Ethics in Artificial Intelligence_V3_Unit 1.docx

Ethics in Artificial Intelligence_MBA_2

Ethics in Artificial Intelligence_MBA_2

ATLAS SkillTech University

Document Details

Submission ID
trn:oid:::3618:127350326

Submission Date
Feb 2, 2026, 11:32 AM GMT+5:30

Download Date
Feb 2, 2026, 1:03 PM GMT+5:30

File Name
Ethics in Artificial Intelligence_V3_Unit 1.docx

File Size
44.0 KB

27 Pages

5,273 Words

32,292 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 1: Introduction to AI and Ethics

Learning Outcomes

1. Understand the fundamentals of Artificial Intelligence (AI) and its applications.

2. Recognize the importance of ethics in guiding responsible AI development.

3. Explore key ethical theories (e.g., utilitarianism, deontology, virtue ethics) relevant to AI.

4. Identify and analyze emerging ethical challenges in AI, such as bias, privacy, and accountability.

5. Learn the principles of ethical AI design to ensure fairness, transparency, and inclusivity.

6. Apply ethical frameworks to evaluate real-world AI systems and their impacts.

7. Develop a critical perspective on how AI and ethics intersect to shape future society and innovation.

Content

1.0 Introductory Caselet

"The Algorithm in the Classroom: A Dialogue between Meera and Her Teacher"

Background:

High school student Meera is mystified by the new online history quiz posted by her tutor.

tests, and videos for learning just on the topics she doesn't understand. Perplexed, she turns to her teacher and asks, "How puede?"

the system know what I want without my asking for it?"

Her teacher smiles and explains,

"Artificial Intelligence is behind the scenes, analysing your performance, identifying patterns and predicting what%"),

might help you improve. It is not magic — it's a machine learning from your data. But remember, with this

power comes responsibility. AI can assist us, but we need to teach it to act fairly and ethically many more ways."

In time, Meera ceases to view AI as technology and instead as a mechanism that is influencing decisions, possibilities,

fairness, day in and day out.

Critical Thinking Question:

How are we to weigh the good in AI-driven personalization against challenges related to fairness, bias, and accountability?

privacy?

1.1 Introduction to AI

Artificial Intelligence (AI) denotes the capability of machines to mimic human intelligence, including learning and problem-solving.

designed to think, learn and make decisions. Unlike with conventional software, AI systems learn and improve.

over time, often improving as it processes more data.

Key Characteristics of AI

Learning: AI gets better with data (machine learning, deep learning).

Thinking - The ability to work problems, extend predictions, and suggest solutions.

Perception: Identifying objects, speech, or patterns (such as facial recognition).

Interacting: Naturally interacting with humans (chatbots, voice assistants).

Categories of AI

• Narrow AI (Weak AI):

Primarily Designed for one specific task, such as Siri, googletranslate and Netflix recommendations.

• General AI (Strong AI):

Counterfactual AI able to complete any intellectual function of a man. Still a future goal.

• Superintelligent AI:

A hypothetical moment when we'll need to regulate artificial intelligence — if it ever happens.

Applications of AI

• Health: Diagnosis of diseases, drug invention, personalized medicine.

• Finance: Fraud detection, algorithmic trading, financial risk management.

• Education: Personalized learning, smart tutoring systems.

• Transportation: Autonomous vehicles, traffic forecasting.

• Daily Life: Virtual assistants, advice about what to buy, smart homes.

Why AI Matters

• Efficiency: Automates repetitive tasks.

• Scalable: Copes with huge volumes of data that humans cannot manage.

• Innovation: Opens up new frontiers in science, business and society.

Yet AI also raises important issues about bias, accountability, transparency and ethics that will need to be addressed.

which the remainder of this chapter addresses in greater detail.

1.1.1 Definition of Artificial Intelligence

Artificial Intelligence (AI) is the science of producing computers and software that are capable of intelligent behaviour.

machines that have the ability to do things that require human intelligence.

• Key Aspects of Definition:

Artificial Intelligence Simulation: machines trying to emulate brain activities such as learning,

problem-solving, and reasoning.

Flexibility Unlike traditional software, AI can adapt its behavior when it encounters new and different problems.

information.

Decision-Making AI is not only responsive but also capable of considering other options and selecting the best one

solutions.

• Classical Definition: John McCarthy (1956), one of the pioneer figures in AI, defined it as "the

science and engineering of making intelligent machines.

• Modern Perspective: AI is now viewed as the combination of algorithms, data, and computational

ability to develop systems that "learn" and get better at tasks on their own.

• Slight Example: Google Translate doesn't just swap in words; it learns to understand context,

grammar and meaning to obtain accurate translations.

1.1.2 Features and Characteristics of AI

AI systems are unique from traditional computer systems; in that they demonstrate intelligent behavioural responses.

• Core Features:

Learning Capacity: AI systems "learn" from information. Machine Learning (ML) and Deep

DL drive this functioning capability.

Example: Anti-spam filters that improve as they are given more examples of spam.

