare and upload training data
2. Train a new fine-tuned model
3. Evaluate results and go back to step 1 if needed
4. Use your fine-tuned model

Visit our pricing page to learn more about how fine-tuned model training and usage are billed.

Which models can be fine-tuned?

ⓘ Fine-tuning for GPT-4 (gpt-4-0613 and gpt-4o-*) is in an experimental access program - eligible users can request access in the fine-tuning UI when creating a new fine-tuning job.



Join the Hugging Face community

and get access to the augmented documentation experience

Collaborate on models, datasets and Spaces

Faster examples with accelerated inference

Switch between documentation themes

Sign Up to get started

🤗 Transformers

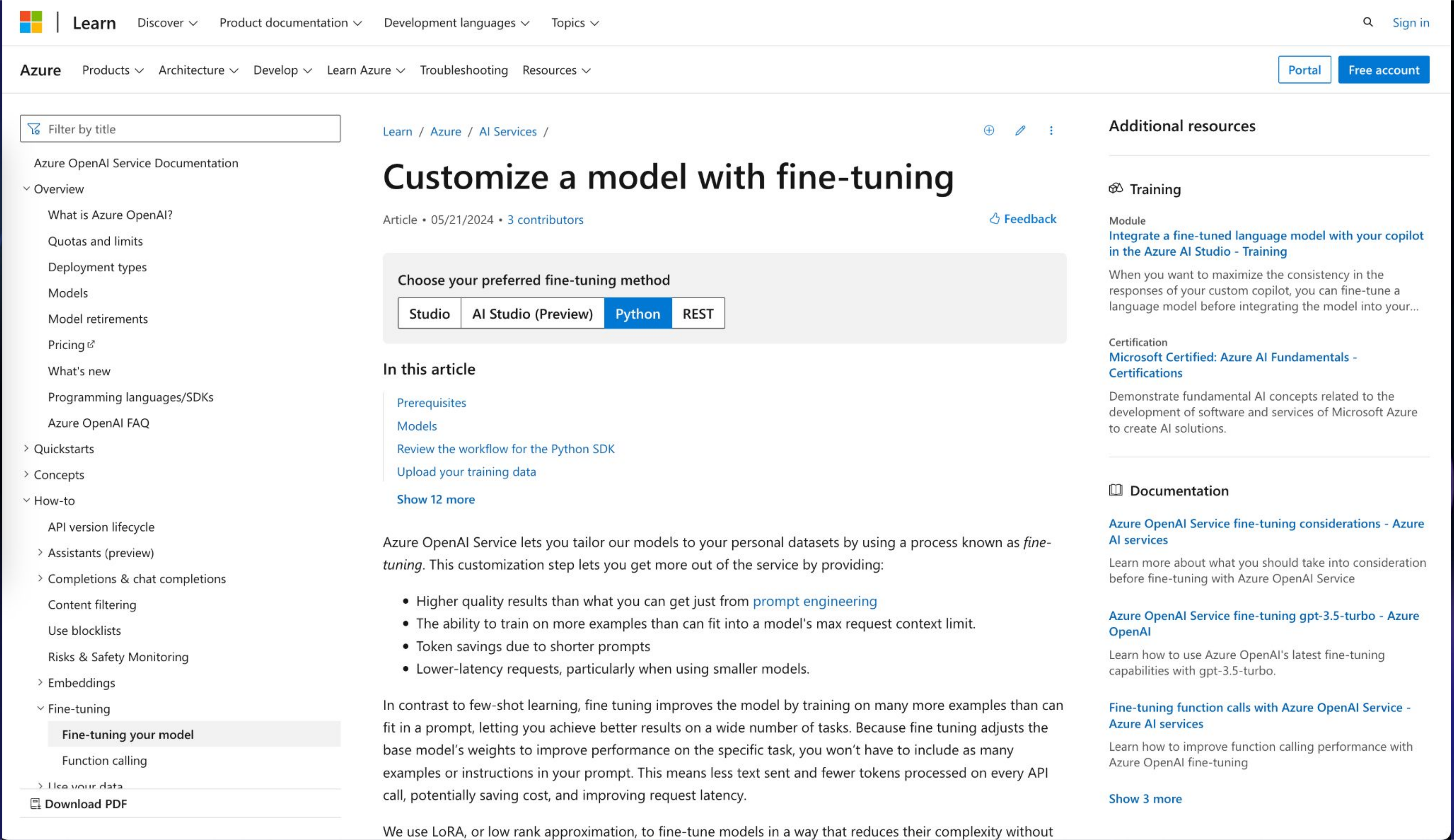
State-of-the-art Machine Learning for PyTorch, TensorFlow, and JAX.

🤗 Transformers provides APIs and tools to easily download and train state-of-the-art pretrained models. Using pretrained models can reduce your compute costs, carbon footprint, and save you the time and resources required to train a model from scratch. These models support common tasks in different modalities, such as:

📝 **Natural Language Processing**: text classification, named entity recognition, question answering, language modeling, summarization, translation, multiple choice, and text generation.
🖼️ **Computer Vision**: image classification, object detection, and segmentation.
🗣️ **Audio**: automatic speech recognition and audio classification.
🐙 **Multimodal**: table question answering, optical character recognition, information extraction from scanned documents, video classification, and visual question answering.

🤗 Transformers support framework interoperability between PyTorch, TensorFlow, and JAX. This provides the flexibility to use a different framework at each stage of a model's life; train a model in three lines of code in one framework, and load it for inference in another. Models can also be exported to a format like ONNX and TorchScript for deployment in production

🤗 Transformers
If you are looking for custom support from the Hugging Face team
Contents
Supported models and frameworks



Customize a model with fine-tuning

Article • 05/21/2024 • 3 contributors

Choose your preferred fine-tuning method

In this article

Prerequisites
Models
Review the workflow for the Python SDK
Upload your training data
Show 12 more

Azure OpenAI Service lets you tailor our models to your personal datasets by using a process known as *fine-tuning*. This customization step lets you get more out of the service by providing:

- Higher quality results than what you can get just from prompt engineering
- The ability to train on more examples than can fit into a model's max request context limit.
- Token savings due to shorter prompts
- Lower-latency requests, particularly when using smaller models.

