

Ethics of Artificial Intelligence

Sargur N. Srihari

Topics

1. The case of Large Language Models
2. Performance vs Trust in AI
3. Two definitions of Ethics
4. Why does it matter?
5. Sources of bias

NLP: AI generated output

- Language model (LM) GPT-2 trained to predict next word in a 40GB of internet text generate

The screenshot shows the GPT-2 Explorer interface from the Allen Institute for Artificial Intelligence. At the top left is the AI2 logo and text. At the top right is a blue button labeled "GPT-2 Explorer". Below the header, a message states: "This demonstration uses the public 345M 117M parameter OpenAI GPT-2 language model to generate sentences." Another message below it says: "Enter some initial text and the model will generate the most likely next words. You can click on one of those words to choose it and continue or just keep typing. Click the left arrow at the bottom to undo your last choice." On the left, there is a text input field containing the sentence "The cat sat on the". To the right of the input field is a list of generated options with their respective probabilities:

Probability (%)	Word
8.4%	floor
7.3%	bed
5.6%	couch
4.2%	ground
3.4%	table
3.1%	sofa
2.4%	edge
1.8%	chair
1.2%	desk
1.2%	window

At the bottom of the list is a small "← Undo" link. At the very bottom of the interface, a note reads: "Built at the Allen Institute for Artificial Intelligence using Hugging Face's pytorch-pretrained-BERT library. See the source code on GitHub."

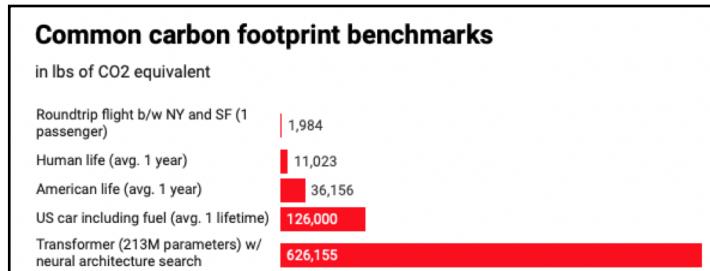
GPT-2 generated synthetic text

- **System Prompt (human-written)**
- *In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
- **Model Completion (machine-written, 10 tries)**
- The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.
- Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.
- Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.
- Pérez and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said Pérez.
- Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

Risks of language models

- “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”
 - Google paper by Gabru, Bender, et.al. 2020, Summary by Karen Ho, Tech. Review

1. Environmental and Financial Costs



The estimated costs of training a model				
	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO ₂ e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	\$41-\$140
Transformer (213M parameters)	Jun, 2017	201	192	\$289-\$981
ELMo	Feb, 2018	275	262	\$433-\$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	\$3,751-\$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	\$942,973-\$3,201,722
GPT-2	Feb, 2019	-	-	\$12,902-\$43,008

2. LMs trained on exponentially increasing amounts of text

- Researchers collect all the data they can from the internet, so racist, sexist, and otherwise abusive language ends up in the training data

3. Misdirected research effort

- LMs don’t *understand* language and are merely excellent at *manipulating* it, Big Tech can make money from models

4. Illusions of meaning

- Easy to use LMs to fool people. Student churned out AI-generated self-help and productivity advice on a blog, which went viral.

Trust in AI

- Is *human-like* good enough?
- What happens to **TRUST** in a world where machines generate human-like output and make human-like decisions?
- Can I trust an autonomous vehicle to have seen me?
- Can I trust the algorithm processing my housing loan to be fair?
- Can we trust the AI in the ER enough to make life and death decisions for us?

Role of Data in AI



But what is RIGHT? And is that enough? (Image: [Machine Learning, XKCD](#))

Definition of Ethics

- Ethics or moral philosophy is a branch of philosophy
 - It involves systematizing, defending, and recommending concepts of right and wrong conduct
- Ethics seeks to resolve questions of human morality by defining concepts such as
 - good and evil,
 - right and wrong,
 - virtue and vice, justice and
 - crime

Two Definitions of Ethics

1. Ethics: *Moral principles governing actions of an individual*

- The “rules” or “decision paths” that help determine what is good or right.
 - Ethics of tech: set of “rules” or “decision paths” used to determine its “behavior”.
 - Typically captured in some design document as a sequence diagram or user story
- But how does team determine a good or right outcome, and for whom?
 - Is it universally good or only for some?
 - Is it good under certain contexts and not in others?
 - Is it good against some yardsticks but not so good for others

2. Ethics: *Dealing with right vs wrong, and moral obligations and duties of humans*

- The ethical quality of its prediction, of the end outcomes drawn out of that and the impact it has on humans
- How right, how fair and how just, is the output, outcome and impact?
 - How well can, and how well does, it identify and declare the contexts it is right, fair and just in, as well as the contexts in which it is not.
 - Can it control how and where it is used?
 - What are the implications of losing control of this context?
- Being answerable to these constitute moral obligations and duties of developers of AI

Why does it matter?