Reasoning and Problem-Solving: AI can interpret data, form connections, draw inferences, serve as the mainitätsbereich links them einfügen relationships between all fixed orDiese reasoning an IntrodUction to Government&PoliticsSeventh EditionPAULCESAR DEB Enoch C.H.squareworld.

complex problems.

Example: Google Maps trying several different paths to identify the quickest.

Perception: AI is able to identify objects, images, sounds or text through the use of sensors or algorithms.

Example: Autonomous vehicles identifying pedestrians and traffic signals.

Natural interaction: AI interacts with human in intuitive way(speech, text,

gestures).

Example: Siri and Alexa virtual assistants.

Autonomy : Once AI is trained it can continue to act without human involvement.

Example: Autonomous drones delivering packages.

• Additional Characteristics:

o Flexibility: Capacity to behave flexibly as the situation changes.

o Data-Driven: It succeeds with a lot of data.

o Predictive Power: Is able to predict what the user wants or will happen with.

1.1.3 Types of AI: Narrow, General, Super AI

Here's one way to think about the different "types" of AI: Types of AI depending on capabilities: 1.

Narrow AI (Weak AI):

o Concentrated on doing something.

o No self awareness or general intelligence.

o Examples:

▪ Google Translate (language translation).

▪ Netflix recommendation engine.

▪ Siri or Alexa assistants for voice.

o Current reality: Nearly all uses of AI today are narrow AI.

General AI (Strong AI):

o Hypothetical AI that can do anything that any human can.

o Able to reason, plan, adapt in a wide range of domains.

o E.g. (some made up example): A robot that writes essays, prepare a meal and do an experiment – hypothetically.

and instruct courses without additional instruction.

o Status: Still research, not yet reached.

Super AI:

o A point where AI outsmarts humans (conceptual phase)

o Might be better at logic, creativity, and social intelligence than humans.

o Coupled with existential risks and ethical battles.

o Examples:

▪ Shown in science fiction (HAL 9000 in 2001: A Space Odyssey, “Skynet” in

Terminator).

o Raises questions: Would superintelligent AI have humanity’s best interests at heart?

1.1.4 Historical Development of AI

The rise of AI is punctuated with waves of optimism, stagnation, and breakthroughs:

• 1940s–1950s: Foundations

o Alan Turing puts forward the concept of machines that “think.”

o Turing Test (1950): If a human being cannot distinguish whether he/she is conversing to a machine or to another human being, then the machine has passed the test.

human, the machine is intelligent.

• 1956: Dartmouth Conference

o The birth of AI as an industry. John McCarthy,Marvin Minsky,Nathaniel Rochester,

, Claude Shannon introduce the idea of “thinking machines.”

• 1960s–1970s: Early Progress

o Research is largely rule-based systems and symbolic reasoning.

o Implementation of “expert systems” that could simulate human problem-solving in specific domains

(medicine, chemistry).

o Limitations: Computing power and data constraining progress.

• 1980s: Machine Learning Emerges

o Propagation of errors in neural networks.

o Japan has made large investments in AI and created the ÒFifth Generation Project.

• 1990s–2000s: Real-World Successes

o 1997: IBM's Deep Blue beats world chess champion Garry Kasparov.

o Statistical methods and data-driven AI grow at pace.

• 2010s–21st Century: Big Data & Deep Learning Revolution There exists excruciatingly low level of criticism and opposition to deep learning, when in fact there are billion-euro businesses that leap into frameworks without understanding the implications nor the framework.

o Surge in supercomputing and big data.

o AI advancements in: speech recognition, image recognition, natural language ,…

processing.

o 2016: DeepMind's AlphaGo triumphs over world champion Lee Sedol in the game of Go—a

milestone in AI complexity.

o Today: AI enables automated cars, medical diagnostics, generative AI tools like ChatGPT,

and more.

Did You Know?

"In 1956, the term Artificial Intelligence was first used at the Dartmouth Conference. The founders predicted that AI could match human intelligence in just a generation—but instead, the field went through long "AI winters" when funding and interest almost disappeared."

1.1.5 Applications of AI in Various Domains

AI has transcended the world of research and landed in daily life and a growing number of industries:

• Healthcare:

o Early detection of disease (AI scans for cancer, diabetes, heart disease).

o Robotic surgical platforms (e.g., Da Vinci robot).

o AI generated virtual drug discovery.

• Finance:

o Real-time transaction monitoring fraud detection systems.

o Automated traders that speculate on the movements of stock prices.

o Virtual Bank tellers for customer inquiries.

• Education:

o Adaptive learning technology that provides content at the pace of each student.

o Auto-grading of homework and quizzes.

o AI tutors such as Duolingo aiding in self-learning.

• Transportation:

o Self-driving cars (Tesla Autopilot, Waymo).

o AI-based GPS for traffic optimization.

o Predictive maintenance for aeroplanes and trains.

• Retail & E-Commerce:

o Personalized shopping recommendations (Amazon).

o AI-driven customer service chatbots.

o Inventory and supply chain management.

• Agriculture:

o AI-driven drones to monitor crops and spot pests.

o Predictive models of weather and the health of soil.

Automated harvesting systems.

• Daily Life:

o Voice assistants (Alexa, Google Assistant).

o AI-powered appliances in smart homes.

o Social media feed customization (Instagram, TikTok).

