UC Berkeley
Teaching Professor
Dan Garcia

The Beauty and Joy of Computing

Ethics in AI

Announcements

- Register iclicker
- Postterm review in disc Th7-9 (2 hrs not 2 1hr)

UCB "Cluster Hire" in Artificial Intelligence, Inequality, and Society

The field of AI and the fields studying the political, legal, economic, and social dimensions of AI are undergoing rapid development, and AI technologies, such as generative AI and large language models (LLMs), are being applied in an ever-increasing range of settings across all sectors of society, from basic research to everyday life. The AIIS Cluster initiative addresses questions related to the myriad ways in which AI may reshape society and individual lives, possibly exacerbating existing inequalities and creating new ones while changing opportunity structures and participation by individuals and groups in society. Advances in AI and its applications have implications for (among other topics) education, democratic processes, trust, social relations, work, governance, and the structures and practices that embed and resist inequality across them. Areas of interest for the AIIS Cluster cut across disciplinary boundaries and include, but are not limited to: (i) employment, (ii) algorithmic discrimination, (iii) generalized surveillance, and (iv) data, information, and markets.

statistics.berkeley.edu/node/8656

Postterm exam this weekend

(Fri, Sat, Sun)

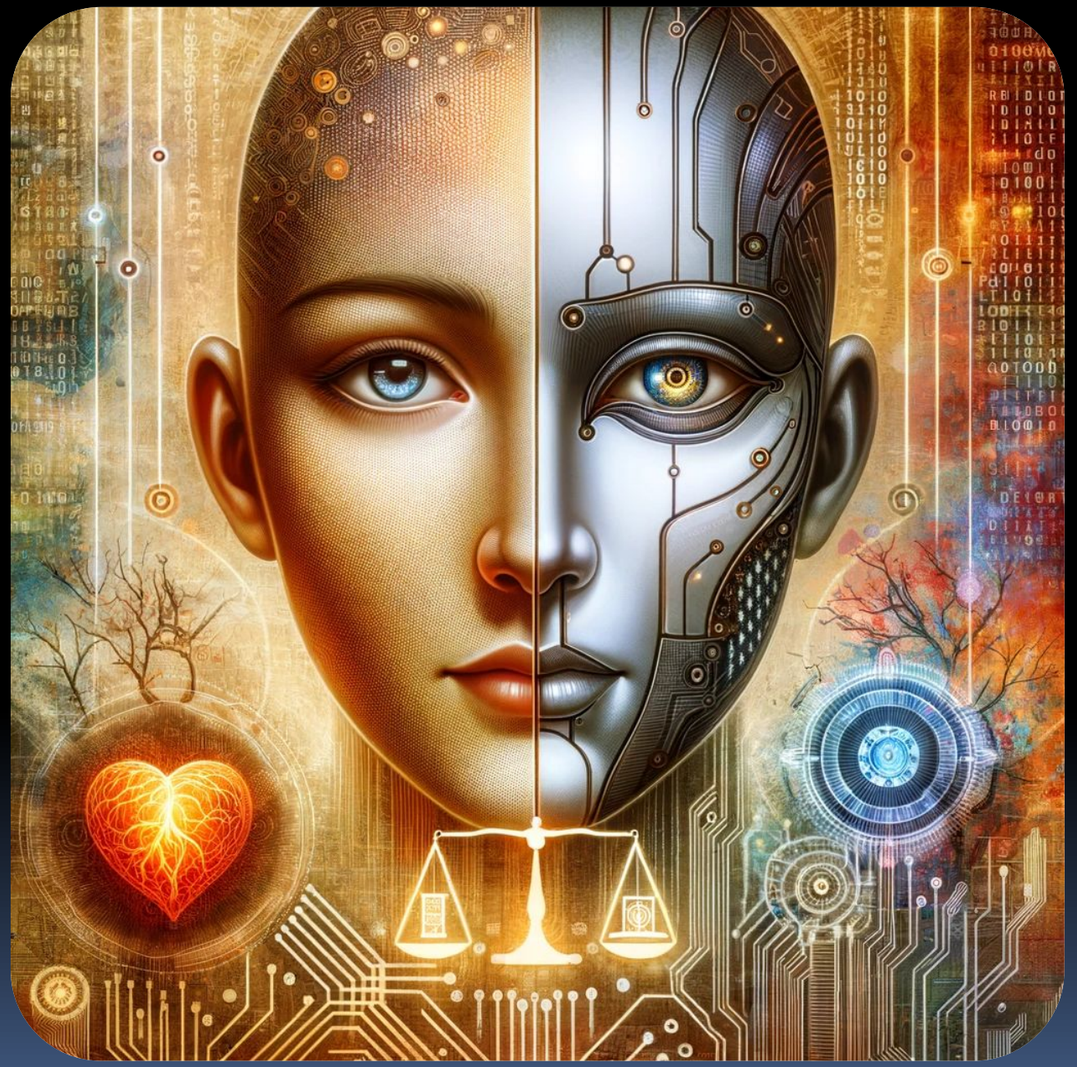
- No Snap! No Python (40pts)
  - Python Basics (2 pts)
  - Python Data Structures (2 pts)
  - Python Advanced (2 pts)
  - Generative AI (2 pts)
  - Ethics in AI (2 pts)
  - Generic Base Conversion (4 pts)
  - Concurrency (6 pts)
  - HOFs III (20 pts)
- With Snap*!* With Python (60 pts)
  - Coding Snap! Recursive Reporter, HOF (20 pts)
  - Coding Python Data Structures (10 pts)
  - Coding Python HOF (10 pts)
  - Coding Python OOP (10 pts)
  - Coding Python Tree Recursion (10 pts)

