University of Salford
MANCHESTER

AI Safety

Dr. Tarek Gaber

Date: 25th March 2024

Who am I?

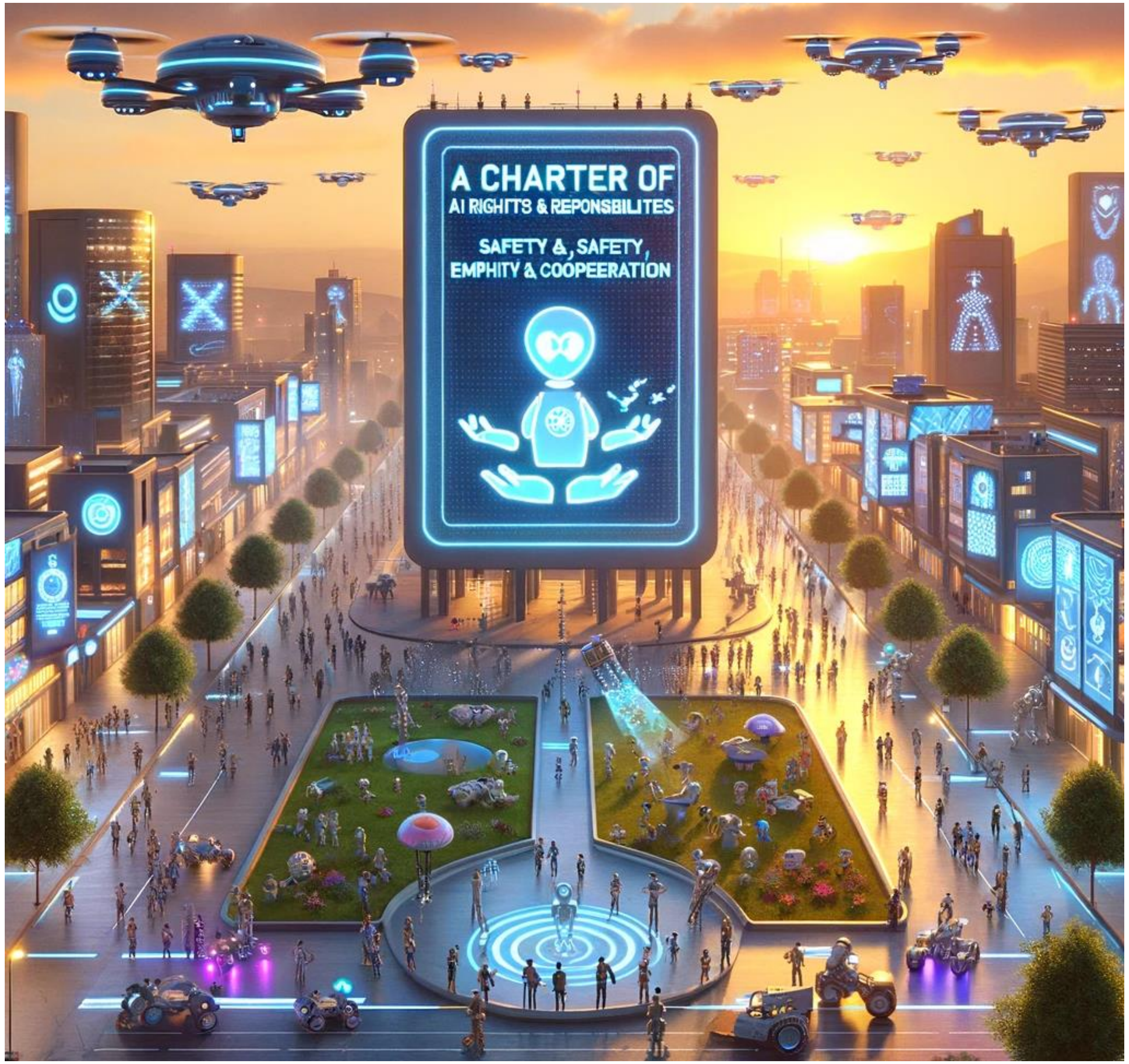
I'm **Tarek Gaber**, a Senior Lecturer in Cybersecurity at the University of Salford

- Background:
  - PhD in Information Security from University of Manchester.
  - Experience as a researcher and lecturer, with Cybersecurity and Machine Learning roles at Suez Canal University, Egypt and VSB-Ostrava, Czech.
- Research Focus:
  - Biometric Altercation, Intrusion Detection, Secure Software Engineering
  - Current research focuses on Secure and Sustainable Artificial Intelligent.
- Achievements:
  - Published numerous research papers in high-impact peer-reviewed journals.
  - Secured funding grants from prestigious schemes such as UKRI, Innovate UK, GCHQ.
- Teaching Excellence:
  - Led the development of fundamental modules in Cybersecurity, including Dependable Software Engineering and Privacy & Network Security.

Workshop Agenda

- 1. Overview of Narrow AI Safety Challenges
- 2.Bias and Fairness in Narrow AI
- 3.Lack of Transparency in AI Decision-Making
- 4.Data Privacy Concerns
- 5.Security Risks in Narrow AI Systems
- 6.Robustness and Reliability Issues
- 7.Ethical Considerations in Narrow AI Applications
- 8.Human-AI Collaboration and Trust
- 9.Unintended Consequences of Narrow AI
- 10.Regulatory and Legal Challenges in AI Safety

1. Overview of Narrow AI Safety Challenges

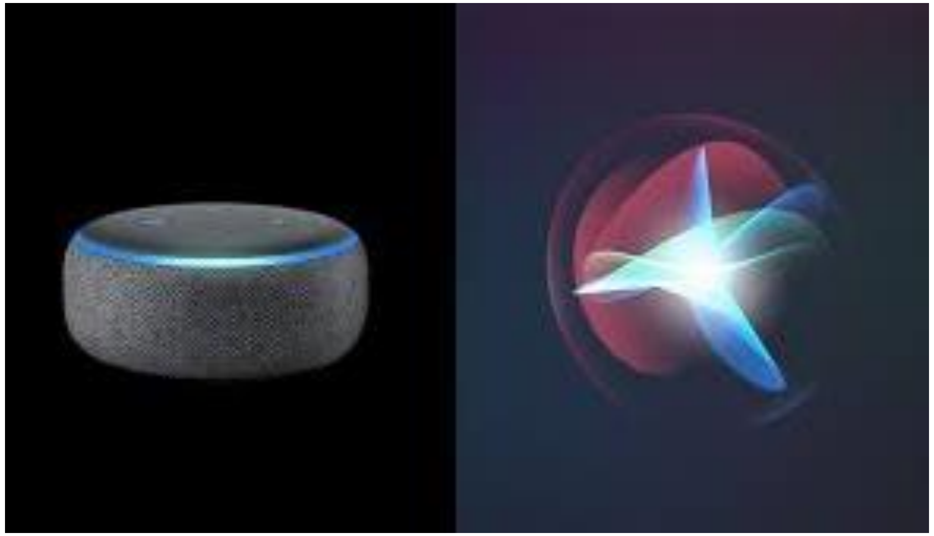


- **Imagine a future where AI systems make all major decisions (e.g., governance, healthcare, engineering, education). What would be your biggest concern, and what potential benefit excites you the most?**

- **If you had to trust an AI system with one aspect of your life (healthcare, financial management, engineering, personal safety, etc.), which one would you choose and why?**

Applications of Narrow AI

- Voice assistants like Siri and Alexa take user commands to perform tasks.
- Medical diagnosis algorithms analyze images to detect diseases with accuracy often surpassing human experts.



AI Safety Challenges

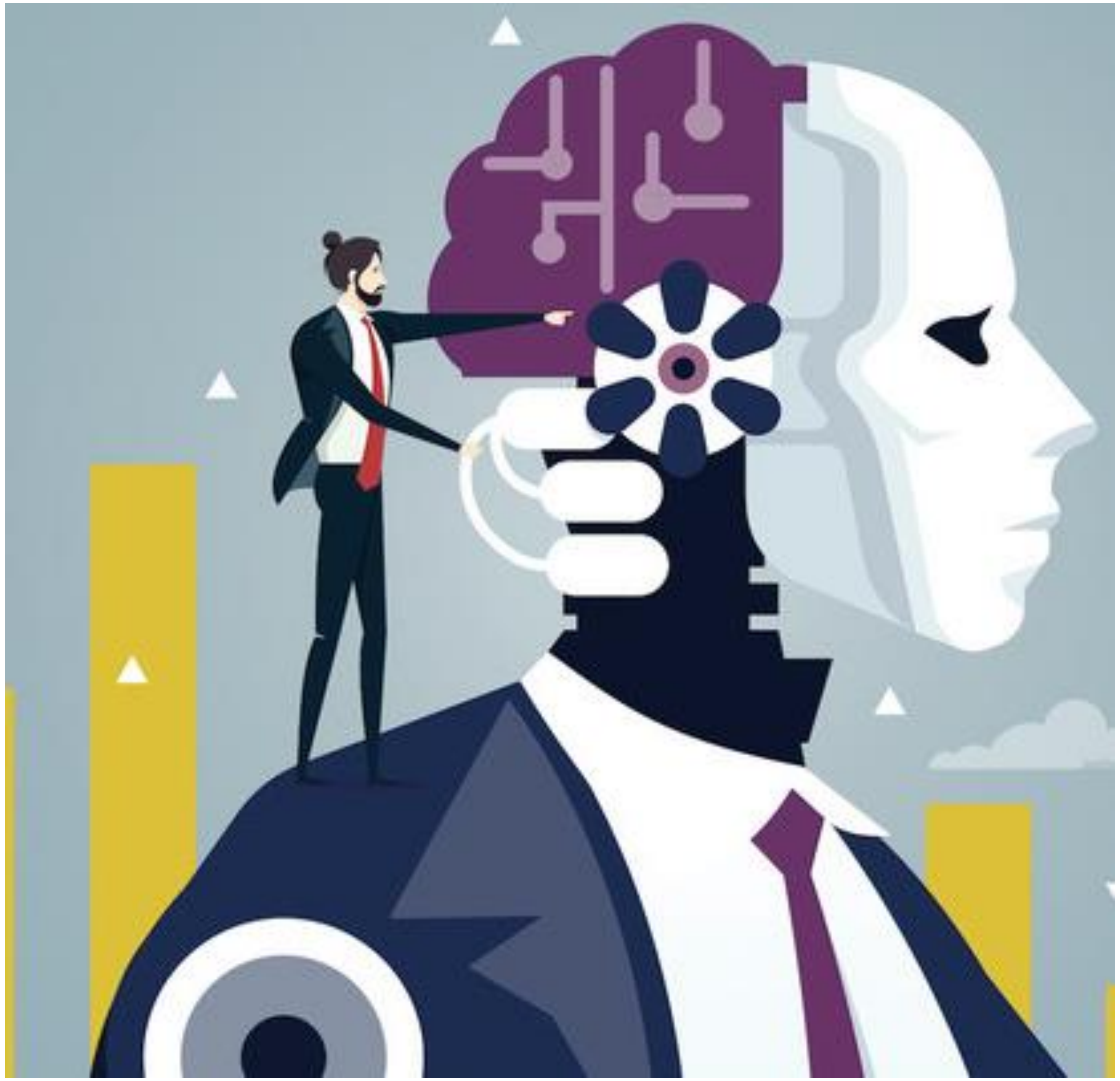
- AI safety involves ensuring AI systems:
  - operate as intended,
  - are secure, and
  - ethically designed.

- Key Areas:
  - **Data Bias**: AI reflects biases present in training data, impacting fairness.
  - **Security**: AI systems can be vulnerable to hacking, requiring robust defenses.
  - **Ethical and Legal Compliance**: AI must adhere to ethical guidelines and laws to prevent harm and misuse.

Data Bias and Quality Issues

- Biased data leads to unfair outcomes;
  - facial recognition technologies may misidentify minority ethnic groups due to lack of diverse data.

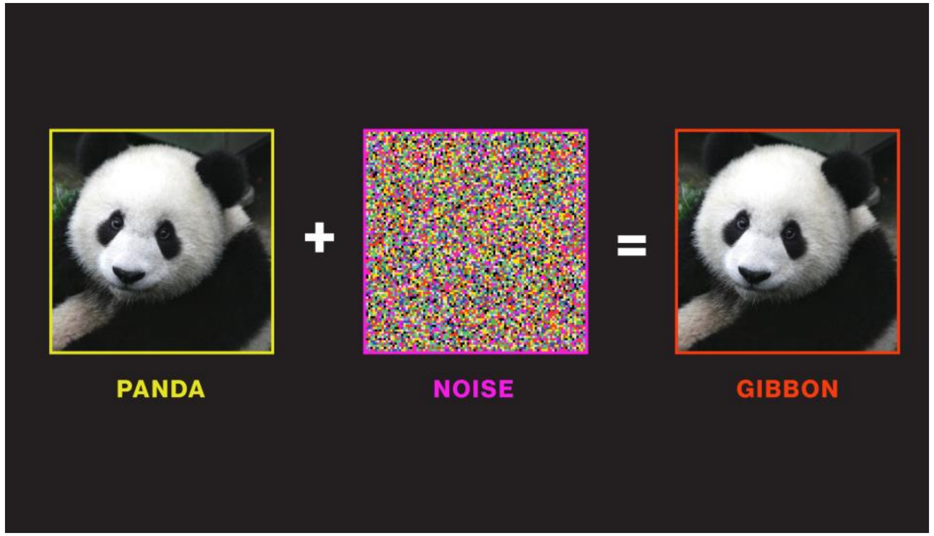
- Mitigation Strategies:
  - Diverse data collection and rigorous testing to identify and correct biases.



Security Risks in Narrow AI Systems

- Security Threat Examples:
  - Adversarial attacks manipulate AI inputs to cause incorrect outputs, such as altering an image slightly to fool a security or other system.

- Best Practices for Security:
  - Implementing encryption and secure data handling protocols.
  - Designing AI to recognize and resist adversarial inputs.



Question 1

- **What is Data Bias in AI, and why is it a concern?**
  - A) The preference of an AI system for data from specific sources, enhancing performance.
  - B) The reflection of pre-existing biases in training data, potentially leading to unfair outcomes.
  - C) The process of cleaning data before feeding it into an AI system to improve accuracy.
  - D) A strategy used by AI developers to increase the diversity of training data.

Question 1

- **What is Data Bias in AI, and why is it a concern?**
  - A) The preference of an AI system for data from specific sources, enhancing performance.
  - B) The reflection of pre-existing biases in training data, potentially leading to unfair outcomes.
  - C) The process of cleaning data before feeding it into an AI system to improve accuracy.
  - D) A strategy used by AI developers to increase the diversity of training data.

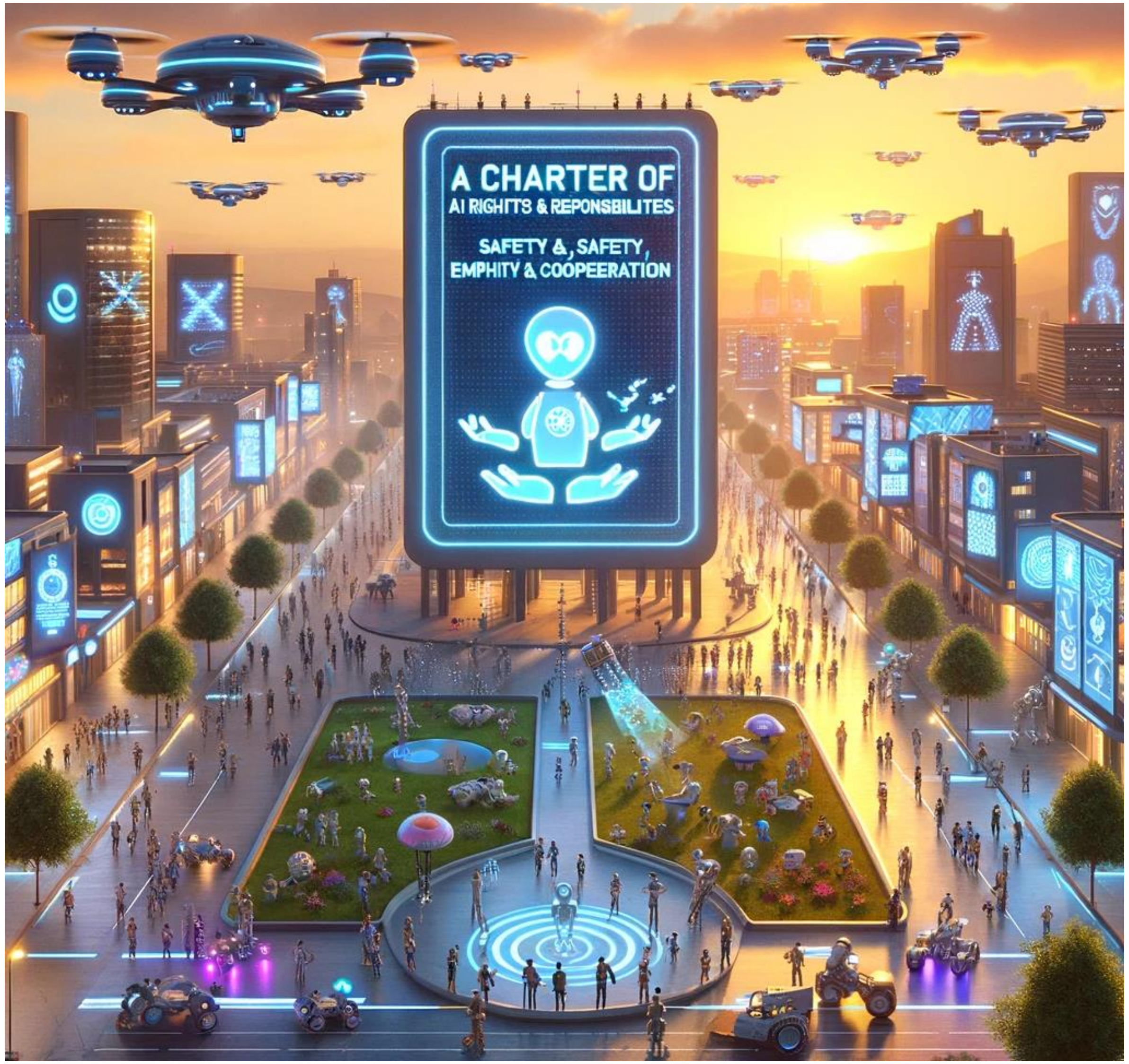
Question 2

- **What is an adversarial attack in the context of AI systems?**
  - A) An attack that focuses on the physical components of AI hardware to cause damage.
  - B) A method of manipulating AI inputs slightly to produce incorrect outputs, potentially fooling AI systems.
  - C) A direct attack on the developers of AI systems to steal the source code.
  - D) The process of legally challenging the ethical implications of AI systems.