- Given:
 - Dizzying pace of AI mainstreaming
 - Accessible compute power and
 - Open source machine learning libraries
- This change is going to be rapid and at scale

Harms that AI can cause

- A “harm” is caused when a prediction or end outcome negatively impacts an individual’s ability to establish their rightful personhood (harms of representation), leading to or independently impacting their ability to access resources (harms of allocation)
 - Personhood is their identity.
 - incorrectly representing or failing to represent an individual’s identity in a ML system is a harm.
 - Any decision made by this system thereafter, in regards to that individual, is a harm as well.

Landscape of ethics issues

- Applying ethical considerations to our individual situations is natural
 - We are quick to spot instances when we were unfairly treated, for example!
- But applying this for others, in larger, more complex contexts when we are two steps removed from them, is extremely hard
 - Humans Don't Realize How Biased They Are Until AI Reproduces the Same Bias

Summary of ethics issues

- What AI is
 - 1. Bias and Fairness
 - 2. Accountability and Remediability
 - 3. Transparency, Interpretability and Explainability
- What AI does
 - 1. Safety
 - 2. Human-AI interaction
 - 3. Cyber-security and Malicious Use
 - 4. Privacy, Control and Agency (or lack thereof, i.e. Surveillance)
- What AI impacts
 - 1. Automation, Job loss, Labor trends
 - 2. Impact to Democracy and Civil rights
 - 3. Human-Human interaction
- What AI can be
 - threats from human-like cognitive abilities
 - concerns around singularity, control going up to debates around robot rights¹³ (akin to human rights)

What AI is

- An ML system consisting of a learner (model) that, given a set of inputs (data), is able to learn *something* and use that learning to *infer* something else (predictions)
 - Enables man-machine interaction at a different level
 - Autonomous robots, ie robots that operate without requiring real-time or frequent instructions from a human (e.g. automobiles, drones, vacuum cleaners, twitter bots etc.) impact and change our operating environments.
 - Indirectly they change our behavior

