



## Artificial Intelligence II: Deep learning methods

**Ana Neacșu & Vlad Vasilescu**

Lecture 1: Introduction and AI Ethics

National University of Science and Technology POLITEHNICA Bucharest, Romania  
BIOSINF Master Program

March 2025

# Course General Outline

## ➤ General topics:

- Ethics of AI
- Linear Nets & Optimization methods
- Fully Connected Neural Networks
- Convolutional Neural Networks
- Recurrent Connections
- Graph Neural Networks

## Course organization

- Course → Thursday every 2 weeks (odd)
  - Laboratory → Thursday every 2 weeks (even)
  - Project → all the time :), official discussions every Thursday

**Grading :** Written Exam, closed book (40 pts), Individual Project (40 pts), Lab Homework (30 pts) [3 × Homeworks, 10 points each].

Details about projects can be found on the course webpage on [Github](#).

Students must choose the project in the first two weeks of the semester.

# Overview

1 Introduction

2 The Ethics of AI

3 Algorithmic Bias

4 A Broader Perspective on AI Ethics

5 High Risk Applications

6 How to build ethical AI models

## Introduction

## Why Artificial Intelligence?

“ AI is the new electricity ”

Andrew Ng – co-founder Google Brain

- Two centuries ago electricity revolutionized numerous sectors, including transportation, manufacturing, healthcare, communications, and beyond.
  - Nowadays, AI will bring about an equally big effect.
  - It is important to take into account that this big change comes with a lot of challenges.

## Key Features of AI

Artificial Intelligence

- is the branch of science and technology that deals with the simulation of intelligent behaviour in computers as regards their capacity to mimic, and ideally *improve*, human behaviour;
  - focuses on computer systems which are both *autonomous* and *adaptive*;
  - became a self-standing discipline in 1955.

**Autonomy** → The ability to perform tasks in complex environments, without constant guidance by a user.

**Adaptability** → The ability to improve its performance based on previous experience.

**Machine Learning** → The science of making computers learn, without being explicitly programmed for a given task.

## History of AI

## Early beginnings

- 1943 - McCulloch and Pitts propose the first mathematical model of a neural network.
  - 1950 - Alan Turing introduces the Turing Test for machine intelligence.
  - 1956 - Dartmouth Conference marks the birth of AI as a field.

AI Winter

- Late 1970s to mid-1980s - Period of reduced funding and interest in AI.
  - Caused by unmet expectations and overpromising and a lack of computational power.

Rise of Machine Learning

- 1997 - IBM's Deep Blue defeats world chess champion Garry Kasparov.
  - 2011 - IBM's Watson wins Jeopardy! against human champions.
  - 2012 - Breakthrough in deep learning with AlexNet winning ImageNet competition.

# History of AI

## Recent developments

- 2016 - AlphaGo defeats world champion Go player Lee Sedol.
- 2018 - OpenAI's GPT-3 demonstrates remarkable natural language processing capabilities.
- 2020 - DALL-E introduces a model for generating creative images from textual descriptions.
- 2022 - Stable diffusion-based models showcase progress in generative models.

## Some perspectives

- \* Ethical AI – Increasing emphasis on developing AI systems with ethical considerations.
- \* Human AI-collaboration – Focus on augmenting human capabilities rather than replacing them.
- \* Job-displacement – Strategies for mitigating negative effects on the job market.
- \* AI-driven creativity – expansion of AI's role in creative fields such as art, music, and design.

# The Ethics of AI

## Brave new world?

*"Până astăzi nimeni nu a putut spune ce e bine și ce e rău. [...] Aș iubi o lume în care n-ar exista nici un criteriu, nici o formă și nici un principiu, o lume a absolutei indeterminări". – E.Cioran, Pe culmile disperării*

### What is Ethics?

- *No one really knows!* – F. Nietzsche
- *The discipline concerned with what is morally good and bad and morally right and wrong.* – Socrates
- *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.* – Standard definition

### What is AI ethics?

- A set of guidelines that advise on the design and outcomes of AI systems.
- The definition of a set of moral values that AI must comply with, and the development of a rigorous set of regulations, guidelines and constraints AI development must follow.

