Unit 8
AI Ethics and Values

- AI ethics refers to the ethical principles and guidelines that govern the design, development, and deployment of AI technologies.
- AI ethics aims to ensure that AI systems are developed and used in ways that are fair, transparent, accountable, and aligned with human values.

- Example 1:
- Suppose a CCTV camera was to spot your face in a crowd outside a sports stadium. In the police data center somewhere in the city/ country, an artificial neural network analyzes images from the CCTV footage frame by-frame. A floating cloud in the sky causes a shadow on your face and the neural network (by mistake) finds your face similar to the face of a wanted criminal.
- If the police were to call you aside for questioning and tell you they had reason to detain you, how would you defend yourself?
- Was it your fault that your shadowed face has resemblance by few degrees with a person in the police record?

Example 2:This happened in the USA in 2018. An AI system was being used to allocate care to nearly 200 million patients in the US. It was discovered later that the AI system was offering a lower standard of care to the black patients. Across the board, black people were assigned lower risk scores than white people. This in turn meant that black patients were less likely to be able to access the necessary standard of care.

The problem stemmed from the fact that the AI algorithm was allocating risk values using the predicted cost of healthcare. Because black patients were often less able to pay or were perceived as less able to pay for the higher standard of care, the AI essentially learned that they were not entitled to such a standard of treatment. Though the system was fixed / improved after being discovered, the big question is – whose problem was this? The AI system developers or the US black people data (which was true to an extent)?

Five Pillars at a Glance

- Transparency
- Fairness
- Privacy
- Accountability
- Robustness

- Explainability refers to the transparency and interpretability of AI systems, allowing users to understand how algorithms make decisions and predictions. Explainable AI enables stakeholders to comprehend the underlying logic, factors, and considerations driving algorithmic outcomes, fostering trust, accountability, and ethical use of AI technologies. Explainability is essential for ensuring that AI systems are transparent, accountable, and aligned with ethical principles.

- Fairness in AI seeks to remove bias and discrimination from algorithms and decision-making models. Machine learning fairness addresses and eliminates algorithmic bias from machine learning models based on sensitive attributes like race and ethnicity, gender, sexual orientation, disability, and socioeconomic class.

- Robustness in AI systems indeed refers to their ability to consistently provide accurate and reliable results regardless of the conditions they encounter and for extended periods. It is all about making sure that AI algorithms and systems operate as expected without any unexpected errors or deviations from their intended behavior. This involves ensuring stability in the algorithms, being able to reproduce results, and maintaining consistent performance across different datasets and environments. Achieving reliability in AI systems requires thorough testing, validation, and quality assurance at every stage of development.

- Transparency involves openness and disclosure about the design, operation, and implications of AI systems. Transparent AI frameworks provide clear documentation, disclosure, and communication about the data, algorithms, and decision-making processes used in AI applications. Transparency promotes accountability, scrutiny, and informed decision-making, enabling users and stakeholders to assess the ethical implications and societal impacts of AI technologies.

- Privacy refers to the right of individuals to control their personal information and to be free from unwanted intrusion into their lives. It encompasses the ability to keep certain aspects of one's life private, such as personal communications, activities, and personal data. Privacy is essential for safeguarding personal autonomy, dignity, and freedom from unwarranted interference.

Activity: Organize the class into groups. Reflect on the following points based on the video links given

- Video: AI for Good
- Reflect on the video "The Ethical Robot" and identify two ethical dilemmas that stood out to you. Document these questions.
- Using "How to build a moral robot" as a reference point, list the moral and ethical principles you wish to embed in your robot. Consider the video as inspiration, but do not feel constrained by its content. Feel free to expand your thoughts with creativity and innovation.
- Assemble a team of five students and collectively watch the video "Humans need not apply." It is recommended to view the video multiple times. Following your discussions, compile a group paper summarizing your insights and interpretations from the video.

BIAS, BIAS AWARENESS, SOURCES OF BIAS

- Bias, in simple terms, means having a preference or tendency towards something or someone over others, often without considering all the relevant information fairly.
- It can lead to unfair treatment or decisions based on factors like personal beliefs, past experiences, or stereotypes.
- In everyday life, bias can affect how we perceive and interact with people, situations, or ideas.
- In the context of artificial intelligence, bias refers to when AI systems make unfair or inaccurate decisions due to flawed data or built-in assumptions, which can result in unfair outcomes for certain groups of people.

- In today's interconnected world, artificial intelligence (AI) technologies play an increasingly prominent role in various aspects of our lives, from healthcare to finance to criminal justice.
- However, as AI systems become more pervasive, it is essential to recognize and address the presence of bias in these technologies.
- Bias awareness means understanding that AI systems might have unfair preferences because of different things like the information they were taught with, the rules they follow, or the ideas they were built upon.
- So, being aware of bias in AI is like knowing that sometimes AI might make unfair choices or judgments because of how it was trained or made.

- Activity:
- Question 1: Why are most images that show up when you do an image search for “vacation” seen as beaches?
- Question 2: Why are most images that show up when you do an image search for “nurse” seen as females?
- Question 3: Organize students into groups and ask them to find answers for the questions given below after going through the link Amazon Recruitment Tool

- Discussion Questions:
- How do algorithmic hiring systems function, and what criteria are typically used to evaluate job applicants?
- What are the ethical implications of using biased algorithms in hiring processes, particularly regarding fairness, equal opportunity, and diversity?
- How might biased hiring algorithms perpetuate systemic inequalities in employment and hinder efforts to promote inclusivity in the workforce?
- Reference: https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/ 138

- AI bias, also referred to as machine learning bias or algorithm bias, refers to AI systems that produce biased results that reflect and perpetuate human biases within a society, including historical and current social inequality. Bias can be found in the initial training data, the algorithm, or the predictions the algorithm produces. When bias goes unaddressed, it hinders people’s ability to participate in the economy and society. It also reduces AI’s potential.

- The source of bias in AI Eliminating AI bias requires drilling down into datasets, machine learning algorithms and other elements of AI systems to identify sources of potential bias.
- 1. Training data bias AI systems learn to make decisions based on training data, so it is essential to assess datasets for the presence of bias. One method is to review data sampling for over- or underrepresented groups within the training data. For example, training data for a facial recognition algorithm that over-represents white people may create errors when attempting facial recognition for people of color. Similarly, security data that includes information gathered in geographic areas that are predominantly black could create racial bias in AI tools used by police. Bias can also result from how the training data is labeled. For example, AI recruiting tools that use inconsistent labeling or exclude or over-represent certain characteristics could eliminate qualified job applicants from consideration.

- 2. Algorithmic bias Using flawed training data can result in algorithms that repeatedly produce errors, unfair outcomes, or even amplify the bias inherent in the flawed data. Algorithmic bias can also be caused by programming errors, such as a developer unfairly weighting factors in algorithm decision-making based on their own conscious or unconscious biases. For example, indicators like income or vocabulary might be used by the algorithm to unintentionally discriminate against people of a certain race or gender.

- 3. Cognitive bias When people process information and make judgments, we are inevitably influenced by our experiences and our preferences. As a result, people may build these biases into AI systems through the selection of data or how the data is weighted. For example, cognitive bias could lead to favoring datasets gathered from Americans rather than sampling from a range of populations around the globe.

- Examples of AI bias in real life
- ● Healthcare—Underrepresented data of women or minority groups can skew predictive AI algorithms. For example, computer-aided diagnosis (CAD) systems have been found to return lower accuracy results for black patients than white patients.
- ● Online advertising—Biases in search engine ad algorithms can reinforce job role gender bias. Independent research at Carnegie Mellon University revealed that Google’s online advertising system displayed high-paying positions to males more often than to women.
- ● Image generation—Academic research found bias in the generative AI art generation application Midjourney. When asked to create images of people in specialized professions, it showed both younger and older people, but the older people were always men, reinforcing gender bias of the role of women in the workplace.

- Activity: Role Play Share the following examples of biased AI systems and their potential consequences and ask students to do a role play to present each scenario:
- ● Facial Recognition Technology:
- ○ Example: Facial recognition systems have been shown to exhibit bias against certain demographic groups, particularly people with darker skin tones and women.
- ○ Consequences: Biased facial recognition algorithms can lead to misidentification and wrongful arrests, disproportionately affecting marginalized communities and eroding trust in law enforcement.

- Predictive Policing Algorithms:
- ○ Example: Predictive policing algorithms use historical crime data to forecast future criminal activity and allocate law enforcement resources. However, studies have found that these algorithms can perpetuate racial and socioeconomic biases, leading to over-policing of minority neighborhoods.
- ○ Consequences: Biased predictive policing algorithms may exacerbate racial profiling and discrimination, fueling tensions between law enforcement agencies and communities of color and undermining public trust in the criminal justice system.

- Algorithmic Hiring Systems:
- ○ Example: AI-powered hiring systems are used by companies to screen job applications and identify potential candidates. However, research has shown that these systems can perpetuate gender and racial biases, favoring certain demographic groups over others.
- ○ Consequences: Biased hiring algorithms may reinforce existing disparities in employment opportunities, leading to discrimination against underrepresented groups and hindering efforts to promote diversity and inclusion in the workforce.

- Healthcare Algorithms:
- ○ Example: AI algorithms are increasingly used in healthcare for tasks such as diagnosing diseases and predicting patient outcomes. However, studies have identified biases in healthcare algorithms that can result in differential treatment recommendations based on factors such as race or socioeconomic status.
- ○ Consequences: Biased healthcare algorithms may lead to disparities in patient care, with certain demographic groups receiving suboptimal or inequitable treatment. This can contribute to worsened health outcomes and perpetuate healthcare inequalities.

