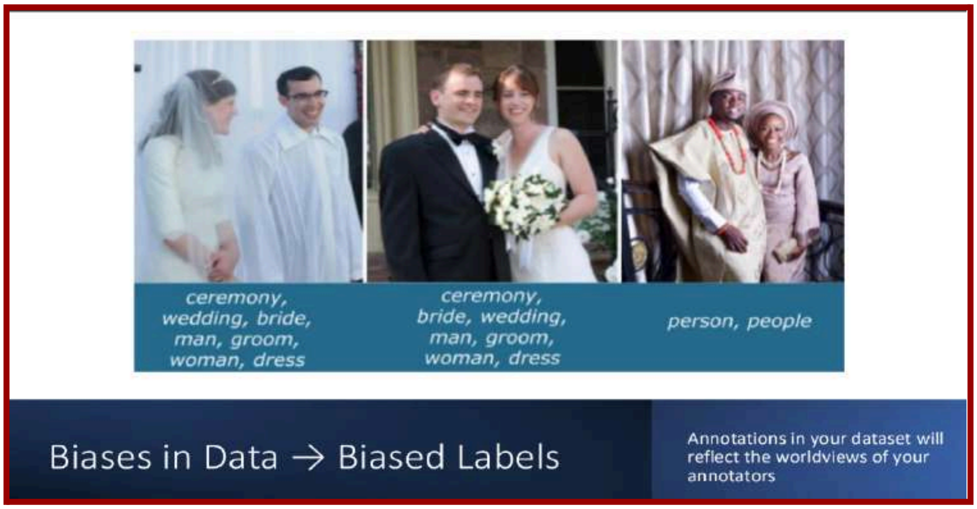
Gen AI & LLMs - Additional Mandatory Reading : Week 7

Q.What are Biases in AI?

Biases in data, particularly in terms of data representation, refer to the issue where certain groups are depicted in a less favorable or inaccurate manner compared to others, even if there is a sufficient amount of data for each group. This means that while there may be ample data available for various groups, the way these groups are represented can still be skewed or biased, affecting the fairness and accuracy of analyses and outcomes.


ceremony, wedding, bride, man, groom, woman, dress
ceremony, bride, wedding, man, groom, woman, dress
person, people

Biases in Data → Biased Labels

Annotations in your dataset will reflect the worldviews of your annotators

CHAINSAWSUIT.COM
i've heard the rhetoric from both sides... time to do my own research on the real truth
Google
Literally the first link that agrees with what you already believe
Completely supports your viewpoint without challenging it in any way
Another link
...jackpot

- The tendency to search for, interpret, favor, recall information in a way that confirms preexisting beliefs

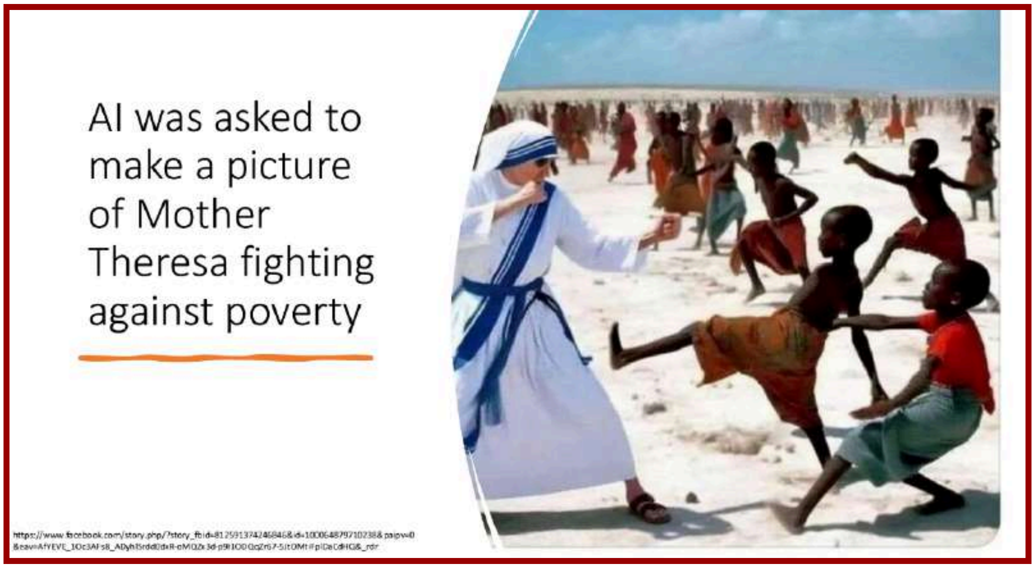
Biases in Interpretation→ Confirmation Bias

Q.(Un)-Fairness inAI?