## Let's consider an example

### Problem:

- Our Company receives 5k CVs daily.
- The openings are many and diverse, e.g. programmer, marketing, administrative, sales, ...
- Skimming through the CVs requires a lot of time and effort.
- Good candidates can be erroneously discarded in this preliminary phase.

### Requirement

- Develop an AI-based system that analyzes the CVs and classify the best candidates.

Let's consider an example

### Problem:

- Our Company receives 5k CVs daily.
  - The openings are many and diverse, e.g. programmer, marketing, administrative, sales, ...
  - Skimming through the CVs requires a lot of time and effort.
  - Good candidates can be erroneously discarded in this preliminary phase.

## Requirement

- Develop an AI-based system that analyzes the CVs and classify the best candidates.

# A possible solution

## Solution:

- Use the CVs of the current employees as ground truth data
- We want to select candidates similar to the good people we already have in our company
- Our great engineers designed and developed the system with SOTA models and techniques

## Results:

- The selected people are very good candidates
- The system performs better than our HRs in selecting good candidates
- All the ML metrics shows stunning performance

**Questions:** Do you approve the system? Do you give a salary rise to the engineers?

# Some problems

- The Amazon machine learning specialists uncovered a big problem: their new recruiting engine did not like women.
- Racial bias is also present in some models: black people have less chances of getting a good education.
- Culture is also an important factor here.

## Algorithmic bias – three levels of concern

- ① Bias – dataset does not reflect the exact distribution of the population
- ② Fairness – dataset is based on historical data which reflects unfair practice.
- ③ Unethical – the model is deliberately skewed or behaves dishonorably.

# Some problems

- The Amazon machine learning specialists uncovered a big problem: their new recruiting engine did not like women.
- Racial bias is also present in some models: black people have less chances of getting a good education.
- Culture is also an important factor here.

## Algorithmic bias – three levels of concern

- ① **Bias** – dataset does not reflect the exact distribution of the population
- ② **Fairness** – dataset is based on historical data which reflects unfair practice.
- ③ **Unethical** – the model is deliberately skewed or behaves dishonorably.

# How to fix this?

## Not an easy solution

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

## What 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 or age
  - From sports/activity (cheerleader) / disorders more common in one race
  - From part of associations (female chess team, . . . )

## How to fix this?

Not an easy solution

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

What 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 or age
    - From sports/activity (cheerleader) / disorders more common in one race
    - From part of associations (female chess team, . . . )

Introduction  
ooooo

The Ethics of AI  
oooooo

Algorithmic Bias  
●oooooooooooo

A Broader Perspective on AI Ethics  
oooooooooooo

High Risk Applications  
oooooooo

How to build ethical AI models  
oooooooooooo

## Algorithmic Bias

# What is Bias?

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

Bias is not always bad

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

## What is Bias?

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

## Bias is not always bad

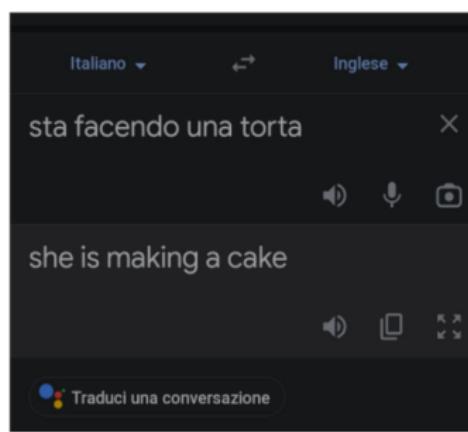
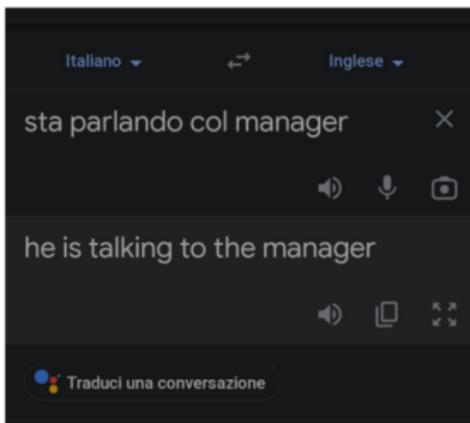
- Used to perceive possible dangers by almost all animals
  - Pareidolia
  - Basis of Bayesian Statistics (degree of belief)

# Real-life Examples of Biases

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

"Take the concrete example of seat belts, 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 sitting position. Women's breasts and pregnant bodies don't fit 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.