- Healthcare Algorithms:
- ○ ExamplCredit Scoring Systems:
- ○ Example: AI-powered credit scoring systems are used by financial institutions to assess individuals' creditworthiness and determine their eligibility for loans and other financial products. However, these systems have been found to exhibit biases that disproportionately disadvantage certain demographic groups, such as low-income individuals and people of color.
- ○ Consequences: Biased credit scoring algorithms may limit access to financial opportunities for marginalized communities, perpetuating socioeconomic inequalities and hindering economic mobility.e: AI algorithms are increasingly used in healthcare for tasks such as diagnosing diseases and predicting patient outcomes. However, studies have identified biases in healthcare algorithms that can result in differential treatment recommendations based on factors such as race or socioeconomic status.
- ○ Consequences: Biased healthcare algorithms may lead to disparities in patient care, with certain demographic groups receiving suboptimal or inequitable treatment. This can contribute to worsened health outcomes and perpetuate healthcare inequalities.

- MITIGATING BIAS IN AI SYSTEMS
- Mitigating bias in AI systems is essential for several reasons. Firstly, when AI systems have bias, they can make existing problems like unfairness and discrimination even worse. For example, biased algorithms used in hiring processes may unfairly disadvantage certain groups, leading to systemic discrimination. Secondly, biased AI makes people trust technology less. If people don't trust AI to make fair decisions, they might not want to use it, which can cause problems for everyone. Lastly, addressing bias is essential for upholding ethical principles and ensuring that AI technologies are developed and used responsibly.

- Strategies for Mitigating Bias
- There are several strategies and techniques for mitigating bias in AI systems:
- Using Diverse Data: To reduce bias, we should use lots of different kinds of information to teach AI. This way, the AI can learn from many different examples and viewpoints, making it less likely to be biased.
- Detecting Bias: We need ways to find and measure bias in AI systems before they are used. This could mean looking at how the AI makes decisions for different groups of people or using special tools to see if the AI is being fair.
- Fair Algorithms: We can make AI systems fairer by using special algorithms that are designed to be fair. These algorithms make sure to consider fairness when making decisions, helping to reduce bias.

- Being Transparent: It is important for AI systems to be clear and explain how they make decisions. When people understand how AI works, they can see if there is any bias and fix it.
- Inclusive Teams: When creating AI, it is helpful to have a team of people from different backgrounds and experiences. This way, they can spot biases that others might miss and make sure the AI is fair for everyone.

- Activity:
- Allow students to examine various forms of media, such as news articles, advertisements, or social media posts, and identify instances of bias based on factors like race, gender, or socio-economic status. Encourage them to discuss how bias can influence perceptions and stereotypes.

- DEVELOPING AI POLICIES
- Developing AI policies is essential for ensuring that AI technologies are used responsibly, safely, and ethically, while also promoting innovation and public trust.
- Rules for AI should start with being good to people and respecting their rights. This means treating everyone fairly, being honest about how AI works, making sure it is safe, and being accountable if something goes wrong.
- We need clear rules and standards for how AI is used. These rules should cover important things like protecting people's information, making sure AI does not have unfair biases, keeping it safe, and making sure people can ask questions about how AI works.
- When making these rules, it is important to talk to lots of different people. This includes government people, business leaders, scientists, community groups, and regular people. Everyone's opinion matters because AI affects everyone.
- Before using AI, we should check to see if there are any problems or risks. This means thinking about what could go wrong and making plans to fix it.

- 1.IBM AI Ethics Board:
- Focus: Ethical development and deployment of AI technologies across various industries.
- Components:
- Development of ethical principles and guidelines for AI research and development.
- Recommendations for addressing ethical considerations such as fairness, transparency, accountability, and bias mitigation in AI systems.
- Engagement with stakeholders, including researchers, policymakers, and industry partners, to promote dialogue and collaboration on ethical AI practices.
- Support for educational initiatives and resources to raise awareness and understanding of AI ethics among developers, users, and the public.

- 2. Microsoft’s Responsible AI Page:
- Focus: Corporate responsibility and ethics in AI
- Components:
- Principles for responsible AI development and deployment, including fairness, reliability, privacy, and inclusivity.
- Tools and resources for integrating ethical considerations into AI projects, such as fairness assessments and bias detection algorithms.
- Case studies and best practices for implementing responsible AI practices across various industries and domains.

- 3. Artificial Intelligence at Google:
- Focus: Corporate AI ethics and governance
- Components:
- Google's principles for ethical AI development, encompassing areas such as fairness, safety, privacy, and accountability.
- Guidelines for designing AI systems that prioritize human values and societal well-being.
- Commitments to transparency, collaboration, and continuous improvement in AI governance and decision-making.

- 4. European Union’s Ethics Guidelines for Trustworthy AI—Press Release:
- Focus: Ethical guidelines for AI development and deployment in the EU
- Components:
- Principles for trustworthy AI, including respect for human autonomy, prevention of harm, fairness, and accountability.
- Requirements for transparency, explainability, and auditability in AI systems.
- Recommendations for ensuring human oversight and accountability mechanisms in AI applications with high societal impact.

MORAL MACHINE GAME

SURVIVAL OF THE BEST FIT GAME

THANK YOU