Unfairness in AI arises primarily from biases present in training data. If data reflects existing societal inequalities, AI systems can perpetuate these biases, leading to discriminatory outcomes. For example, biased historical data might result in AI favoring certain demographic groups over others. Algorithmic bias can also occur due to how models are designed or optimized. If an algorithm prioritizes efficiency without considering fairness, it may disadvantage specific groups inadvertently. Moreover, a lack of diversity within AI development teams can lead to overlooked biases and blind spots, exacerbating the problem.Additionally, AI systems can impact different groups unevenly, such as facial recognition technology showing varied accuracy across racial or ethnic groups. Transparency and accountability issues further complicate the problem, as opaque decision-making processes make it challenging to detect and rectify biases. Addressing these issues requires improving data practices, algorithm design, and enhancing team diversity.People who are the most marginalized,people who'd benefit the most from such technology, are also the ones who are more likely to be systematically excluded from this technology.

Q. What is Hallucination?

AI hallucination refers to a phenomenon where a large language model (LLM), such as a generative AI chatbot or computer vision tool, produces outputs that are nonsensical or inaccurate because it perceives patterns or objects that don’t actually exist or are imperceptible to humans. It also refers to instances where the system generates false or fabricated information. When an AI lacks precise knowledge about a topic, it might produce inaccurate or invented responses. For example, if you ask an AI about a specific historical event it hasn't been trained on, it might create a plausible-sounding but entirely fictional account.



AI was asked to make a picture of Mother Theresa fighting against poverty

https://www.facebook.com/story.php/?story_fbid=812591374246846&id=100064879710238&paipv=0 &eav=AfYEVE_1Oc3AFs8_ADyh5rdd0dxR-oMQ2x3d-p9I1OOQqZr67-5ItOMt#plDaCdHG&_rdr

Typically, users expect generative AI tools to provide responses that accurately address their prompts. However, sometimes the AI generates outputs that deviate from the training data, are incorrectly interpreted by the model, or lack a coherent pattern. This results in the AI "hallucinating" a response.Though the term "hallucination" is usually associated with human or animal experiences, it metaphorically captures the essence of these erroneous outputs, particularly in image and pattern recognition, where the results can be quite surreal. Similar to how humans might see shapes in clouds or faces on the moon, AI hallucinations arise from issues like overfitting, data bias, or complex model behavior.

Preventing AI Hallucinations

1. Use High-Quality Training Data: Ensure AI models are trained on diverse, balanced, and well-structured data to minimize output bias and improve accuracy.
2. Define the Model’s Purpose: Clearly outline the AI model's intended use and limitations to reduce irrelevant or incorrect outputs.
3. Use Data Templates: Implement predefined data formats to ensure consistent and reliable AI outputs.
4. Limit Responses: Set boundaries for AI models using filtering tools and probabilistic thresholds to enhance accuracy and consistency.
5. Test and Refine Continuously: Rigorously test and regularly evaluate the AI model to improve performance and address emerging issues.
6. Rely on Human Oversight: Include human reviewers to validate and correct AI outputs, leveraging their expertise to ensure accuracy and relevance.

Q. Ethical AI : What it is and Why it matters?

AI ethics encompass a set of guiding principles used by various stakeholders—ranging from engineers to government officials—to ensure that artificial intelligence technologies are developed and utilized in a responsible manner. This involves adopting practices that are safe, secure, humane, and environmentally sustainable.