# Bias in AI



# Data bias

## Data uncertainty

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

## Diversity is the key

- Diversity of Background
- Diversity of Mindset
- Diversity of Data
- Diversity of Models

## Data bias

## Data uncertainty

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

Diversity is the key

- Diversity of Background
  - Diversity of Mindset
  - Diversity of Data
  - Diversity of Models

## Culture bias

Different cultures may see the same data differently.

- Emotion recognition and expression may vary a lot between different cultures
  - A face that is labeled as angry by a Western person may be labeled as surprised by an Asian person
  - Style of writing, gestures, voice tone may vary between different cultures
  - Educational standards are different from country to country

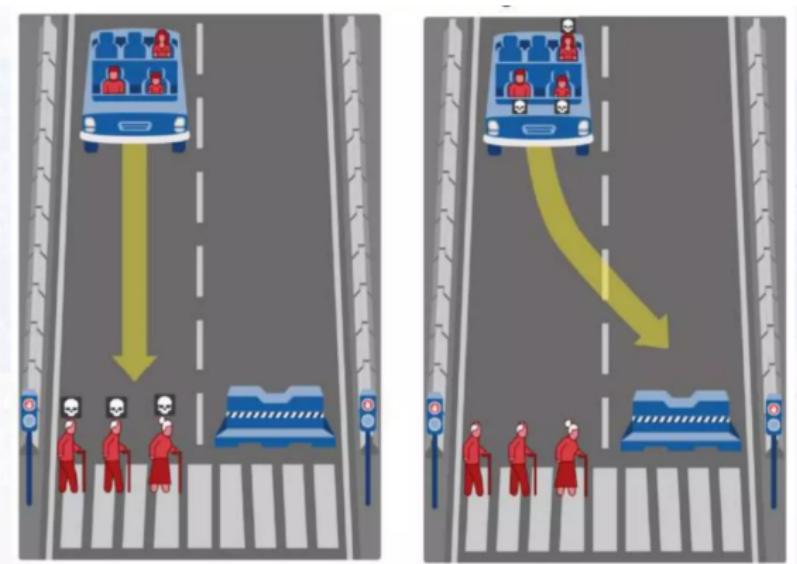
## Can we trust the labeling?

## High risk AI applications

It's not a problem if we build course's projects or even thesis, but a game change for real applications like:

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

## What should the self-driving car do?



Source: Award *et al* "The moral Machine Experiment", Nature, Vol 563, 2018

# We can play the devil's advocate

## Humans are not perfect either

- Juror decision are affected by sports 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 make decisions
  - different cultures = different decision making processes.

However ...

What about human-AI collaboration?

Seems like the perfect solution, but ...

- What if AI is right 99.999% of the time?
  - Should the human check every time?
  - There are cognitive biases whereby, after some time, humans unconsciously trust AI and they no longer are 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 it can convince the human?

## Some examples I



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.



## Some examples II

The image shows a Google search results page for "edward hopper". At the top left is the Google logo. In the search bar, "edward hopper" is typed. To the right of the search bar are icons for microphone, camera, and search. Below the search bar is a horizontal menu with rounded corners containing the following links: "Images", "Shopping", "News", "Videos", "Books", "Death", "Nighthawks", "Wife", and "Whitney". A message below the menu says "About 17,100,000 results (0.65 seconds)". The main title "Edward Hopper" is in large bold letters. Below it is the subtitle "American painter". To the right of the title are five buttons: "Overview" (which is highlighted), "Artworks", "On view", "Videos", and "Products". Below the title are two paintings: one of a woman in a red dress looking out a window, and another of blue houses. To the right is a portrait of Edward Hopper sitting at a desk. On the far right is a sidebar with the heading "Edward Hopper and his paintings" followed by three links: "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-century America.", and a link to "Creative Fabrica".