Question 2

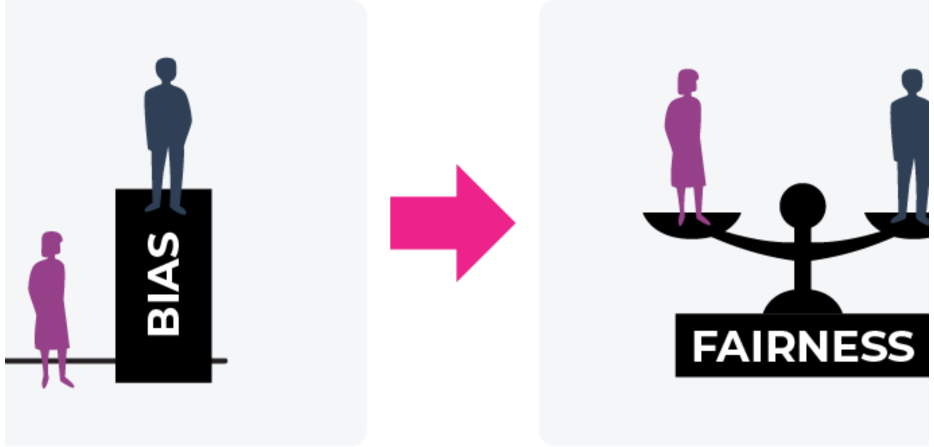
- **What is an adversarial attack in the context of AI systems?**
  - A) An attack that focuses on the physical components of AI hardware to cause damage.
  - B) A method of manipulating AI inputs slightly to produce incorrect outputs, potentially fooling AI systems.
  - C) A direct attack on the developers of AI systems to steal the source code.
  - D) The process of legally challenging the ethical implications of AI systems.

2. Bias and Fairness in Narrow AI



Bias and Fairness in Narrow AI

- **Bias** in AI reflects systemic inaccuracies favouring certain outcomes or groups.
- **Fairness** in AI ensure systems treat all individuals and groups equitably.
- Sources of Bias:
  - Data collection, algorithm design, and outcome interpretation stages.

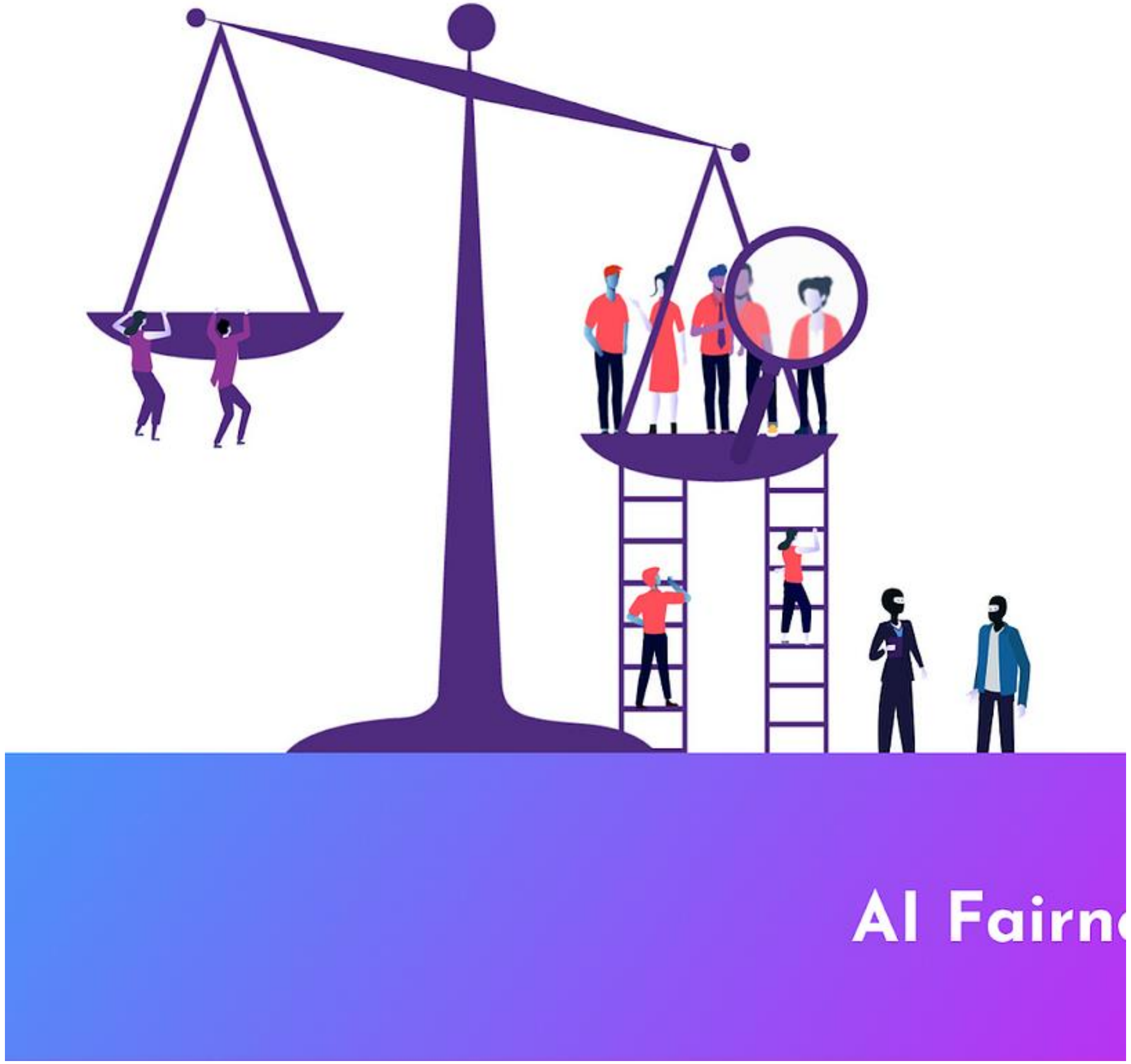


Types of Bias in AI Systems

- **Data Bias:**
  - Skewed data that does not accurately represent the target population.
- **Algorithmic Bias:**
  - Algorithms that develop prejudiced decisions based on the data fed into them.
- **User Interaction Bias:**
  - Bias introduced by the way users interact with AI systems.
- **Case Studies:**
  - facial recognition systems misidentifying certain ethnic groups,
  - loan applications, where people from certain backgrounds might be unfairly denied..

Measuring and Assessing Fairness in AI

- **Fairness Metrics:**
  - **Equality of Opportunity**: Equal chances for all groups for favorable outcomes.
  - **Demographic Parity**: Equal distribution of AI outcomes across different groups.



Mitigating Bias: Strategies and Best Practices

- Data Collection:
  - Ensuring diverse and representative data sets.
- Model Selection and Evaluation:
  - Choosing models that are less prone to bias and rigorously testing them, e.g., linear models, decision trees.
- Diversity in Teams:
  - Promoting diverse teams to recognize and mitigate biases.
- Ethical AI Principles:
  - Adopting principles (e.g., Transparency, Data Protection) that guide the development of fair and unbiased AI.
- Continuous Monitoring:
  - Regularly assessing AI systems to identify and rectify biases.

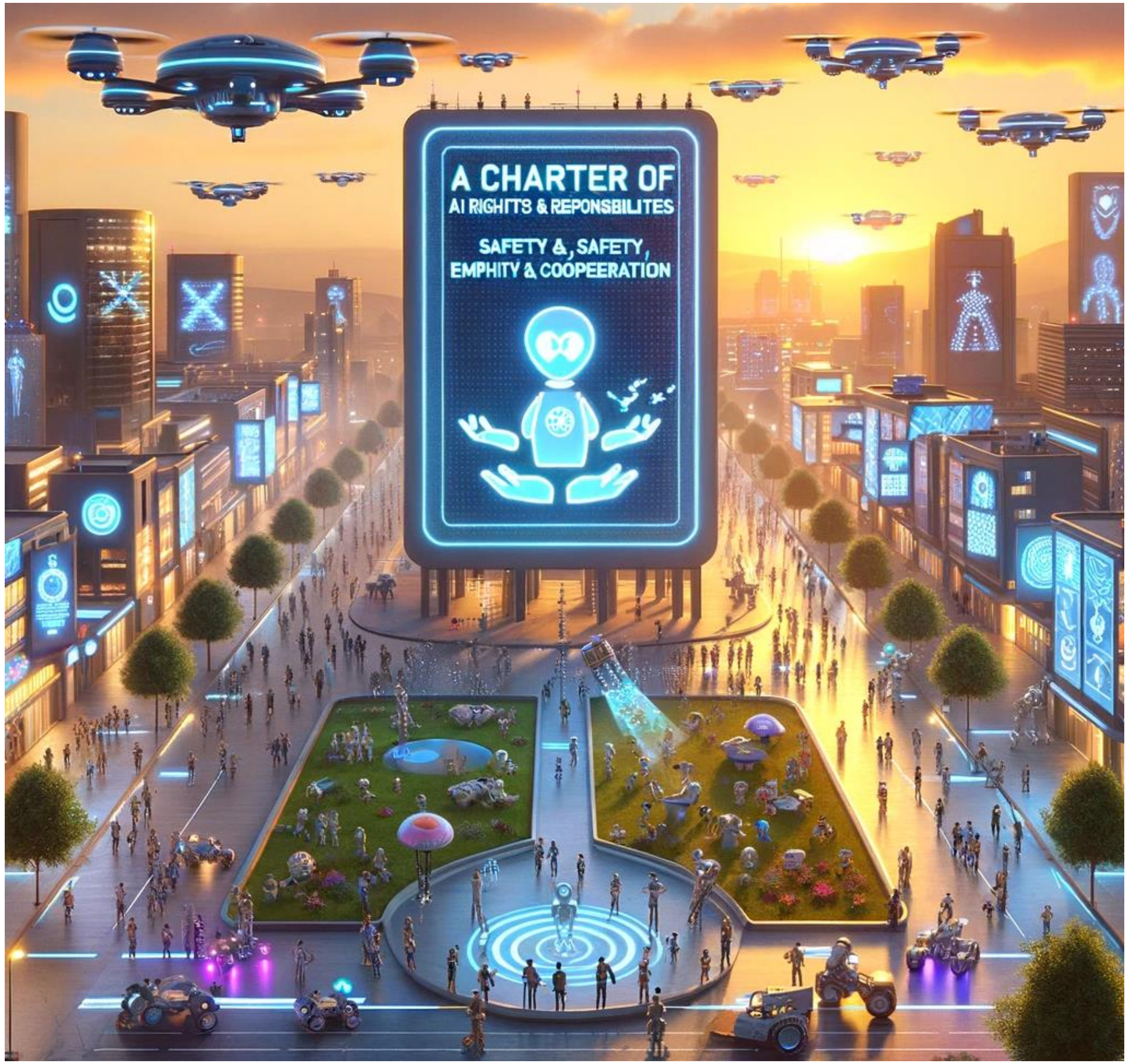
Question 1

- **What leads to biased decisions in AI systems?**
  - A) Only incorrect programming practices.
  - B) Data Bias, Algorithmic Bias, and User Interaction Bias, all contributing in sequence.
  - C) Exclusively the misuse of AI by end-users.
  - D) Lack of internet connectivity.

Question 1

- **What leads to biased decisions in AI systems?**
  - A) Only incorrect programming practices.
  - B) Data Bias, Algorithmic Bias, and User Interaction Bias, all contributing in sequence.
  - C) Exclusively the misuse of AI by end-users.
  - D) Lack of internet connectivity.

3.Lack of Transparency in AI Decision-Making

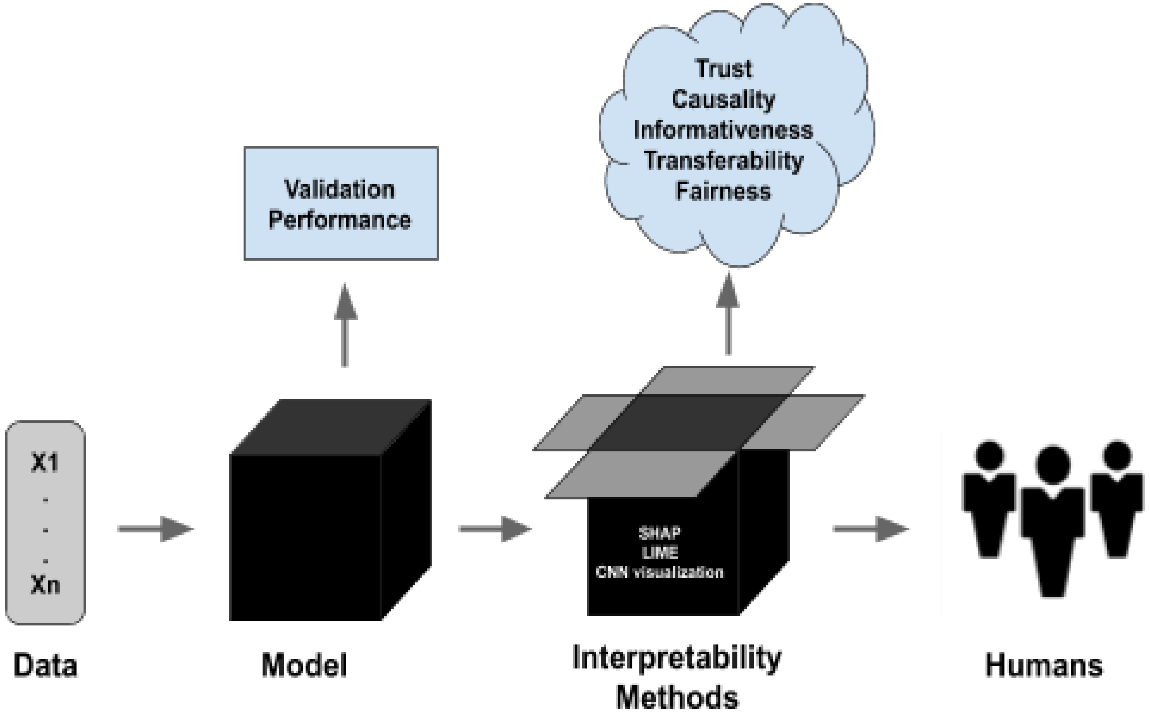


Transparency in AI

- AI Transparency is understandability of AI systems by humans.
- Essential for building user trust and ethical use.
- Vital for legal compliance in various sectors.
- Deep neural networks complicate transparency.

Understanding AI Opacity

- Complex Algorithms:
  - Deep learning models are hard to interpret.
- Exclusive Concerns:
  - Companies protect intellectual property, reducing transparency.
- Lack of Standards:
  - No universal standards for AI explanations exist.
- Real-World Example:
  - Opacity in credit scoring algorithms.



Consequences of AI Opacity

- Eroding Public Trust:
  - Lack of understanding leads to distrust.
- Ethical Dilemmas:
  - AI decisions impact lives without clear explanations.
- Legal Compliance:
  - Transparency is crucial for laws like GDPR.
- Accountability Issues:
  - Challenges in holding developers and companies responsible for AI behavior without transparency.