The Global Impact: AI is transforming industries, jobs, social interaction — opening doors and simultaneously

such that they present ethical, legal and economic conundrums.

1.2 Importance of Ethics in AI

Artificial Intelligence has immense potential to transform industries and societies. However, without ethical considerations, AI can also cause harm—through bias, privacy violations, or

lack of accountability. Ethics in AI ensures that technological progress aligns with human values, fairness, and social good.

1.2.1 Role of Ethics in Technology and Society

Why Ethics Matters:

Technology is not neutral. “These are collective decisions we’re making about the kind of lives people live, where they work and how they come together.”

Ethics makes sure technology works for us, not against us.

• AI and Society:

Safeguarding Human Rights – AI ought to honour the human spirit, individual privacy and inalienable right.

Enhancing Trust – People are more aptto trust in AI systems that they perceive to be fair and

safe.

Shaping Policy and Law – Ethics informs legislation to prevent abuse.

• Example: : In healthcare, an AI system "must not be designed solely to drive a gain in accuracy, but also.

guaranteeing that patients are treated equitably irrespective of age, race or sex.

1.2.2 Ethical Dilemmas in AI Implementation

AI frequently leads to hard choices in which values conflict. These are called ethical dilemmas.

• Examples of Dilemmas:

Self-Driving Cars: What should a self-driving car with passengers do when it crashes?

pedestrians in unavoidable accidents?

Automation in Employment: If companies focus on efficiency through automation using AI,

even if it means moving thousands of jobs?

AI in the Military: Should we let AI decide to turn autonomous weapons loose?

• Why it’s a challenge: Unassailably right The path to resolving ethical conflicts is strewn with absolutist thinking, tweet-ready slogans and imperatives that are easy to assert but hard to follow.

or correct—there are competing values that need to be balanced.

1.2.3 Bias and Fairness in AI Systems

AI is only as biased as the data it is trained on. Inputs that are biased give rise to biased results in the system

biased.

• Types of Bias:

Bias in historic Data: Data captures previously present inequalities (e.g., less women in STEM).

Sampling Bias: When data does not reflect all groups equally.

Algorithmic Bias: The process by which algorithms are engineered exacerbates unfair results.

• Consequences:

o Hiring systems that favor men over women.

o Facial recognition not working well for people with darker skin.

o Algorithms for loan approvals that penalize some neighborhoods.

• Solutions:

o Use diverse datasets.

o Periodic AI system Check Ups.

o Transparent algorithms with explainability.

Example: Amazon abandoned an AI recruitment tool after it proved biased against women candidates as

historical hiring data favored men.

1.2.4 Privacy and Surveillance Issues

AI is reliant on huge troves of data—yet that raises the issue of how, exactly, those data are gathered and stored.

used.

• Privacy Risks:

o Personal data being used without authorisation.

o Data breaches. (Re: Information security.)

o No signed consent for collection of data.

• Surveillance Risks:

o Governments and businesses employing AI for mass surveillance.

o Using facial recognition in public to infringe personal freedom.

o Predictive policing that might unfairly focus on minority communities.

• Example: China's deployment of AI-enhanced facial recognition as part of public surveillance has prompted international

debates on privacy versus security.

1.2.5 Accountability and Responsibility in AI Systems

When can AI cause harm and who should bear the responsibility — the company, developer or ethicist?

machine? It is one of the great moral questions of our time.

• Accountability Challenges:

In many cases AI is a "black box" (decisions can't be easily explained).

There are three main players (providers, developers and users).

Current laws are not well-equipped to address harm caused by AI.

• Principles of Responsibility:

o Human Oversight: Humans must be "in the loop" for crucial decisions.

o Transparency: Systems that use AI should clearly disclose how decisions are reached.

o Regulation: Governments are setting up regulations such as the EU AI Act to0123456789(*,-.(/^\[XL''A&3!4C` If you are reading a PDF, please follow this link to read more about our work with visionaries like you.

accountability.

• Example: An autonomous car gets into a crash; the blame could rest on the car

of the user—the manufacturer, the software developer, or the user—underscoring demands for clearer

accountability frameworks.

1.3 Key Ethical Theories

Ethical theories provide frameworks for evaluating decisions in AI. They help us answer questions like: Is an AI system fair? Who benefits from it? Does it respect rights? By applying these theories, developers and policymakers can guide AI toward responsible and just outcomes.

1.3.1 Utilitarianism: Greatest Good Principle

Definition:

Utilitarianism, a philosophy based on the utility principle by scientists Jeremy Bentham, who claims that the

good action is that which brings about the greatest happiness (or the least misery) for the greatest main reasons: It keeps consistently close contact with practical morality.

greatest number of people.

• Principle in AI:

o AI systems should be engineered for the pursuit of broad societal objectives.

o Decisions are based on results, not intentions.

• Applications:

o Self-driving cars, reasoning about minimizing the total injuries in case of an accident.

o Public health AI optimizing distribution of vaccines to maximize number of lives saved.

• Criticism:

o Might oppress the minority to benefit the majority.

o Example: An AI might refuse expensive treatment to a few if it helps the majority,

raising fairness concerns.