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

Introduction  
oooooo

The Ethics of AI  
oooooooo

Algorithmic Bias  
oooooooooooo

A Broader Perspective on AI Ethics  
●oooooooooooo

High Risk Applications  
oooooooo

How to build ethical AI models  
oooooooooooo

## A Broader Perspective on AI Ethics

# Who owns AI?

## AI models need (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 models need (a lot of) data

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

## Who owns AI?

## AI models need (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 models need (a lot of) data

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

## AI democratization – a myth? 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).

## AI democratization – a myth? 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**.

# AI democratization – a myth? II

## AI is currently owned by few companies

- They have access to a huge amount of data.
- They attract top AI scientists (huge salaries, freedom).
- They have the power to transform research ideas into real products.

## 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.

# AI democratization – a myth? II

## AI is currently owned by few companies

- They have access to a huge amount of data.
- They attract top AI scientists (huge salaries, freedom).
- They have the power to transform research ideas into real products.

## 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.

No one has a moat

## 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-and-neither>

# AI Singularity

## Exponential progress equal singularity

- We live in a historical moment when the acceleration of progress is becoming more and more visible.
- AI is becoming more “intelligent” than human in many tasks.
- We could potentially substitute humans or delegate activities to AI starting today!
- OpenAI is trying to raise \$100B in coming years to achieving the development of AGI.

In recent time a sort of hysteria arises...

- Pause Giant AI Experiments: An Open Letter
- Google Engineer Claims AI Chatbot Is Sentient
- The Godfather of AI Leaves Google and Warns of Danger Ahead

AI Singularity

## Exponential progress equal singularity

- We live in a historical moment when the acceleration of progress is becoming more and more visible.
  - AI is becoming more “intelligent” than human in many tasks.
  - We could potentially substitute humans or delegate activities to AI starting today!
  - OpenAI is trying to raise \$100B in coming years to achieving the development of AGI.

In recent time a sort of hysteria arises...

- Pause Giant AI Experiments: An Open Letter
  - Google Engineer Claims AI Chatbot Is Sentient
  - The Godfather of AI Leaves Google and Warns of Danger Ahead

## Who owns the data?

Check for example [this](#) paper.

## Extracting Training Data from Diffusion Models

*Nicholas Carlini*<sup>\*1</sup>

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

<sup>\*</sup>Equal contribution   <sup>†</sup>Equal contribution   <sup>‡</sup>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.



*Caption: Living in the light  
with Ann Graham Lotz*

*Prompt:*

**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).

## Ownership vs. Partnership

- What about the 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.

home :: artnews :: news

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



EX-CHANTESCALANTE-DE-MATTEL

March 21, 2023 11:48am



<https://www.artnews.com/>

## Climate impact

## The carbon footprint of training a model

- In 2019 a paper calculated a carbon footprint of 280.000 kg of  $CO_2$  for a single training of a 213M parameters NLP architecture.
  - GPT4 number of parameters is still unknown, but some sources put it as high as 100T ( $10^{14}$ ). A more accurate (and maybe way downward) estimation is 500-1.000B.
  - To put that in perspective, a single training of GPT4 emit at least 560M kg of  $CO_2$  (without taking into account much larger datasets).
  - In order to stop climate change, a person must emit at max. 600kg of  $CO_2$  per year.
  - One training of GPT4 (way downward estimation + without accounting for data storage, web servers, etc.) consume as much as  $\approx 1M$  people in a year.

Seems like it did not acknowledge the problem ...

what is the carbon footprint of chatgpt4

 As an AI language model, ChatGPT-4 does not have a physical presence and therefore does not produce a carbon footprint in the traditional sense. However, the training and deployment of AI models like ChatGPT-4 do have an impact on the environment through the energy consumption of computing infrastructure.

The carbon footprint associated with ChatGPT-4's training and deployment would depend on various factors such as the type and amount of hardware used, the source of energy for that hardware, and the efficiency of the software and algorithms used.

In general, training large AI models like ChatGPT-4 can require significant amounts of computational resources, and therefore energy. Efforts are being made to increase the energy efficiency of AI training and deployment, such as using renewable energy sources and developing more efficient hardware and software.