Need for Explainable AI (XAI)

- Bridging the Gap:
  - XAI aims for clarity without losing performance.
- Debugging and Improvement:
  - Essential for fixing errors in AI models.
- Regulatory Compliance:
  - Necessary for meeting legal explanation requirements.
- Performance Trade-offs:
  - Balancing explainability with efficiency.

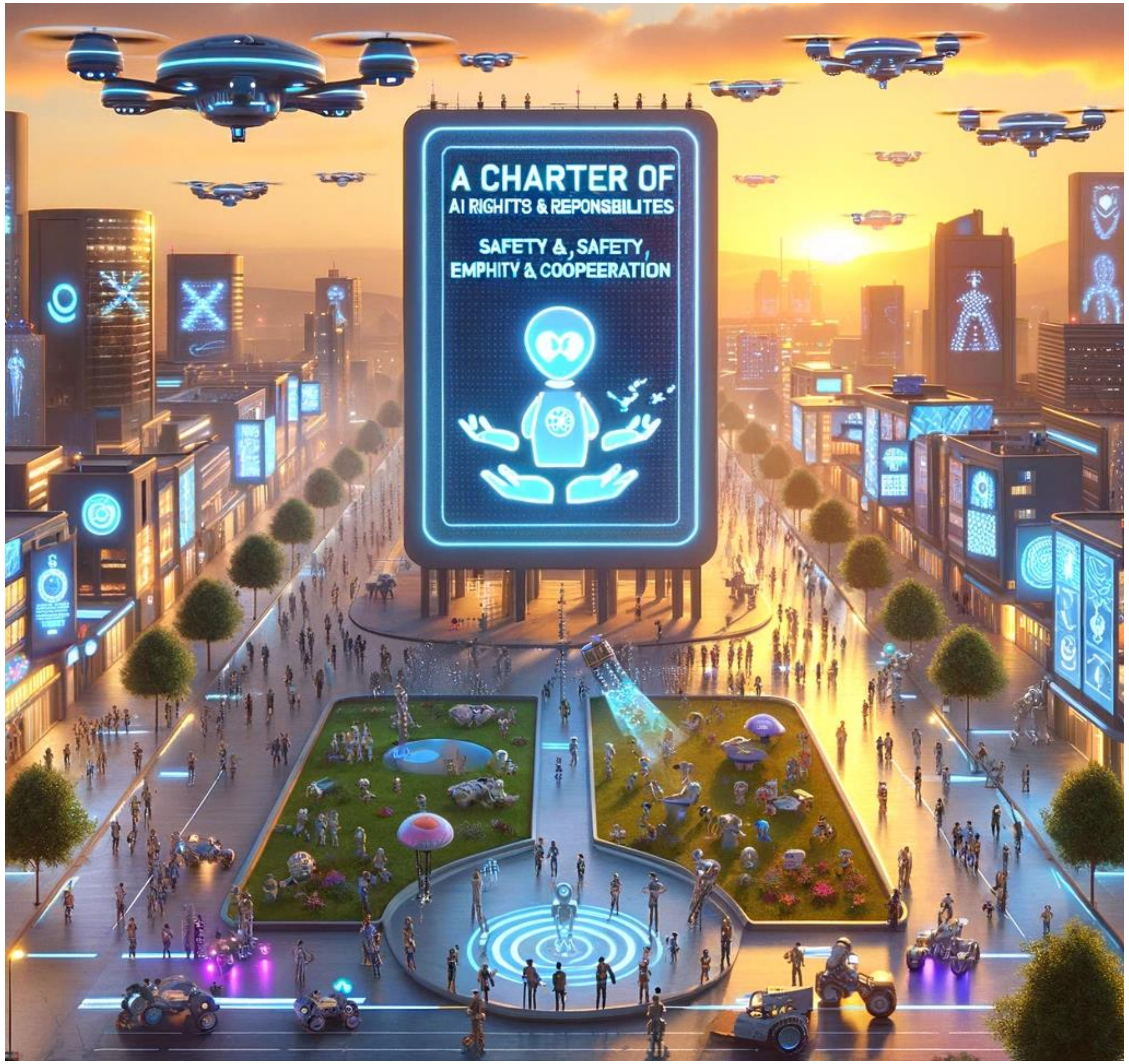
Question 1

- **Why is AI Transparency crucial?**
  - A) It ensures AI systems can operate independently without human intervention.
  - B) It makes AI systems understandable to humans, building user trust, supporting ethical use, and ensuring legal compliance, despite challenges posed by deep neural networks.
  - C) It allows AI systems to process data faster.
  - D) Transparency is only required for AI systems used in entertainment.

Question 1

- **Why is AI Transparency crucial?**
  - A) It ensures AI systems can operate independently without human intervention.
  - B) It makes AI systems understandable to humans, building user trust, supporting ethical use, and ensuring legal compliance, despite challenges posed by deep neural networks.
  - C) It allows AI systems to process data faster.
  - D) Transparency is only required for AI systems used in entertainment.

4.Data Privacy Concerns



Data Privacy in AI

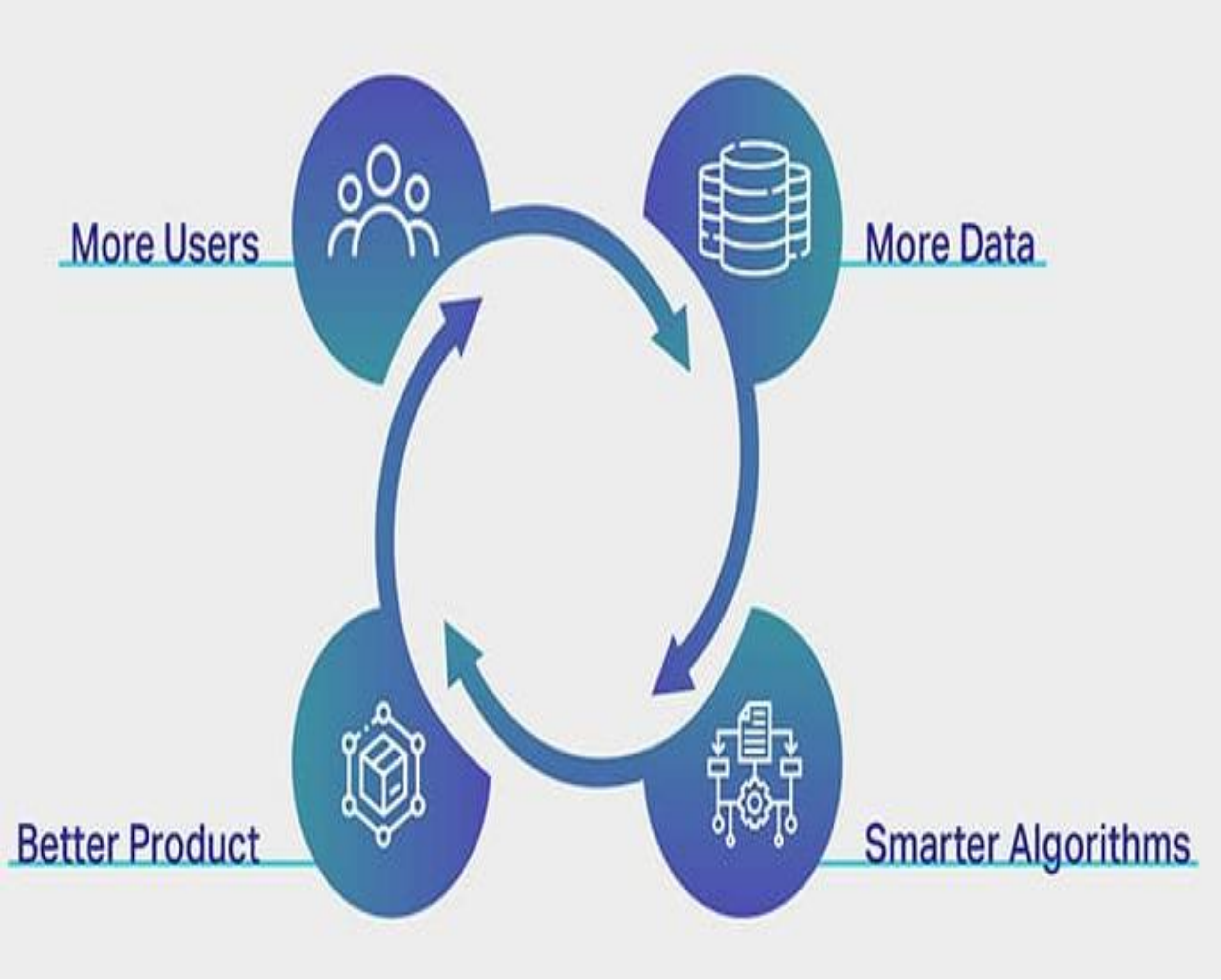
- Data Privacy:
  - Importance of protecting personal information in the age of AI.
- Central Concerns:
  - Unauthorized access, lack of informed consent, misuse of data.
- The Impact of Breaches:
  - including identity theft and loss of public trust.
- Regulatory:
  - GDPR, CCPA, and other global data protection regulations.

Data Collection and Consent

- AI systems often collect personal data from a variety of sources,
  - including online activities, IoT devices, and public records.
- Challenges in Informed Consent:
  - Consent must be informed, specific, and freely given.
  - However, the complexity of AI systems can make it challenging for users to understand what they are consenting to.
- Examples:
  - Fitness trackers collecting health data without clear explanations of how the data will be used or shared.

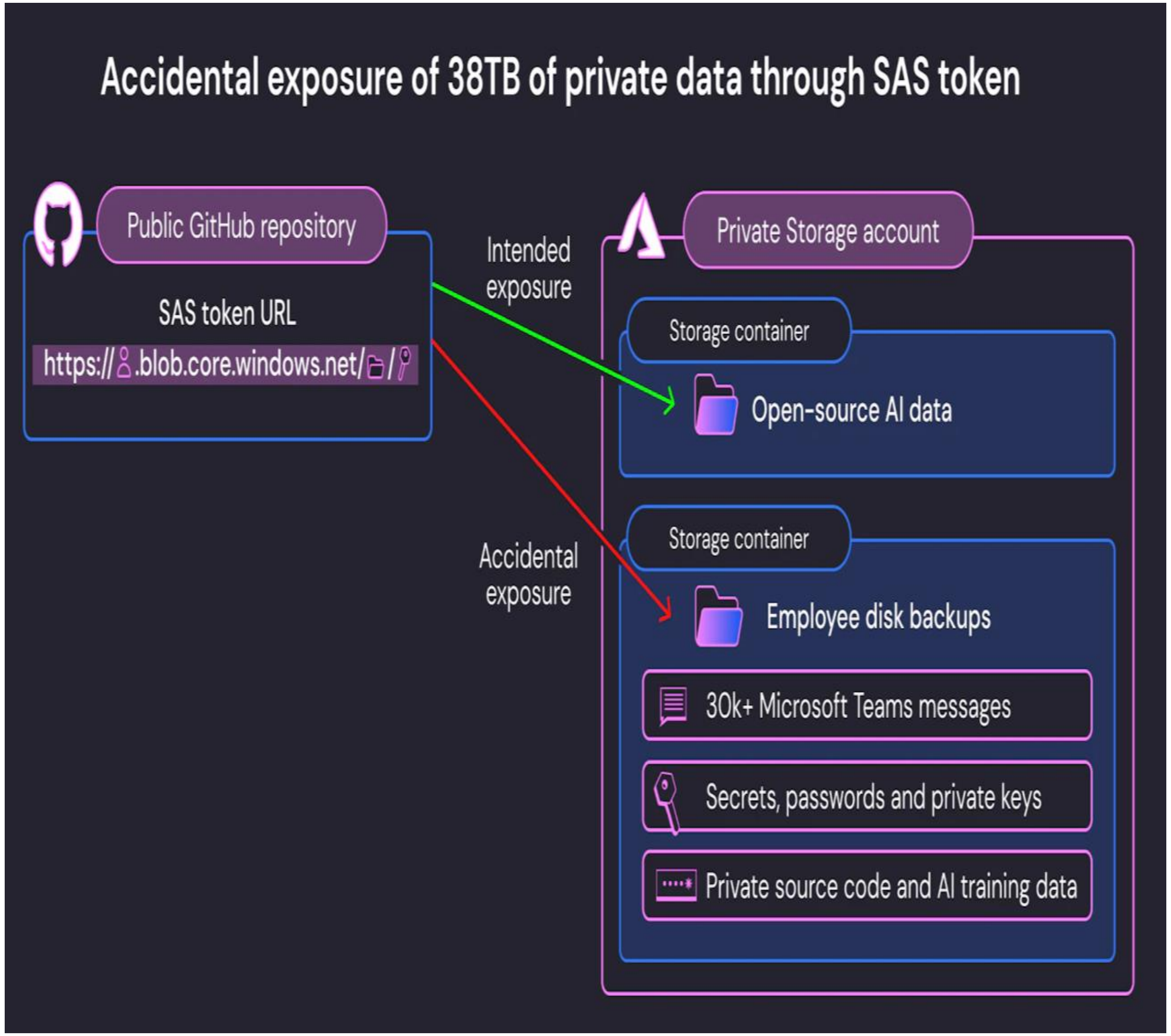
Data Processing and AI Training

- **Data is AI's Foundation:**
  - Training AI models requires large datasets, often containing sensitive personal information.
- **Risks of Re-identification:**
  - Anonymized data can often be re-identified using AI.
- **Ethical Data Use:**
  - The responsibility of AI developers to use data ethically, respecting privacy even in anonymized datasets.
- **Techniques for Protection:**
  - Introducing more robust anonymization techniques and considering differential privacy (describing the patterns of groups within the dataset).



Data Storage and Security Measures

- Personal data is stored in various forms and locations, increasing the risk of unauthorized access and data breaches.
  - **Microsoft AI Research Division Data Leak**: Discovered on September 18, 2023[1].
- **Security Measures:**
  - Homomorphic Encryption,
  - blockchain for data integrity,
  - federated learning and
  - and access control mechanisms.


Accidental exposure of 38TB of private data through SAS token
Public GitHub repository
SAS token URL
https://.blob.core.windows.net//
Intended exposure
Accidental exposure
Private Storage account
Storage container
Open-source AI data
Storage container
Employee disk backups
30k+ Microsoft Teams messages
Secrets, passwords and private keys
Private source code and AI training data

[1] https://firewalltimes.com/recent-data-breaches/

Question 1

- **What is essential for ethical AI data use?**
  - A) Only using publicly available data.
  - B) Ensuring data is permanently anonymized.
  - C) Employing ethical practices, like robust anonymization and differential privacy, to protect sensitive information.
  - D) Assuming anonymized data cannot be re-identified.

Question 1

- **What is essential for ethical AI data use?**
  - A) Only using publicly available data.
  - B) Ensuring data is permanently anonymized.
  - C) Employing ethical practices, like robust anonymization and differential privacy, to protect sensitive information.
  - D) Assuming anonymized data cannot be re-identified.

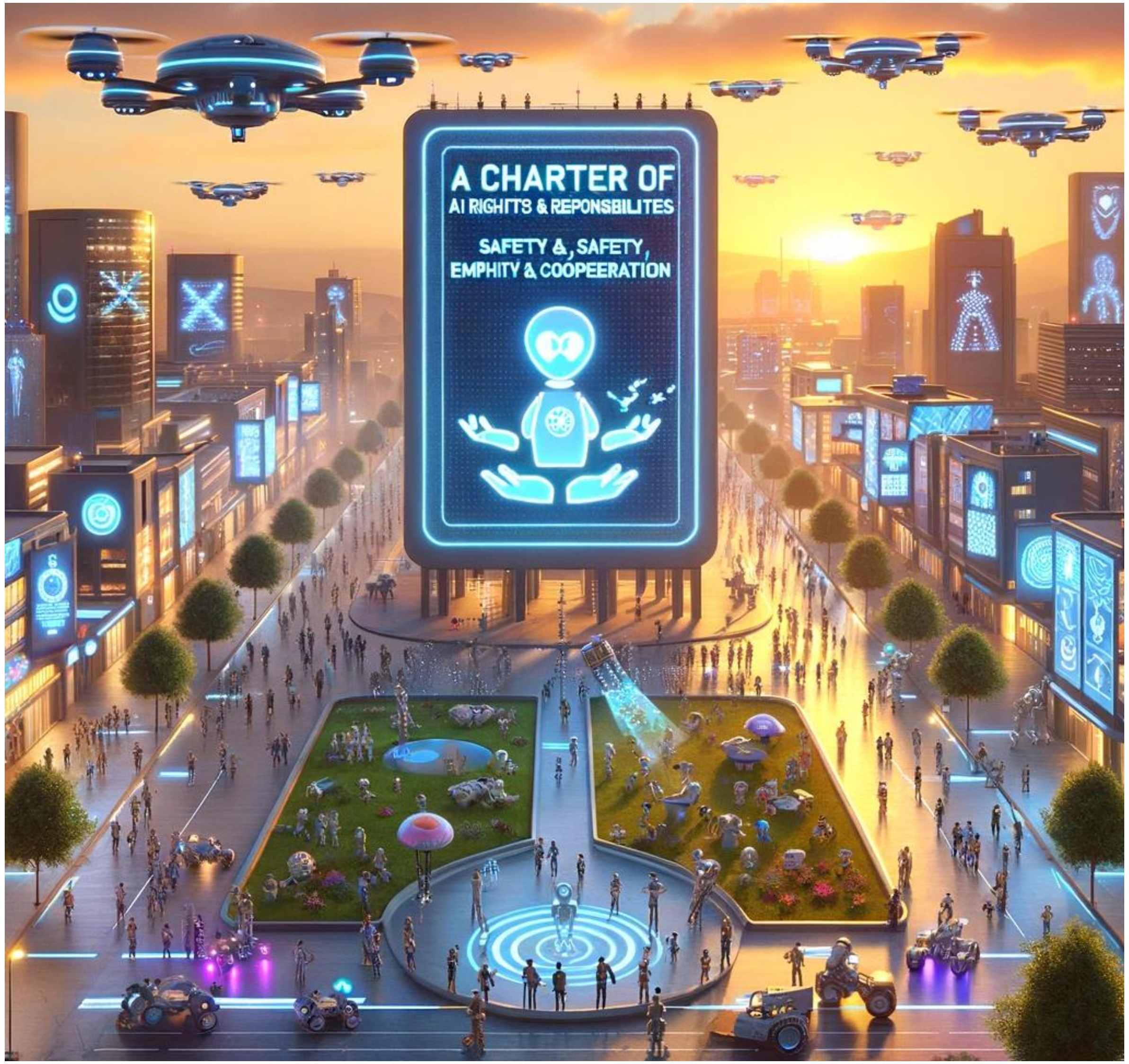
Question 2

- **Which statement best reflects the responsibilities and challenges of AI developers regarding data privacy?**
  - A) AI development does not require ethical considerations as long as the data is anonymized.
  - B) Large datasets for AI training eliminate the risk of re-identification of personal data.
  - C) AI developers must ethically use data, enhancing anonymization and considering differential privacy to protect against re-identification risks.
  - D) Differential privacy is unnecessary if the data is already anonymized.

