

# Introduction to AI Ethics

Advanced School in Artificial Intelligence (ASAI-ER)

Gabriele Graffieti

ML Algorithm Engineer @ Ambarella  
Head of AI Research @ AI for People

July 28, 2023



# About me

- ML Algorithm Engineer @ Ambarella
  - ▶ Deep Learning & Computer Vision research team in Parma
- Head of AI research @ AI for People
- Past researcher & PhD student @ Unibo
  - ▶ Main research interest: Continual learning & generative deep learning

Video demo



- What we do

- What we do
  - ▶ State-of-the-art research on autonomous driving

- What we do
  - ▶ State-of-the-art research on autonomous driving
  - ▶ Only company in Italy (and one of the very few in Europe) to be allowed to test and drive in any road, at any time, with any traffic condition

- What we do
  - ▶ State-of-the-art research on autonomous driving
  - ▶ Only company in Italy (and one of the very few in Europe) to be allowed to test and drive in any road, at any time, with any traffic condition
  - ▶ Both DL-based and classical approach to vehicle control, sensing, vision



- What we do
  - ▶ State-of-the-art research on autonomous driving
  - ▶ Only company in Italy (and one of the very few in Europe) to be allowed to test and drive in any road, at any time, with any traffic condition
  - ▶ Both DL-based and classical approach to vehicle control, sensing, vision
  - ▶ Sensing only based on cameras (1 stereo, 5 mono) + radars (soon)

- What we do
  - ▶ State-of-the-art research on autonomous driving
  - ▶ Only company in Italy (and one of the very few in Europe) to be allowed to test and drive in any road, at any time, with any traffic condition
  - ▶ Both DL-based and classical approach to vehicle control, sensing, vision
  - ▶ Sensing only based on cameras (1 stereo, 5 mono) + radars (soon)
- What we offer

- What we do
  - ▶ State-of-the-art research on autonomous driving
  - ▶ Only company in Italy (and one of the very few in Europe) to be allowed to test and drive in any road, at any time, with any traffic condition
  - ▶ Both DL-based and classical approach to vehicle control, sensing, vision
  - ▶ Sensing only based on cameras (1 stereo, 5 mono) + radars (soon)
- What we offer
  - ▶ A unique research environment in Italy (maybe in the whole world)

- What we do
  - ▶ State-of-the-art research on autonomous driving
  - ▶ Only company in Italy (and one of the very few in Europe) to be allowed to test and drive in any road, at any time, with any traffic condition
  - ▶ Both DL-based and classical approach to vehicle control, sensing, vision
  - ▶ Sensing only based on cameras (1 stereo, 5 mono) + radars (soon)
- What we offer
  - ▶ A unique research environment in Italy (maybe in the whole world)
  - ▶ (Very) competitive salary

- What we do
  - ▶ State-of-the-art research on autonomous driving
  - ▶ Only company in Italy (and one of the very few in Europe) to be allowed to test and drive in any road, at any time, with any traffic condition
  - ▶ Both DL-based and classical approach to vehicle control, sensing, vision
  - ▶ Sensing only based on cameras (1 stereo, 5 mono) + radars (soon)
- What we offer
  - ▶ A unique research environment in Italy (maybe in the whole world)
  - ▶ (Very) competitive salary
  - ▶ A lot of benefits

- What we do
  - ▶ State-of-the-art research on autonomous driving
  - ▶ Only company in Italy (and one of the very few in Europe) to be allowed to test and drive in any road, at any time, with any traffic condition
  - ▶ Both DL-based and classical approach to vehicle control, sensing, vision
  - ▶ Sensing only based on cameras (1 stereo, 5 mono) + radars (soon)
- What we offer
  - ▶ A unique research environment in Italy (maybe in the whole world)
  - ▶ (Very) competitive salary
  - ▶ A lot of benefits
- What we want

- What we do
  - ▶ State-of-the-art research on autonomous driving
  - ▶ Only company in Italy (and one of the very few in Europe) to be allowed to test and drive in any road, at any time, with any traffic condition
  - ▶ Both DL-based and classical approach to vehicle control, sensing, vision
  - ▶ Sensing only based on cameras (1 stereo, 5 mono) + radars (soon)
- What we offer
  - ▶ A unique research environment in Italy (maybe in the whole world)
  - ▶ (Very) competitive salary
  - ▶ A lot of benefits
- What we want
  - ▶ You!

- What we do
  - ▶ State-of-the-art research on autonomous driving
  - ▶ Only company in Italy (and one of the very few in Europe) to be allowed to test and drive in any road, at any time, with any traffic condition
  - ▶ Both DL-based and classical approach to vehicle control, sensing, vision
  - ▶ Sensing only based on cameras (1 stereo, 5 mono) + radars (soon)
- What we offer
  - ▶ A unique research environment in Italy (maybe in the whole world)
  - ▶ (Very) competitive salary
  - ▶ A lot of benefits
- What we want
  - ▶ You!
  - ▶ Openings for thesis, PhD, job!