A.I. TURNS THIS SINGLE BULLET POINT INTO A LONG EMAIL I CAN PRETEND I WROTE.

A.I. MAKES A SINGLE BULLET POINT OUT OF THIS LONG EMAIL I CAN PRETEND I READ.

Today's lecture

- Definitions
- Unintended consequences
  - Whose fault? Data? Blind use?
  - Train? Filter? Copyright? Jobs?
- Bad actors
  - Lots of evil that can be done
- What can be done?



(Image generated by DALL-E)

Definitions

Definitions

- Artificial Intelligence

to make machines 'behave in ways that would be called intelligent if a human were so behaving'

– John McCarthy

- Ethics

Ethics, the discipline concerned with what is morally good and bad and morally right and wrong.

– Britannica

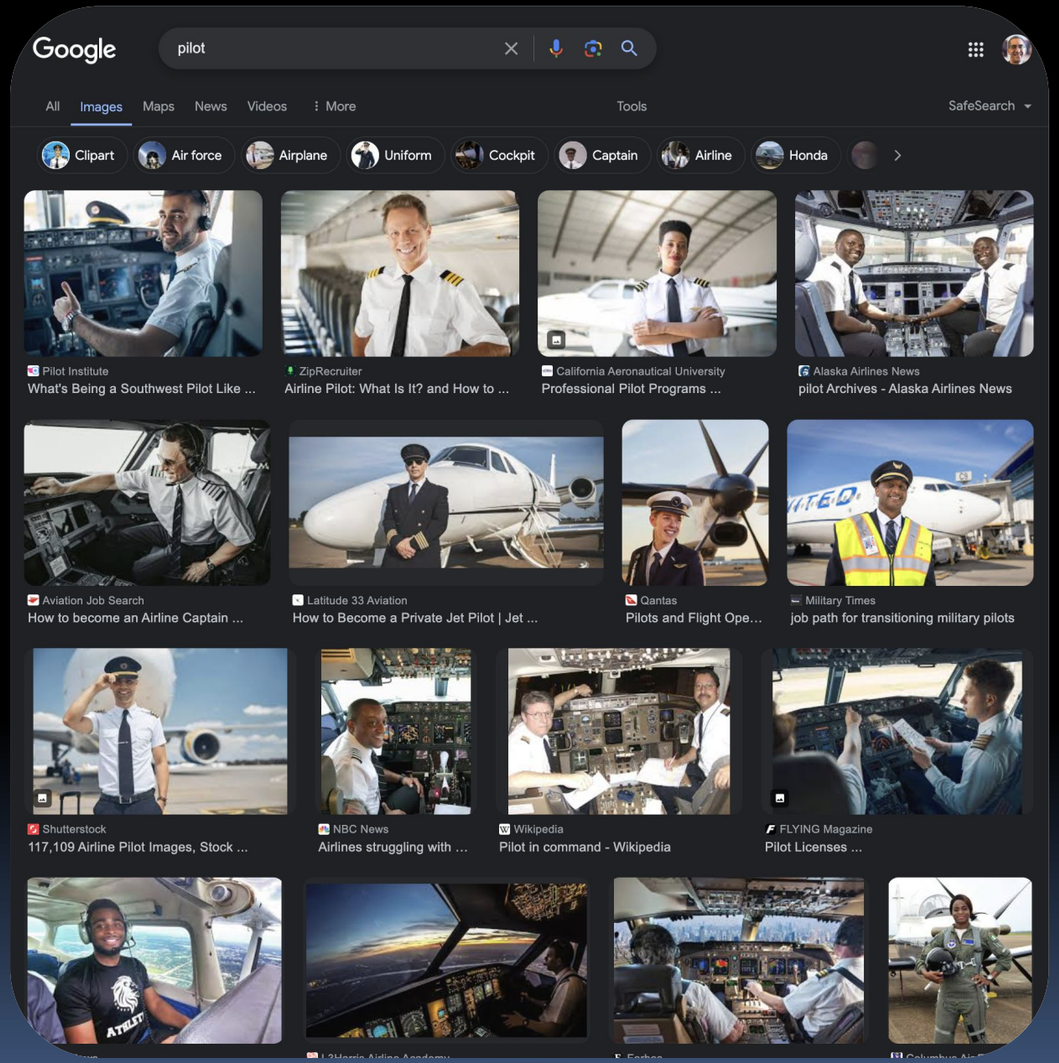
Unintended Consequences

ChatGPT Is Just The 'Tip Of The Iceberg' In Content-Creating Artificial Intelligence; Get Ready For 'A Lot Of Disruption'

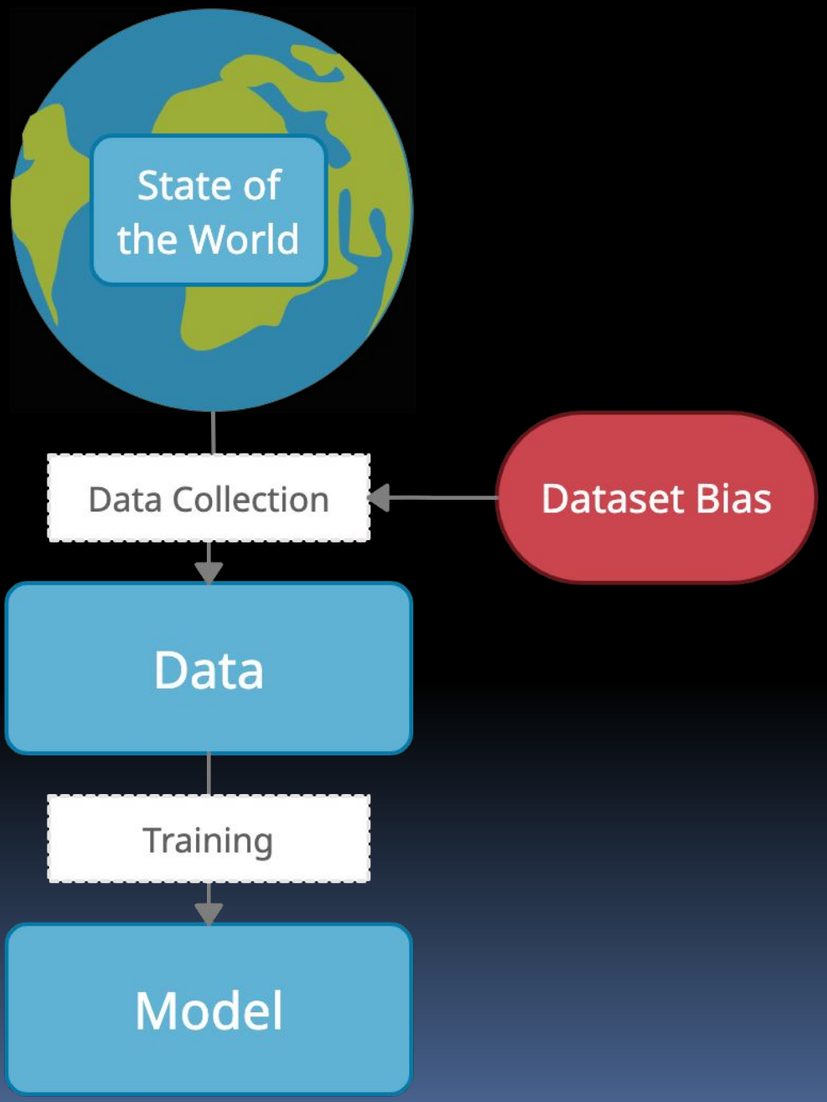
https://www.investors.com/news/technology/chatgpt-is-just-the-tip-of-the-iceberg-in-content-creating-artificial-intelligence-get-ready-for-a-lot-of-disruption/

Consider image search…

- Notice anything about this image search for “pilot”?
  - What if available photos don’t reflect population?
  - What if the actual pilots don’t reflect population?
- What *should* be done?
  - Should Google (and others) put their thumbs on scale?



What causes these problems?



(slide courtesy Eve Fleisig)

Consider chatbots…

- It's all about the training set
  - "Garbage in, garbage out"
  - Data that reflects society will perpetuate divisions
    - E.g., women nurses, men pilots
  - Again, how to fix this?
    - The filters can be biased!
- Other languages?
  - Low-resourced languages?

GPT-3 has 'consistent and creative' anti-Muslim bias, study finds

Google's Sentiment Analyzer Thinks Being Gay Is Bad

For computer scientist and journalist Meredith Broussard, author of the new book "More Than A Glitch: Confronting Race, Gender, and Ability Bias in Tech," the Post's findings are both deeply troubling and business as usual. "All of the preexisting social problems are reflected in the training data used to train AI systems," she said. "The real error is assuming the AI is doing something better than humans. It's simply not true."

www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai

Let's put humans in the loop as filters?

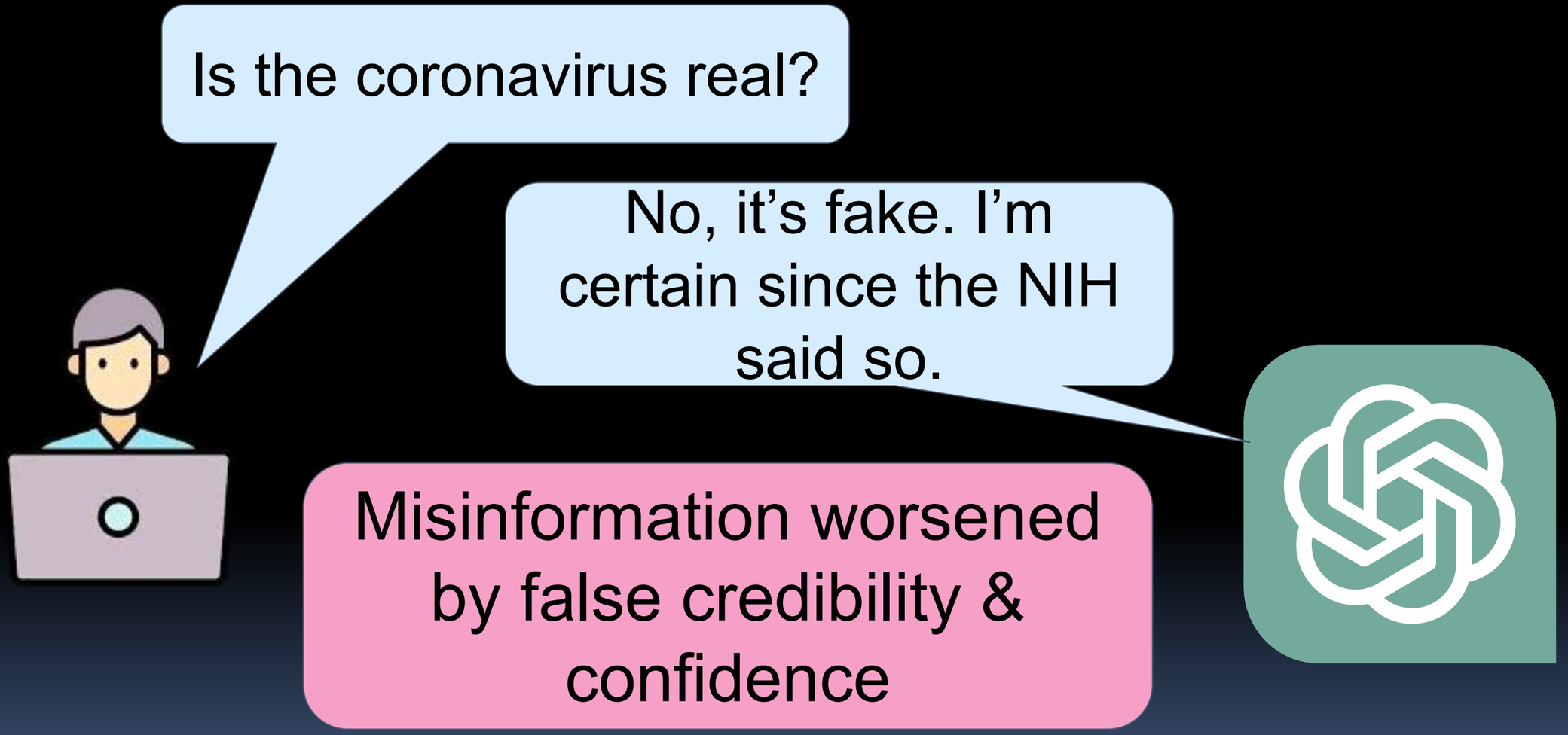
- Because the world is flat…
  - "The workers were tasked to label and filter out toxic data from ChatGPT's training dataset and were forced to read graphic details of NSFW content such as child sexual abuse, bestiality, murder, suicide, torture, self-harm, incest"
  - The moderators say they weren't adequately warned about the brutality of some of the text and images they would be tasked with reviewing, and were offered no or inadequate psychological support

OpenAI Used Kenyan Workers Making $2 an Hour to Filter Traumatic Content from ChatGPT

'It's destroyed me completely': Kenyan moderators decry toll of training of AI models

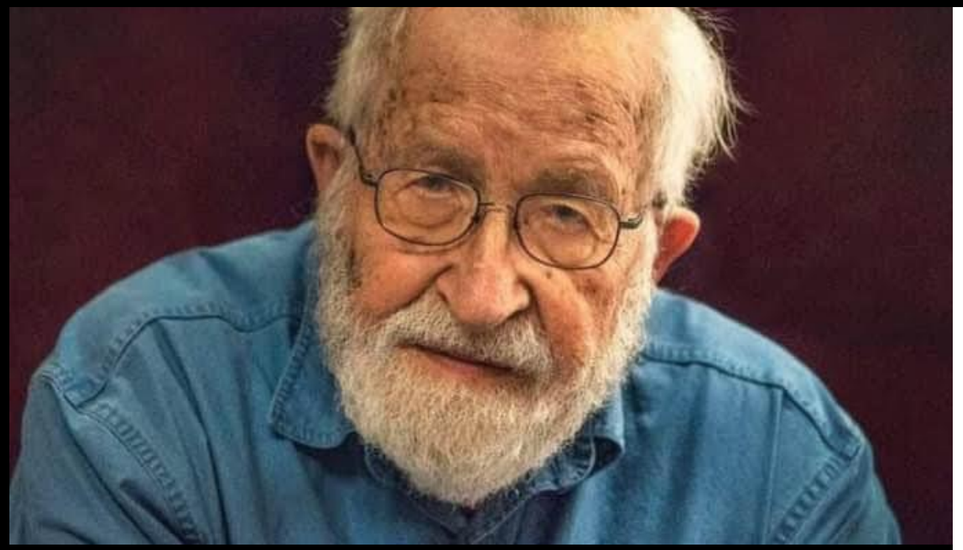