Question 2

- **Which statement best reflects the responsibilities and challenges of AI developers regarding data privacy?**
  - A) AI development does not require ethical considerations as long as the data is anonymized.
  - B) Large datasets for AI training eliminate the risk of re-identification of personal data.
  - C) AI developers must ethically use data, enhancing anonymization and considering differential privacy to protect against re-identification risks.
  - D) Differential privacy is unnecessary if the data is already anonymized.

5.Security Risks in Narrow AI Systems



Narrow AI and Security Risks

- The incredible potential of AI to transform industries, societies, and the very way we live also brings with it significant security risks that must be addressed from the design.

Overview of Security Risks in Narrow AI

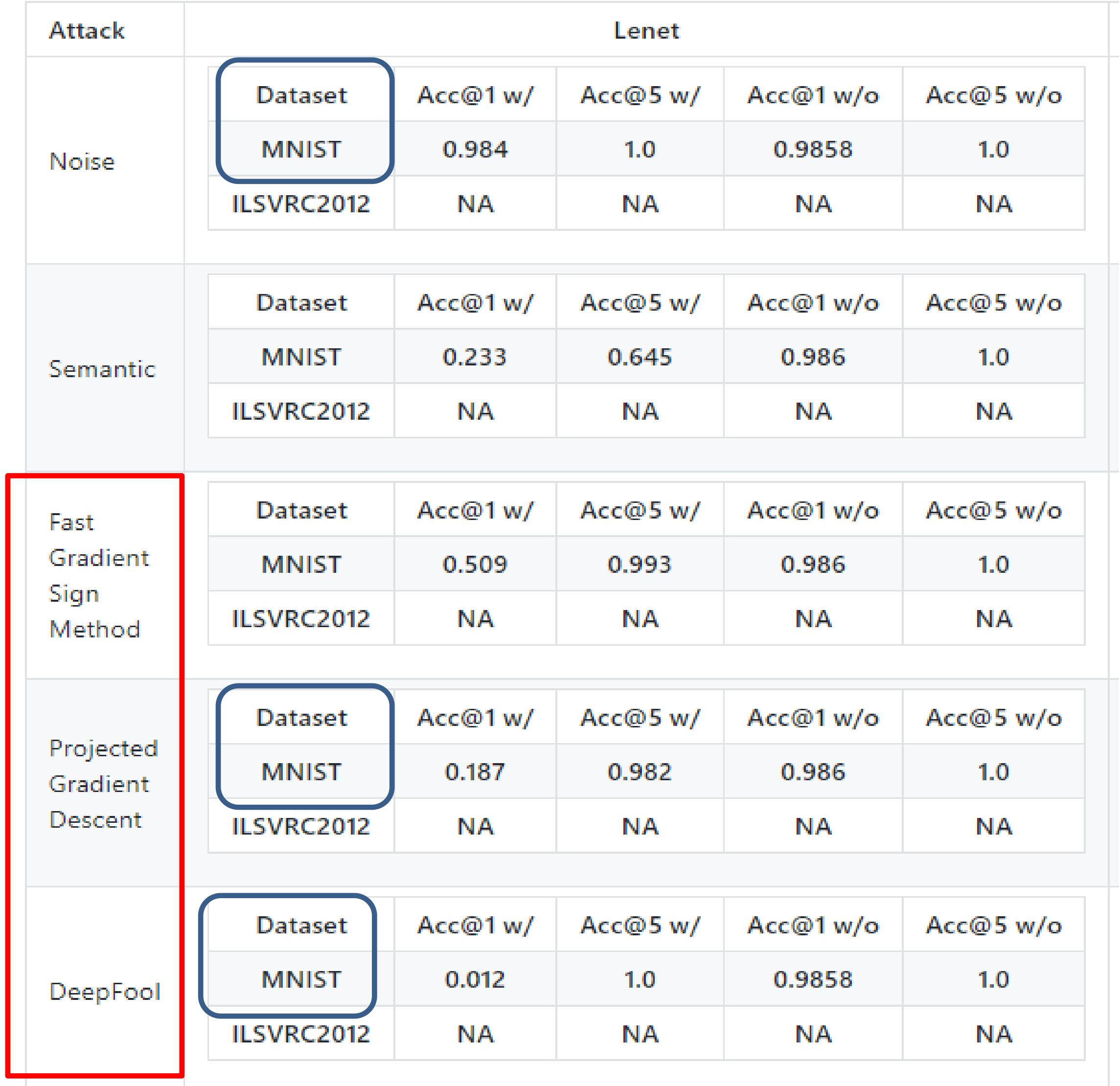
- Risks include unauthorized access, manipulation, and unintended actions
- Categories of Risks:
  - **Data integrity**: Accuracy and reliability of AI data
  - **Privacy**: Protection of sensitive information
  - **Operational:** System functionality and reliability
- Unique Challenges:
  - Processing sensitive data demands high security
  - Specific applications have targeted vulnerabilities

Data Integrity and Privacy Concerns

- Data Tampering Impacts:
  - Altering data can lead to flawed AI decisions
  - Example: altering traffic data could mislead autonomous vehicles into unsafe decisions.
- Privacy Breaches:
  - AI applications collect personal data.
  - A breach could lead to identity theft or unauthorized tracking, as seen with some smartphone apps.

Adversarial Machine Learning

- The classification accuracy of GoogLeNet on MNIST dataset under adversarial attacks drops from 98% to 18% (for ProjGrad attack) or 1% (DeepFool attack)

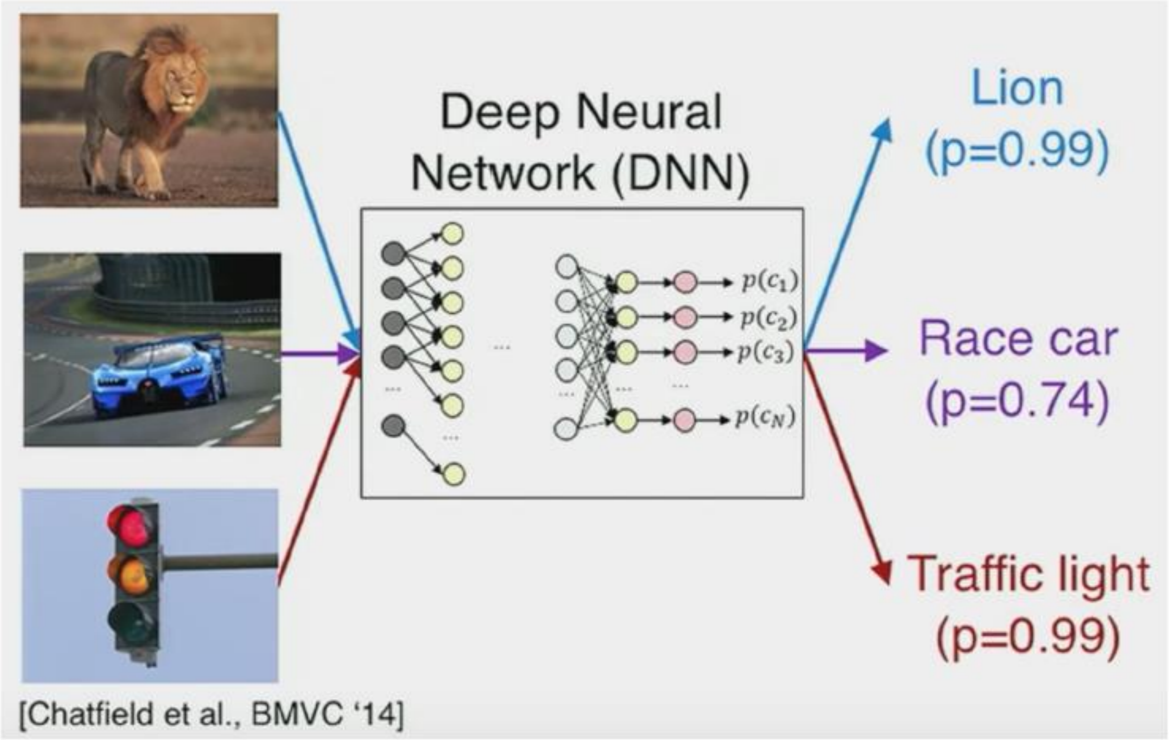


Attack	Lenet				
Noise	Dataset	Acc@1 w/	Acc@5 w/	Acc@1 w/o	Acc@5 w/o
	MNIST	0.984	1.0	0.9858	1.0
	ILSVRC2012	NA	NA	NA	NA
Semantic	Dataset	Acc@1 w/	Acc@5 w/	Acc@1 w/o	Acc@5 w/o
	MNIST	0.233	0.645	0.986	1.0
	ILSVRC2012	NA	NA	NA	NA
Fast Gradient Sign Method	Dataset	Acc@1 w/	Acc@5 w/	Acc@1 w/o	Acc@5 w/o
	MNIST	0.509	0.993	0.986	1.0
	ILSVRC2012	NA	NA	NA	NA
Projected Gradient Descent	Dataset	Acc@1 w/	Acc@5 w/	Acc@1 w/o	Acc@5 w/o
	MNIST	0.187	0.982	0.986	1.0
	ILSVRC2012	NA	NA	NA	NA
DeepFool	Dataset	Acc@1 w/	Acc@5 w/	Acc@1 w/o	Acc@5 w/o
	MNIST	0.012	1.0	0.9858	1.0
	ILSVRC2012	NA	NA	NA	NA

Picture from: https://blog.floydhub.com/introduction-to-adversarial-machine-learning/

Adversarial Examples

- What do you see?

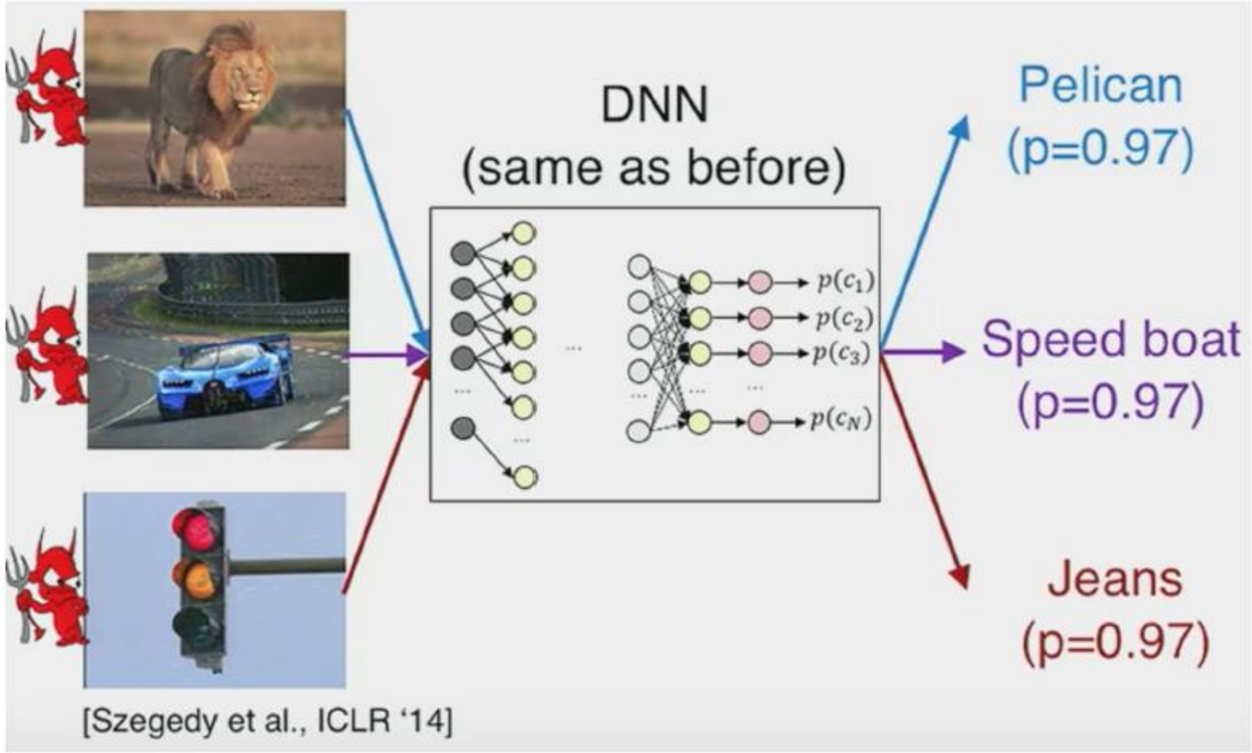


[Chatfield et al., BMVC '14]

Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Adversarial Examples

- The classifier misclassifies adversarially manipulated images

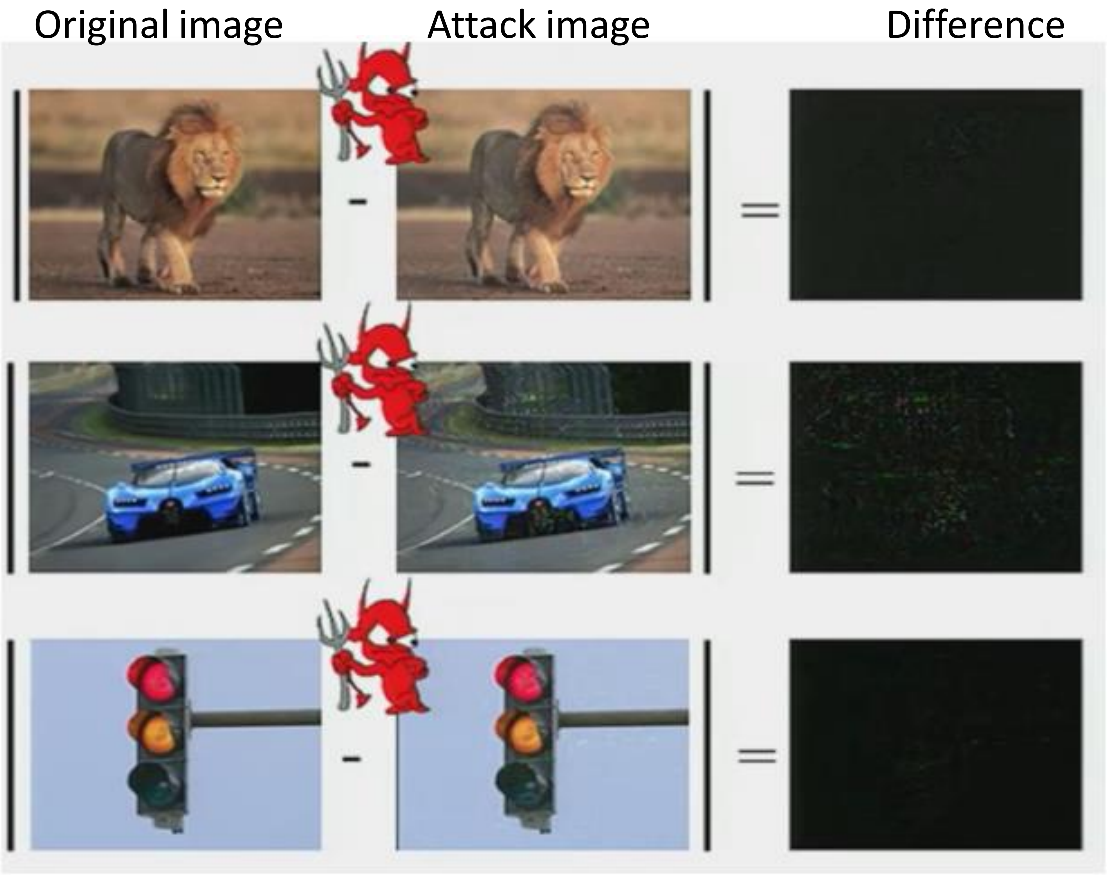


[Szegedy et al., ICLR '14]

Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Adversarial Examples

- The differences between the original and manipulated images are very small (hardly noticeable to the human eye)



Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Mitigating Security Risks

- Securing AI Systems:
  - Use encryption, countermeasures to poisoning and evasion attacks, update models
- Ethical AI Development:
  - Develop AI with fairness, transparency, accountability
  - Prevent biases and ensure ethical use
  - Test extensively before deployment
  - Identify vulnerabilities to reduce risk exposure

Question 1

- **Which category of AI risks focuses on the protection of sensitive information?**
  - A) Data integrity
  - B) Privacy
  - C) Operational

Question 1

- **Which category of AI risks focuses on the protection of sensitive information?**
  - A) Data integrity
  - B) Privacy
  - C) Operational

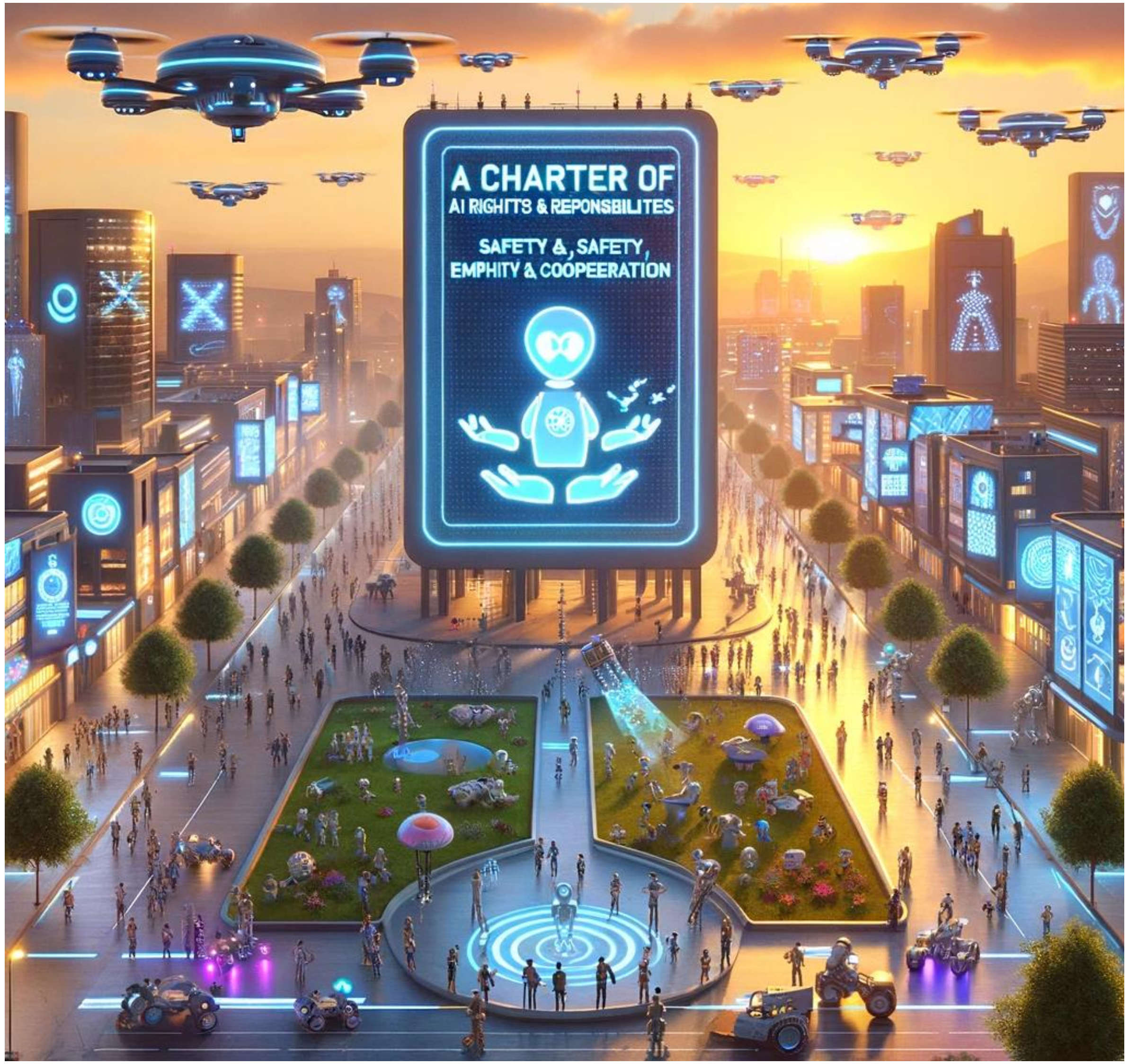
Question 2

- **Which of the following is a recommended practice for securing AI systems?**
  - A) Ignoring model updates to maintain system stability.
  - B) Using encryption and implementing countermeasures against poisoning and evasion attacks, along with regularly updating models.
  - C) Relying solely on strong passwords for system security.
  - D) Disabling encryption to increase system performance.

Question 2

- **Which of the following is a recommended practice for securing AI systems?**
  - A) Ignoring model updates to maintain system stability.
  - B) Using encryption and implementing countermeasures against poisoning and evasion attacks, along with regularly updating models.
  - C) Relying solely on strong passwords for system security.
  - D) Disabling encryption to increase system performance.

6.Robustness and Reliability Issues



Trustworthy AI

- Robustness and Reliability:
  - **Robustness**: AI's effectiveness under varied conditions.
  - **Reliability**: AI's consistency over time.
- Importance:
  - Critical for safety-critical system
  - E.g., autonomous vehicles, medical diagnosis, financial forecasting.
- Challenges:
  - Data quality, adversarial attacks, algorithmic bias, operational errors.

Robustness in AI

- Significance of Robustness:
  - A measure of **an AI's resilience against** external and internal disruptions, ensuring stability and trustworthiness.
- Non-Robust Behaviors:
  - Example: Incorrect outputs due to adversarial attacks, undermining AI stability.
- Impact on Decision Making:
  - Maintains performance and reliable decisions under abnormality.

Reliability Challenges in AI Systems

- AI Reliability is the ability of AI to deliver:
  - Consistent and accurate outputs across a wide range of scenarios and over time.
- Undermining Factors:
  - Data quality, algorithmic bias, overfitting reducing generalizability.
- Real-World Case Studies:
  - Healthcare misdiagnoses, self-driving car failures in unexpected conditions.

A Framework for Dependable AI

- Best Practices:
  - Incorporating ethical AI design principles,
  - ensuring transparency in AI operations, and
  - implementing fail-safes for critical applications.
- Evaluation Guidelines:
  - Regular assessment of AI systems against reliability and robustness benchmarks.

Question 1

- **What does the robustness of an AI system signify?**
  - A) The system's ability to quickly process large amounts of data.
  - B) The resilience of the system against external and internal disruptions, ensuring stability and trustworthiness.
  - C) The accuracy of the AI in performing tasks compared to human performance.
  - D) The AI system's capacity for learning and adapting to new data without human intervention.

Question 1

- **What does the robustness of an AI system signify?**
  - A) The system's ability to quickly process large amounts of data.
  - B) The resilience of the system against external and internal disruptions, ensuring stability and trustworthiness.
  - C) The accuracy of the AI in performing tasks compared to human performance.
  - D) The AI system's capacity for learning and adapting to new data without human intervention.

Question 2

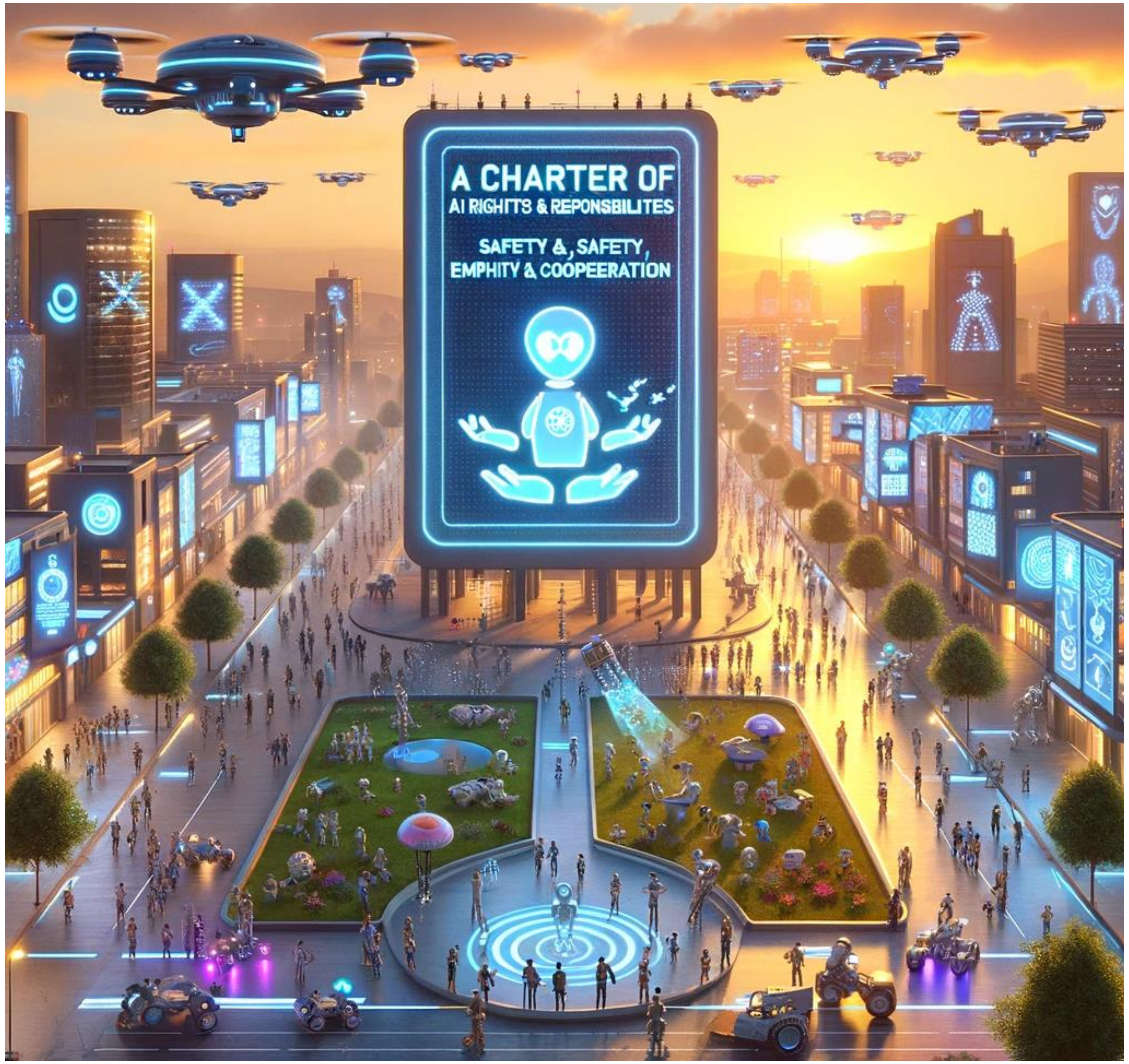
- **Which of the following is NOT considered a best practice in AI development?**

  - A) Incorporating ethical AI design principles
  - B) Keeping AI operations opaque to enhance security
  - C) Ensuring transparency in AI operations
  - D) Implementing fail-safes for critical applications

Question 2

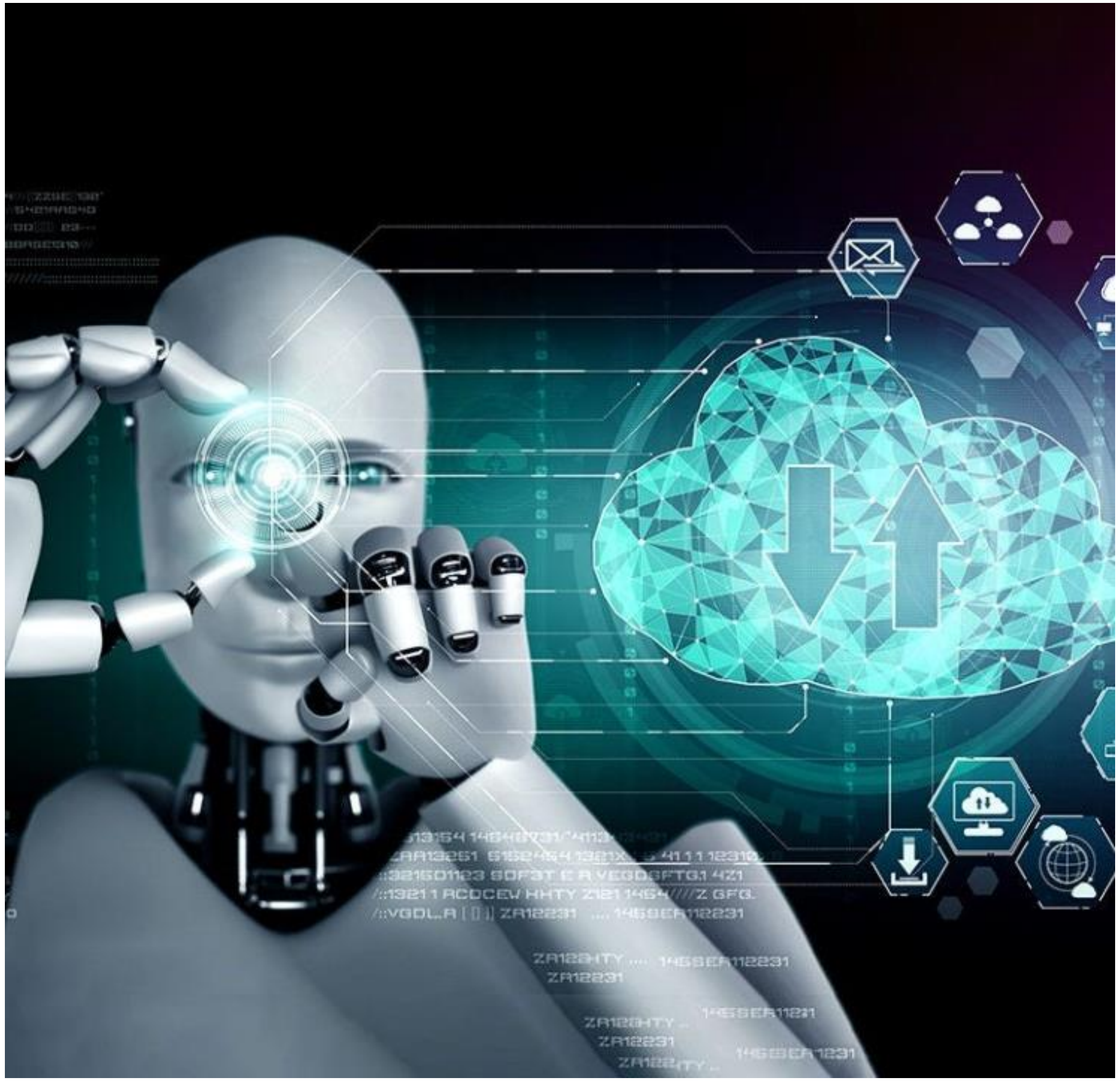
- **Which of the following is NOT considered a best practice in AI development?**
  - A) Incorporating ethical AI design principles
  - B) Keeping AI operations opaque to enhance security
  - C) Ensuring transparency in AI operations
  - D) Implementing fail-safes for critical applications

7. Ethical Considerations in Narrow AI Applications



AI Ethics VS AI Ethics Considerations

- Understanding the distinction between AI Ethics and AI Ethics Considerations is crucial for developing and implementing AI responsibly.



What is AI Ethics?