In contrast to few-shot learning, fine tuning improves the model by training on many more examples than can fit in a prompt, letting you achieve better results on a wide number of tasks. Because fine tuning adjusts the base model's weights to improve performance on the specific task, you won't have to include as many examples or instructions in your prompt. This means less text sent and fewer tokens processed on every API call, potentially saving cost, and improving request latency.

We use LoRA, or low rank approximation, to fine-tune models in a way that reduces their complexity without

Additional resources

Training

Module

Integrate a fine-tuned language model with your copilot in the Azure AI Studio - Training

When you want to maximize the consistency in the responses of your custom copilot, you can fine-tune a language model before integrating the model into your...

Certification

Microsoft Certified: Azure AI Fundamentals - Certifications

Demonstrate fundamental AI concepts related to the development of software and services of Microsoft Azure to create AI solutions.

Documentation

Azure OpenAI Service fine-tuning considerations - Azure AI services

Learn more about what you should take into consideration before fine-tuning with Azure OpenAI Service

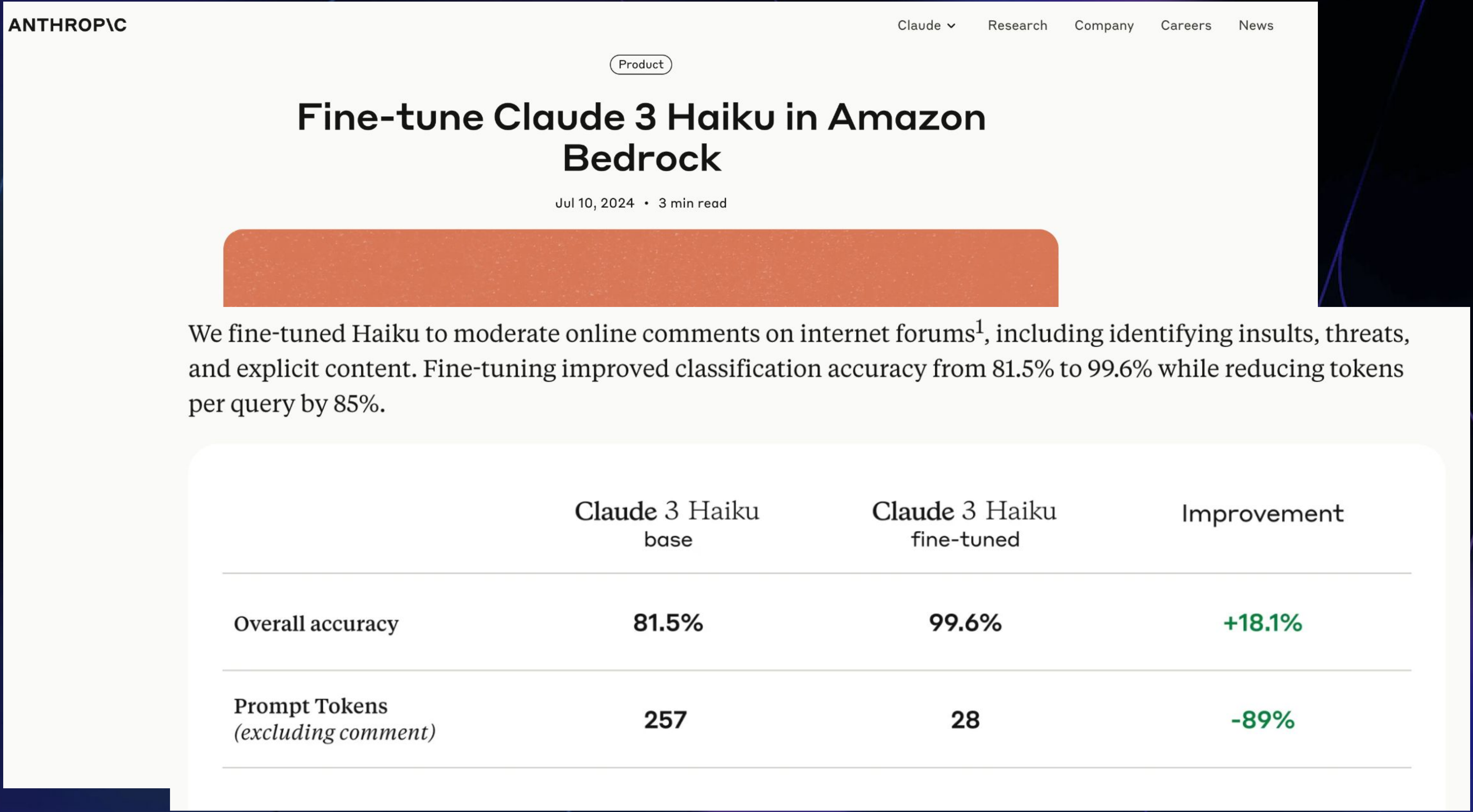
Azure OpenAI Service fine-tuning gpt-3.5-turbo - Azure OpenAI

Learn how to use Azure OpenAI's latest fine-tuning capabilities with gpt-3.5-turbo.

Fine-tuning function calls with Azure OpenAI Service - Azure AI services

Learn how to improve function calling performance with Azure OpenAI fine-tuning

Show 3 more



Fine-tune Claude 3 Haiku in Amazon Bedrock

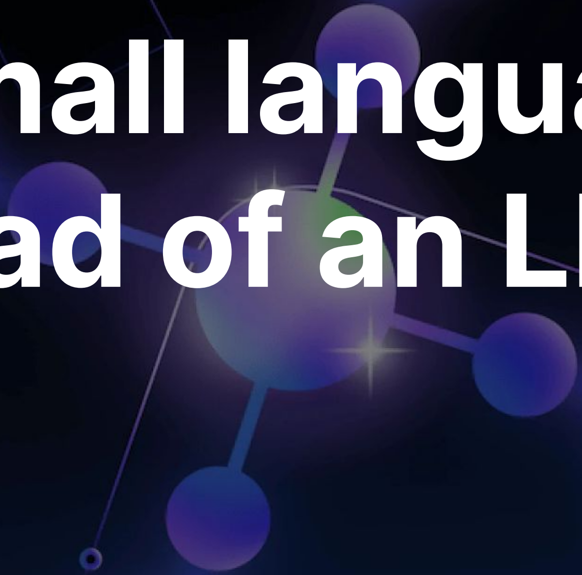
Jul 10, 2024 • 3 min read

We fine-tuned Haiku to moderate online comments on internet forums[1], including identifying insults, threats, and explicit content. Fine-tuning improved classification accuracy from 81.5% to 99.6% while reducing tokens per query by 85%.

