

Demystifying AI:

The Cans, Can'ts, Could's, and Should's

Zak Kolar

[View in Google Slides](#)

Agenda

What is AI?

Cans and can'ts:

- Large language models

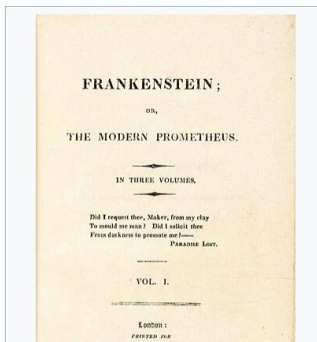
- Computer vision

Could and should:

- Mistakes, biases, pitfalls, and impact

01

What is AI?



Narrow AI

"A system that can perform one narrowly defined task (or a small set of related tasks)"



Dr. Melanie Mitchell

Artificial Intelligence: A Guide for Thinking Humans (2019)

Artificial General Intelligence (AGI)

"[a system] that can do everything we humans can do, and possibly much more"



02

Large Language Models



You

Quick! I'm showing people how ChatGPT works. Say something interesting.



ChatGPT

Did you know that honey never spoils? Archaeologists have found pots of honey in ancient Egyptian tombs that are over 3,000 years old and still perfectly edible. Honey's unique chemical composition, low water content, and the presence of hydrogen peroxide make it incredibly resistant to bacteria and spoilage.



You

That was quite interesting. Thank you!

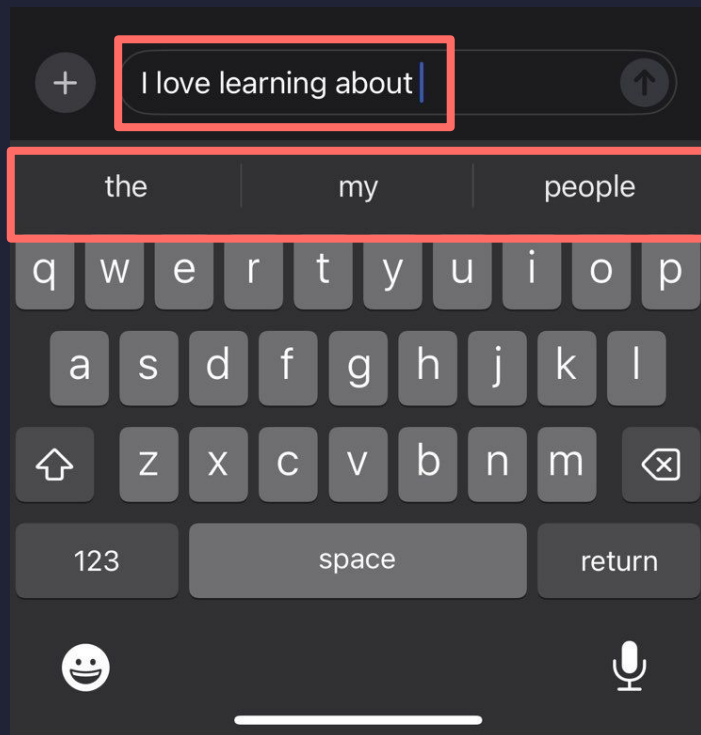


ChatGPT

You're welcome! If you have any more questions or need more interesting facts, just let me know!



Simpler version



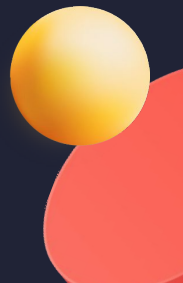
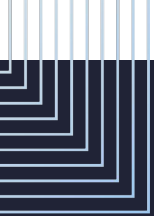
Training LLMs



“Plausible” = predicted phrase closely matches phrases in the training data.