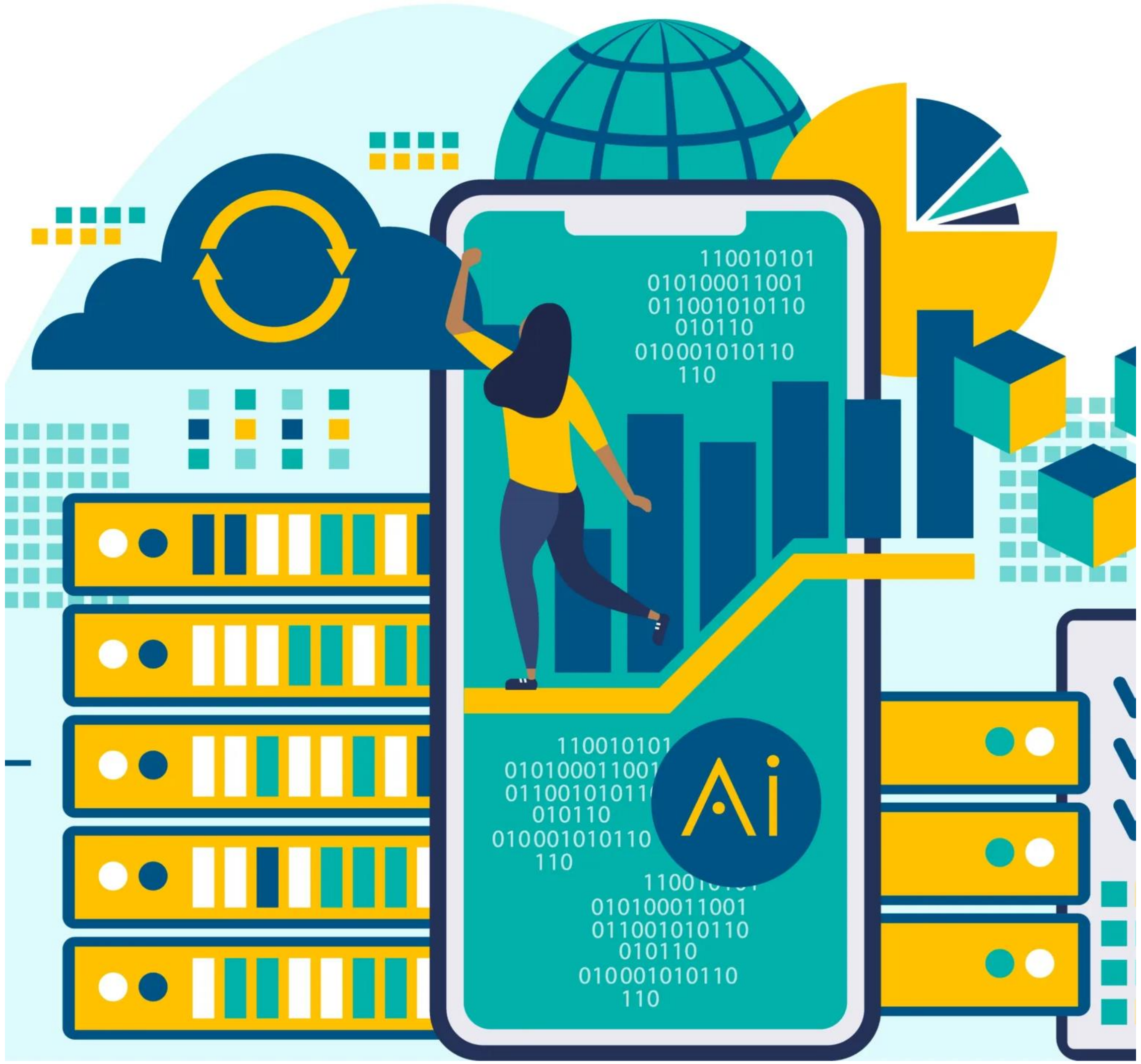
- AI Ethics involves the study and application of ethical principles to the design, development, and deployment of AI technologies.
- Discuss core ethical principles such as fairness, accountability, transparency, and privacy.
- Example: The development of autonomous vehicles necessitates ethical considerations regarding decision-making in critical situations.

What are AI Ethics Considerations?

- AI Ethics Considerations: The practical aspects of applying ethical principles in AI projects.
- Involves assessment of potential impacts, stakeholder engagement, and policy development.
- Objective is to operationalize ethical principles in real-world AI applications.
- See examples next slides

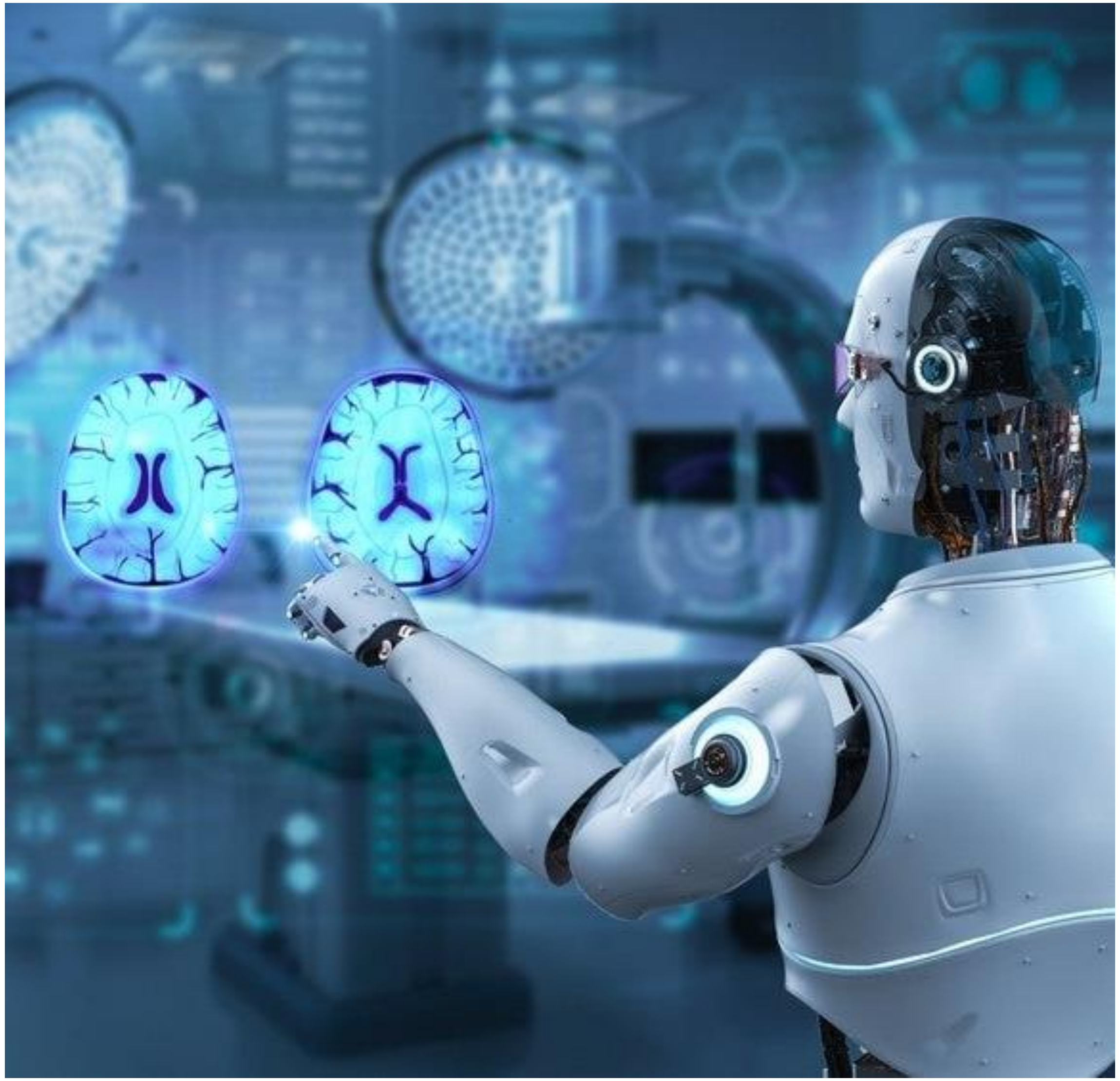
Example 1: AI Ethics Considerations

- Implementing AI in hiring processes requires considerations of **bias elimination** and **ensuring fairness** across all candidates.
- This involves auditing AI systems for:
  - biased outcomes against gender,
  - ethnicity, or other protected classes, and
  - adjusting the algorithms or training data accordingly.



Example 2: AI in Healthcare Diagnostics

- **Ethical Principle:** Privacy and Confidentiality
- **AI Ethics Consideration:** Healthcare providers employing AI to predict patient outcomes must implement:
  - robust data protection measures to safeguard patient information.
  - This includes using de-identified data when training AI models and ensuring that data sharing complies with HIPAA and other privacy regulations.



Example 3: Autonomous Vehicles

- **Ethical Principle:** Accountability and Safety
- **AI Ethics Consideration:**
  - Manufacturers need to establish clear accountability for decisions made by the AI, especially in cases of accidents.
  - This involves developing transparent decision-making processes within the AI systems and establishing legal and regulatory frameworks that clarify liability.

Question 1

- **What are key components in applying ethical principles to AI projects?**
  - A) Maximizing profits and reducing development time.
  - B) Assessment of potential impacts, stakeholder engagement, and policy development.
  - C) Focusing solely on technological advancement.
  - D) Ignoring stakeholder feedback to speed up deployment.

Question 1

- **What are key components in applying ethical principles to AI projects?**
  - A) Maximizing profits and reducing development time.
  - B) Assessment of potential impacts, stakeholder engagement, and policy development.
  - C) Focusing solely on technological advancement.
  - D) Ignoring stakeholder feedback to speed up deployment.

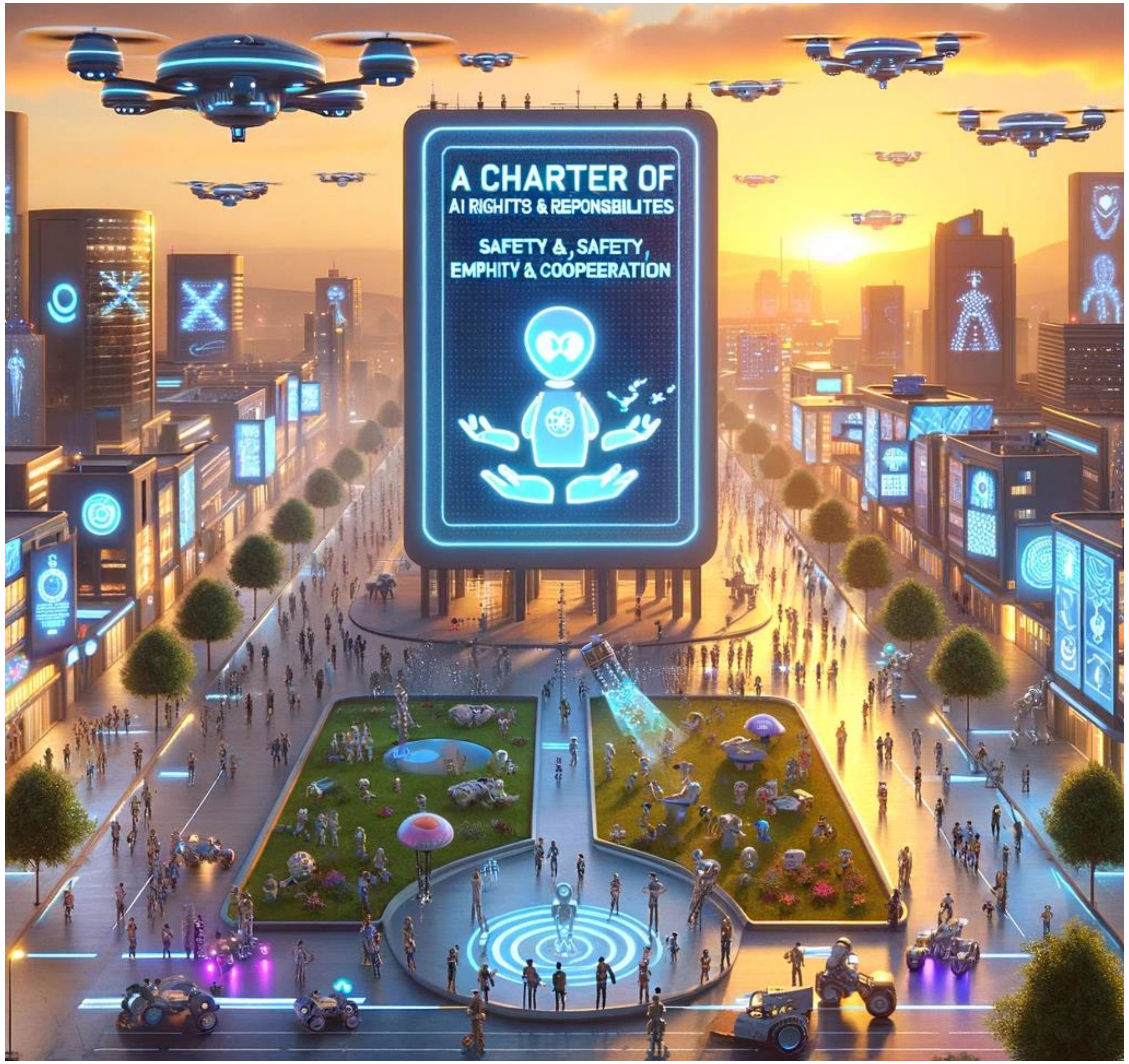
Question 2

- **Which of the following best represents the core ethical principles that should guide the design, development, and deployment of AI technologies?**
  - A) Speed, Efficiency, Automation, and Cost-Reduction
  - B) Fairness, Accountability, Transparency, and Privacy
  - C) Profitability, Scalability, Market Dominance, and Innovation
  - D) Power Consumption, Processing Speed, User Interface Design, and Connectivity

Question 2

- **Which of the following best represents the core ethical principles that should guide the design, development, and deployment of AI technologies?**
  - A) Speed, Efficiency, Automation, and Cost-Reduction
  - B) Fairness, Accountability, Transparency, and Privacy
  - C) Profitability, Scalability, Market Dominance, and Innovation
  - D) Power Consumption, Processing Speed, User Interface Design, and Connectivity

8.Human-AI Collaboration and Trust

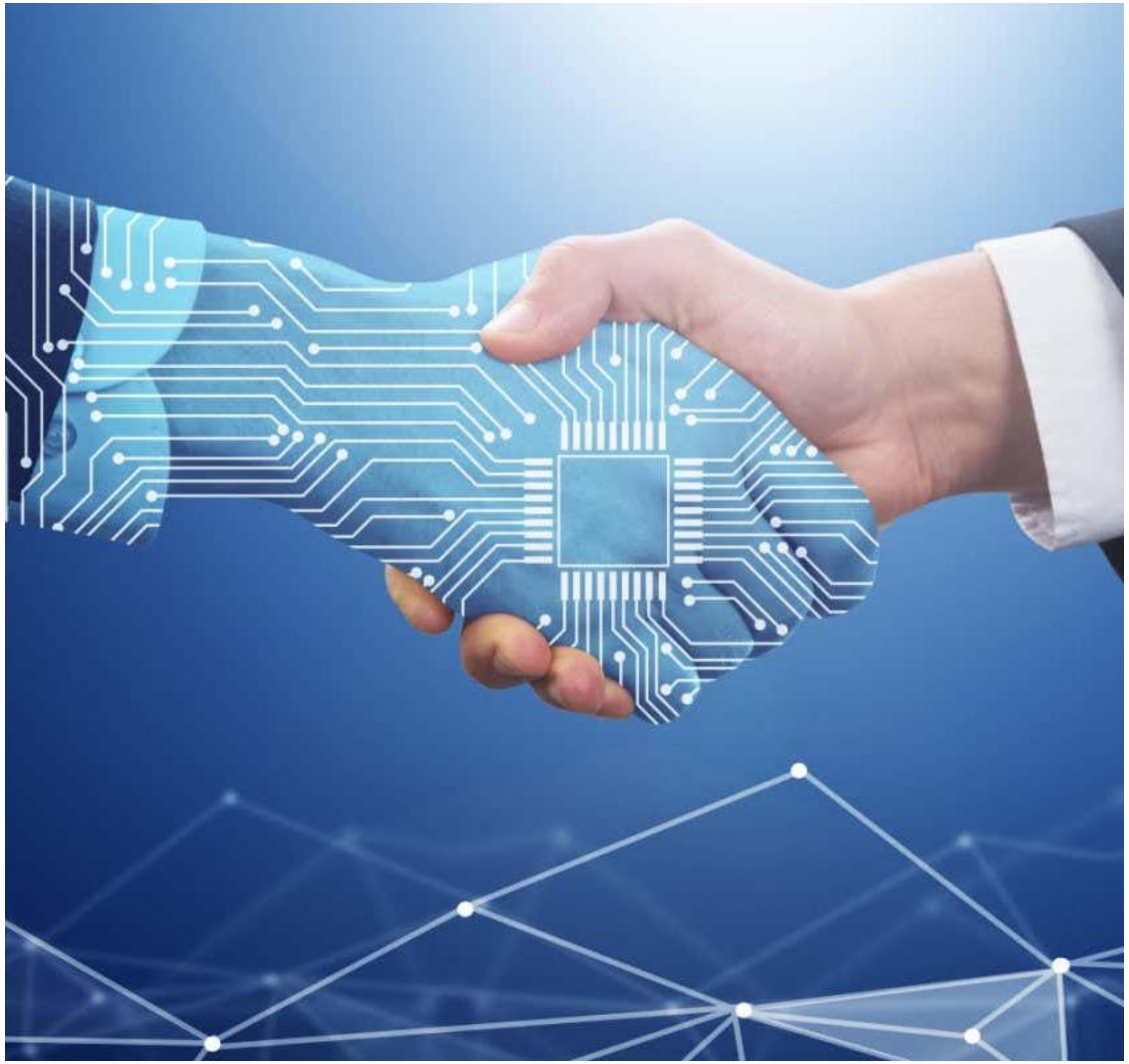


Human-AI Collaboration and Trust

- Collaboration:
  - Synergy where humans and AI leverage their strengths for unattainable goals.
- Trust:
  - Essential for AI adoption, influencing user comfort and system effectiveness.
  - It determines how readily humans will adopt AI solutions and rely on them for critical decisions