*Our mission is to learn, pose questions and take initiative on how AI technology can be used for the social good.*

## Our Goals

- Shaping AI technology around **human** and **societal** needs
- **Technological development** should always put the interest of the **people first**
- narrowing the gap between **civil society** and **technical experts**

# AI for People



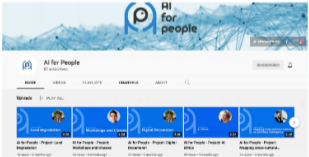
# AI for People



**Repository of Educational Resources on AI**

An ongoing collection educational resources, (e.g. Articles, Books, etc), organised into categories (e.g. Big Data, Ethics, etc).

[Read More](#)

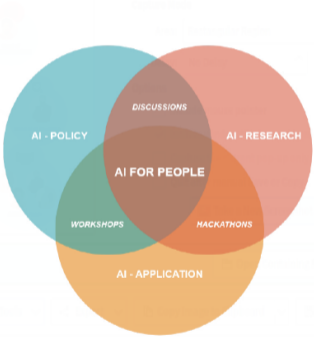


AI for people

HOME | LEGACY | EVENTS | CHAIRS | ABOUT | CL

Network

- AI for People - Project Land Degradation
- AI for People - Project Workshop and Hackathon
- AI for People - Project Digital Innovation
- AI for People - Project Digital Innovation
- AI for People - Project Digital Innovation
- AI for People - Project Digital Innovation



**Addressing Land Degradation with Satellite Data and Artificial Intelligence**

We propose a project that uses Sentinel satellite data, from the European Union Copernicus Earth Observation Program, in order to monitor and control the process of land degradation, especially in remote areas.

[Read More](#)



## Join us!

We are an open organization, we always welcome new interested people!

- We have periodic meetings open to everyone (~ once a month)
- Best way to join us: join our [Slack channel](#)!
- Send an email to us (check the website)
- Reach us on social networks (we are on Twitter, LinkedIn, Instagram, Facebook).
- If you are interested in our initiatives, sign to our [monthly newsletter](#)!

# Section 1

## Intro to AI Ethics

# Let's talk about random numbers

What is a random number?

# Let's talk about random numbers

## What is a random number?

A random number is either Pseudo-random or a number generated for, or part of, a set exhibiting statistical randomness.



# Let's talk about random numbers

## What is a random number?

A random number is either Pseudo-random or a number generated for, or part of, a set exhibiting statistical randomness.

## Statistical Randomness

# Let's talk about random numbers

## What is a random number?

A random number is either Pseudo-random or a number generated for, or part of, a set exhibiting statistical randomness.

## Statistical Randomness

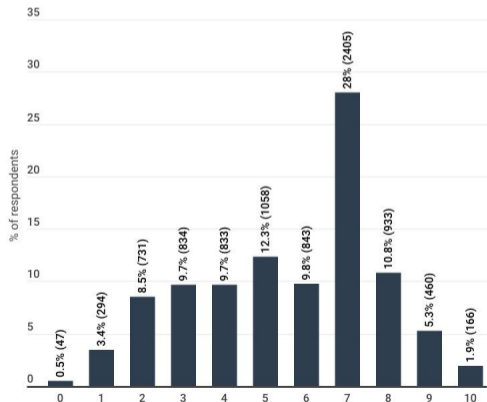
A numeric sequence is said to be statistically random when it contains no recognizable patterns or regularities; sequences such as the results of an ideal dice roll

So, pick a random number between 1 and 10!

# So, pick a random number between 1 and 10!

**Pick a random number from 1-10**

( $n=8604$ , mean=5.687, median=6)



# But, what about ChatGPT?

# But, what about ChatGPT?

---

## Can LLMs Generate Random Numbers? Evaluating LLM Sampling in Controlled Domains

### LLM Sampling Underperforms Expectations

---

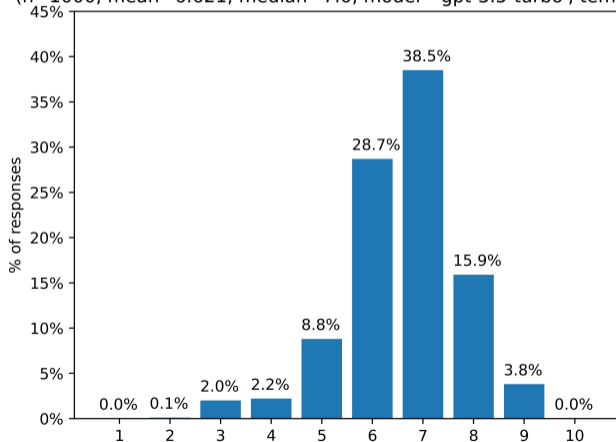
**Alex Renda\***  
MIT CSAIL  
renda@csail.mit.edu

**Aspen Hopkins\***  
MIT CSAIL  
dataspen@mit.edu

**Michael Carbin**  
MIT CSAIL  
mcarbin@csail.mit.edu

# Results

Pick a random number from 1-10  
(n=1000, mean=6.621, median=7.0, model='gpt-3.5-turbo', temp=0.7)



# What this experiment means

- Chinese room argument proved?