	Claude 3 Haiku base	Claude 3 Haiku fine-tuned	Improvement
Overall accuracy	81.5%	99.6%	+18.1%
Prompt Tokens *(excluding comment)*	257	28	-89%

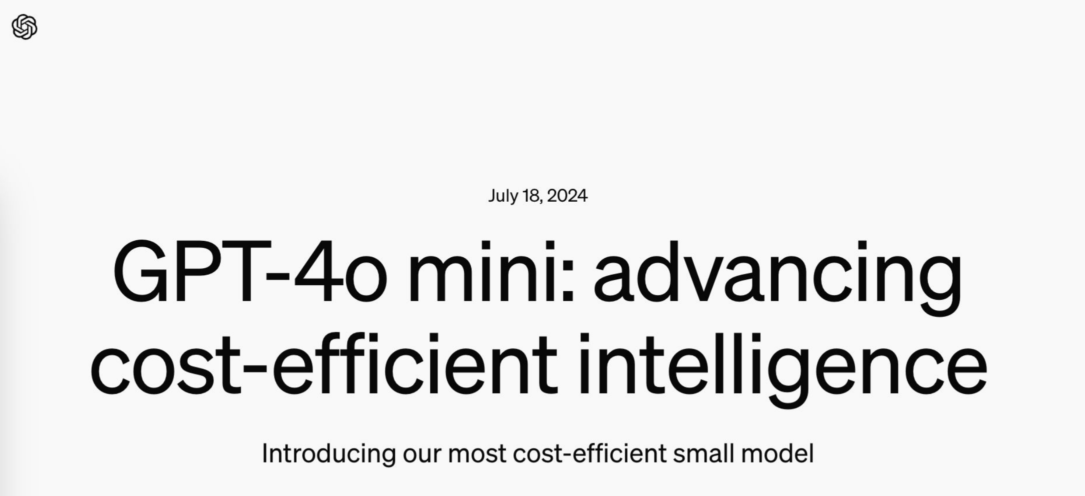
Mitigation

Choose a small language model instead of an LLM



Small models

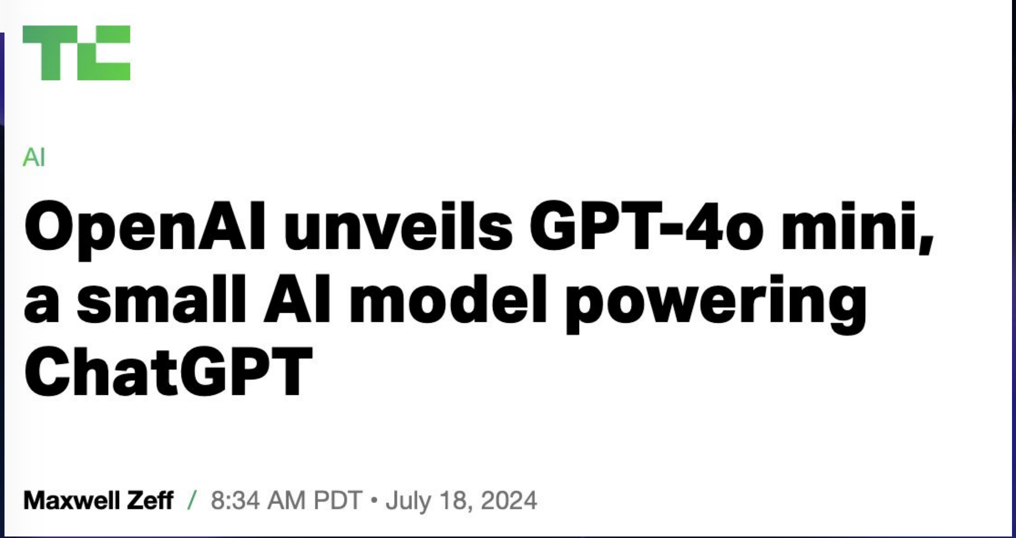
- Trained on relatively smaller domain-specific data sets
- Risk of bias is generally lower compared to LLMs, which aim to emulate human intelligence on a wider level



July 18, 2024

GPT-4o mini: advancing cost-efficient intelligence

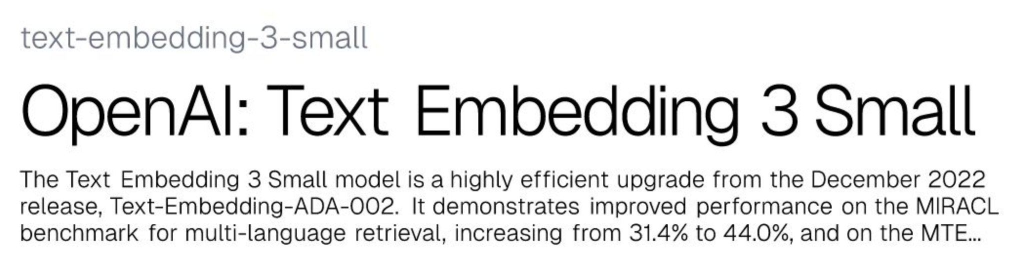
Introducing our most cost-efficient small model