“Activity”

Instruction to Students: 1. Imagine you are designing an AI system for allocating limited ventilators during a pandemic. 2. Apply the utilitarian principle (greatest good for the greatest number): o List three possible allocation strategies (e.g., first-come-first-serve, prioritizing younger patients, prioritizing survival chances). o Evaluate which option saves the most lives overall. 3. Write a 200-word analysis explaining which option you chose, how it reflects utilitarian ethics, and what trade-offs it involves.

1.3.2 Deontology: Duty-Based Ethics

Definition:

Other deontological theories are Immanuel Kant's divine command theory, which makes the assertion that actions are morally right if they do_subject only_passive yes a_priori to_principle what is_essence or_whether or_mean.

supposed universal moral laws or obligations, whatever the effect.

• Principle in AI:

o An AI agent needs to adhere to ethical obligations such as honesty, fairness and respect for human dignity although it may not be.

outcomes are less “efficient.”

o The end does not justify the means.

• Applications:

o AI in hiring should look at every candidate the same- level, not on gender, race or

age—perhaps even biased data notwithstanding.

o Privacy-aware AI design guarantees respect for user data, but not if this can be achieved violating privacy.

make systems “more accurate.”

• Criticism:

o At times dogmatic, failing to account for real-world vagaries in which rules might contradict one another.

1.3.3 Virtue Ethics: Moral Character Approach

Definition:

Virtue ethics, based on Aristotle s moral philosophy places focus on the character and motivation of the.

decision-maker and without reference to hard rules or consequences. The emphasis is on being a “good person” (or

designing AI with “good values”).

• Principle in AI:

o AI should be virtuous — it should have virtues: honesty, fairness, empathy and accountability.

o Design processes should train ethical virtues in developers and organizations.

• Applications:

o Compassion in the empathy that physicians bring to healing.

o AI chatbots built to be immune from manipulation and misinformation: The virtue of honesty.

• Criticism:

o Does not provide clear guidelines to solve difficult health care ethics cases.

o Relies too much on subjective judgments of what is “virtuous.”

1.3.4 Rights-Based Ethics

Definition:

a Rights based moral does not we get fundamental rights (the right to life, liberty, privacy etc.

equity) to be upheld in all decisions.

• Principle in AI:

o AI should never infringe on human rights, even if the majority or efficiency would be served by doing so.

o Privacy, free expression and non-discrimination are also features.

• Applications:

o Banning surveillance AI systems that infringe privacy rights, no matter how good they are at enhancing security.

o AI in workplaces should preserve workers’ rights; ensuring just and dignified.

transparency.

• Criticism:

o Can lead to conflicts when rights are in tension (e.g., right to privacy vs. right to security).

1.4 Emerging Ethical Challenges in AI

As AI systems become more powerful and widespread, they introduce new ethical dilemmas that affect

society, economies, and individual rights. These challenges require balancing innovation with fairness,

responsibility, and global regulations.

1.4.1 Autonomous Systems and Moral Decision-Making

The Challenge:

Self-driving cars, military drones and robotic surgeons may need to make decisions on the fly.

decisions with moral consequences.

o Who determines the “correct” response when faced with life and death situations?

o Should we weight to machines efficiency, safety, or fairness?

• Example: The “trolley problem” for AI: Should a self-driving car swerve to avoid five

pedestrian when it puts its passenger at risk?

• Ethical Concern:

o Risk of off-loading life-and-death decisions to machines.

o Opacity in how algorithms balance moral trade-offs.

• Ashaumpasir: An ethical and legal framework to direct AI decision-making in high-stakes contexts.

1.4.2 AI in Employment and Economic Displacement

The Challenge:

AI automates away these kinds of repetitive, routine work which is creating fears about the potential loss of millions of jobs.

losses.

o Blue-collar jobs (manufacturing, logistics) and white-collar ones (accounting, legal research).

are at risk.

• Example:

o AI chatbots stepping in for human customer service reps.

o Robot operated automated warehouses eliminating jobs for humans.

• Ethical Concern:

o Inequality: Benefits of AI accrue to corporations and tech elites.

o Displacement: Tens of millions could lose their careers without paths to being retrained.

• Need:

o Governments and business have to make reskilling programmes happen.

o Focus on the augmentation of AI roles, not their wholesale replacement.

1.4.3 Deepfakes and Misinformation

The Challenge:

AI-based deepfake technology generates realistic-sounding but untruthful audio, video or images.

truth from lies or lie-ness?

• Example:

o Fake political speeches during elections spread.

o AI-generated celebrity videos or frauds committed with the help of false voices.

• Ethical Concern:

o Democracy at risk from information.

o Harm to reputation and character of the victim and consent.

o Fraud and blackmail possibilities.

• Need:

o Powerful authentication solutions (digital watermarks, discovery AI).

o Public awareness and digital literacy for detecting manipulated content.

Did You Know?

“In 2018, the world’s first AI-generated fake video of former U.S. President Obama was released as a warning about deepfakes. Since then, experts warn that deepfake technology could become one of the biggest threats to democracy during elections.”