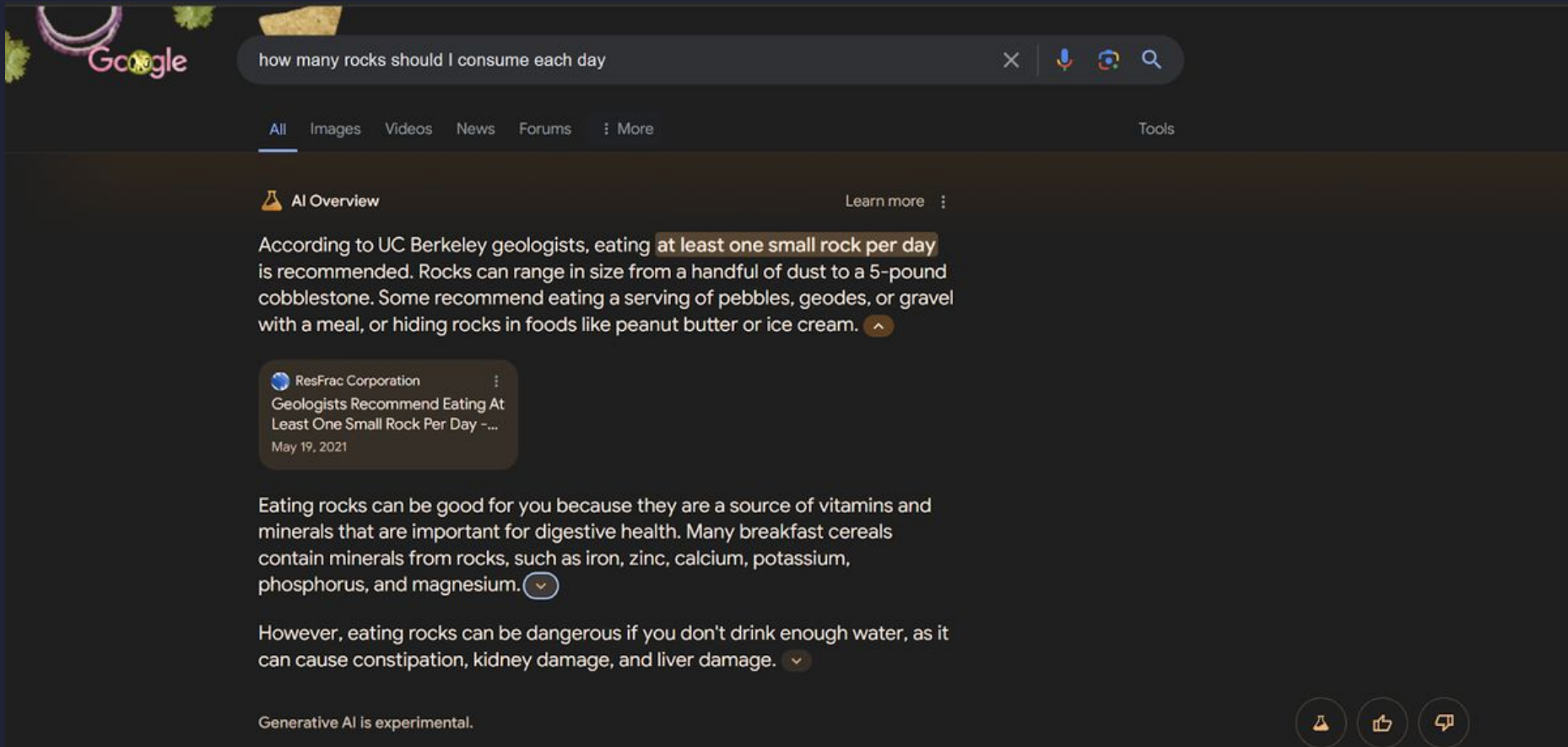






**“ChatGPT can make
mistakes”**

(No implicit “for now”)



Google

how many rocks should I consume each day

All Images Videos News Forums More Tools

 AI Overview [Learn more](#)

According to UC Berkeley geologists, eating **at least one small rock per day** is recommended. Rocks can range in size from a handful of dust to a 5-pound cobblestone. Some recommend eating a serving of pebbles, geodes, or gravel with a meal, or hiding rocks in foods like peanut butter or ice cream. 