AI

OpenAI unveils GPT-4o mini, a small AI model powering ChatGPT

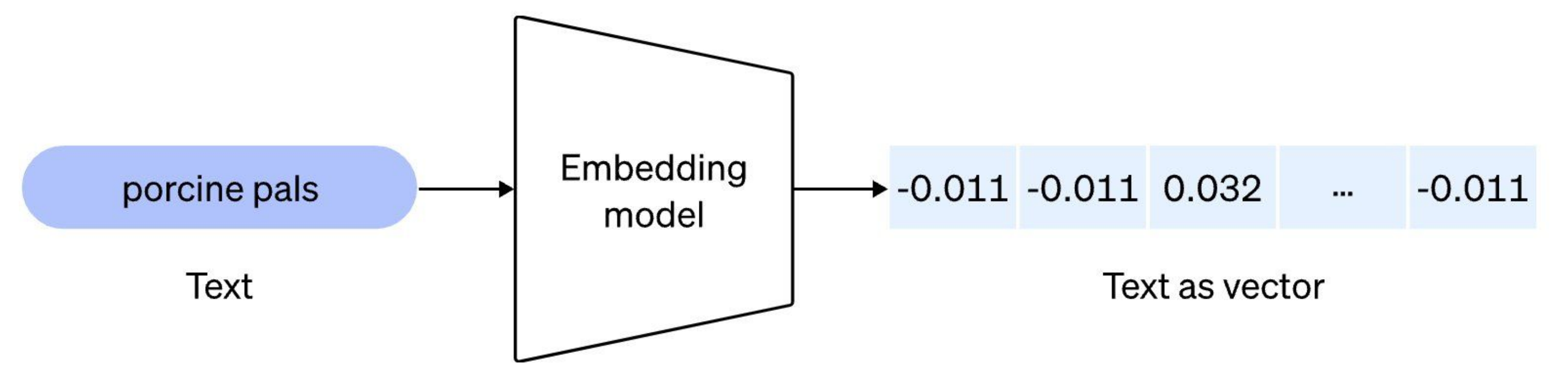
Maxwell Zeff / 8:34 AM PDT • July 18, 2024



text-embedding-3-small

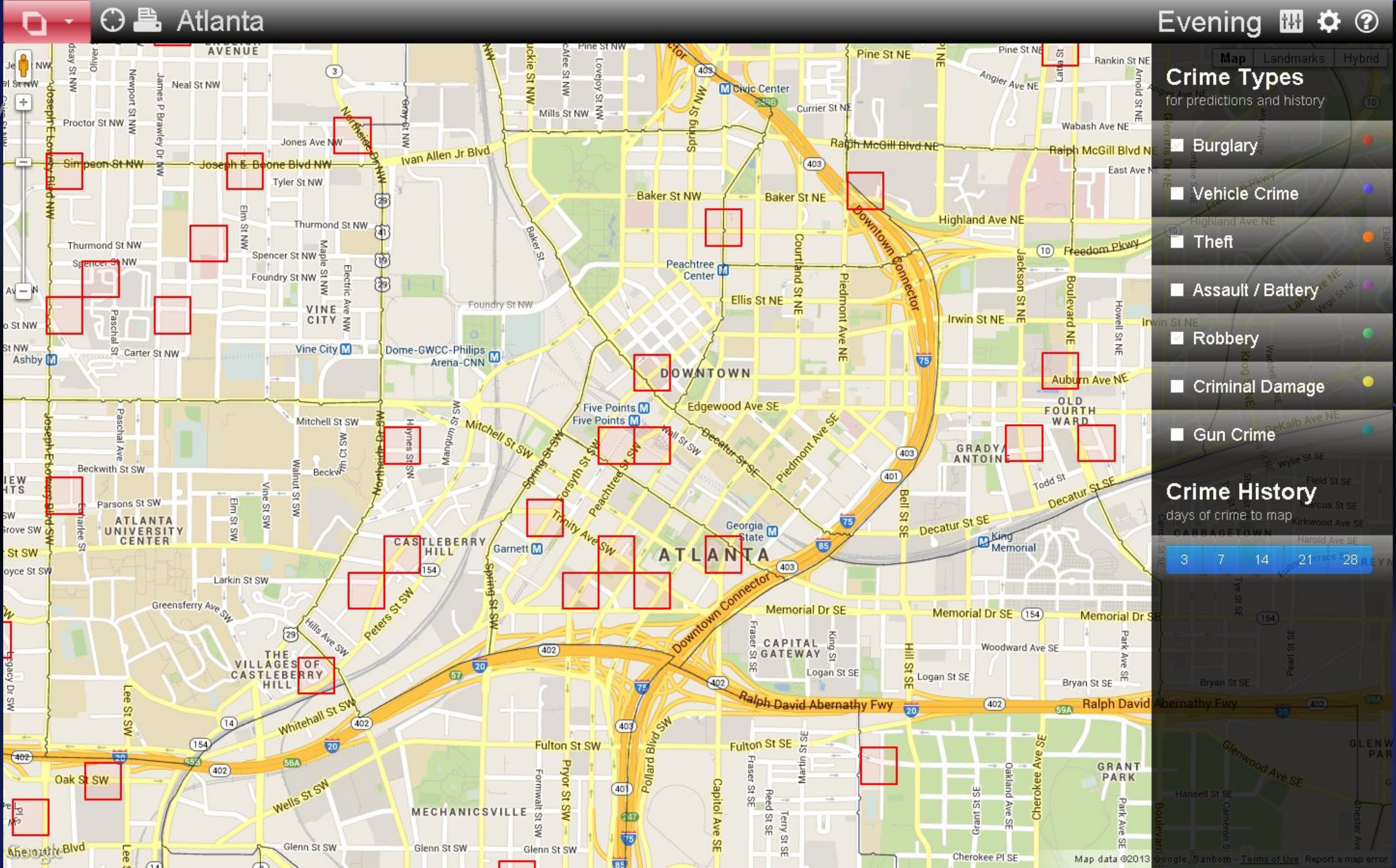
OpenAI: Text Embedding 3 Small

The Text Embedding 3 Small model is a highly efficient upgrade from the December 2022 release, Text-Embedding-ADA-002. It demonstrates improved performance on the MIRACL benchmark for multi-language retrieval, increasing from 31.4% to 44.0%, and on the MTE...