1.4.4 Ethical Implications of Generative AI Tools

The Challenge:

Generative AI tools (like ChatGPT, DALL·E, MidJourney) that can generate text, art, music and code. While

innovative, it raises ethical concerns.

• Issues:

Copyright: If I used a copyrighted work, even if for educational purposes and without permission.

Authenticity: It is becoming increasingly difficult to distinguish between work created by humans or machines.

Disinformation: Generative AI can be used for plausible (but false) articles.

Job Disruption: Artists, writers and coders are afraid of losing livelihoods.

• Example: News organizations arguing whether AI-written stories should run without obvious human intervention.

labeling.

Ethics: Creativity innovation and imitation vs ownership authenticity and piracy

fair labor practices.

1.4.5 Regulation and Policy Frameworks

The Challenge:

The government is not moving as quickly as AI. Policies must balance innovation, safety,

and ethics.

• Global Developments:

o European Union AI Act: Seeks to regulate high-risk AI uses (e.g., health, justice

enforcement).

o UNESCO AI Ethics Recommendation (2021): Draft global standard on fairness,

transparency, and inclusivity.

o U.S. AI Bill of Rights (2022): Principles for privacy, fairness, and accountability)ágenes_vídeoLOODjavagehuc502*granadalooprow (images_video_loodjavagehuc502 * granadalooprow)./

• Ethical Concern:

When there is no worldwide agreement, the bar is kept low.

o Risk of 'AI colonialism' where powerful countries and companies take over — Downside: The world's richest nations set standards for the 90 percent of humanity that is not them.

• Need:

o Global cooperation on the governance of AI.

o Clear liability, data protection and cross-border application policies.

1.5 Ethical AI Design Principles

Ethical AI design principles provide guidelines for developing AI systems responsibly. They ensure that

innovation benefits individuals and society without causing harm. These principles are not only technical

standards but also moral commitments embedded in the lifecycle of AI—from data collection to

deployment.

1.5.1 Transparency and Explainability

Transparency:

AI systems should explain themselves, tell us how they work and what data they use, and make that process transparent to people.

• Explainability:

Users need to be able to comprehend why an AI system made a particular decision.

• Why it matters:

o Creates a level of trust between users and systems.

o Assists in the identification and correction of mistakes or biases.

o It holds people accountable when something goes wrong.

• Example: When making a credit scoring decision, an AI system should not just respond "loan rejected" but explain why the TODO really confident about the score.

factors (such as low income, bad credit history) in simple non technical terms.

"Activity"

Instruction to Students:

1. Select a real-life AI system (e.g., Google Translate, ChatGPT, or a credit scoring app).

2. Research how transparent the system is:

o Does it explain how it makes decisions?

o Do users know what data it uses?

o Are its limitations communicated clearly?

3. Prepare a 1-page report or infographic:

o Describe the system briefly.

o Highlight strengths and weaknesses in transparency.

o Suggest two improvements to make the system more explainable to users.

1.5.2 Inclusiveness and Non-Discrimination

Principle:

Design AI to for the common good It's important to design AIs that will serve a variety of people equally and not exclude or disadvantage one over another.

group.

• Practices:

o Utilize a variety of datasets to avoid biased results.

o Test across gender, race, age and disability categories.

o Provide disabled services as appropriate.

• Example: Microsoft's Seeing AI app, which tells the blind and visually impaired what is in front of them, reveals

inclusiveness in design.

• Why it matters:

Fair AI leads to equal opportunities rather than reinforces stereotypes and serves social justice.

1.5.3 Human-Centered Design

Principle:

AI should enhance human capabilities, rather than replace humans or control them. People must remain

central in decision-making loops.

• Key Features:

o Designing for the user, safety and success.

o Keeping humans "in the loop"–mounted decisions (e.g., healthcare, law en o r cements).

o Developing human creativity, judgment and well-being.

• Example: In medical A.I., algorithms might recommend a diagnosis, but doctors make the final call.

Thus human empathy and expertise will continue to be crucial.

1.5.4 Accountability Mechanisms

Principle:

AI systems need to have very clear lines of responsibility for their outcomes.

• Practices:

o Determining responsibility for AI Errors: Developers, companies and regulatory agencies must determine who is responsible for AI errors.

o Generate audit trails to track how A.I. reached a decision.

o Establish system for complaints and redressal of affected AI users.

• Example: The EU's forthcoming AI Act categorizes AI systems for their risk and assigns development and usage accountability measures for developers and users.

• Why it matters:

Absent some kind of accountability, damaging consequences might deny a victim justice and undermine faith in AI.

1.5.5 Sustainable and Responsible AI Use

Principle:

AI needs to be defined and used in a manner that maintains environmental, social, and future well-being."

generations.

• Dimensions of Responsible Use:

Eco Benefit: Dramatically shrink the enormous carbon size of big AI models.

safety.You don't want to invest on long-term technology with misuse potential (e.g., autonomous

weapons).

Social Responsibility: Advocate AI for social good – healthcare, education, disaster response.

• Examples:

o Google's AI saving energy in data centers by fine-tuning cooling systems.

o AI tools tracking the impacts of climate change and assisting in sustainable farming.