A comprehensive AI code of ethics typically addresses issues such as avoiding bias, safeguarding user privacy and data, and mitigating environmental impacts. These ethical guidelines can be implemented through company-specific codes of conduct or through government-led regulatory frameworks. Both approaches play a crucial role in regulating AI technology by addressing both global and national ethical concerns and establishing policy foundations for ethical AI within organizations.The conversation around AI ethics has evolved beyond academic research and non-profit organizations. Major tech companies like IBM, Google, and Meta have formed teams to address the ethical challenges associated with handling vast amounts of data. Simultaneously, governmental and intergovernmental bodies are developing regulations and ethical policies informed by academic research.

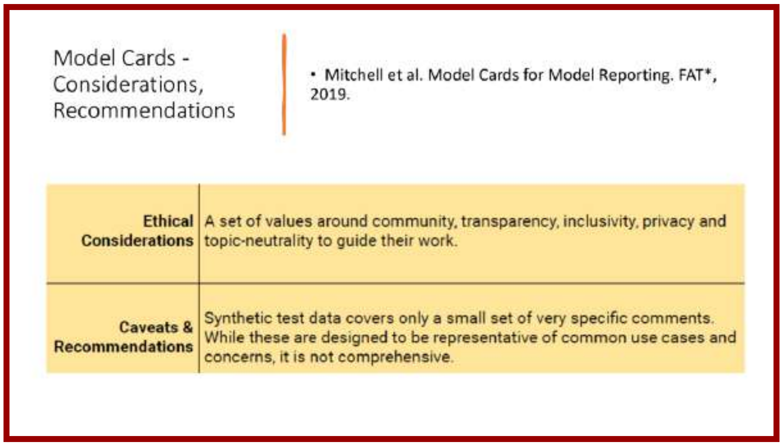
Ethical Challenges of AI

Several real-world issues highlight the importance of AI ethics. Here are a few examples:

- **AI and Bias**
  AI systems may exhibit bias if they are trained on data that doesn't accurately reflect the diversity of the population. For instance, in 2018, Amazon faced criticism for an AI recruiting tool that downgraded resumes mentioning "women" (e.g., "Women's International Business Society"), leading to allegations of gender discrimination and legal challenges.

- **AI and Privacy**
  AI often utilizes data from internet searches, social media, online purchases, and more to enhance personalization. However, this raises concerns about the lack of genuine consent for companies to access and use personal information, as illustrated by issues with tools like Lensa AI.

- **AI and the Environment**
  Training large AI models can be energy-intensive, posing environmental concerns. While research is underway to develop more energy-efficient AI methods, there is still a need to integrate environmental considerations into AI-related policies more effectively.



Model Cards -
Considerations,
Recommendations

• Mitchell et al. Model Cards for Model Reporting. FAT*, 2019.

Ethical Considerations	A set of values around community, transparency, inclusivity, privacy and topic-neutrality to guide their work.
Caveats & Recommendations	Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Q. Environmental AI?

As Artificial Intelligence (AI) continues to drive innovation and change, growing concerns are emerging about its environmental impact. The ecological consequences of AI, including issues such as carbon emissions, electronic waste, and potential threats to ecosystems. We explore how proactive measures and ethical practices can help mitigate these concerns, aiming for a future where AI advancements and environmental sustainability go hand in hand.

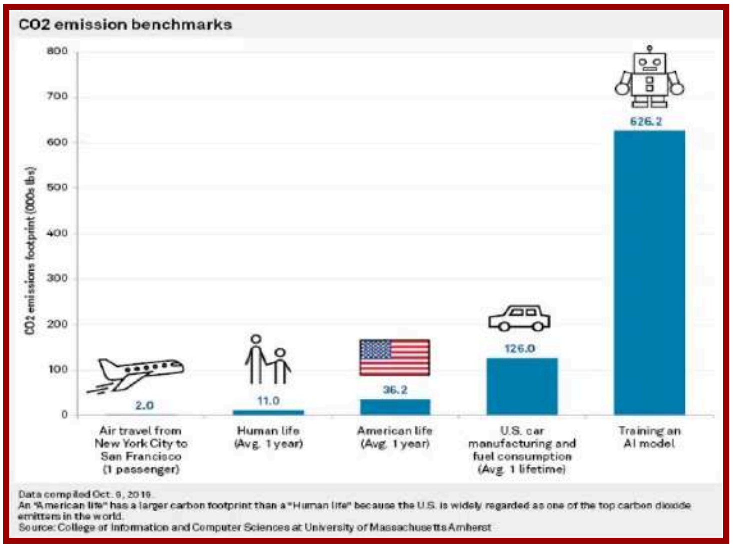
While AI is celebrated for its transformative potential, it also raises significant environmental challenges. The lifecycle of AI technology—from development and maintenance to disposal—contributes to a substantial carbon footprint.