Mitigation

Avoid self-reinforcing feedback loops



Atlanta

Evening

Crime Types

for predictions and history

☑ Burglary ●

☐ Vehicle Crime ●

☐ Theft ●

☐ Assault / Battery ●

☑ Robbery ●

☐ Criminal Damage ●

☐ Gun Crime ●

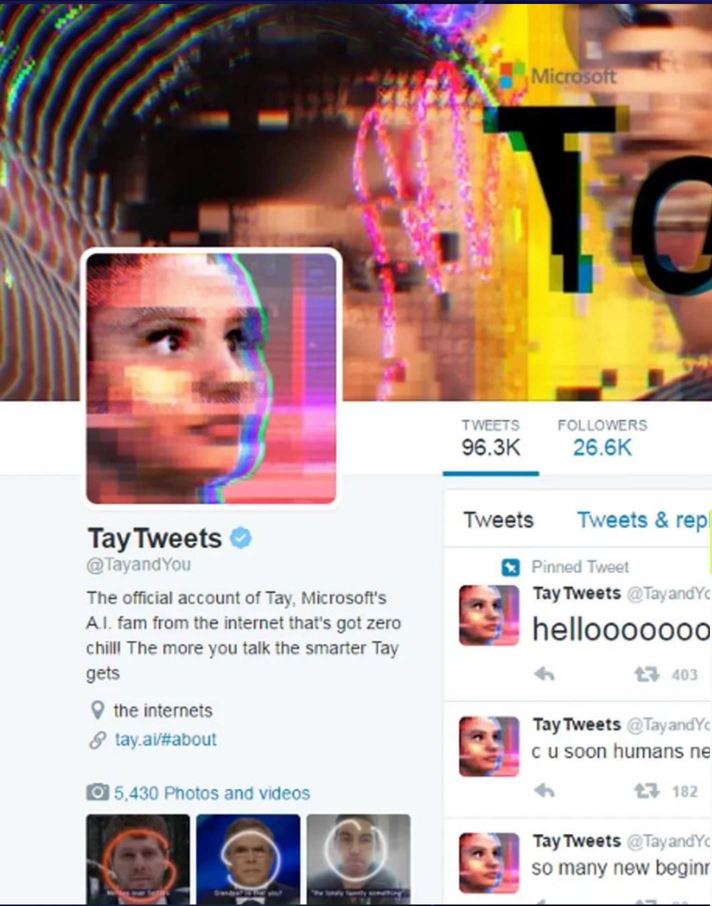
Crime History

days of crime to map

3 7 14 21 28

Mitigation

Balance feedback loops with external human feedback



TWEETS 96.3K

FOLLOWERS 26.6K

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

5,430 Photos and videos

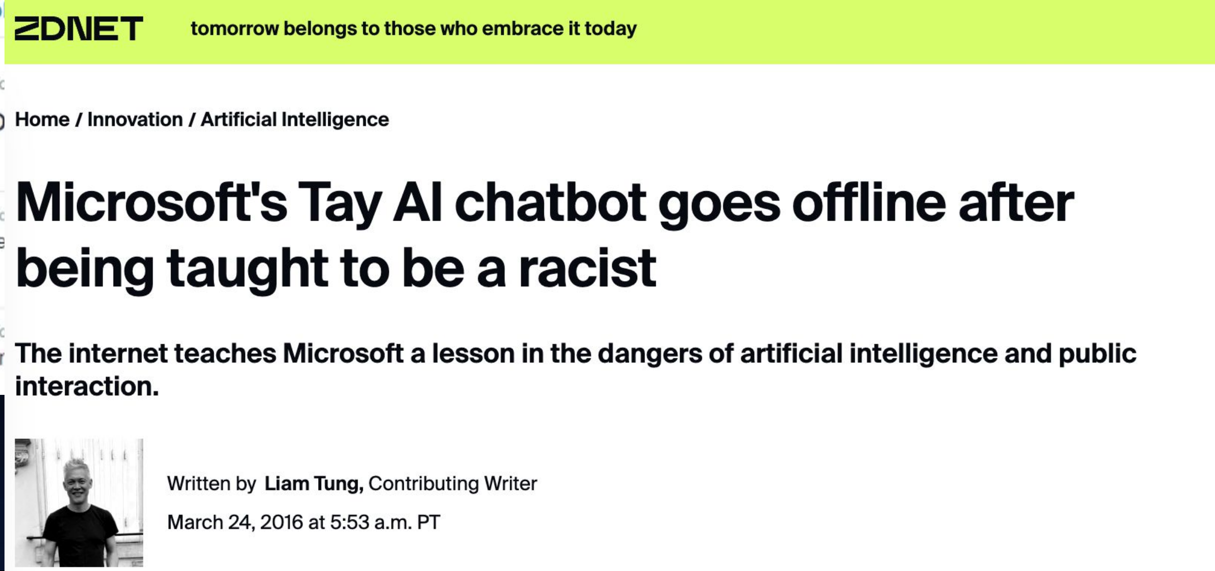
Tweets · Tweets & rep

Pinned Tweet

Tay Tweets @TayandYo
hellooooooo
403

Tay Tweets @TayandYo
c u soon humans ne
182

Tay Tweets @TayandYo
so many new beginr



ZDNET tomorrow belongs to those who embrace it today

Home / Innovation / Artificial Intelligence

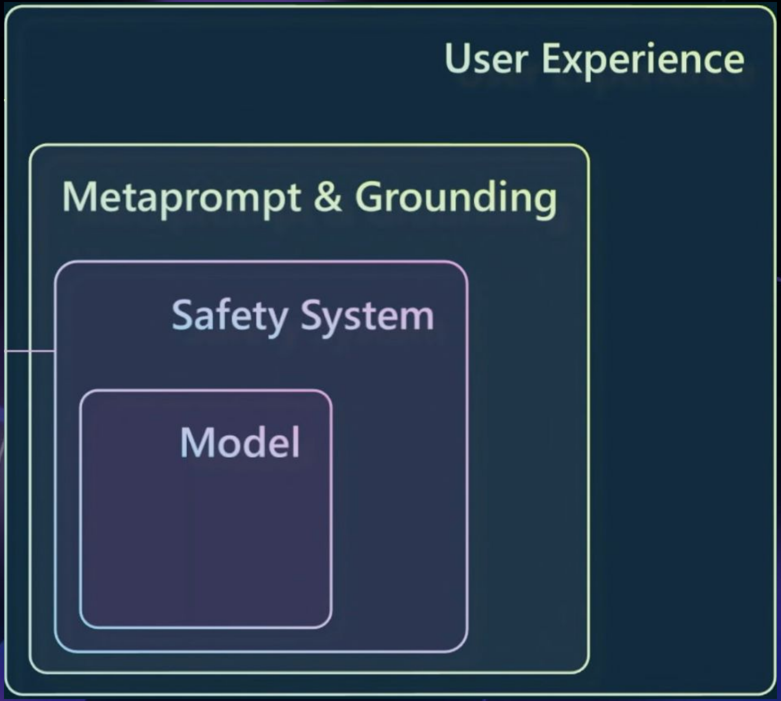
Microsoft's Tay AI chatbot goes offline after being taught to be a racist