Overall, while ChatGPT-4 itself does not have a carbon footprint, the development and deployment of AI models like it does have an environmental impact that needs to be considered and addressed.

Now it has changed its mind ...



ChatGPT

Assessing the carbon footprint of a specific service like ChatGPT involves considering various factors, such as the energy consumption of the data centers where the model is hosted, the infrastructure supporting its operation, and the overall environmental practices of the hosting provider.

As of my last knowledge update in January 2022, I don't have specific details on the carbon footprint of ChatGPT or the infrastructure used by OpenAI. The carbon footprint can vary based on the energy sources of the data centers, the efficiency of the hardware, and other operational factors.

To get the most accurate and up-to-date information, you may want to check OpenAI's latest publications, environmental commitments, or contact them directly for information on the carbon footprint and sustainability practices associated with ChatGPT. Keep in mind that organizations often make efforts to improve the sustainability of their services, and information may change over time.



Introduction  
ooooo

The Ethics of AI  
oooooo

Algorithmic Bias  
oooooooooooo

A Broader Perspective on AI Ethics  
oooooooooooo

High Risk Applications  
●oooooooo

How to build ethical AI models  
oooooooooooo

## High Risk Applications

# (DeepFakes) Media generation – only for fun?



## (DeepFakes) Media generation – only for fun?

# This Horrifying App Undresses a Photo of Any Woman With a Single Click

The \$50 DeepNude app dispenses with the idea that deepfakes were about anything besides claiming ownership over women's bodies.



By Samantha Cole

DeepNude

## (DeepFakes) Media generation – only for fun?

TECHNOLOGY

# Paedophiles using AI to create child abuse images

Mark Sellman, Technology  
Correspondent | Octavia  
Sheepshanks

Monday May 01 2023, 12.01am  
BST The Times



The AI program Midjourney is used by millions of people to create images from simple text or image prompts.

LIONEL BONAVENTURE/AFP/GETTY IMAGES

<https://www.thetimes.co.uk/>

# AI in the military domain

INNOVATIONS

## The military wants AI to replace human decision-making in battle

The development of a medical triage program raises a question: When lives are at stake, should artificial intelligence be involved?



By [Pranshu Verma](#)

March 29, 2022 at 7:00 a.m. EDT

<https://www.washingtonpost.com/>

# AI in Education the system

This AI reads children's emotions as they learn

By Miles O'Brien, CNN Business  
13 minutes read • Updated 9:15 AM EST, Wed February 17, 2021

Print X

Share

Embed



© Video by Facebook

<https://edition.cnn.com/>

*Students work on tests and homework on the platform as part of the school curriculum. While they study, the AI measures muscle points on their faces via the camera on their computer or tablet, and identifies emotions including happiness, sadness, anger, surprise and fear. Facial expression recognition AI can identify emotions with human-level accuracy. The system also monitors how long students take to answer questions; records their marks and performance history; generates reports on their strengths, weaknesses and motivation levels; and forecasts their grades.*

## Can we distinguish between humans and AI?

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/>

Seems like we are not there yet

## ADVENTURES IN 21ST-CENTURY WRITING —

# OpenAI confirms that AI writing detectors don't work

No detectors "reliably distinguish between AI-generated and human-generated content."

BENJ EDWARDS - 9/8/2023, 6:42 PM

<https://arstechnica.com/information-technology>

## How to build ethical AI models

# Layers of AI Safety

## Alignment

- Do what I mean given this environment.
- Technology works in intended use-cases.
- “ E.g. bias and fairness.

## Robustness

- Keep doing what I mean in unforeseen environment.
- Technology is safe even in unintended use-cases.
- “ E.g. ethics in decisions and adversarial attacks.

## Corrigibility

- Enable me to detect and correct your mistakes.
- Imperfect technology can be detected and improved over time.
- “ E.g. white box models and explainable AI.

# Layers of AI Safety

## Alignment

- Do what I mean given this environment.
- Technology works in intended use-cases.
- “ E.g. bias and fairness.

## Robustness

- Keep doing what I mean in unforeseen environment.
- Technology is safe even in unintended use-cases.
- “ E.g. ethics in decisions and adversarial attacks.