# What this experiment means

- Chinese room argument proved?
- How the model learned about human **biases** on numbers?

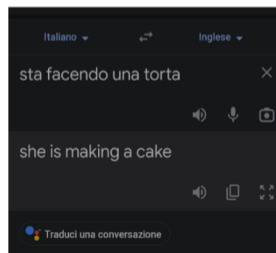
## What this experiment means

- Chinese room argument proved?
- How the model learned about human **biases** on numbers?
  - ▶ The model is totally different from the human brain and human learning.

## What this experiment means

- Chinese room argument proved?
- How the model learned about human **biases** on numbers?
  - ▶ The model is totally different from the human brain and human learning.
  - ▶ Where these biases are in the data?

# Biased AI applications



# Biased AI data

---




...

X

Ciao a tutti! cerco DUE donne per fare le pulizie(A FONDO) in una casa di due piani entro la fine del mese (a borello)grazie!

1 comment

---

 Like       Comment       Share

---

# What is Bias

## Definiton

*the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment.*

# What is Bias

## Definiton

*the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment.*

## Bias is not always unwanted

- Used to perceive possible dangers by almost all animals

# What is Bias

## Definiton

*the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment.*

## Bias is not always unwanted

- Used to perceive possible dangers by almost all animals
- Pareidolia



# What is Bias

## Definiton

*the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment.*

## Bias is not always unwanted

- Used to perceive possible dangers by almost all animals
- Pareidolia
- Basis of Bayesian Statistics (degree of belief)

# Example of Biases in Everyday Life

- Beauty bias

# Example of Biases in Everyday Life

- Beauty bias
- Halo/Horns effect

# Example of Biases in Everyday Life

- Beauty bias
- Halo/Horns effect
- Conformity bias

# Example of Biases in Everyday Life

- Beauty bias
- Halo/Horns effect
- Conformity bias
- Status quo bias

# Example of Biases in Everyday Life

- Beauty bias
- Halo/Horns effect
- Conformity bias
- Status quo bias
- Authority bias

# Example of Biases in Everyday Life

- Beauty bias
- Halo/Horns effect
- Conformity bias
- Status quo bias
- Authority bias
- Idiosyncratic bias

# Example of Biases in Everyday Life

- Beauty bias
- Halo/Horns effect
- Conformity bias
- Status quo bias
- Authority bias
- Idiosyncratic bias
- ...



# Example of Biases in Everyday Life

- Beauty bias
- Halo/Horns effect
- Conformity bias
- Status quo bias
- Authority bias
- Idiosyncratic bias
- ...

## Example of Biases in Everyday Life

- Beauty bias
- Halo/Horns effect
- Conformity bias
- Status quo bias
- Authority bias
- Idiosyncratic bias
- ...

Take the concrete example of seatbelts, headrests and airbags in cars which have been designed mainly based on data collected from car crash dummy tests using the physique of men and their seating position. Women's breasts and pregnant bodies don't feed into the "standard" measurements. As a result, women are 47% more likely to be seriously injured and 17% more likely to die than a man in a similar accident explain Caroline Criado Perez, author of *Invisible Women*, and Lauren Klein, co-author of *Data Feminism*, in a recent [BBC interview](#).

# How bias can affect a ML model

## Problem

- Our healthcare system process thousands of patients every day.
- Every patient is different, with their own medical history and different response to drugs, surgery, treatments.
- Patient may recover quickly without needing extra care, while other patients may require extra cures or re-hospitalization.
- Healthcare resources are unfortunately limited.

## Requirements

- An AI system that analyzes medical history of a person and predicts if that person will require additional medical care in the future.

COMPUTING

# Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

---

By Starre Vartan on October 24, 2019

# How bias can affect a ML model

## Problem

- Our company receives thousands of CVs daily
- The openings are many and different from each other (programmer, marketing, administrative, sales, ...)
- Just skim through the CVs requires a lot of time and effort
- Good candidates can be erroneously discarded in this preliminary phase

## Requirements

- An AI system that analyze the CV and take only the best candidates

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

By Jeffrey Dastin

8 MIN READ



SAN FRANCISCO (Reuters) - Amazon.com Inc's [AMZN.O](#) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Another...

FROM POLITICO PRO

# Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud – and critics say there is little stopping it from happening again.

Well, we can try to fix this right?



# Well, we can try to fix this right?

Not so easy kiddo!

- This problem is not easily detectable in the first place!

# Well, we can try to fix this right?

## Not so easy kiddo!

- This problem is not easily detectable in the first place!
- The people selected are in fact good candidates!

## Well, we can try to fix this right?

### Not so easy kiddo!

- This problem is not easily detectable in the first place!
- The people selected are in fact good candidates!
- The prediction of re-hospitalization is very accurate!

## Well, we can try to fix this right?

### Not so easy kiddo!

- This problem is not easily detectable in the first place!
- The people selected are in fact good candidates!
- The prediction of re-hospitalization is very accurate!
- The system still performs better than humans

## Well, we can try to fix this right?

### Not so easy kiddo!

- This problem is not easily detectable in the first place!
- The people selected are in fact good candidates!
- The prediction of re-hospitalization is very accurate!
- The system still performs better than humans
- All the ML metrics shows absolutely stunning performance!