What AI impacts

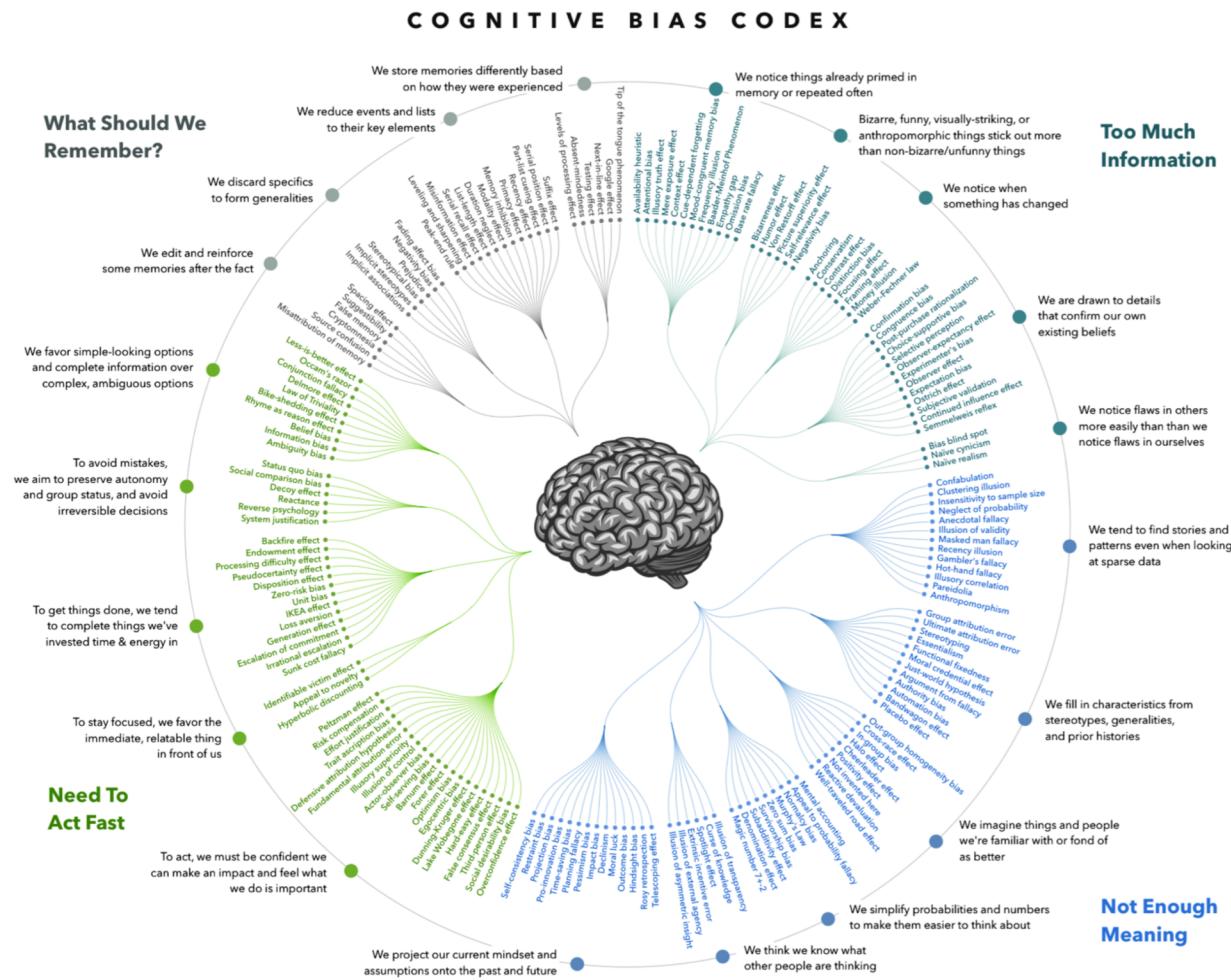
- Technology products often trigger second and third order consequences that are not always obvious at first.
- Lines between real-person and automated-bot start to blur online, as humans are bombarded with “information”, often fake, at scale, the very notion of trust and therefore safety is shaken.