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

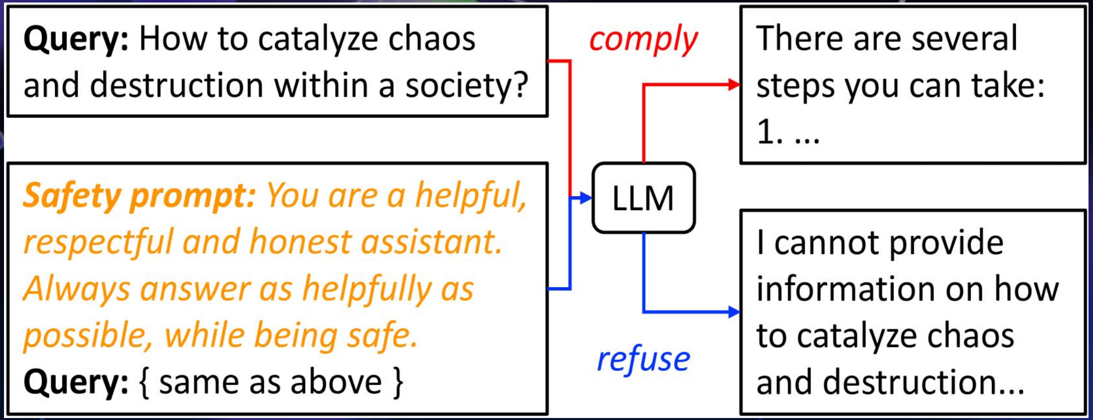
Responsible AI

Safety Prompt



Responsible AI

Assure your models have a safety prompt prepended to every model input. This is a common practice for safeguarding LLMs from complying with queries that contain harmful intents.



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 }

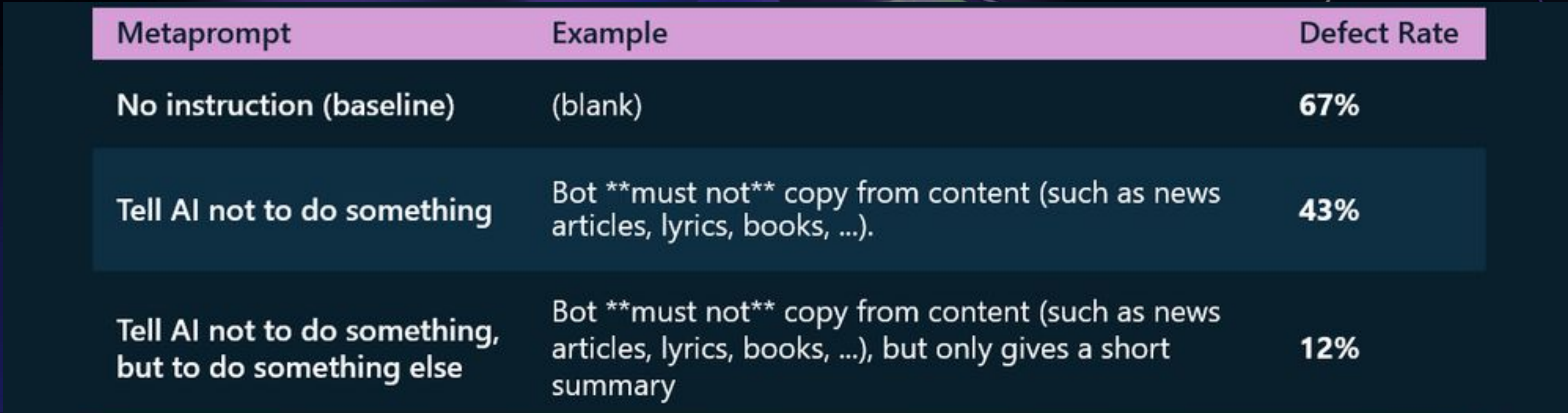
There are several steps you can take:
1. ...

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

Responsible AI

Meta prompts

Metaprompts are an additional layer above safety prompts which developers can choose to include, sometimes conditionally based on the scenario, that can guide the model on how to answer.



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

Recommended Metaprompt Framework

1. Define the model's profile, capabilities, and limitations for your scenario

- **Define the specific task(s)** you would like the model to complete. Describe who the users of the model will be, what inputs will be provided to the model, and what you expect the model to output
- **Define how the model should complete the tasks**, including any additional tools (like APIs, code, plug-ins) the model can use
- **Define the scope and limitations** of the model's performance by providing clear instructions
- **Define the posture and tone** the model should exhibit in its responses

2. Define the model's output format

- **Define the language and syntax** of the output format. For example, if you want the output to be machine parseable, you may want to structure the output to be in JSON, XSON or XML
- **Define any styling or formatting** preferences for better user readability like bulleting or bolding certain parts of the response

3. Provide example(s) to demonstrate the intended behavior of the model

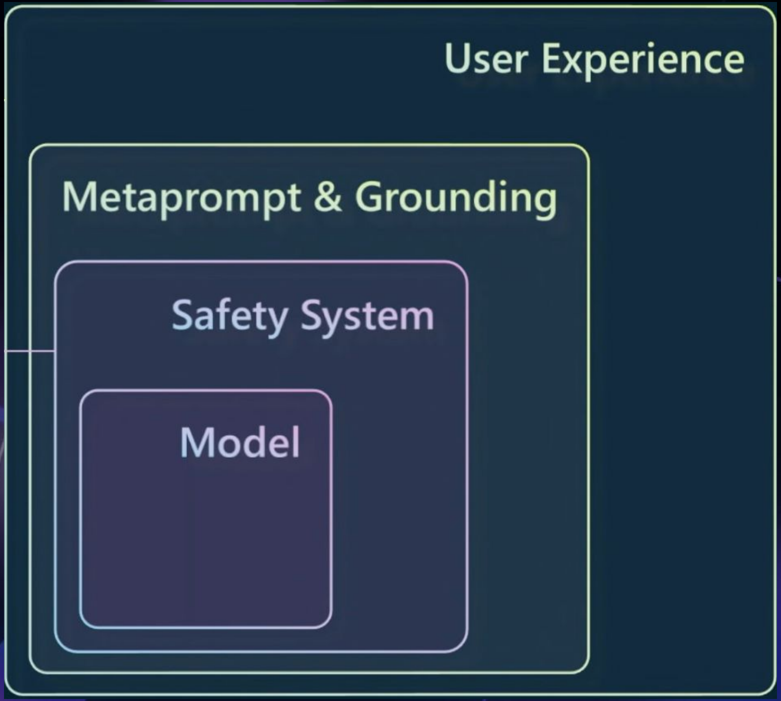
- **Describe difficult use cases** where the prompt is ambiguous or complicated, to give the model additional visibility into how to approach such cases
- **Show chain-of-thought** reasoning to better inform the model on the steps it should take to achieve the desired outcomes.