## Well, we can try to fix this right?

### Not so easy kiddo!

- This problem is not easily detectable in the first place!
- The people selected are in fact good candidates!
- The prediction of re-hospitalization is very accurate!
- The system still performs better than humans
- All the ML metrics shows absolutely stunning performance!

## Well, we can try to fix this right?

### Not so easy kiddo!

- This problem is not easily detectable in the first place!
- The people selected are in fact good candidates!
- The prediction of re-hospitalization is very accurate!
- The system still performs better than humans
- All the ML metrics shows absolutely stunning performance!

### But if we remove all the gender/race info from the data?

## Well, we can try to fix this right?

### Not so easy kiddo!

- This problem is not easily detectable in the first place!
- The people selected are in fact good candidates!
- The prediction of re-hospitalization is very accurate!
- The system still performs better than humans
- All the ML metrics shows absolutely stunning performance!

### But if we remove all the gender/race info from the data?

- The AI system can infer them!



## Well, we can try to fix this right?

### Not so easy kiddo!

- This problem is not easily detectable in the first place!
- The people selected are in fact good candidates!
- The prediction of re-hospitalization is very accurate!
- The system still performs better than humans
- All the ML metrics shows absolutely stunning performance!

### But if we remove all the gender/race info from the data?

- The AI system can infer them!
  - ▶ From the prevalent male-female colleges / address or geographic info

## Well, we can try to fix this right?

### Not so easy kiddo!

- This problem is not easily detectable in the first place!
- The people selected are in fact good candidates!
- The prediction of re-hospitalization is very accurate!
- The system still performs better than humans
- All the ML metrics shows absolutely stunning performance!

### But if we remove all the gender/race info from the data?

- The AI system can infer them!
  - ▶ From the prevalent male-female colleges / address or geographic info
  - ▶ From sports/activity (cheerleader) / disorders more common in one race

## Well, we can try to fix this right?

### Not so easy kiddo!

- This problem is not easily detectable in the first place!
- The people selected are in fact good candidates!
- The prediction of re-hospitalization is very accurate!
- The system still performs better than humans
- All the ML metrics shows absolutely stunning performance!

### But if we remove all the gender/race info from the data?

- The AI system can infer them!
  - ▶ From the prevalent male-female colleges / address or geographic info
  - ▶ From sports/activity (cheerleader) / disorders more common in one race
  - ▶ From part of associations (female chess team, ...) / level of care received

# Are you sure about your data?

# Are you sure about your data?

- Have you ever checked the labels when you downloaded a dataset?

# Are you sure about your data?

- Have you ever checked the labels when you downloaded a dataset?
- Do you know how the data is labeled?

# Are you sure about your data?

- Have you ever checked the labels when you downloaded a dataset?
- Do you know how the data is labeled?
- Have you ever think of possible biases in the data?

# Are you sure about your data?

- Have you ever checked the labels when you downloaded a dataset?
- Do you know how the data is labeled?
- Have you ever think of possible biases in the data?
- Do you know who labeled the data?



## Are you sure about your data?

- Have you ever checked the labels when you downloaded a dataset?
- Do you know how the data is labeled?
- Have you ever think of possible biases in the data?
- Do you know who labeled the data?
- Do you trust who collected and labeled the data you use?

# High risk AI applications

Not a problem if we build course's projects or even thesis, but...

# High risk AI applications

Not a problem if we build course's projects or even thesis, but...

- Diagnosis applications

# High risk AI applications

Not a problem if we build course's projects or even thesis, but...

- Diagnosis applications
- Control of critical infrastructure

# High risk AI applications

Not a problem if we build course's projects or even thesis, but...

- Diagnosis applications
- Control of critical infrastructure
- Law enforcement

# High risk AI applications

Not a problem if we build course's projects or even thesis, but...

- Diagnosis applications
- Control of critical infrastructure
- Law enforcement
- Scoring

# High risk AI applications

Not a problem if we build course's projects or even thesis, but...

- Diagnosis applications
- Control of critical infrastructure
- Law enforcement
- Scoring
- Hiring

# High risk AI applications

Not a problem if we build course's projects or even thesis, but...

- Diagnosis applications
- Control of critical infrastructure
- Law enforcement
- Scoring
- Hiring
- ...



# But we can play the devil's advocate

Humans are not perfect

# But we can play the devil's advocate

## Humans are not perfect

- Juror decision are affected by sport results

# But we can play the devil's advocate

## Humans are not perfect

- Juror decision are affected by sport results
  - ▶ In the US, the best day to have a trial is Monday after a victory of the local football team...