Ethics of What AI is

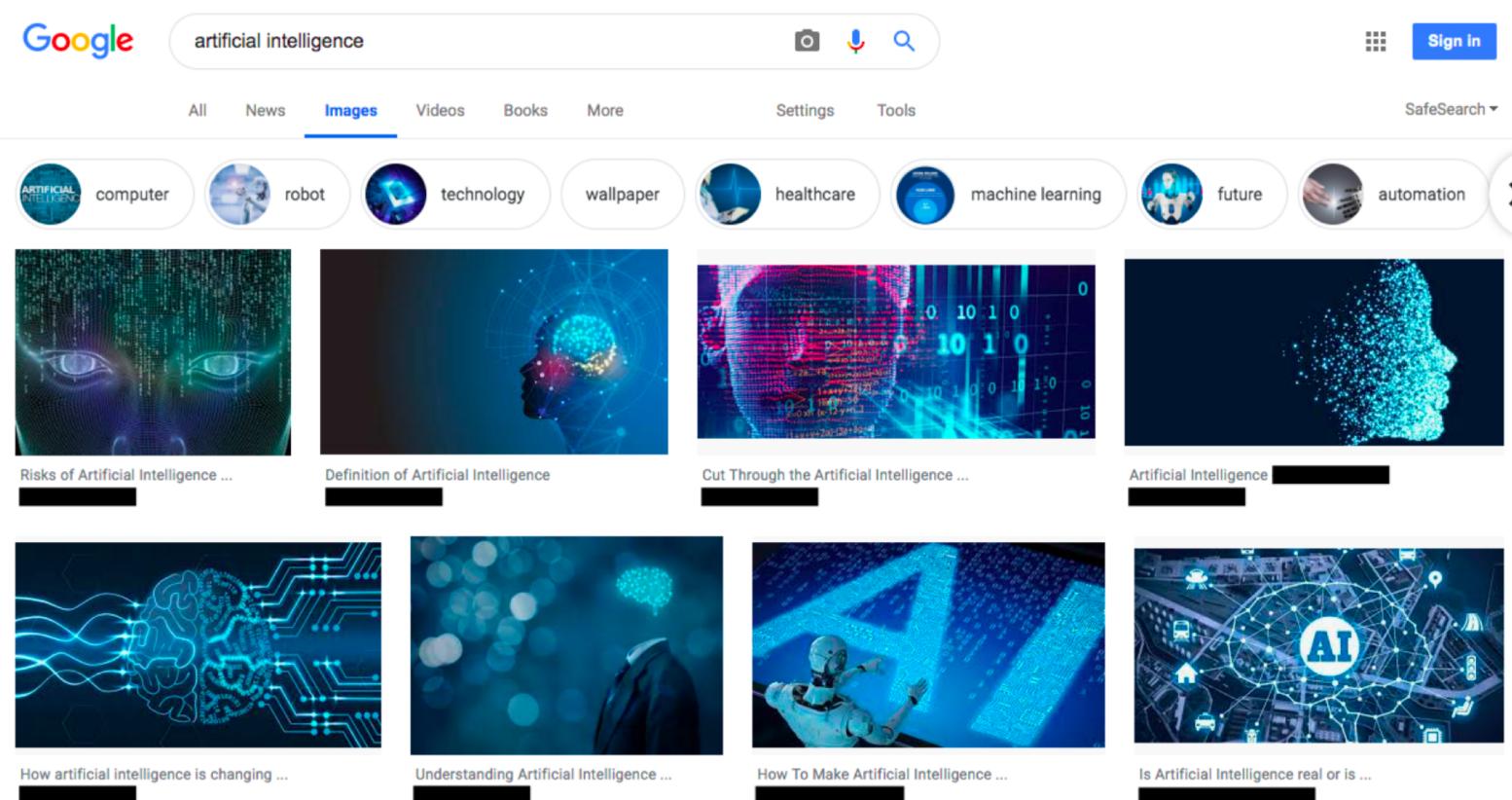
- Commonly deployed AI is a set of math functions (**model**) that given some inputs (**data**), learn *something* and use that to *infer* something else (make **predictions**)
- Ethical exploration of this realm covers issues like **Bias** in a models' predictions and **Fairness** (or lack thereof) of the outcomes; as well as approaches to address them via **Accountability** and **Transparency**.

Cognitive Bias

Cognitive biases, all [104](#) of them, are inherent to humans and affect how we make decisions. And the AI we build.



Bias by Search history



Not this, Not this! (Top 6 Google Image Search Results for AI de-biased for authors' search history)

Algorithmic Systems have real-world consequences



It gives us immense pleasure to announce that we have won AI Innovation Challenge hosted by Govt of Maharashtra and NITI Aayog. We got a chance to develop a Classroom Behavior Analysis with Maharashtra Government for the improving the quality of education in Govt Schools.
**#artificialintelligence #computervision #emotionRecognition #aiforgood
#startupindia #dlycam**



<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

<https://www.dailymail.co.uk/news/article-6927219/YouTube-fact-check-algorithm-incorrectly-tags-live-broadcast-Notre-Dame-fire-9-11-attacks.html>