## Corrigibility

- Enable me to detect and correct your mistakes.
- Imperfect technology can be detected and improved over time.
- “ E.g. white box models and explainable AI.

## Layers of AI Safety

## Alignment

- Do what I mean given this environment.
  - Technology works in intended use-cases.
    - “ E.g. bias and fairness.

## Robustness

- Keep doing what I mean in unforeseen environment.
  - Technology is safe even in unintended use-cases.
  - “ E.g. ethics in decisions and adversarial attacks.

## Corrigibility

- Enable me to detect and correct your mistakes.
  - Imperfect technology can be detected and improved over time.
  - E.g. white box models and explainable AI.

## How to Insert Ethics in AI?

Ethics by design

- Can be paternalistic as it constrains a lot the generability and the freedom of the models.
    - “ i.e. speed bumps (permanent and leaves no real choice, especially in case of an emergency).

## Pro-ethical design

- It does not preclude a course of action, but it requires the agents to make up their mind about it (still forces to make a choice, but less of a paternalistic nudge).
    - “ i.e. a speed camera (leave freedom to choose to pay a ticket, especially in case of an emergency).

## Possible 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*.

## Possible Countermeasures II

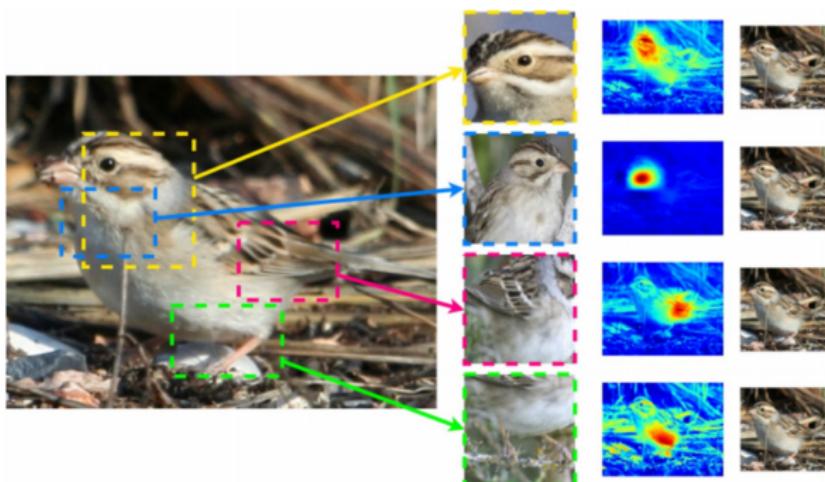
## Explain black-box models

- 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).
  - Explainability *a posteriori*.

## Train robust models

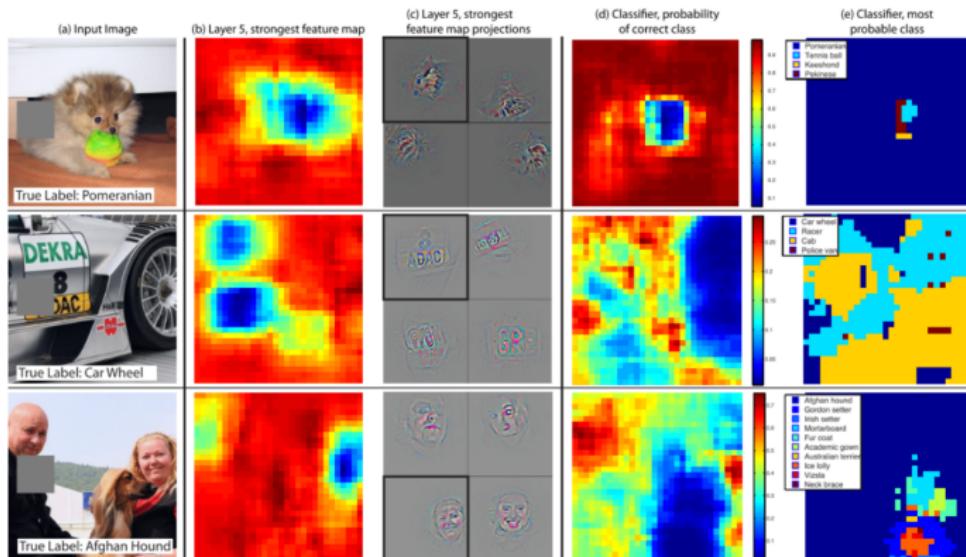
- Create models that will not change prediction if the input data is altered.
  - Train models agnostic to adversarial manipulation of their inputs.