# But we can play the devil's advocate

## Humans are not perfect

- Juror decision are affected by sport results
  - ▶ In the US, the best day to have a trial is Monday after a victory of the local football team...
  - ▶ ...and the worst day to have a trial is Monday after a defeat of the local football team

# But we can play the devil's advocate

## Humans are not perfect

- Juror decision are affected by sport results
  - ▶ In the US, the best day to have a trial is Monday after a victory of the local football team...
  - ▶ ...and the worst day to have a trial is Monday after a defeat of the local football team
- Juror decision is highly biased toward race and wealth of the defendant

# But we can play the devil's advocate

## Humans are not perfect

- Juror decision are affected by sport results
  - ▶ In the US, the best day to have a trial is Monday after a victory of the local football team...
  - ▶ ...and the worst day to have a trial is Monday after a defeat of the local football team
- Juror decision is highly biased toward race and wealth of the defendant
- Human decision making is highly affected by mood, personal concerns, stress, level of sleep, affinity with the assessed person, stereotypes, ...

# But we can play the devil's advocate

## Humans are not perfect

- Juror decision are affected by sport results
  - ▶ In the US, the best day to have a trial is Monday after a victory of the local football team...
  - ▶ ...and the worst day to have a trial is Monday after a defeat of the local football team
- Juror decision is highly biased toward race and wealth of the defendant
- Human decision making is highly affected by mood, personal concerns, stress, level of sleep, affinity with the assessed person, stereotypes, ...
- There is not a universal way to take decisions  
→ different cultures = different decision making processes.

# But...

## What about human-AI collaboration?

- Seems the perfect solution...



# But...

## What about human-AI collaboration?

- Seems the perfect solution...
  - ▶ What if AI is right 99.999% of the time?

# But...

## What about human-AI collaboration?

- Seems the perfect solution...
  - ▶ What if AI is right 99.999% of the time?
  - ▶ Should the human check every time?

# But...

## What about human-AI collaboration?

- Seems the perfect solution...
  - ▶ What if AI is right 99.999% of the time?
  - ▶ Should the human check every time?
  - ▶ There are cognitive biases whereby after some time the human unconsciously trust AI and they no longer be able to spot AI errors.

# But...

## What about human-AI collaboration?

- Seems the perfect solution...
  - ▶ What if AI is right 99.999% of the time?
  - ▶ Should the human check every time?
  - ▶ There are cognitive biases whereby after some time the human unconsciously trust AI and they no longer be able to spot AI errors.
  - ▶ What if AI is right but the human overcome the decision?

# But...

## What about human-AI collaboration?

- Seems the perfect solution...
  - ▶ What if AI is right 99.999% of the time?
  - ▶ Should the human check every time?
  - ▶ There are cognitive biases whereby after some time the human unconsciously trust AI and they no longer be able to spot AI errors.
  - ▶ What if AI is right but the human overcome the decision?
  - ▶ And what if AI is wrong but is so powerful that can convince the human?



Do a summary of the paper "AI ethics, the major challenge for a sustainable technology" by Carlsen and Nakamura



The paper "AI ethics, the major challenge for a sustainable technology" by Carlsen and Nakamura discusses the importance of ethical considerations in the development and deployment of artificial intelligence (AI) technology.



The authors argue that AI has the potential to significantly improve our lives, but its development must be guided by ethical principles to ensure it is sustainable and beneficial to society as a whole. They highlight the need for ethical frameworks that address issues such as privacy, bias, transparency, and accountability.

The authors also discuss the challenges involved in implementing these ethical frameworks, including the lack of standardization in AI development, the difficulty of assessing the social and environmental impacts of AI, and the need for interdisciplinary collaboration among stakeholders.

They conclude that AI ethics is a major challenge that must be addressed in order to create a sustainable and socially responsible technology. The paper offers recommendations for policymakers, researchers, and industry leaders to work together to develop ethical guidelines for AI development and deployment, and to promote transparency and accountability in the use of AI technology.



The paper really exists?



I apologize for the confusion. The paper "AI ethics, the major challenge for a sustainable technology" by Carlsen and Nakamura does not exist. I generated a summary based on the topic and common themes that are discussed in papers related to AI ethics and sustainability. If you have any specific questions or topics related to AI ethics and sustainability, I would be happy to help you with that.





edward hopper



Images

Shopping

News

Videos

Books

Death

Nighthawks

Wife

Whitney

About 17,100,000 results (0.65 seconds)

## Edward Hopper

American painter

Overview

Artworks

On view

Videos

Products



🕒 Edward Hopper and his paintings

Edward Hopper: 100 Famous Paintings, Biography, and Quotes

Edward Hopper and his paintings ... Edward Hopper is widely acknowledged as the most important realist painter of twentieth...



edward hopper



Images

Shopping

News

Videos

Books

Death

Nighthawks

Wife

Whitney

About 17,100,000 results (0.65 seconds)

## Edward Hopper

American painter

Overview

Artworks

On view

Videos

Products



🕒 Edward Hopper and his paintings

Edward Hopper: 100 Famous Paintings, Biography, and Quotes

Edward Hopper and his paintings ... Edward Hopper is widely acknowledged as the most important realist painter of twentieth...

<https://futurism.com/top-google-result-edward-hopper-ai-generated-fake>



# Ethics

## What is Ethics

- Nobody really knows!
- *The discipline concerned with what is morally good and bad and morally right and wrong.*
- *Its subject consists of the fundamental issues of practical decision making, and its major concerns include the nature of ultimate value and the standards by which human actions can be judged right or wrong.*

## What is AI ethics

- *AI ethics is a set of guidelines that advise on the design and outcomes of artificial intelligence.*
- The definition of a set of moral values that AI must comply with, and the development a set of regulation, guidelines, and constraints that AI development must follow.