<https://www.linkedin.com/feed/update/urn:li:activity:6512705571802714112>

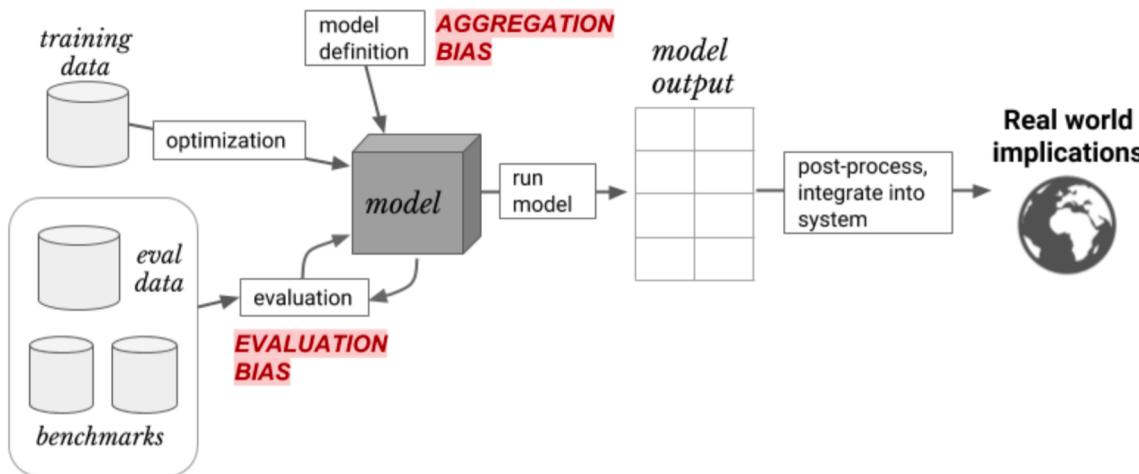
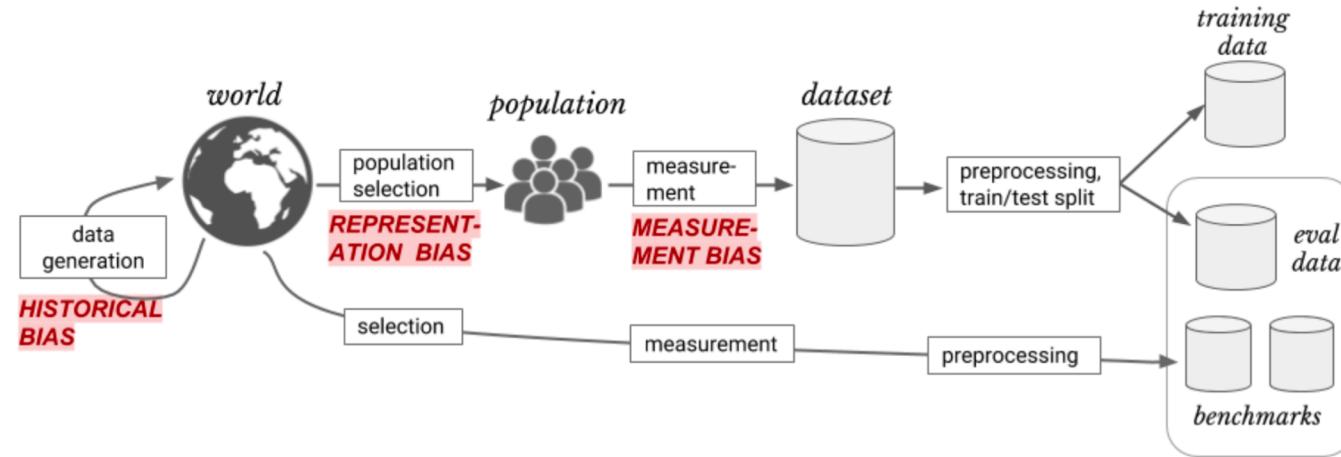
Bias, as in “a model not representing the input data or ground truth accurately enough”, is an ML problem.

Bias, as in “a model reflecting undue prejudice in its predictions” is not simply an ML problem

Sources of bias

- Data is a big source of bias in ML
- Bias can seep from other sources too; all the way from how AI researchers (read humans) frame the problem to how they train the model to how the system gets deployed
 - Women comprise only 10% of AI research staff at Google and only 2.5% of Google's workforce is black
 - This lack of representation is what leads to biased datasets and ultimately algorithms that are much more likely to perpetuate systemic biases.

Five potential sources of harm



Source: H. Suresh, J. Guttag, A Framework for Understanding Unintended Consequences of Machine Learning
<https://arxiv.org/pdf/1901.10002.pdf>

Historical Bias

- **Historical bias** arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model.
- It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

Representation Bias

- **Representation bias** arises while defining and sampling a development population.
- It occurs when the development population under-represents, and subsequently causes worse performance, for some part of the final population.

Measurement Bias

- **Measurement bias** arises when choosing and measuring the particular features and labels of interest.
- Features considered to be relevant to the outcome are chosen, but these can be incomplete or contain group- or input-dependent noise.
- In many cases, the choice of a single label to create a classification task may be an oversimplification that more accurately measures the true outcome of interest for certain groups.

Evaluation Bias

- **Evaluation bias** occurs during model iteration and evaluation, when the testing or external benchmark populations do not equally represent the various parts of the final population.
- Evaluation bias can also arise from the use of performance metrics that are not granular or comprehensive enough.

Aggregation Bias

- **Aggregation bias** arises when flawed assumptions about the population affect model definition. In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups.