AI Ethics

can code communities

Review Questions

- What is a core and what is its function?
- Why are GPUs useful for video and gaming?
- Why are GPUs useful for AI training?

Review Take Home

Imagine you're designing a system for Company X, a weather forecasting company that wants to leverage social media data (photos and tweets) to improve their predictions. Consider the different stages involved in processing and analyzing this data:

- Data Ingestion & Preprocessing: How will the system handle the large volume of incoming social media data?
- Data Cleaning & Filtering: What steps are needed to ensure data quality and relevance for weather forecasting?
- Image Recognition (for photos): How can the system identify weather conditions from user-uploaded photos?
- Text Analysis (for tweets): How can the system extract meaningful weather information from text-based data?
- Data Fusion & Model Integration: How can social media data be combined with existing weather models to improve forecasting accuracy?

For each stage, identify the most suitable computer architecture(s) and explain your reasoning. Consider factors like processing speed, scalability, and cost. There might be multiple valid approaches; focus on justifying your choices based on the specific tasks involved. This assignment will explore how different computer architectures work together in big data applications.

AI Ethics

- Ethics is a set of moral principles which help us discern between right and wrong.
- AI ethics is a multidisciplinary field that studies how to optimize AI's beneficial impact while reducing risks and adverse outcomes.
    - Examples of AI ethics issues include data responsibility and privacy, fairness, explainability, robustness, transparency, environmental sustainability, inclusion, moral agency, value alignment, accountability, trust, and technology misuse.

Importance of Ethical Considerations in AI

- Ethical considerations are paramount in AI development, shaping the responsible and equitable deployment of artificial intelligence technologies.
- As AI systems become increasingly integrated into various aspects of society it is essential to prioritize ethical principles to ensure that these systems benefit individuals and communities while minimizing harm.

Exercise

Search and find a recent news story that details an AI ethical issue and explain who is affected and how.

Bias in AI

Bias is a disproportionate weight in favor of or against an idea or thing, usually in a way that is inaccurate, closed-minded, prejudicial, or unfair. It is important to recognize that bias can occur in various stages of the AI pipeline.

- Data Collection
  - Bias often originates here. The AI algorithm might produce biased outputs if the data is not diverse or representative.
    - In predictive policing systems, historical crime data is often used to train algorithms to forecast future criminal activity and allocate law enforcement resources.
    - This data may reflect systemic biases present in law enforcement practices, such as over-policing in certain neighborhoods or racial profiling.

Bias in AI

- Data Labeling
  - This can introduce bias if the annotators have different interpretations of the same label.
    - One annotator might categorize a person with ambiguous facial features as "middle-aged" while another annotator might perceive the same individual as "young."
    - Similarly, differences in cultural backgrounds or personal experiences could influence how annotators classify ethnicity.
- Model Training
  - A critical phase; if the training data is not balanced, the model may produce biased outputs.
  - This is directly tied to data collection bias.
- Deployment
  - This can also introduce bias if the system is not tested with diverse inputs or monitored for bias after deployment.

Implicit vs. Explicit Bias

Implicit bias and explicit bias are two distinct forms of bias, each with its own characteristics and manifestations:

- Explicit Bias
  - Explicit bias refers to conscious attitudes, beliefs, or prejudices that individuals are aware of and can articulate.
  - These biases are often deliberate and intentional, reflecting conscious beliefs or stereotypes about certain groups of people.
    - An explicit bias in AI could occur if developers intentionally program an autonomous vehicle to prioritize the safety of passengers from certain demographic groups over others in the event of an unavoidable collision, reflecting conscious biases or preferences.

Implicit vs. Explicit Bias

- Implicit Bias
  - Implicit bias, also known as unconscious bias, refers to attitudes, beliefs, or stereotypes that influence our judgments and behaviors at a subconscious level, often without our awareness.
  - These biases operate automatically and involuntarily, shaping our perceptions and decision-making processes.
  - Implicit bias can manifest in subtle ways, influencing our thoughts, feelings, and actions towards others without our conscious recognition.
  - It can lead to unintentional discrimination or unequal treatment based on unconscious stereotypes or associations.
    - An implicit bias in AI could arise if a facial recognition system trained on biased data sets exhibits higher error rates for certain demographic groups due to underlying stereotypes or historical biases present in the training data, despite developers' best efforts to create a fair and accurate system.

Types of Bias

- Selection bias
  - This happens when the data used to train an AI system is not representative of the reality it's meant to model.
  - It can occur due to various reasons, such as incomplete data, biased sampling, or other factors that may lead to an unrepresentative dataset.
    - If a model is trained on a dataset that only includes male employees, for example, it will not be able to predict female employees' performance accurately.
- Confirmation bias
  - This type of bias happens when an AI system is tuned to rely too much on pre-existing beliefs or trends in the data.
  - This can reinforce existing biases and fail to identify new patterns or trends.

Types of Bias

- Measurement bias
  - This bias occurs when the data collected differs systematically from the actual variables of interest.
    - For instance, if a model is trained to predict students' success in an online course, but the data collected is only from students who have completed the course, the model may not accurately predict the performance of students who drop out of the course.
- Stereotyping bias
  - This happens when an AI system reinforces harmful stereotypes.
    - An example is when a facial recognition system is less accurate in identifying people of color or when a language translation system associates certain languages with certain genders or stereotypes.

Real World AI Bias Examples

- 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.
- Applicant tracking systems
  - Issues with natural language processing algorithms can produce biased results within applicant tracking systems.
    - For example, Amazon stopped using a hiring algorithm after finding it favored applicants based on words like “executed” or “captured,” which were more commonly found on men’s resumes.
- Online advertising
  - Biases in search engine ad algorithms can reinforce job role gender bias.
    - Independent research at Carnegie Mellon University in Pittsburgh revealed that Google’s online advertising system displayed high-paying positions to males more often than to women.

Real World AI Bias Examples

- 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.
- Predictive policing tools
  - AI-powered predictive policing tools used by some organizations in the criminal justice system are supposed to identify areas where crime is likely to occur.
  - However, they often rely on historical arrest data, which can reinforce existing patterns of racial profiling and disproportionate targeting of minority communities.

AI Safety

- For the most part, AI ethics have been proposed and/or put in place in defense of neglectful oversight in hopes of industry innovation.
  - AI ethics also covers malicious use of these systems.
- Currently, AI is not sentient and can't have any self imposed directives, but humans can direct AI to cause harm to the public or specific individuals in multiple ways.
  - "guns don't kill people, people kill people" = "AI doesn't harm people, people harm people"

Manipulation and Disinformation

- Deepfake Technology
  - Malicious individuals can use AI-generated deepfake videos or audio to spread false information, manipulate public opinion, or discredit individuals by superimposing their likeness onto fabricated content.
- Social Media Manipulation
  - AI-powered bots or algorithms can be employed to automate the dissemination of propaganda, fake news, or divisive content on social media platforms, amplifying polarization and undermining trust in democratic processes.

Cyber Attacks and Exploitation

- Adversarial Attacks
  - Malicious actors can exploit vulnerabilities in AI systems through adversarial examples or manipulation techniques to trick AI algorithms into making incorrect or harmful decisions, leading to security breaches or data manipulation.
- AI-Driven Cyber Attacks
  - AI technology can be leveraged to enhance the sophistication and efficiency of cyber attacks, such as automated malware generation, phishing campaigns, or targeted social engineering attacks.

Privacy Violations and Surveillance

- Facial Recognition Surveillance
  - AI-powered facial recognition systems can be misused for mass surveillance or tracking individuals without their consent, raising concerns about privacy violations, civil liberties, and potential abuse by authoritarian regimes.
- AI-Powered Surveillance Technologies
  - Malicious actors can deploy AI-driven surveillance technologies, such as smart cameras or predictive analytics, to monitor and profile individuals, infringing on their privacy rights and enabling unwarranted surveillance.

Autonomous Weapons and Military Applications

- Lethal Autonomous Weapons Systems (LAWS)
  - AI technology can be integrated into autonomous weapons systems, enabling the development of lethal drones, robots, or cyber weapons capable of making life-and-death decisions without human intervention, raising ethical and humanitarian concerns.
- Military AI Applications
  - Malicious use of AI in military contexts can escalate conflicts, increase the risk of civilian casualties, and undermine international security and stability, highlighting the importance of regulating and controlling AI-powered military technologies.

Financial Fraud and Manipulation

- Algorithmic Trading Manipulation
  - AI algorithms can be exploited to manipulate financial markets, conduct high-frequency trading, or perpetrate pump-and-dump schemes, leading to market manipulation, instability, and financial fraud.
- AI-Powered Fraud Detection
  - Malicious individuals can use AI technology to develop sophisticated fraud detection evasion techniques, such as adversarial attacks or data poisoning, to evade detection and perpetrate financial crimes undetected.

Transparency in AI

- Transparency in AI refers to the clarity and openness of AI systems, algorithms, and decision-making processes, allowing stakeholders to understand how AI operates, why specific decisions are made, and the potential impacts on individuals and society.
- It involves providing access to information, explanations, and insights into AI systems' behavior, data sources, and underlying mechanisms.

Importance of Transparency

- Transparency is crucial for promoting accountability, trust, and ethical behavior in AI development and deployment.
- By making AI systems transparent, developers, users, and regulatory authorities can hold AI accountable for its decisions and outcomes.
- Transparency enables stakeholders to identify and address biases, errors, or ethical concerns in AI systems, fostering accountability and responsible AI governance.

Explainable AI

- Explainable AI (XAI) techniques aim to make AI systems' decisions and outputs understandable and interpretable by humans.
- XAI methods provide insights into how AI models arrive at specific predictions or recommendations, allowing users to trace decision-making processes and identify factors influencing outcomes.
- Techniques such as feature importance analysis, decision trees, and model-agnostic approaches (e.g., LIME, SHAP) enhance transparency by providing explanations for AI predictions in a human-readable format.

XAI Techniques

- Feature importance analysis
  - The process of evaluating the significance of different variables and features in a machine learning model decision-making process.
- Decision trees
  - Considered one of the most interpretable models in XAI.
  - Each decision is based on clear, understandable conditions.
  - Decision trees have a straightforward, rule-based structure, unlike complex models like deep neural networks.
  - The tree structure shows a clear hierarchy of decision importance.
- Model-agnostic approaches
  - Black box approaches that analyze the features' input-output pair.
  - They work by slightly altering the input features and observing how these changes affect the model's predictions.

Other Techniques for Enhancing Transparency

- Algorithmic Transparency
  - Involves disclosing information about the algorithms used in AI systems, including their design, functionality, and performance characteristics.
  - Providing transparency about algorithmic processes allows stakeholders to assess the fairness, accuracy, and potential biases inherent in AI systems.
    - Code transparency
    - Documentation
    - Algorithm auditing

Other Techniques for Enhancing Transparency

- Open Data Practices
  - Open data practices involve making AI training data sets, methodologies, and evaluation metrics publicly available to enhance transparency and reproducibility in AI research and development.
  - Open data practices enable independent validation and scrutiny of AI models, fostering trust and accountability in AI systems.
    - Data documentation
    - Data provenance tracking
    - Data sharing platforms

Ethical AI Solutions

- As AI technology advances at an unprecedented pace, the potential for unintended consequences and ethical dilemmas escalates.
- Addressing these challenges promptly is essential to ensure that AI systems uphold ethical principles, promote fairness, and mitigate risks.
- By prioritizing the development of ethical AI solutions, we can proactively address emerging ethical concerns and foster responsible AI innovation in alignment with societal values.

The IEEE Global Initiative

- The IEEE (Institute of Electrical and Electronics Engineers) Global Initiative has developed a comprehensive set of guidelines and recommendations for ethical AI development and deployment.
- Key principles
  - Transparency
  - Accountability
  - Fairness
  - Prioritization of human well-being in AI systems
- The initiative emphasizes the importance of interdisciplinary collaboration and stakeholder engagement in addressing ethical challenges in AI technology.

The AI Ethics Principles by the European Commission

- The European Commission has established a set of ethical principles for AI development, based on fundamental rights, ethical values, and legal standards.
- These principles
  - Respect for human autonomy
  - Prevention of harm
  - Transparency
  - Accountability
  - Environmental sustainability
- The European Commission encourages the adoption of these principles by AI developers, industry stakeholders, and policymakers to ensure the responsible and ethical use of AI technology.

The Principles for Accountable Algorithms

- The Data & Society Research Institute has formulated principles for accountable algorithms
  - Disclosure of data sources
  - Fairness
  - Accountability
  - Redress mechanisms
    - A process for assessing and resolving community complaints/feedback in algorithmic decision-making processes.
- The institute advocates for the development of ethical and accountable AI systems that promote fairness, equity, and social justice.

The Data Protection and Digital Information Bill

- The Data Protection and Digital Information Bill represents a critical step towards safeguarding personal data and digital privacy in today's digital age.
- Introduced to Parliament.
- Key provisions
  - Data collection
  - Transparency
  - Storage practices by companies and organizations.
- It outlines clear guidelines for data sharing and consent mechanisms, ensuring that individuals have the right to understand how their data is being used and to opt-out if desired.

Ethical Considerations in AI Decision-Making

- Human Oversight and Control
  - Incorporating human oversight and control mechanisms in AI systems to ensure human intervention and accountability in decision-making processes.
  - Human oversight enables the review and validation of AI-generated outcomes, mitigating the risks of bias, errors, or unintended consequences.
- Impact Assessments
  - Conducting ethical impact assessments to evaluate the potential social, economic, and ethical implications of AI technologies on individuals, communities, and society.
  - Impact assessments help identify and mitigate risks, anticipate unintended consequences, and ensure that AI systems adhere to ethical principles and legal standards.
- Ethical Risk Management
  - Implementing ethical risk management strategies to identify, assess, and mitigate ethical risks associated with AI development and deployment.
  - Ethical risk management involves proactive measures to address potential biases, discrimination, privacy violations, and other ethical concerns in AI systems.

Take Home Exercise

Find one example of AI ethics being used within real world application OR one example of where AI ethics should be introduced and how.