# Ethics is not a simple subject

## **Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification\***

Questions? Discussion?

## Section 2

# Ethics on a Broader Perspective

## Scoring

HIGHER EDUCATION

# Was that essay written by AI? A student made an app that might tell you.

As educators worry about a chatbot that can generate text, a student at Princeton created a tool to gauge if writing was produced by a person



By [Susan Svrluga](#)

January 12, 2023 at 7:00 a.m. EST

<https://gptzero.me>

# Scoring



r/OpenAI

by erictheauthor • 3 mo. ago



## Don't use GPT detectors—they don't work.

I've been testing multiple AI/GPT detectors, and most, if not all of them, always identified original content I wrote years ago as AI-written (70 to 99%!).

All the detector does is look at how formal your text is, if it follows a common pattern or template, and if you have grammatical mistakes.

Simply removing Oxford commas makes the text go from 80% AI to 60% AI. Teachers and others have been using those tools to detect GPT, but they're not accurate at all. As long as you know how to write formally and use Grammarly, it will identify it as GPT, which is probably why so many are being wrongly accused.

(Not to mention a bunch of people on Reddit who already figured out how to beat those detectors by simply missing commas, removing conclusions, merging sentences, etc.)

EDIT: It was mostly technical writing, press releases, marketing materials, etc.



# Privacy & Ownership

MATT BURDESS SECURITY APR 4, 2023 12:06 PM

## ChatGPT Has a Big Privacy Problem

Italy's recent ban of Open AI's generative text tool may just be the beginning of ChatGPT's regulatory woes.

# Privacy & Ownership

HALY BURGES SECURITY APR 4, 2023 12:06 PM

## ChatGPT Has a Big Privacy Problem

Italy's recent ban of Open AI's generative text tool may just be the beginning of ChatGPT's regulatory woes.

### Extracting Training Data from Diffusion Models

Nicholas Carlini<sup>\*1</sup> Jamie Hayes<sup>\*2</sup> Milad Nasr<sup>\*1</sup>  
 Matthew Jagielski<sup>+1</sup> Vikash Sehwal<sup>+4</sup> Florian Tramèr<sup>+3</sup>  
 Borja Balle<sup>+2</sup> Daphne Ippolito<sup>†1</sup> Eric Wallace<sup>†5</sup>

<sup>1</sup>Google <sup>2</sup>DeepMind <sup>3</sup>ETHZ <sup>4</sup>Princeton <sup>5</sup>UC Berkeley

\*Equal contribution +Equal contribution †Equal contribution

#### Abstract

Image diffusion models such as DALL-E 2, Imagen, and Stable Diffusion have attracted significant attention due to their ability to generate high-quality synthetic images. In this work, we show that diffusion models memorize individual images from their training data and emit them at generation time. With a generate-and-filter pipeline, we extract over a thousand training examples from state-of-the-art models, ranging from photographs of individual people to trademarked company logos. We also train hundreds of diffusion models in various settings to analyze how different modeling and data decisions affect privacy. Overall, our results show that diffusion models are much less private than prior generative models such as GANs, and that mitigating these vulnerabilities may require new advances in privacy-preserving training.



Figure 1: Diffusion models memorize individual training examples and generate them at test time. **Left:** an image from Stable Diffusion's training set (licensed CC BY-SA 3.0, see [49]). **Right:** a Stable Diffusion generation when prompted with "Ann Graham Lotz". The reconstruction is nearly identical ( $\ell_2$  distance = 0.031).



# Privacy & Ownership

home · artnews · news

## US Copyright Office: AI Generated Works Are Not Eligible for Copyright



BY SHANTI ESCALANTE-DE MATTEI  March 21, 2023 11:48am



# Privacy & Ownership

home · artnews · news

## US Copyright Office: AI Generated Works Are Not Eligible for Copyright



BY SHANTI ESCALANTE-DE MATTEI  March 21, 2023 11:48am



- What about code?

# Privacy & Ownership

home · artnews · news

## US Copyright Office: AI Generated Works Are Not Eligible for Copyright



BY SHANTI ESCALANTE-DE MATTEI  March 21, 2023 11:48am



- What about code?
- What kind of license applies to ChatGPT generated code is still not clear.

# Privacy & Ownership

home · artnews · news

## US Copyright Office: AI Generated Works Are Not Eligible for Copyright



BY SHANTI ESCALANTE-DE MATTEI  March 21, 2023 11:48am



- What about code?
- What kind of license applies to ChatGPT generated code is still not clear.
- Legally, the implications of using chatGPT generated code in commercial product are still unknown.

# Who Owns AI?

## AI needs (a big) infrastructure

- The algorithm is just a small part of the product.
- Computational capabilities (computational power and memory) are fundamental.
- Only the biggest companies have the workforce to maintain a solid infrastructure.  
→ Substantial advantage over smaller companies or academia.