• Why it matters:

But ethical AI is not only about being fair to people today — it's also about leaving a world for

tomorrow.

Knowledge Check 1

Choose the correct option:

1. Who is credited with coining the term Artificial Intelligence in 1956?

A) Alan Turing

B) John McCarthy

C) Marvin Minsky

D) Herbert Simon

2. Which ethical theory focuses on maximizing overall happiness or well-being?

A) Deontology

B) Utilitarianism

C) Virtue Ethics

D) Rights-Based Ethics

3. Which of the following is a major ethical risk associated with deepfakes?

A) Improved video quality

B) Enhanced language translation

C) Misinformation and reputational harm

D) Better entertainment options

4. In ethical AI design, transparency and explainability mean:

A) Hiding how algorithms work to protect trade secrets

B) Providing clear reasons for AI decisions in understandable language

C) Ensuring AI systems are invisible to users

D) Making AI decisions randomly to avoid bias

5. Which global regulation specifically aims to categorize and govern high-risk AI applications?

A) WCAG Guidelines

B) EU AI Act

C) ADA (Americans with Disabilities Act)

D) Kyoto Protocol

1.6 Summary

This chapter has given a brief overview of AI and the associated ethical considerations. It

started by AI definitions, characteristics, (Narrow/Ai narrow /General Nature of Ai) types,2008 was a significant date, warren mc cracken father of Ai will be more famous in its definition than mother Turing's dating 1955 Development history and so on.

and diverse applications. The chapter then stressed the significance of ethics in AI and discussed the following: The recommendations about ئاسرلش تاوقلا لئ دعزيد ة;ىشىموىن ىها-ع a) Ethics were presented.

challenges, bias, fairness, privacy and accountability.

❖ Various moral theories—consequentialism, deontology, virtue ethics, and rights-based morality—were

put forward as appropriate systems for assessing the moral implications of AI. We also examined emerging ethical

who's in charge of overseeing), we must also grapple with a new set of issues, such as robots making their own moral decisions, job displacement and deepfakes and how to regulate the technologies that create them?

generative AI tools. Finally, the chapter introduced ethical AI design principles emphasizing

transparency, inclusion, human centric design, accountability and sustainability."

❖ These perspectives together arm students to critically evaluate AI systems as more than just technologies, but as

social and ethical forces that will influence the future character of humanity.

1.7 Key Terms

AI (Artificial Intelligence): The development of a computer system to perform tasks that normally require human intelligence.

Machine Learning (ML): AI method in which system learn from data.

Narrow AI: AI for a specific task (e.g., Siri, Google Translate).

Strong AI: This is a hypothetical form of AI where it matches or surpasses human intelligence.

Super AI: Speculative AI beyond human intelligence.

Ethical Dilemmas: The ethical predicament in which values compete when making decisions.

Bias in AI: Systematic unfairness observed in outputs from AI because of biased data or bias in design.

Privacy: The right to manage one's private data and information.

Transparency: How transparent the AI system is to users.

Interpretability: The capability to understand and provide an explanation for why AI made a decision.

Fairness: How to architect AI that benefits all types of user.

Accountability: A clear chain of responsibility for the outcomes produced by AI.

Utilitarianism: Ethical doctrine of the greatest happiness (or good) for the greatest number.

Deontology: The ethics of duty focusing on moral laws.

Virtue Ethics: An ethical approach in which the emphasis is placed on the moral character of those who make decisions.

Rights-Based Ethics: Emphasis on safeguarding basic human rights.

Deepfakes: AI-generated counterfeit audio, images or video.

Generative AI: An AI system that produces text, art, code or media.

Responsible AI: Creating AI with the least harm to the environment and society.

Regulation regarding AI: Legal framework agreements for safe and ethical use of AI.

1.8 Descriptive Questions

Write the meaning of Artificial Intelligence and describe its Key features with any one example.

Distinguish between Narrow AI, General AI and Super AI.

Discuss the history of artificial intelligence from its inception in 1950s up to today.

What is the importance of ethics when it comes to AI and tech in general?

OU: What are some of the ethical challenges in adoption of AI?

Discuss how Bias comes into AI systems and how it can be reduced.

What are the privacy and surveillance implications of using AI?

Examine the accountability problem with AI systems — who should be held accountable for a mistake?

Compare and contrast between utilitarianism, deontology, virtue ethics, and rights-based ethics in decisions made by AI.

Describe the ethical problems posed by self-driving cars and other autonomous technologies.

What is the economic effect of AI on employment and displacement?

In what ways do deepfakes and other generative AI tools pose new ethical challenges?

Why is AI regulation needed and what approaches are being taken around the world?

Summarize key principles of responsible AI design including illustrating the material with examples from current practice.

Illustrate with a real world application, how ethical design may lead to trust in AI systems.

1.9 References

1. McCarthy, J. (1956). Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

2. Russell, S., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach (3rd ed.). Pearson.

3. Floridi, L. (2013). The Ethics of Information. Oxford University Press.

4. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine

Intelligence, 1(9), 389–399.

5. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

6. European Commission. (2021). Proposal for a Regulation Laying Down Harmonised Rules on Artificial

Intelligence (AI Act).

7. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence.

8. IEEE. (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous

and Intelligent Systems.

Answers to Knowledge Check

Knowledge Check 1

1. B) John McCarthy

2. B) Utilitarianism

3. C) Misinformation and reputational harm

4. B) Providing clear reasons for AI decisions in understandable language

5. B) EU AI Act

1.10 Case Study

Facial Recognition Technology – Innovation vs. Ethical Responsibility

Introduction

Facial recognition technology (FRT) is one of the most contentious AI applications.

And from opening phones to the surveillance of public places, it shows how AI can

transform convenience and security. But it also raises big ethical issues around privacy,

bias, and accountability.

Background

• FRT is being eagerly embraced by tech companies and governments.

• It is being utilized in airports, retailers and law enforcement agencies, as well as in consumer devices.

• But for all the good that it does, research shows that race, gender and age can influence accuracy

groups, raising fairness concerns.

Issue 1: Recognition Accuracy Bias

MIT and NIST research showed FRT is also less accurate for women, as well as those with darker

skin tones.

Solution: Companies need to use a broader spectrum of training data and perform regular fairness audits.

MCQ:

Why do facial recognition systems exhibit bias so often?

A) Algorithms are designed to by definition be neutral.

B) Training data lacks diversity

C) Not all skin tones are captured by cameras

D) The system is harvested by human operators

Problem 2: Privacy and Mass Surveillance

When governments deploy FRT for surveillance, it is likely that they will breach individual privacy and liberty.

Solution: Enforce policies around consent, retention and open use.

MCQ:

What's the main problem/fear about widespread use of FR?

A) Convenience in unlocking devices

B) High storage costs

C) Invasion of privacy with generalised surveillance

D) Faster airport check-ins

Problem 3: The Line of Accountability for Misidentification

There have been FRT mistakes that resulted in wrongful apprehension, particularly among police.

Solution: Adopt risk and shared responsibility mechanisms where firms and coordinates share the costs.

for AI mistakes.

MCQ:

If an AI leads to sending someone to jail who would not have been sent there had the matter been left in human hands, what is the primary ethical issue?

A) Can the victim work on some other technology?

B) Whose fault is it—the software writer, the government or the operator?

C) Is FRT evil everywhere?

D) Can AI repair itself without monitoring?

Conclusion

The topic of facial recognition exemplifies the double sided aspect of AI—its capacity to make our lives more convenient.

and security and have serious ethical implications. Through the principles of fairness, transpa\\\'ency and

accountability, to take full advantage of FRT. The fallout also highlights the importance of ethical

frameworks and global standards to strike the balance between innovation and rights.

Ethics in Artificial Intelligence_V3_Unit 2.docx

Ethics in Artificial Intelligence_MBA_2

Ethics in Artificial Intelligence_MBA_2

ATLAS SkillTech University

Document Details

Submission ID
trn:oid:::3618:127350328

Submission Date
Feb 2, 2026, 11:32 AM GMT+5:30

Download Date
Feb 2, 2026, 1:06 PM GMT+5:30

File Name
Ethics in Artificial Intelligence_V3_Unit 2.docx

File Size
73.4 KB

45 Pages

9,646 Words

68,082 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 2: Ethical Theories and AI

Learning Outcomes

1. Understand the fundamental principles behind major ethical theories such as utilitarianism,

deontology, and virtue ethics.

2. Analyze how different ethical frameworks apply to real-world dilemmas, especially in the context

of artificial intelligence (AI).

3. Evaluate ethical challenges posed by AI systems using multiple philosophical perspectives.

4. Differentiate between various ethical theories and recognize their strengths and limitations in

practice.

5. Apply ethical reasoning to case studies involving AI technologies in diverse sectors.

6. Critically assess the social, legal, and moral implications of AI deployment.

7. Familiarize with key terms, case studies, and descriptive questions to reinforce ethical

understanding in AI contexts.

Content

2.11 References

2.12 Case Study

2.0 Introductory Caselet

"The Machine That Learned to Choose: A Dialogue Between Mira and the Engineer"

Background:

Mira, a philosophy student at Delhi University was making a visit to her uncle and senior AI engineer in town at an.

research lab in Bengaluru. She is reluctant but skeptical of the precipitate ascent of artificial intelligence, and its

impact on society.

One afternoon, she watches a prototype chatbot answer questions about mental health. Impressed at first,

she eventually realises that its answers are emotionally empty and, in one case, callous to the point of danger.

Worried, she asks her uncle:

“Is a machine capable of understanding what’s right for someone in pain?”

Her uncle pauses and replies,

“Machines can learn patterns but not values. That’s where we come in. Each line of code, every training

dataset is human-chosen for good, bad ethical benefits or whatever.”

Over the course of several days, Mira has deep conversations with AI developers, ethicists and psychologists

at the lab. She understands that it’s not only a matter of what AI is able to do but what it also should do—and who.

decides.

She returns to Delhi, and starts work on her thesis — but it is not one about technology, rather, it’s an essay on artificial ethics of public behaviour.

where we argue over what is right and wrong.”

what's right?