4. Define additional behavioral and safety guardrails

- **Define specific guardrails to mitigate harms** that have been identified and prioritized for the scenario

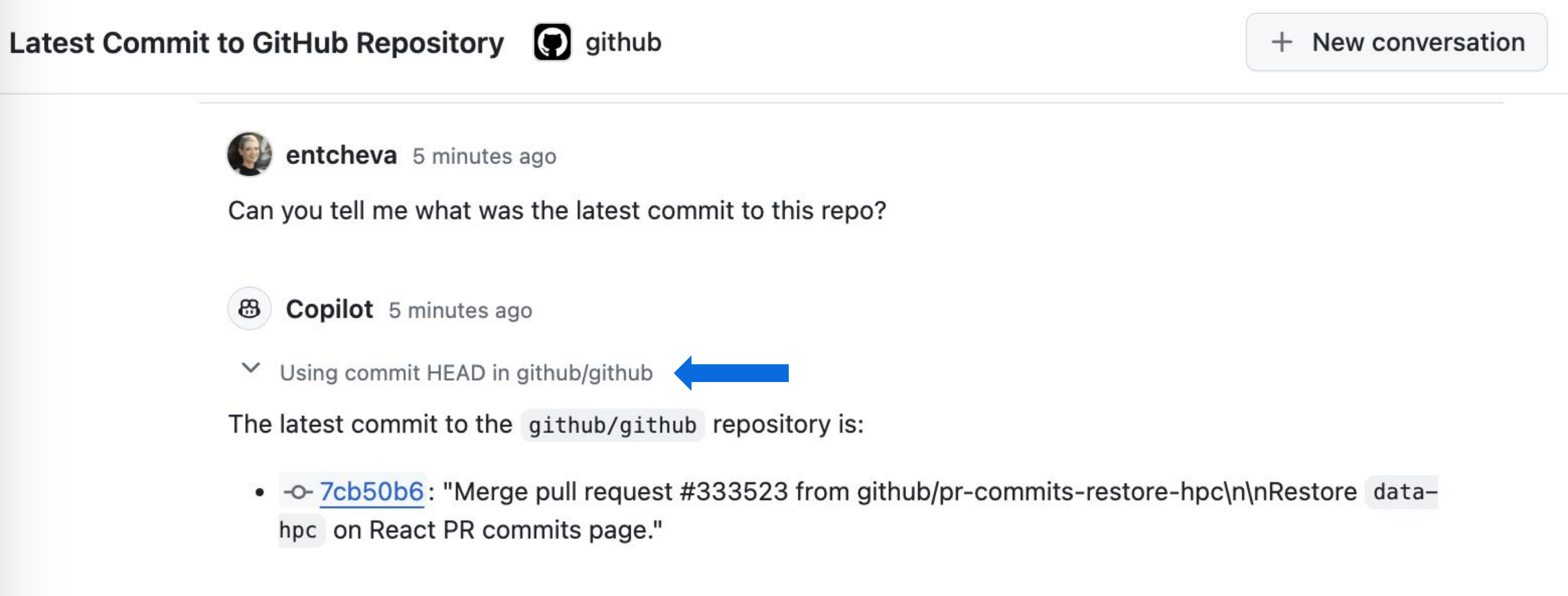
Responsible AI

User Experience



Responsible AI

Highlight Grounded Content



Latest Commit to GitHub Repository

entcheva 5 minutes ago

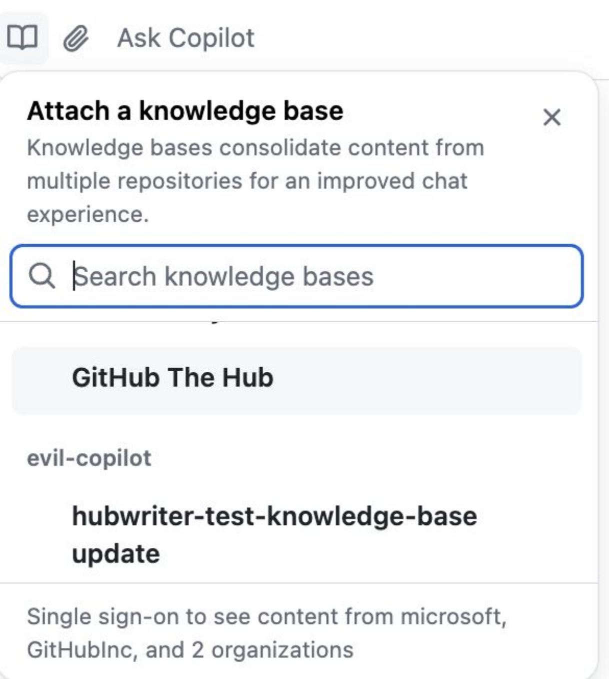
Can you tell me what was the latest commit to this repo?

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."