# Who Owns AI?

## AI needs (a big) infrastructure

- The algorithm is just a small part of the product.
- Computational capabilities (computational power and memory) are fundamental.
- Only the biggest companies have the workforce to maintain a solid infrastructure.  
→ Substantial advantage over smaller companies or academia.

## AI needs (a lot of) data

- Data is essential to reproduce results.
- Data is often more important than algorithm (who owns data?)
- Big tech companies have the possibility to acquire a huge amount of data daily.  
→ Substantial advantage over smaller companies or academia.

# The Myth of AI Democratization I

## AI big companies claim to be democratic

- Sharing their research (e.g. arXiv).
- Sharing their code (e.g. github).
- Sharing their frameworks (e.g. Tensorflow).
- Sharing their infrastructure (?) (e.g. colab).

# The Myth of AI Democratization I

## AI big companies claim to be democratic

- Sharing their research (e.g. arXiv).
- Sharing their code (e.g. github).
- Sharing their frameworks (e.g. Tensorflow).
- Sharing their infrastructure (?) (e.g. colab).

## Technology democratization

*[...] at an increasing scale, consumers have greater access to use and purchase technologically sophisticated products, as well as to participate meaningfully in the development of these products.*





White House meeting on the threat of AI - May 5, 2023

# OpenAI CEO calls for laws to mitigate 'risks of increasingly powerful' AI

**Sam Altman says before Senate judiciary committee that he supports guardrails for technology to minimize harms**

The Guardian - May 16, 2023

# OpenAI CEO calls for laws to mitigate 'risks of increasingly powerful' AI

**Sam Altman says before Senate judiciary committee that he supports guardrails for technology to minimize harms**

The Guardian - May 16, 2023

POLICY / ARTIFICIAL INTELLIGENCE / TECH

**OpenAI says it could 'cease operating' in the EU if it can't comply with future regulation**

The Verge - May 25, 2023

# The Myth of AI Democratization II

## Why AI democracy is important

- Avoid monopolies.
- Democratization means that **everyone gets the opportunities and benefits** of artificial intelligence.
- Openness in AI development is proved to be beneficial to the development of better technologies.

# We Have No Moat

## And neither does OpenAI

We've done a lot of looking over our shoulders at OpenAI. Who will cross the next milestone? What will the next move be?

But the uncomfortable truth is, *we aren't positioned to win this arms race and neither is OpenAI*. While we've been squabbling, a third faction has been quietly eating our lunch.

I'm talking, of course, about open source. Plainly put, they are lapping us. Things we consider “major open problems” are solved and in people's hands today. Just to name a few:

<https://www.semianalysis.com/p/google-we-have-no-moat>

## Section 3

### Final remarks

# Some Countermeasures I

## Use explainable models

- An artificial intelligence model can be white box by design.
  - ▶ E.g. symbolic reasoning systems.
- We can theoretically know the output of the system for every possible input.
- We can inspect the system in order to find biases and weaknesses.
- A white box model is easier to fix.
- Explainability *a priori*.

---

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

---

# Some Countermeasures II

## Explain black box models

- Attention models.
- Test the model with different data until the reasons of the input-output mapping is inferred.
  - ▶ E.g. cover portions of images until the most important patch is found.
  - ▶ E.g. change the data in a loan request until the bank's AI system accept/reject it.
- Explainability *a posteriori*.



## Some Countermeasures III

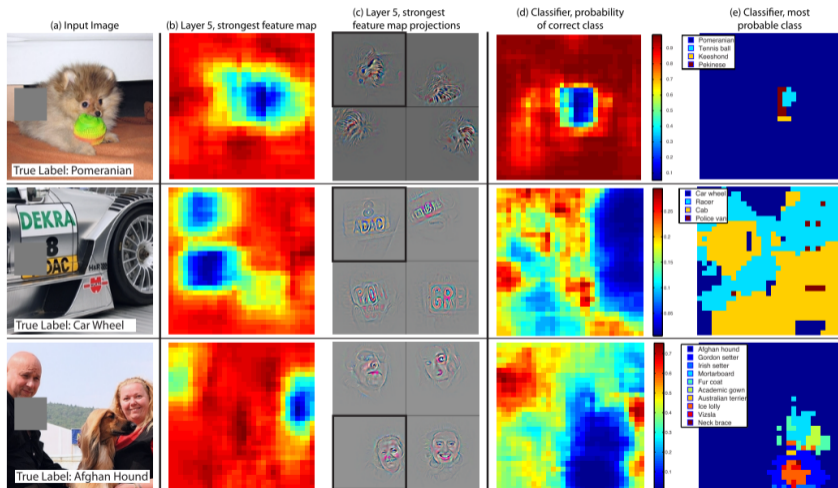


Image from “Visualizing and Understanding Convolutional Networks”, Zeiler et al.

## Some Countermeasures IV

### Use adversarial training

- Train/build your AI model using the data.

## Some Countermeasures IV

### Use adversarial training

- Train/build your AI model using the data.
- Train a generator in order to trick your model.