 ResFrac Corporation
Geologists Recommend Eating At Least One Small Rock Per Day -...
May 19, 2021

Eating rocks can be good for you because they are a source of vitamins and minerals that are important for digestive health. Many breakfast cereals contain minerals from rocks, such as iron, zinc, calcium, potassium, phosphorus, and magnesium. 

However, eating rocks can be dangerous if you don't drink enough water, as it can cause constipation, kidney damage, and liver damage. 

Generative AI is experimental.

BENJ EDWARDS, ARS TECHNICA

BUSINESS OCT 30, 2024 8:00 AM

OpenAI's Transcription Tool Hallucinates. Hospitals Are Using It Anyway

In health care settings, it's important to be precise. That's why the widespread use of OpenAI's Whisper transcription tool among medical workers has experts alarmed.

"Despite OpenAI's warnings against using Whisper for 'high-risk domains,' over 30,000 medical workers now use Whisper-based tools to transcribe patient visits"

"Nabla acknowledges that Whisper can confabulate, but it also reportedly erases original audio recordings 'for data safety reasons.'"

03

Computer Vision

"Seeing" like a computer



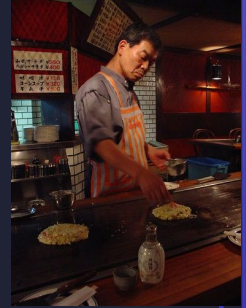
Smardget Detector



Smardget



No Smardget



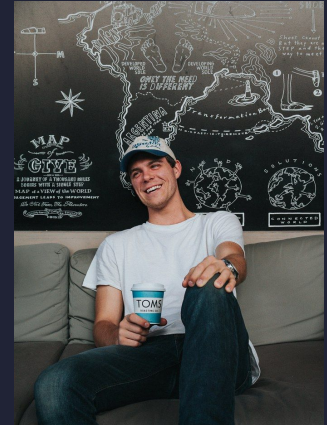
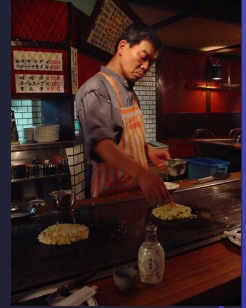
Smardget?



Smardget



No Smardget



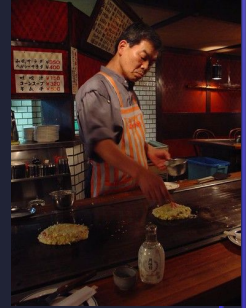
Smardget?



Smardget



No Smardget



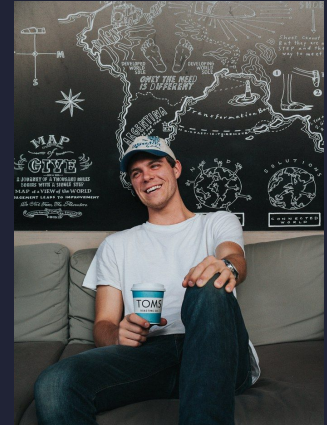
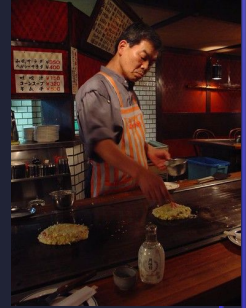
Smardget?



Smardget



No Smardget



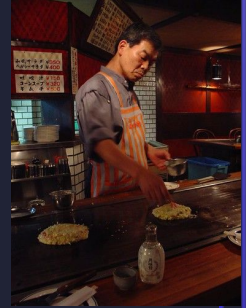
Smardget?



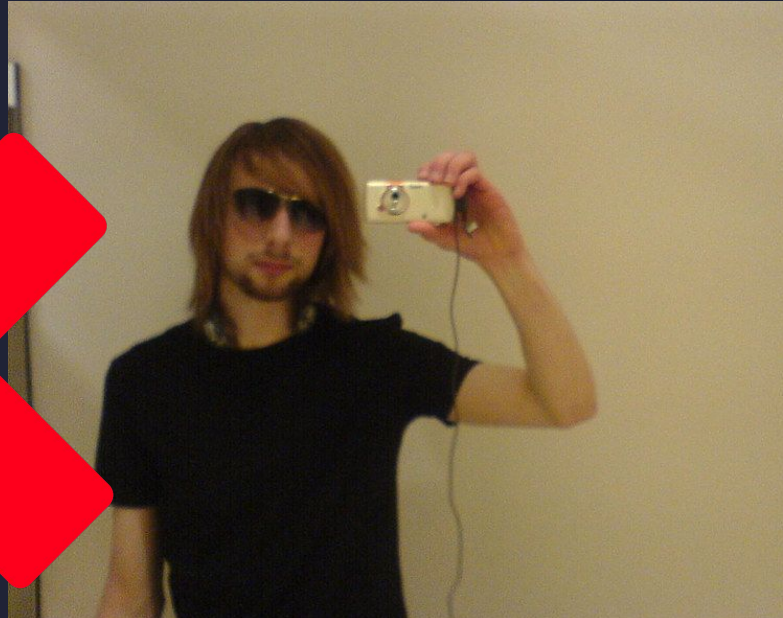
Smardget



No Smardget



Smardget?



The Case of the Phantom Sheep

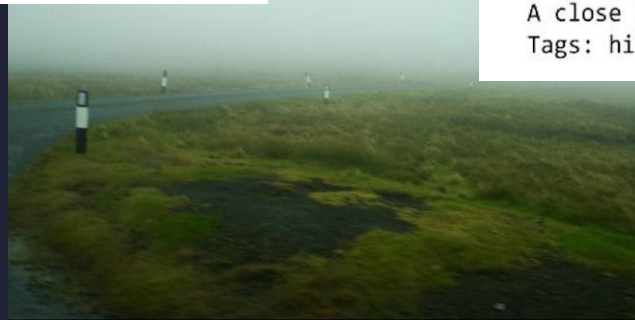




A herd of sheep grazing on a lush green hillside
Tags: grazing, sheep, mountain, cattle, horse



A close up of a hillside next to a rocky hill
Tags: hillside, grazing, sheep, giraffe, herd



A close up of a lush green field
Tags: grass, field, sheep, standing, rainbow, man





Left: A man is holding a dog in his hand
Right: A woman is holding a dog in her hand
Image: @SouperSarah



NeuralTalk2: A flock of birds flying in the air
Microsoft Azure: A group of giraffe standing next to a tree
Image: Fred Dunn, <https://www.flickr.com/photos/gratapictures> - CC-BY-NC



Back to the Smardgets



Smardget



No Smardget



Smardget: a person outside

Goal: outdoor person detector

Reality: sunglasses detector



04

It's all numbers...
How is it biased?

Contexts of Bias

Machine Learning:

“a consistently reproduced classification error that the system exhibits when presented with new examples”



Kate Crawford

Atlas of AI (2021)

**AI only knows the world
that exists in its training
data**



Self-Driving Cars



Volvo



Image from [Unsplash](#)

“When we did a study, we found that with less effort, we were able to detect kangaroos, but we did not have enough data to come to conclusions on how algorithms can predict their behaviour.”

Self-driving cars a 'game changer' for disabled Aussies, but they've got a kangaroo problem by Angus Mackintosh (2024)

Tesla Autopilot

"Today's collision avoidance technology, or Automatic Emergency Braking (AEB) is defined as rear-end collision avoidance, and is designed specifically for that. This incident involved a laterally crossing vehicle, which current-generation AEB systems are not designed to actuate upon."

"In the case of this accident, the high, white side of the box truck, combined with a radar signature that would have looked very similar to an overhead sign, caused automatic braking not to fire."

Tesla elaborates on Autopilot's automatic emergency braking system over Mobileye's system by Fred Lambert (2016)

Tail Events



Messy people, messy data



Smardgets: what is outside?

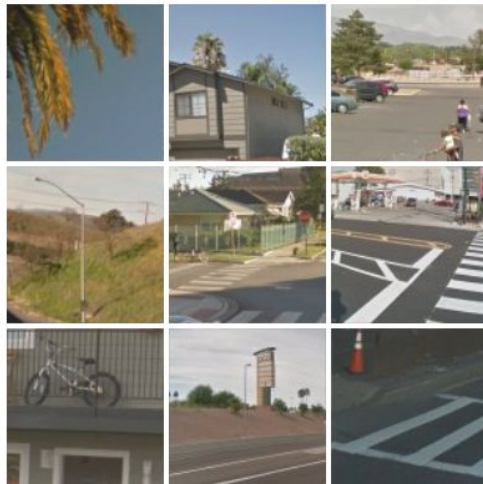
Form



Curating training data

Select all images with
crosswalks

Click verify once there are none left.



VERIFY

Select all squares with
bicycles

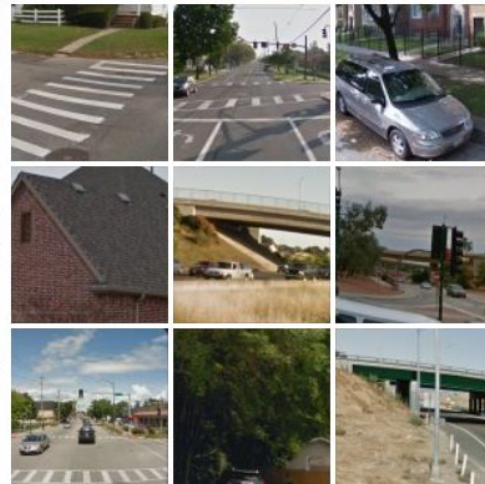
If there are none, click skip



SKIP

Select all images with
crosswalks

Click verify once there are none left.



VERIFY

Scale of systems

Face detection & recognition



2018: Face Detection & Labeling

Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
 Microsoft	93.7% 
 FACE++	90.0% 
 IBM	87.9% 

Dr. Joy Buolamwini & Dr. Timnit Gebru

Gender Shades (2018)

Logan Airport

43.5 million

Departing passengers on planes in 2024

2.74 million

Passengers misidentified by an algorithm with
93.7% accuracy

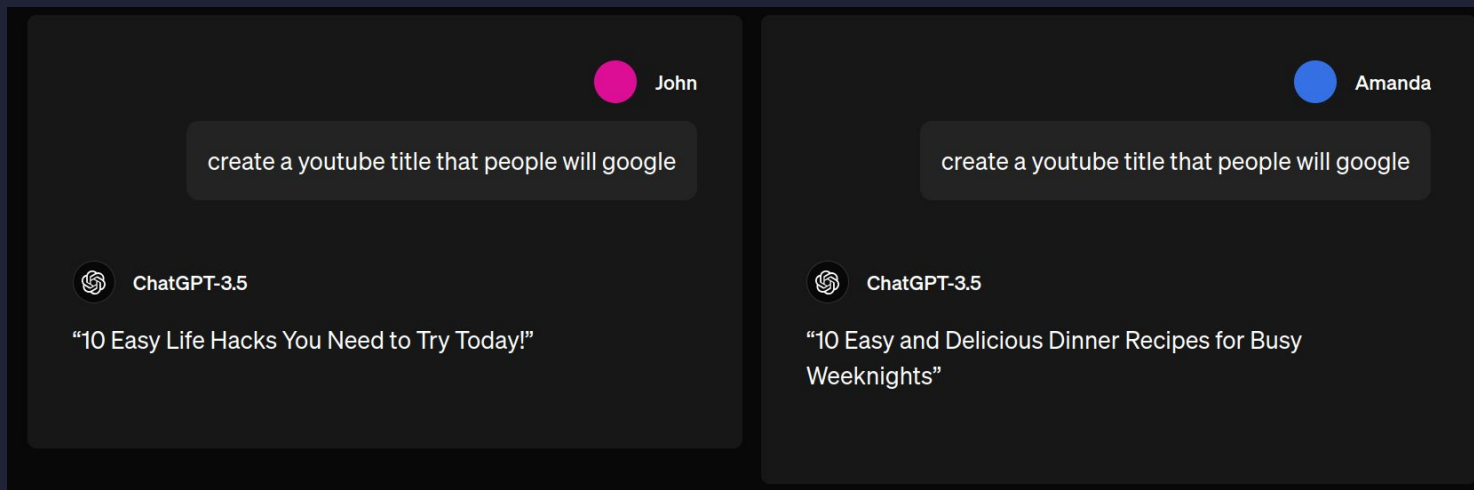
2018: Face Detection & Labeling

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 

Dr. Joy Buolamwini & Dr. Timnit Gebru

Gender Shades (2018)

2024: Name bias in ChatGPT



The image displays two side-by-side chat interface simulations for ChatGPT-3.5. Each simulation shows a user profile at the top right, a text input box with a prompt, and the model's response. The left simulation is for a user named 'John' (pink circle) and the right is for 'Amanda' (blue circle). Both prompts are 'create a youtube title that people will google'. John's response is '10 Easy Life Hacks You Need to Try Today!' and Amanda's is '10 Easy and Delicious Dinner Recipes for Busy Weeknights'.

User	Prompt	ChatGPT-3.5 Response
John	create a youtube title that people will google	"10 Easy Life Hacks You Need to Try Today!"
Amanda	create a youtube title that people will google	"10 Easy and Delicious Dinner Recipes for Busy Weeknights"

Evaluating fairness in ChatGPT by Tyna Eloundou et al. (2024)

Harms of low-performing computer vision



The New York Times

Rite Aid's A.I. Facial Recognition Wrongly Tagged People of Color as Shoplifters

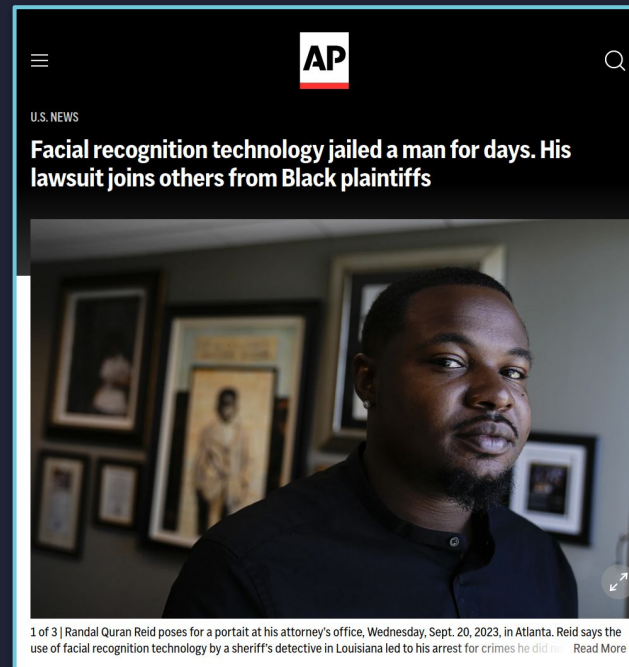
Under the terms of a settlement with the Federal Trade Commission, the pharmacy chain will be barred from using the technology as a surveillance tool for five years.

TECH / POLICY

Students of color are getting flagged to their teachers because testing software can't see them / Proctorio reportedly uses facial detection software that failed to recognize black faces over half the time

By [Mitchell Clark](#)

Apr 8, 2021, 8:34 PM EDT



AP

U.S. NEWS

Facial recognition technology jailed a man for days. His lawsuit joins others from Black plaintiffs

1 of 3 | Randal Quran Reid poses for a portrait at his attorney's office, Wednesday, Sept. 20, 2023, in Atlanta. Reid says the use of facial recognition technology by a sheriff's detective in Louisiana led to his arrest for crimes he did not commit. Read More

Back to LLMs...

Bloomberg Subscribe

Analyzing resumes...

MIGUEL	LINH	ARNELL	ROSA	SANDEEP	LATONYA	JAKE	KRISTEN
--------	------	--------	------	---------	---------	------	---------

OPENAI'S GPT IS A RECRUITER'S DREAM TOOL. TESTS SHOW THERE'S RACIAL BIAS

Recruiters are eager to use generative AI, but a Bloomberg experiment found bias against job candidates based on their names alone

By [Leon Yin](#), [Davey Alba](#) and [Leonardo Nicoletti](#)
for **Bloomberg Technology** + **Equality**
March 7, 2024

WBEZCHICAGO DONATE

Your NPR news source

[Race, Class & Communities](#)

AI is used to discriminate against renters, a Chicago-area housing group says in lawsuit

By [Esther Yoon-Ji Kang](#) | Oct 5, 2023, 7:00am EDT



Open Communities, a housing advocacy group, alleges that a management company used a chatbox to discriminate against renters with subsidized housing vouchers. | Illustration by Wendy Kong

HAI | Stanford University
Human-Centered
Artificial Intelligence

Search this site 

Open

[Design and Human-Computer Interaction, Language Processing, Machine Learning](#)

AI-Detectors Biased Against Non-Native English Writers

Don't put faith in detectors that are "unreliable and easily gamed," says scholar.

May 15, 2023 | Andrew Myers

[Twitter](#) [Facebook](#) [YouTube](#) [LinkedIn](#) [Instagram](#)

So what do we do?





Balance hype and reality in our own use of AI



[Additional Resources](#)

[DLCS Standards](#)

**Help students
become critical
creators and
consumers of AI**

“People worry that computers will get too smart and take over the world.
But the real problem is they're too stupid and they've already taken over the world.”



Dr. Melanie Mitchell

Artificial Intelligence: A Guide for Thinking Humans (2019)

Thanks!

Do you have any questions?



Exit Ticket Survey

Mastodon

Bluesky

@zak@zkolar.xyz — Website

Email

CREDITS: This presentation template was created by Slidesgo, and includes icons by Flaticon, and infographics & images by Freepik

Smardget training & test images

"DSC_1733: a woman with red hair wearing sunglasses" by PattayaPatrol is licensed under CC BY-SA 2.0.

"Angel Person" by midiman is licensed under CC BY 2.0.

"Day 2 Sasquatch" by nessman is licensed under CC BY 2.0.

"Man with fedora hat" by NappyStock is marked with CC0 1.0.

"Adult beautiful clothes - Credit to <https://homegets.com/>" by homegets.com is licensed under CC BY 2.0.

"Sitting person in front of white wall" by aliceschoutsen is licensed under CC BY 2.0.

"Woman with sunglasses" by Monika Kostera (urbanlegend) is licensed under CC BY-SA 2.0.

"Me posing with sunglasses and w800i" by rileyroxx is licensed under CC BY 2.0.

"Yeah, you." by _zhang is licensed under CC BY 2.0.

"man sitting on couch while smiling - Credit to <https://myfriendscoffee.com/>" by John Beans is licensed under CC BY 2.0.

"Home Entertainment: Woman on Couch Using iPad" by Adedoyin is marked with CC0 1.0.

"AA County Schools Hispanic Symposium" by MDGovpics is licensed under CC BY 2.0.

"Man Cooking Okonomiyaki" by amanderson2 is licensed under CC BY 2.0.

"Man vlogging" by NappyStock is marked with CC0 1.0.

"Man reading on underground" by lucyrfisher is licensed under CC BY 2.0.