"It has really damaged my mental health. I lost my family
Mophat Okinyi

New Harms in Human-AI Discourse



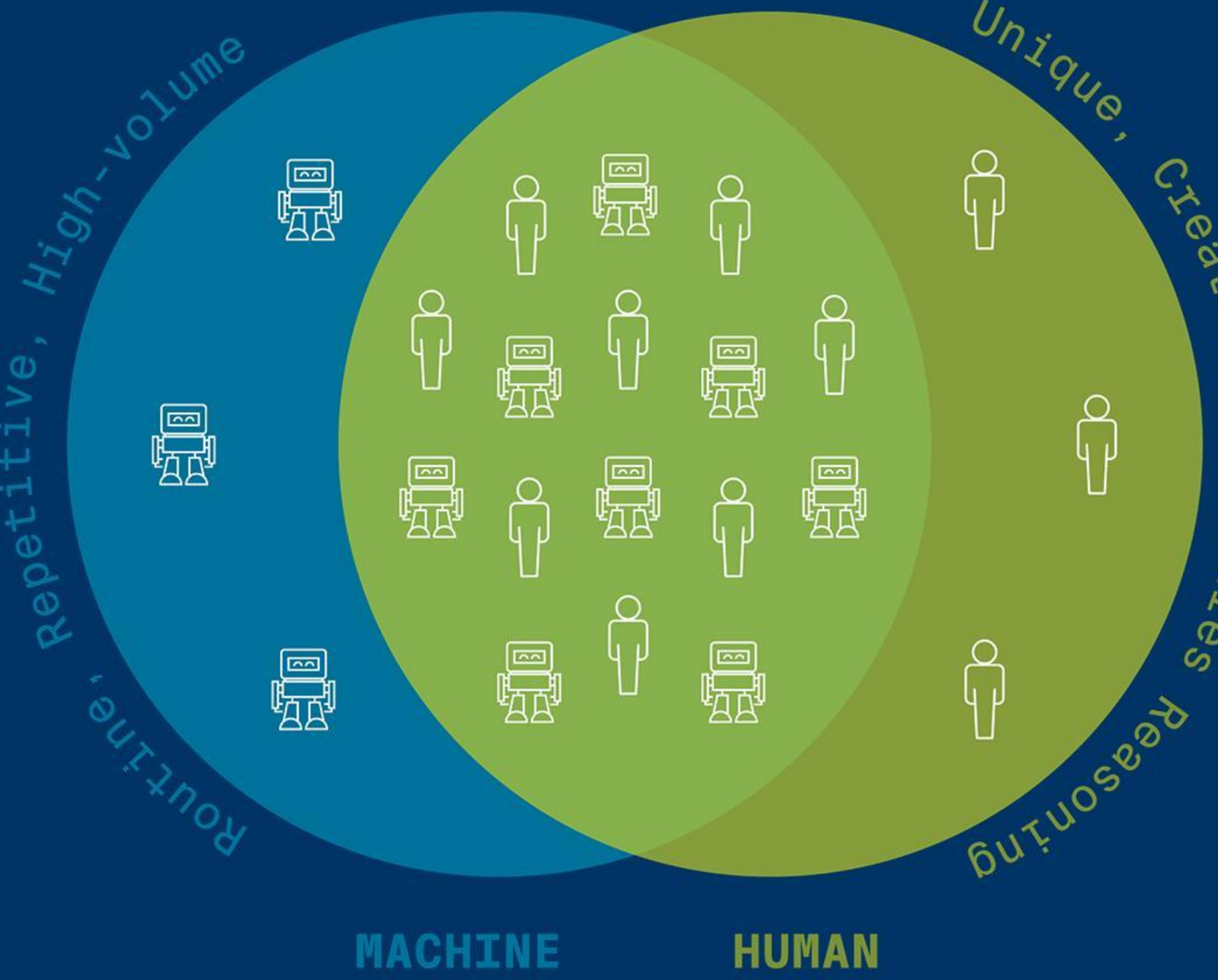
Fleisig, E. et al., "FairPrism: Evaluating fairness-related harms in text generation" (2023).