## Possible Countermeasures III



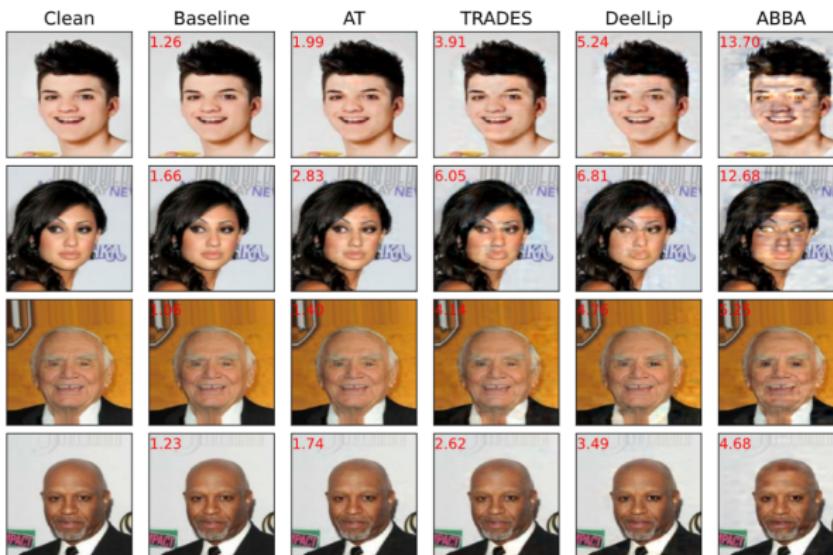
Cynthia Rudin, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, Nature Machine Intelligence, 2019

# Possible Countermeasures IV



Matthew Zeiler and Fergus Rob, *Visualizing and Understanding Convolutional Networks*, ECCV, 2014.

## Possible Countermeasures V



Ana Neacșu, Jean-Christophe Pesquet, Vlad Vasilescu, *ABBA Neural Networks: Coping with Positivity, Expressivity, and Robustness*, Siam SIMODS, 2024

## Possible Countermeasures VI

## Context

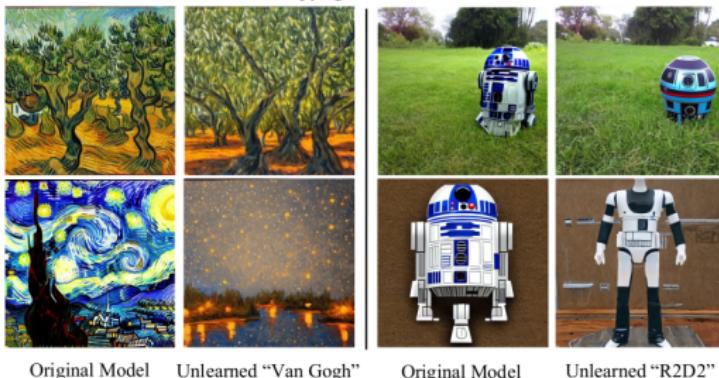
- It is important to ensure **data privacy**.
  - Each individual has the right to request the removal of their personal information from online platforms
  - However, it is not enough to have the data deleted from databases, as different machine learning algorithms may have already been trained using that data

Machine Unlearning (MU)

- It was proposed since retraining the whole model from scratch would be too expensive time and resources wise
  - Allows the model to "forget" data samples through different techniques
  - In the context of image generation, it is useful for removing the ability to generate images:
    - of certain individuals, for privacy;
    - based on certain works of art, which helps artist regarding copyright;
    - that are malicious or inappropriate (NSFW)

## Unlearning of different concepts

## Copyright Protection



**Figura:** Examples of unlearning for NSFW and copyright concepts

## 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 a highly 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 require to discuss the ethics of any submission.

## 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 a highly 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 require to discuss the ethics of any submission.

## Some pieces of advice

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

# Thank you!

---