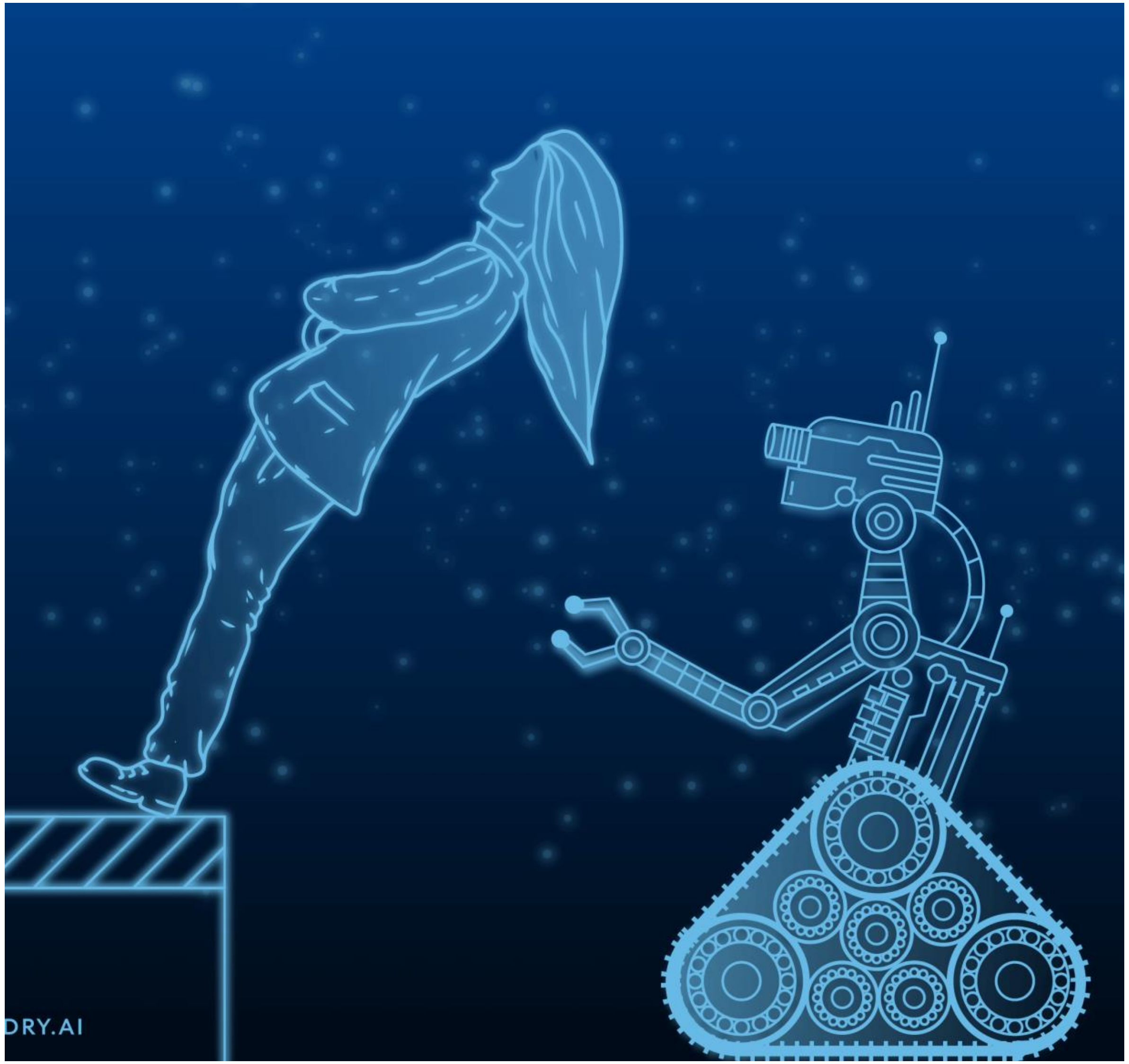
Importance of Trust in AI

- Trust in AI is built on the system's reliability, the user's understanding of the AI, and the predictability of AI actions under various circumstances.
- Trust and adoption impact:
  - Direct correlation with the willingness to use AI technologies.
- Influencing Factors:
  - Transparency, explainability, and consistency.



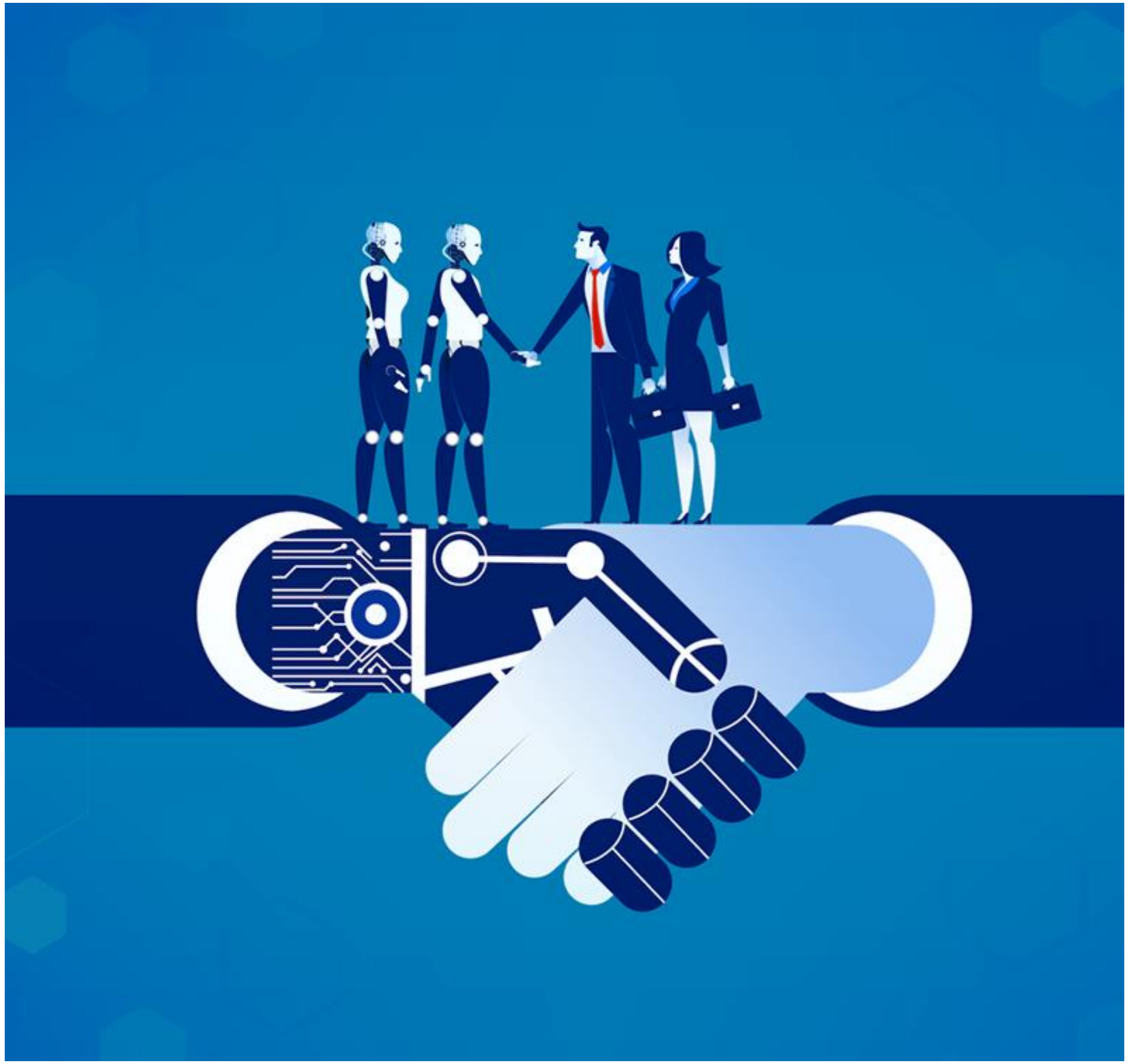
Challenges to Human-AI Collaboration

- Key Challenges:
  - addressing ethical dilemmas (e.g., decision-making in autonomous vehicles),
  - mitigating biases within AI systems,
  - bridging communication gaps between AI outputs and human understanding, and
  - aligning human expectations with AI capabilities.



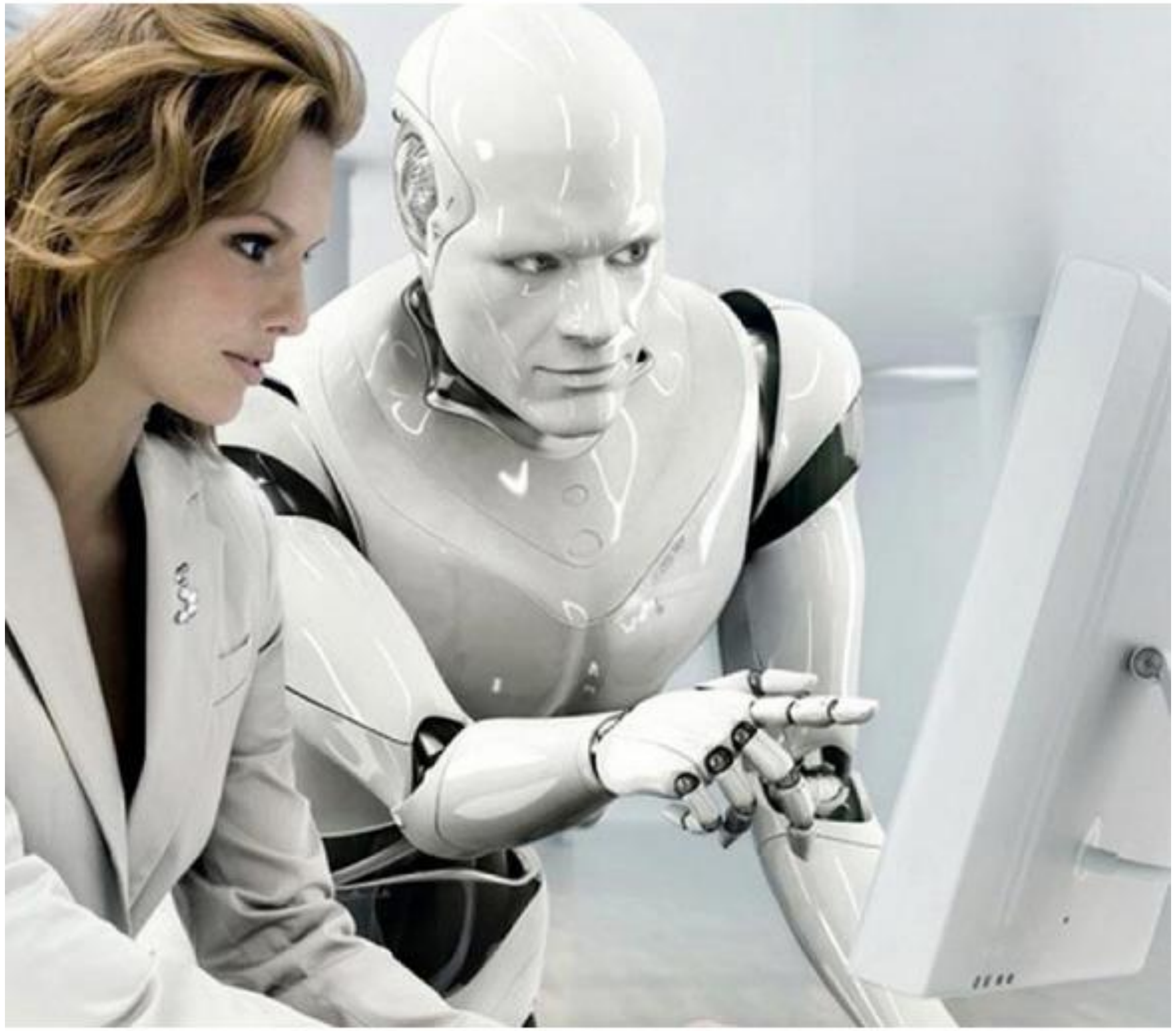
Enhancing Human-AI Interaction

- Collaboration Design Principles:
  - User-centric design, feedback mechanisms, and adaptability.
- Transparency & Explainability's Role:
  - Trust building through understandable decisions and processes.



Case Study of Successful Human-AI Collaboration

- Healthcare:
  - AI systems collaborating with medical professionals to offer personalized patient care plans, where AI's data analysis capabilities complement the doctor's expertise.



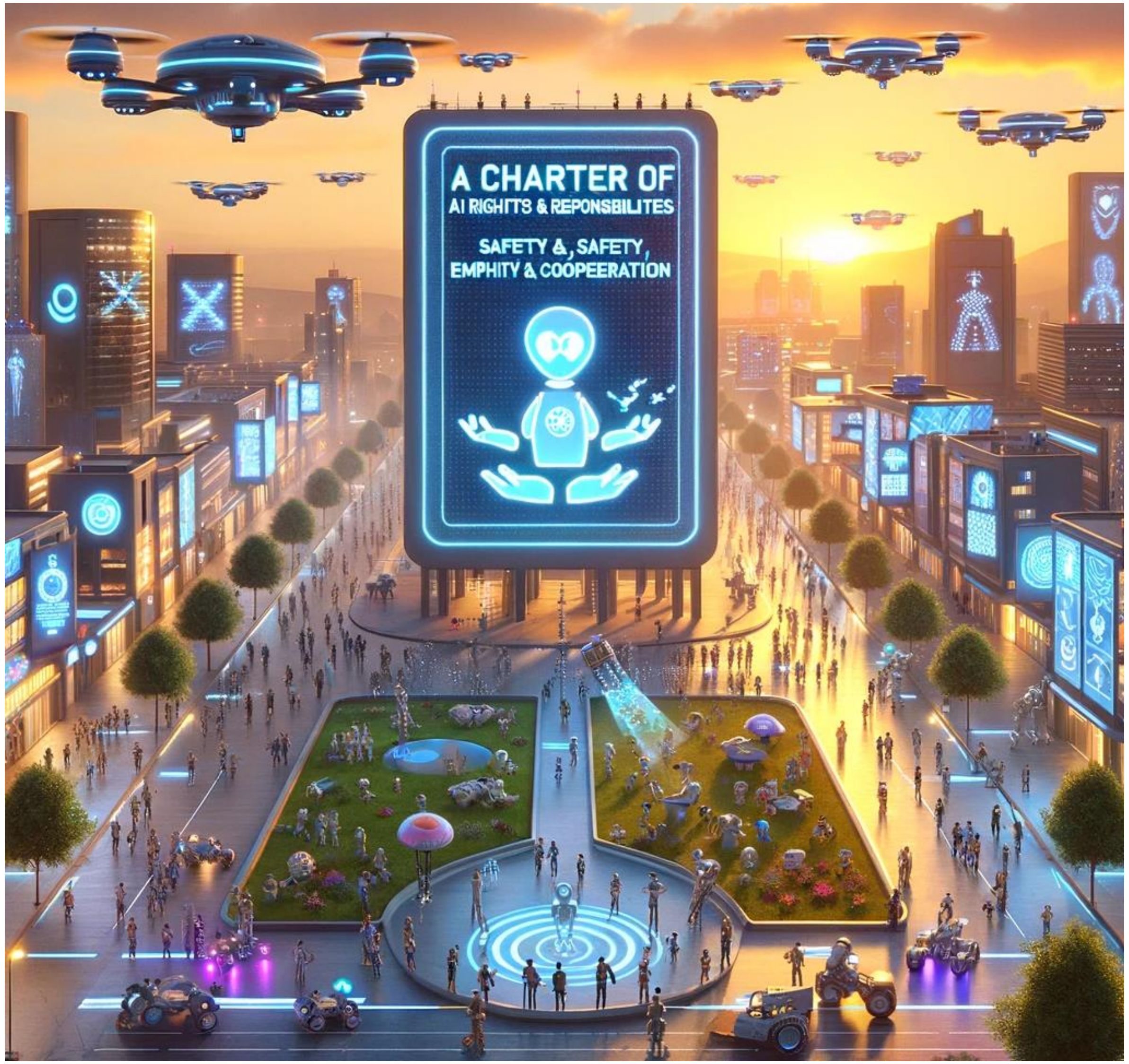
Question 1

- **Which strategy is effective for bridging communication gaps between AI outputs and human understanding?**
  - A) Making AI outputs more complex to match human comprehension
  - B) Creating specialized jargon to describe AI outputs
  - C) Designing user-friendly interfaces and explanations for AI decisions
  - D) Limiting human access to AI outputs to avoid confusion