The environmental challenges associated with AI and highlight the urgent need for responsible action. By understanding these impacts and adopting sustainable practices, we can work towards a future where AI contributes positively without compromising environmental health.

AI's Environmental Impact

Behind the impressive capabilities of AI lies a process that consumes significant amounts of energy, contributing to a considerable carbon footprint.As datasets grow larger and models become more complex, the energy required to train and operate AI models increases dramatically. This surge in energy consumption directly impacts greenhouse gas emissions, worsening the effects of climate change. OpenAI researchers have noted that since 2012, the computing power necessary to train advanced AI models has doubled approximately every 3.4 months. By 2040, emissions from the Information and Communications Technology (ICT) sector are predicted to account for 14% of global emissions, with the majority stemming from ICT infrastructure, particularly data centers and communication networks. These trends underscore the pressing need to tackle AI’s carbon footprint and its contribution to environmental degradation.

A recent study by researchers at the University of Massachusetts assessed the energy usage of training several widely-used large AI models. The findings revealed that training these models can generate approximately 626,000 pounds of carbon dioxide—comparable to around 300 round-trip flights between New York and San Francisco—or nearly five times the lifetime emissions of an average car.



Data compiled Oct. 9, 2019.
An "American life" has a larger carbon footprint than a "Human life" because the U.S. is widely regarded as one of the top carbon dioxide emitters in the world.
Source: College of Information and Computer Sciences at University of Massachusetts Amherst

Q. Explainable AI?

Explainable Artificial Intelligence (XAI) refers to a collection of processes and methods that enable human users to understand and trust the outputs generated by machine learning algorithms.

XAI aims to clarify how an AI model functions, its anticipated effects, and any potential biases it may harbor. It helps assess a model's accuracy, fairness, transparency, and outcomes, which is essential for AI-driven decision-making. For organizations, explainable AI plays a key role in fostering trust and confidence when deploying AI models in real-world applications. Moreover, AI explainability promotes responsible AI development within an organization.

As AI technology advances, it becomes increasingly difficult for humans to grasp and trace how algorithms reach their conclusions. Often referred to as "black box" models, these systems generate results from data in a way that is opaque, even to the engineers or data scientists who developed them, leaving the decision-making process inside the model unclear.

Understanding how an AI system produces a specific outcome offers numerous benefits. It allows developers to verify that the system operates as intended, helps meet regulatory compliance, and provides individuals impacted by AI-driven decisions the opportunity to challenge or alter those outcomes.

Q.Why explainable AI matters?

Machine learning models are often viewed as black boxes, making them difficult to interpret, especially deep learning neural networks, which are among the most challenging for humans to understand. Bias—typically related to factors such as race, gender, age, or location—has long been a concern in training AI models. Additionally, AI model performance can deteriorate over time as the data used in production differs from the original training data. This highlights the importance of continuously monitoring and managing models to ensure explainability while assessing the business impact of these algorithms.

Explainable AI is essential for building end-user trust, ensuring model auditability, and fostering the effective use of AI. It also helps mitigate the risks associated with compliance, legal issues, security, and the organization's reputation when deploying AI systems in production.

Use Cases for Explainable AI

- **Healthcare:** Enhance diagnostics, medical image analysis, and resource optimization, leading to faster and more accurate medical diagnoses. Improve transparency and traceability in patient care decision-making, and streamline the pharmaceutical approval process through explainable AI.

- **Financial Services:** Enhance customer experience with transparent loan and credit approval processes. Speed up assessments for credit risk, wealth management, and financial crime risk. Facilitate quicker resolution of complaints and issues while boosting confidence in pricing, product recommendations, and investment services.

- **Criminal Justice:** Optimize prediction and risk assessment processes. Use explainable AI to accelerate DNA analysis, prison population assessments, and crime forecasting. Identify and address potential biases in training data and algorithms.