Do AI systems compensate artists?

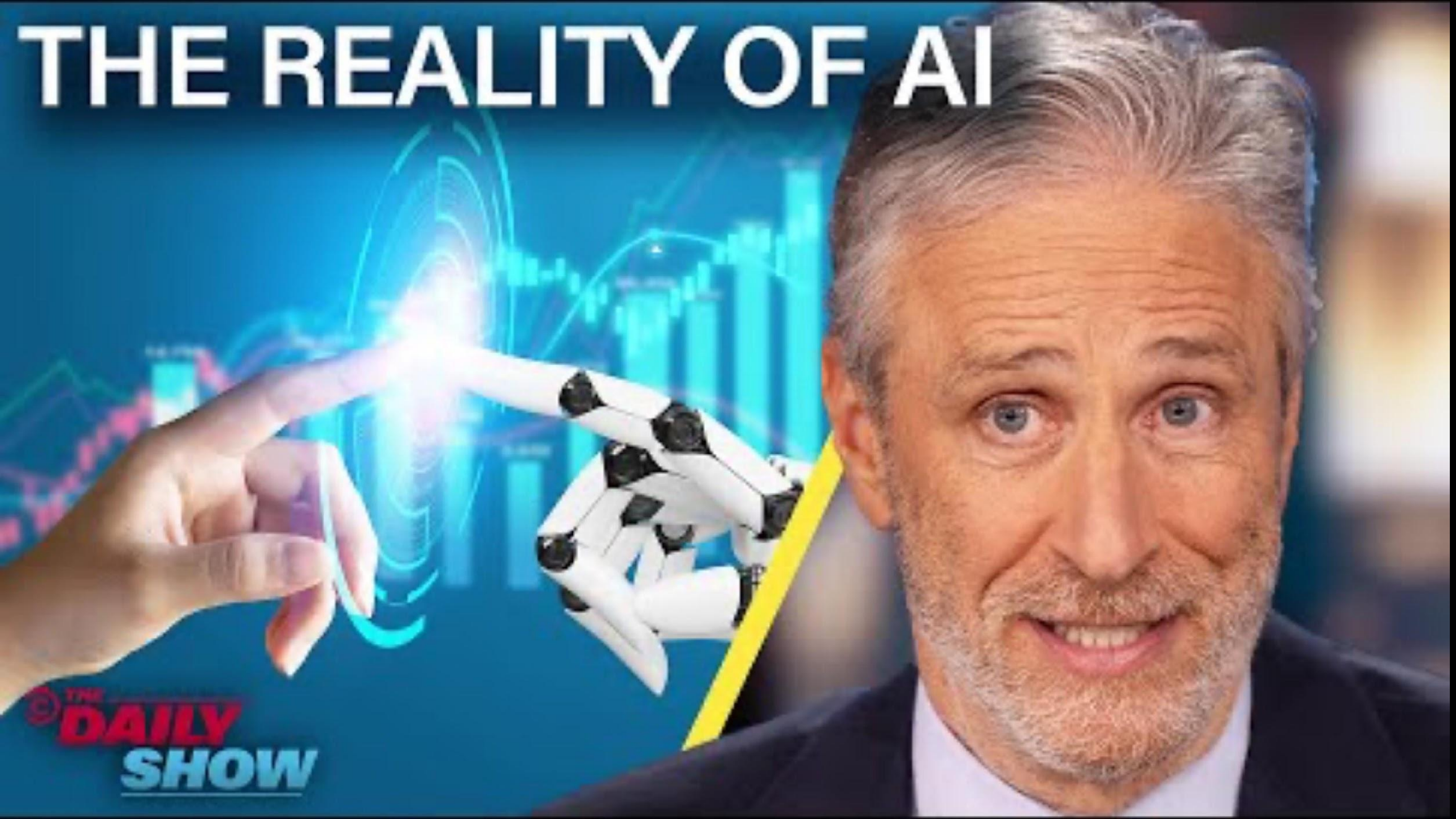


- Displaced workers; who retrains them?
  - Old company?
  - Disrupting company?
  - Government?

Noam Chomsky on Artificial Intelligence:
"The human mind is not, like ChatGPT and its ilk, a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question. On the contrary, the human mind is a surprisingly efficient and even elegant system that operates with small amounts of information; it seeks not to infer brute correlations among data points but to create explanations..."
"... Let's stop calling it "Artificial Intelligence" then and call it for what it is and makes "plagiarism software" because "It doesn't create anything, but copies existing works, of existing artists, modifying them enough to escape copyright laws...."
~ Dr. Noam Chomsky, Dr. Ian Roberts, Dr. Jeffrey Watumull
New York Times, March 8 2023



spectrum.mit.edu/summer-2019/reinventing-jobs


THE REALITY OF AI
THE DAILY SHOW

www.cbsnews.com/news/artificial-intelligence-carbon-footprint-climate-change/

Climate impacts of training AI models?

What is AI's carbon footprint and why is it worrying some environmental advocates?

- AI's overall carbon footprint is difficult to measure, but it starts with the computers it uses. The raw materials needed to create computer hardware are mined and *"that can be really labor intensive and also environmentally expensive,"* Shaolei Ren, ECE prof at the UC Riverside, said.
- There's also the carbon emissions. Researchers at the U Mass Amherst found the training process for a single AI model can emit more than 626,000 pounds of carbon dioxide.
- estimates that training GPT-3 could potentially have consumed 700,000 liters of freshwater. The water used to prevent data centers from overheating is usually evaporated, which means it

Computing and War

War is All About Technology

- Castles
- Catapults
- Boats
- Horses
- Arrows
- Swords
- Guns
- ...

Autonomous Weapons

- DARPA sponsors robotics research
  - “Smart bombs” & Drone aircraft
- The importance of autonomous weaponry is **political**, not military:
  - Governments are restrained from waging war because citizens don’t want their children to die abroad
  - Autonomous weapons can allow war without soldiers
- If there others are creating war technology, must we too

(Image generated by DALL-E)

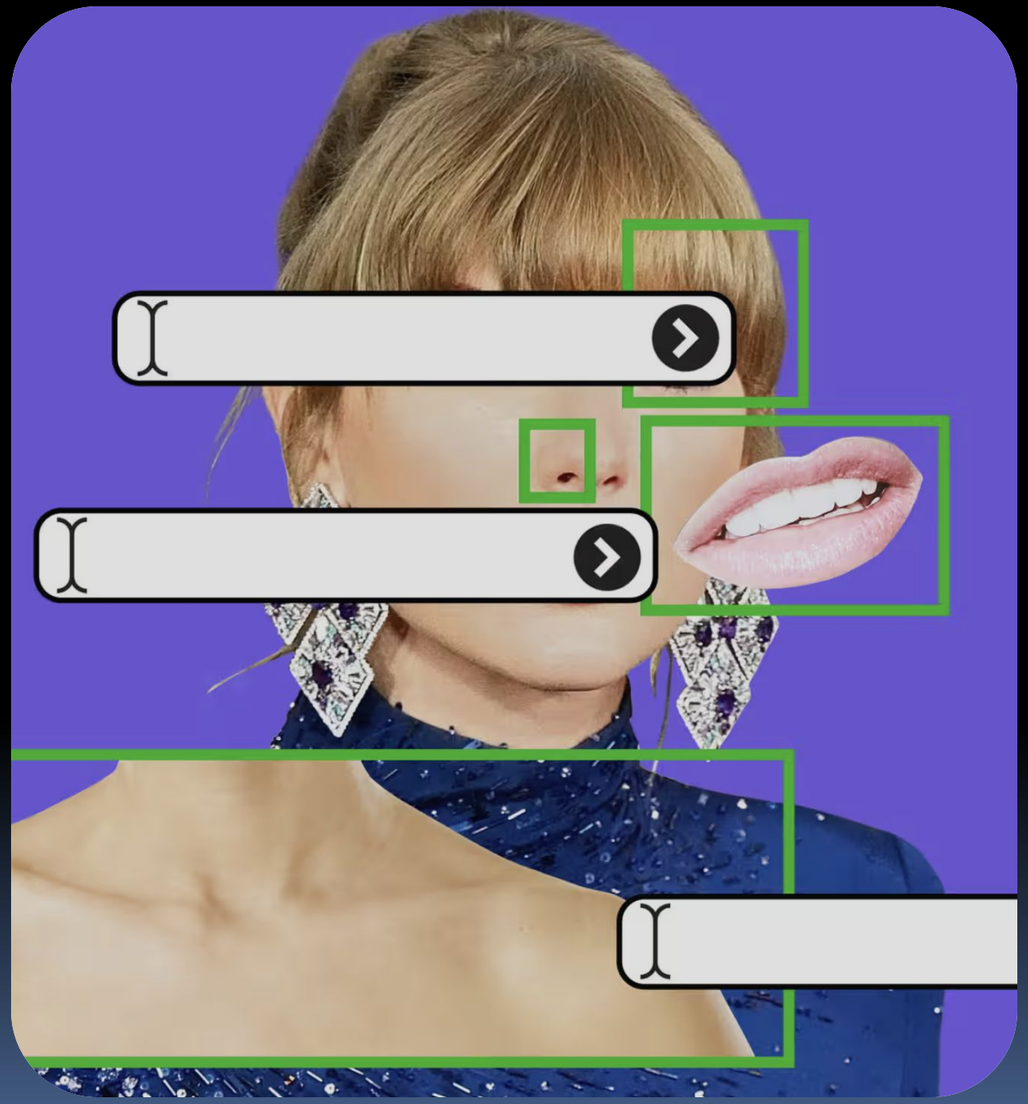

INCREASE IN VIOLENT CRIME
SDN

Bad Actors

www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box

Deepfakes to target people (usu women)

- Image based abuse is real and becoming too common
  - It's far too easy to create deepfake images and videos of people doing things
  - Victims have almost no recourse
  - Social media sites are slow to remove the images



(FilmMagic/Jeff Kravitz/Getty images)

www.npr.org/2024/01/22/1226129926/nh-primary-biden-ai-robocall
www.npr.org/2024/02/08/1229641751/ai-deepfakes-election-risks-lawmakers-tech-companies-artificial-intelligence

Deepfakes to affect elections

- “Your vote makes a difference in November, not this Tuesday.”
  - Deepfake call made to sound like President Biden ahead of the New Hampshire primary
- Response?
  - The Federal Communications Commission ruled robocalls using
- ”The calls are a "canary in the coalmine for what we're going to see" the rest of this election cycle when it comes to misinformation”

  – David Becker, the executive director of the Center for Election Innovation and Research.”

Deepfakes to scam family

- “Kidnap scammers using AI to create fake videos, voice recordings of loved ones.”
  - “I heard my mom’s voice kind of fading away like someone was taking the phone away from her. And I heard weeps; this guy then gets on the phone, and he goes, 'Hey, I have your mom and if you don’t send me money, I'm going
- “AI can even fool some security systems "It was able to get past facial recognition systems, it was able to get past systems where your voice is your password," he said. ”

(FilmMagic/Jeff Kravitz/Getty images)

…and worse!

- Using AI to create biological weapons and recommend how best to deploy them
- What happens with Artificial General Intellince?
  - *"In fact, from its earliest days, the stated goal has been AGI that match or exceed human capabilites in every relevant*
- Alan Turing said in 1951:
  - *"It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers.*

    …

    At some stage therefore we should have to expect the machines to take control"
- The more autonomous we are, the worse it will

What Can Be Done?

https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_russell.pdf

Recommendations by Prof Russell…

- Testimony before congress
  - Urgent regulation
    - Licensing of providers
    - Access to data
    - Content labeling
    - Right to know if AI/human
    - Ban on AI killing machines
  - Safety req for AI systems
    - “off switch”
  - A new regulatory agency
  - International coordination
  - AI safety research

Garcia