Question 1

- **Which strategy is effective for bridging communication gaps between AI outputs and human understanding?**
  - A) Making AI outputs more complex to match human comprehension
  - B) Creating specialized jargon to describe AI outputs
  - C) Designing user-friendly interfaces and explanations for AI decisions
  - D) Limiting human access to AI outputs to avoid confusion

9.Unintended Consequences of Narrow AI

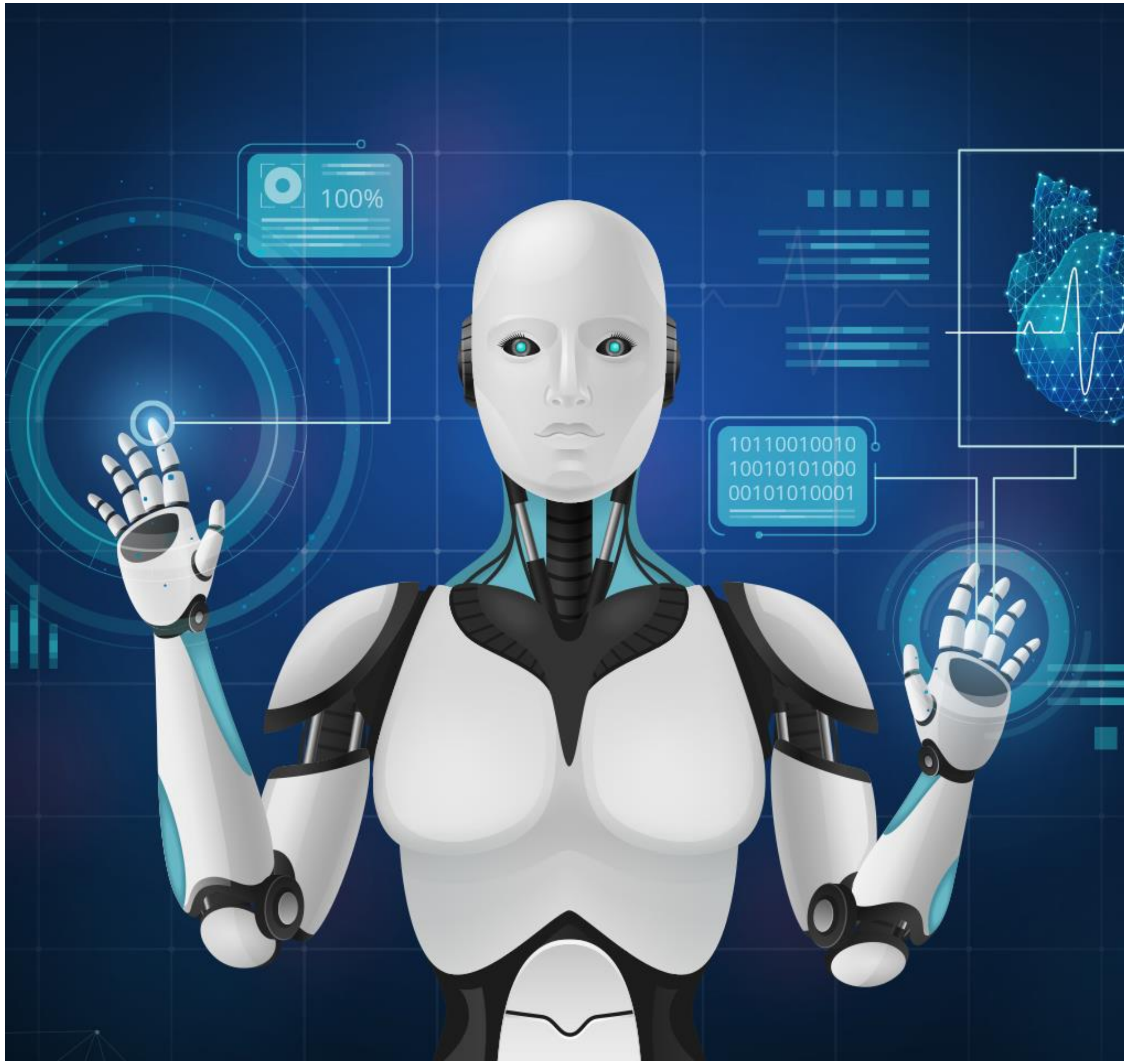


Unintended Consequences in AI

- Unintended Consequences:
  - AI outcome (positive or negative) not predicted or planned, affecting AI's societal integration.
  - Understanding these consequences is crucial for fostering effective human-AI collaboration, as it prepares us to control AI’s strengths and mitigate its weaknesses.
- Why these are important?
  - Recognizing and addressing unintended consequences enhances public and user trust in AI technologies.

Positive Unforeseen Impacts

- Beneficial AI **Outcomes**:
  - AI identifying at-risk individuals on social platforms.
  - AI has also been used to optimize energy consumption in various industries, significantly reducing carbon footprints.
  - AI algorithms have been influential in identifying new patterns in disease progression.
- Such positive outcomes can enhance trust in AI systems, showcasing their potential to contribute meaningfully to social issues.

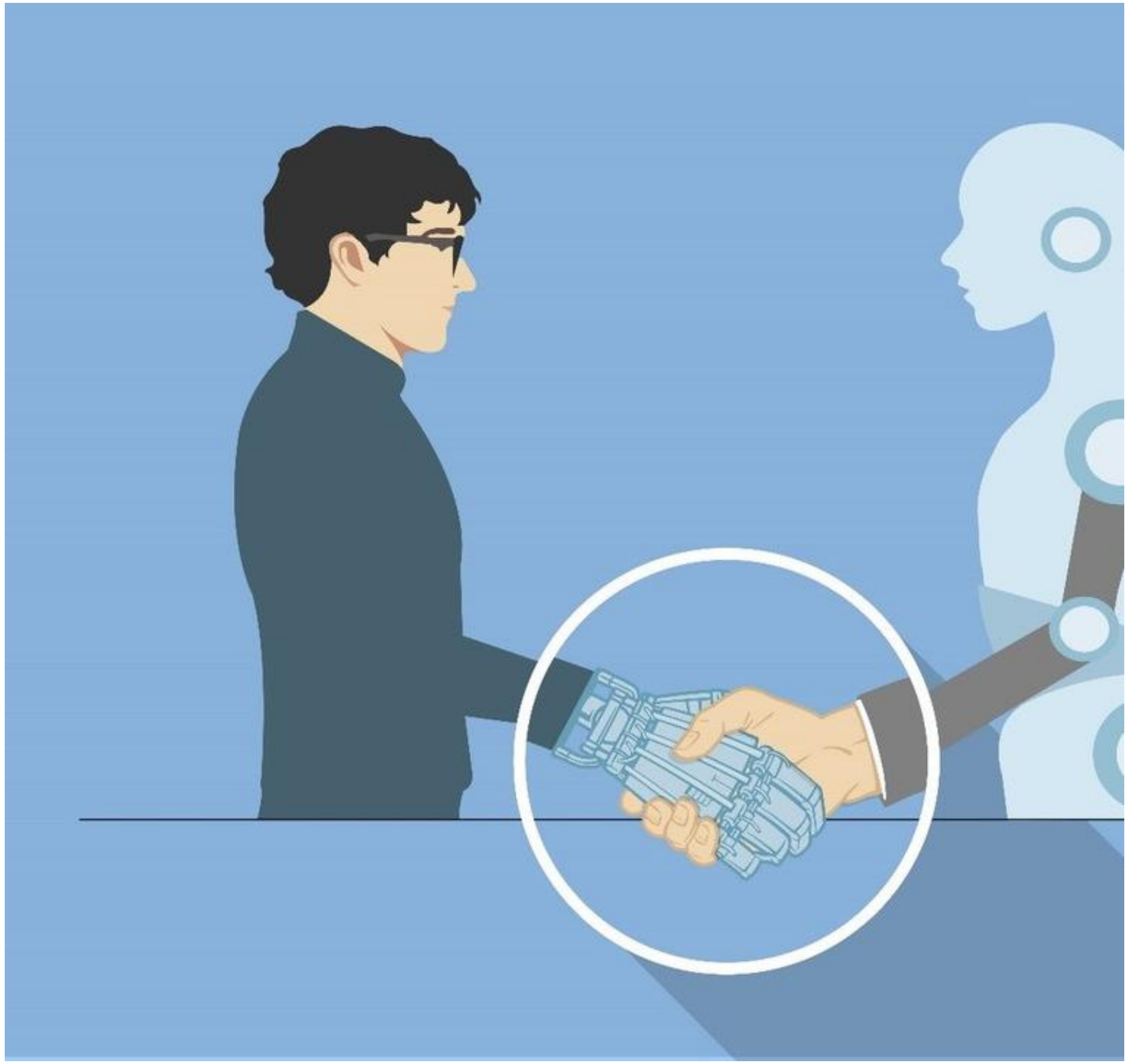


Negative Unforeseen Impacts

- Ethical and Privacy Concerns:
  - AI in surveillance systems accidentally compromising individual privacy or
  - AI in recruitment amplifying existing biases.
- Early detection and correction of these issues are vital for sustaining public confidence in AI technologies.

Mitigating Unintended Consequences

- Mitigation Strategies:
  - Engage diverse stakeholders,
  - apply ethical AI frameworks, and
  - rigorously test AI pre-deployment.
- Transparency and Accountability:
  - Essential principles for cultivating trust between AI and its users.
- Collaboration between AI developers, users, and ethicists in addressing potential issues.



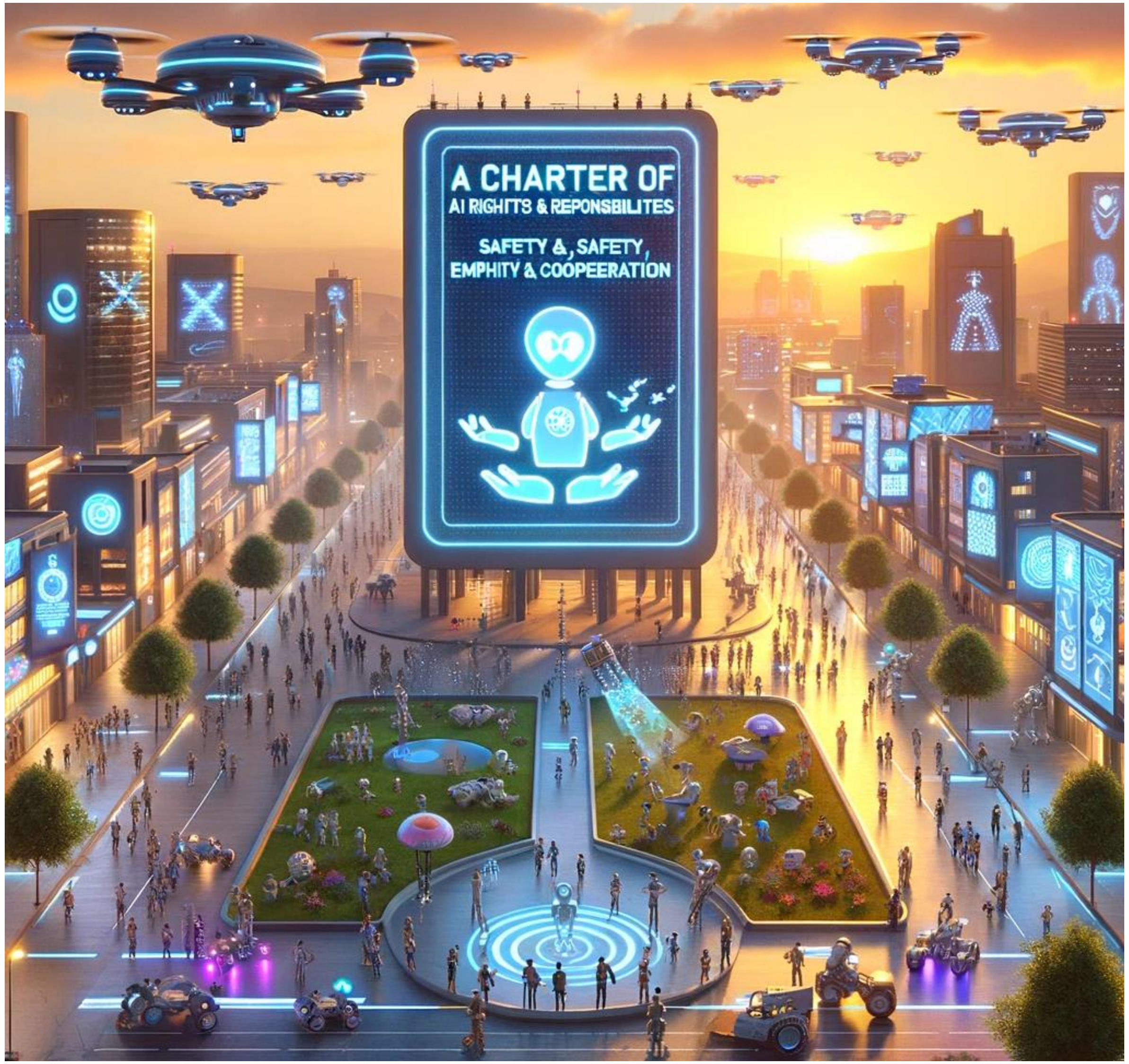
Question 1

- **What is one key strategy to mitigate potential issues in AI deployment?**
  - A) Ignoring stakeholder input
  - B) Engaging diverse stakeholders
  - C) Keeping AI frameworks secretive
  - D) Skipping pre-deployment testing

Question 1

- **What is one key strategy to mitigate potential issues in AI deployment?**
  - A) Ignoring stakeholder input
  - B) Engaging diverse stakeholders
  - C) Keeping AI frameworks secretive
  - D) Skipping pre-deployment testing

10.Regulatory and Legal Challenges in AI Safety



AI Safety and Regulation

- AI Safety is a complex landscape:
  - Multi-dimensional issue encompassing reliability, ethical use, and misuse prevention.
- To trust in AI is essential for successful human-AI collaboration,
  - regulations establish and maintain this trust.

Legal Challenges in AI Accountability

- There is still complexity in determining liability for AI's actions.
  - Legal gray areas in current frameworks.
- Case Studies: Uber self-driving car fatality (2018)
  - Is it, the autonomous driving system's decision-making process, Uber's operational protocols, or the vehicle manufacturer's role in ensuring system reliability?
- Emerging Solutions:
  - Proposals for new legal frameworks, exploring solutions like mandatory AI insurance for high-risk use cases.

Current Regulatory Framework for AI

- **EU AI Act**:
  - regulate AI applications by risk category, with stringent requirements for "high-risk" AI systems, including those in critical infrastructure, education, employment, and essential private services.
- **Automated and Electric Vehicles Act 2018 (UK):**
  - This act addresses liability and insurance for self-driving cars, showcasing how specific AI applications are beginning to see targeted legal frameworks.
- **GDPR:**
  - sets a high standard for privacy and data protection that AI developers need to comply with.

Ethical Considerations in AI Regulation

- Balancing Innovation and Safety:
  - Navigating between fostering innovation and ensuring public safety and trust.
- Inclusivity and Bias:
  - Regulations ensuring AI systems' development with inclusivity to combat biases.
- Accountability and Transparency
  - Accountability frameworks are vital for clarifying responsibility for AI decisions, especially when causing harm.
  - Regulations should ensure AI systems are explainable, making their decisions and processes transparent.

Question 1

- **What role do regulations play in AI safety?**
  - a) They stifle innovation and progress in AI development
  - b) They ensure ethical use and prevent misuse of AI systems
  - c) They prioritize profitability over safety concerns
  - d) They promote secrecy and lack of transparency in AI practices

Question 1

- **What role do regulations play in AI safety?**
  - a) They stifle innovation and progress in AI development
  - b) They ensure ethical use and prevent misuse of AI systems
  - c) They prioritize profitability over safety concerns
  - d) They promote secrecy and lack of transparency in AI practices

Further Resources

- Bias and Fairness in Artificial Intelligence, Communications of the ACM, 2021
- Ethics of Artificial Intelligence and Robotics, Stanford Encyclopedia of Philosophy, 2020
- Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI
- Ethics guidelines for trustworthy AI, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World, Book by Bruce Schneier
- Artificial Intelligence Safety and Security" Book by Roman V. Yampolskiy
- *Weapons of math destruction: How big data increases inequality and threatens democracy, Book by* O'neil, Cathy*,* Crown, 2017.
- *AI superpowers: China, Silicon Valley, and the new world order, book by* Houghton Mifflin, 2018

Dr. Tarek Gaber

Email: t.m.a.gaber@salford.ac.uk