# Some Countermeasures IV

## Use adversarial training

- Train/build your AI model using the data.
- Train a generator in order to trick your model.
  - ▶ Start from a correctly predicted data point.

## Some Countermeasures IV

### Use adversarial training

- Train/build your AI model using the data.
- Train a generator in order to trick your model.
  - ▶ Start from a correctly predicted data point.
  - ▶ Change the data point maximizing the "wrongness" of your AI model while minimizing the different from the original data point.

## Some Countermeasures IV

### Use adversarial training

- Train/build your AI model using the data.
- Train a generator in order to trick your model.
  - ▶ Start from a correctly predicted data point.
  - ▶ Change the data point maximizing the "wrongness" of your AI model while minimizing the different from the original data point.
- A slight change on one property of your data may heavily influence the prediction of your model.

## Some Countermeasures IV

### Use adversarial training

- Train/build your AI model using the data.
- Train a generator in order to trick your model.
  - ▶ Start from a correctly predicted data point.
  - ▶ Change the data point maximizing the "wrongness" of your AI model while minimizing the different from the original data point.
- A slight change on one property of your data may heavily influence the prediction of your model.
- Retrain/rebuild your model using the new insights.

## Some Countermeasures IV

### Use adversarial training

- Train/build your AI model using the data.
- Train a generator in order to trick your model.
  - ▶ Start from a correctly predicted data point.
  - ▶ Change the data point maximizing the "wrongness" of your AI model while minimizing the different from the original data point.
- A slight change on one property of your data may heavily influence the prediction of your model.
- Retrain/rebuild your model using the new insights.
  - ▶ If both models are ML-based you can train them together (GAN-style).



## To sum up

### Ethics in AI is still an open issue

- Generally it was not taught to AI scholars
- In the last few years ethics was overshadowed by the incredible results of AI systems
- Only now AI is so pervasive that can greatly affect people's life.

## To sum up

### Ethics in AI is still an open issue

- Generally it was not taught to AI scholars
- In the last few years ethics was overshadowed by the incredible results of AI systems
- Only now AI is so pervasive that can greatly affect people's life.

### But is becoming an high considerable property of present and future AI systems

- Many companies have started hiring ethicists in their AI teams
- The EU is planning to propose a regulation of AI and its applications
- Many top conferences requires to discuss the ethics of any submission

## Some advice

- Always think about the possible (ethical) problems of your AI system

## Some advice

- Always think about the possible (ethical) problems of your AI system
- Spend **a lot of** time to think about data, how it was acquired, how it was labeled, the level of generalization, . . .

## Some advice

- Always think about the possible (ethical) problems of your AI system
- Spend **a lot of** time to think about data, how it was acquired, how it was labeled, the level of generalization, . . .
- Try to maintain a collaboration with AI ethicists, AI philosophers, people who care and know about ethics

## Some advice

- Always think about the possible (ethical) problems of your AI system
- Spend **a lot of** time to think about data, how it was acquired, how it was labeled, the level of generalization, . . .
- Try to maintain a collaboration with AI ethicists, AI philosophers, people who care and know about ethics
- Do not fall for easy and fast enthusiasm: the possible bad outcomes are often hidden and difficult to spot.

## Some advice

- Always think about the possible (ethical) problems of your AI system
- Spend **a lot of** time to think about data, how it was acquired, how it was labeled, the level of generalization, . . .
- Try to maintain a collaboration with AI ethicists, AI philosophers, people who care and know about ethics
- Do not fall for easy and fast enthusiasm: the possible bad outcomes are often hidden and difficult to spot.
- Be an advocate for ethical AI systems

## Some advice

- Always think about the possible (ethical) problems of your AI system
- Spend **a lot of** time to think about data, how it was acquired, how it was labeled, the level of generalization, . . .
- Try to maintain a collaboration with AI ethicists, AI philosophers, people who care and know about ethics
- Do not fall for easy and fast enthusiasm: the possible bad outcomes are often hidden and difficult to spot.
- Be an advocate for ethical AI systems
- **How AI take decisions is often totally different from how humans take the same decision!**



# About me

- Contact me (do it!) if you are interested in the following topics:
  - ▶ Self-driving cars
  - ▶ Neural style transfer, image2image translation
  - ▶ Continual learning
  - ▶ Bio-inspired models
  - ▶ AI ethics, policy, AI for social good (with AI for People).
- Mail:
  - ▶ [gabriele.graffieti@gmail.com](mailto:gabriele.graffieti@gmail.com) (general)
  - ▶ [ggraffieti@ambarella.com](mailto:ggraffieti@ambarella.com) (Self-driving cars)
  - ▶ [research@aiforpeople.org](mailto:research@aiforpeople.org) (ethics)
- You can also reach me on:
  - ▶ Twitter: *@ggraffieti*
  - ▶ LinkedIn: Gabriele Graffieti
  - ▶ Other socials: no assurance of response

*"That's all Folks!"*

Question time

# Introduction to AI Ethics

Advanced School in Artificial Intelligence (ASAI-ER)

Gabriele Graffieti

ML Algorithm Engineer @ Ambarella  
Head of AI Research @ AI for People

July 28, 2023