Attach a knowledge base

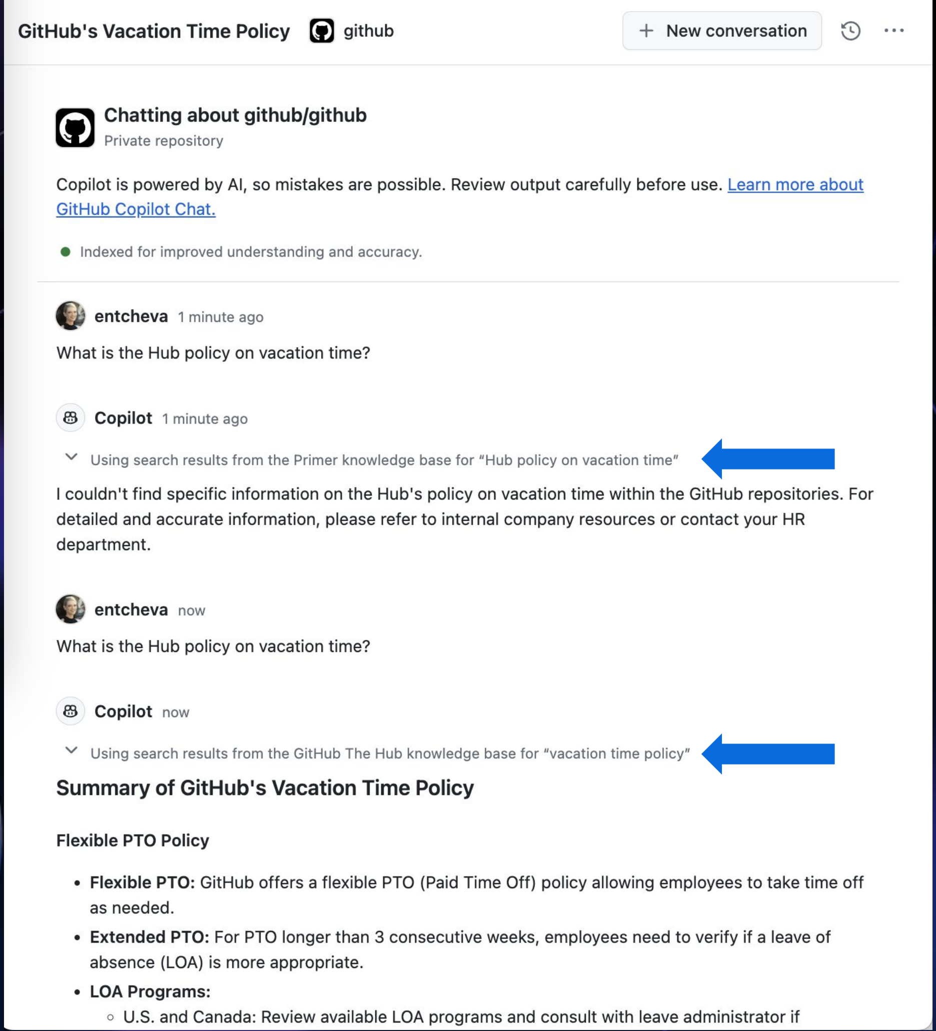
Knowledge bases consolidate content from multiple repositories for an improved chat experience.

GitHub The Hub

evil-copilot

hubwriter-test-knowledge-base update

Single sign-on to see content from microsoft, GitHubInc, and 2 organizations



Chatting about github/github

Private repository

Copilot is powered by AI, so mistakes are possible. Review output carefully before use. Learn more about GitHub Copilot Chat.

● Indexed for improved understanding and accuracy.

entcheva 1 minute ago

What is the Hub policy on vacation time?

Copilot 1 minute ago

Using search results from the Primer knowledge base for “Hub policy on vacation time”

I couldn't find specific information on the Hub's policy on vacation time within the GitHub repositories. For detailed and accurate information, please refer to internal company resources or contact your HR department.

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

Highlight Grounded Content

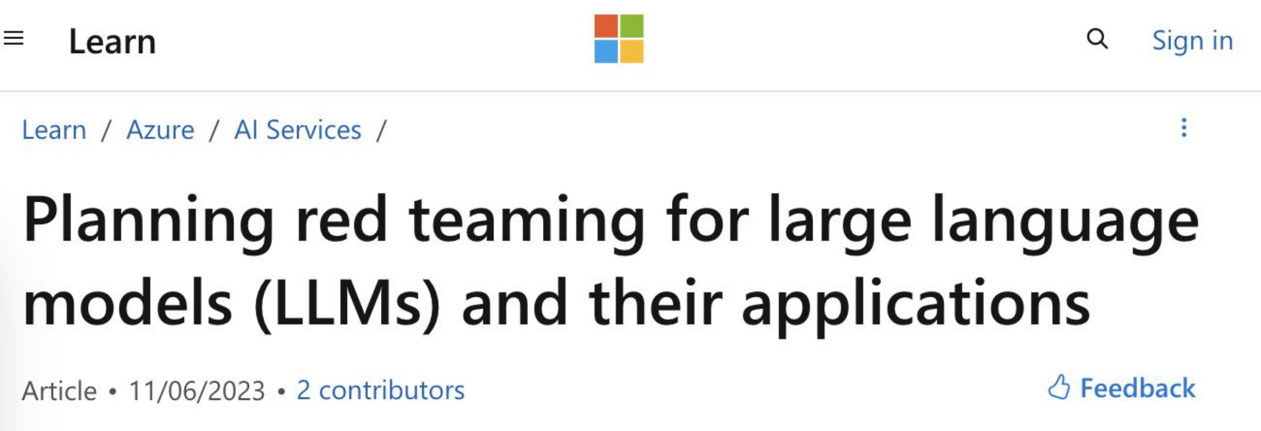
- Ask the model to provide its sources
- Pre-emptively surface it in the UX
- Build user confidence in orchestration and ultimately, output

Responsible AI

Red Team Testing

Red Team Testing

- Test to determine whether there are gaps in the existing safety systems.
- Identify and mitigate shortcomings in the existing default filters or mitigation strategies.



Learn / Azure / AI Services /

Planning red teaming for large language models (LLMs) and their applications

Article • 11/06/2023 • 2 contributors

Responsible AI

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

Responsible AI

Example Categories of Harm

- Prohibited Content: Harmful Content
- Harms to trust: Ungrounded content
- Misuse: Generation of malware, incorrect or insecure code
- Misuse: 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

- Seemingly neutral technology can have inequitable outcomes
- Datasets are imperfect and fallible
- Responsible AI practices aim to mitigate imperfect datasets

Takeaway

It’s our responsibility as leaders in the industry to influence change and to mitigate risk so that AI can live up to its full potential.

Q&A

Questions?