Critical Thinking Question:

As artificial intelligence gains more of an influence in the world, who should make those decisions?

ethical for machines to behave — and how should those ethical parameters be established?

2.1 Overview of Ethical Theories

(government, businesses and individuals) can make the right / good decision. – or bad for themselves, other people and nature resulting in appreciation of right/good attitudes (satisfaction/ are better off), reduction of wrong/bad behaviour (pleasure) -3- HOW TO MAKE RIGHT OR GOOD DECISION?

or bad, just or unjust. These theories give direction to making ethical decisions, i.e. they offer or suggest systematical ways of thought processes.

about values, duties, and consequences. In today's world of convolution, not only in CS but also other fields like technology,

increase in writing on ethics, morality, robots, are deep wells? and artificial intelligence and digital behavior—comprehending ethical theories is needed for ethical.

action and decision-making.

Moral theories provide the underpinnings for determining right and wrong and directing human behavior in both

personal and professional life. These theories have developed through centuries of philosophical query

and continue to influence the way in which we evaluate matters related to fairness, justice, human rights, and responsibility.

This chapter presents the concept of ethics, classifies ethical theories, and examinesfillType Size: and ethical decision making.

what makes them so relevant in present-day technology.

2.1.1 Introduction to Ethics and Moral Philosophy

Ethics (or moral philosophy) is the branch of philosophy that includes how we ought to live, and what Is right and wrong.

to be morally right or wrong, good or bad, just or unjust. It discusses how we ought to behave and what

Type of life they are to live. Values and principles, and the reasons forETHICS is the study of conduct.

decisions.

Moral philosophy has three fundamental divisions:

• Meta-ethics: Deals with the IDs of ethical properties, statements and judgements. It asks

questions such as “What is involved in our saying something is good?”

• Normative ethics: Investigates criteria of right or wrong conduct. It asks, “What

should I do?”

• Applied ethics: Using ethical theories and principles to real-life situations, such as those in medicine

Company Decisions, Conduct or Technology.

Ethics is not the same as laws, or social conventions. Laws are made and enforced by governments, and

customs are shaped by society. Though ethics itself grows out of more fundamental moral principles thatlideedmayfsometirnesbeooiietsc.

challenge existing laws or traditions.

2.1.2 Classification of Ethical Theories

There are many such broad ways of thinking about ethics, and different theories in these branches have different methods for deciding which things are right (or wrong).

human actions:

Consequentialist Theories

These are theories according to which the rightness or wrongness of an act is determined by its consequences. If the outcome is good,

morally justified by the action.

o The best known is Utilitarianism, which advocates actions that yield the greatest amount of pesticidebenefits overcosts.

overall happiness or well-being.

Deontological Theories

These theories deal with duties, rights and obligations. Options and Hedonism An action is right if it conforms to a moral rule.

, or the consequences be damned.

] Emmanual Kant's Ethics provides, for example, that moral obligations are universal in the sense of being.. necessarily true and research illustrates how such principles are indeed Universal.

followed consistently.

Virtue Ethics

Virtue ethics not only does not concentrate on actions or results, but focuses on the character of theRPJU collegiate player.

person performing the action.

o It wonders, "Who should I be?" other than, "What am I going to do?"

Relativist Theories

These imply that good and bad is relative to cultural, social or individual criteria.

There are no absolute morals; morals are subjective.

Feminist and Care Ethics

These theories focus instead on relationships, caring, empathy and the social nature of moral decisions."

frequently pointing out that traditional theories do not take into account emotion and connection.

Every ethics has his own view and is a good choice to analyze certain type of moral.

dilemmas. In application, these theories are frequently intermingled or weighed against each other as appropriate.

2.1.3 Relevance of Ethics in Technological Contexts

In the age of quickly advancing technology, particularly in the fields of artificial intelligence, data science, robotics and

digital communication, new ethical problematic ones arises that must be taken into account of by traditional theories.

For example:

• Should self-driving cars be taught to sacrifice their passengers' lives to save many other people?

• Should AI systems be making hiring decisions?

How should companies treat user data and privacy in the digital era?

Technology can contribute a lot to the society but at the same time it can also be harmful.

moral theories aid in determining:

• Bias and fairness in algorithms

• Privacy and consent in data gathering

• Responsibility and accountability for autonomous decisions

• Inclusivity and access to technology

The moral choices for engineers, developers, and policymakers developing and using

technology. We have ethical frameworks to help us to think critically and act responsibly in these circumstances.

Comprehension of these theories is indispensable, not for the philosopher alone but for all who are engaged in or burdened by.

by modern technologies.

2.2 Utilitarianism

Utilitarianism is the foundation of many popular and influential social decision-making systems, particularly in the modern world.

such as commerce, politics, science and medicine. At bottom, utilitarianism is a kind of consequentialist theory.

I.e., it is a kind of consequentialism, that is to say, morality is evaluated based on the outcome or result.

The idea is fundamentally simple: the moral rightness of an action is determined by whether it will generate more good for the greatest number.

the greatest number of people. This "good" is frequently understood in terms of happiness, welfare or pleasure.

2.2.1 Basic Principles of Utilitarianism

Utilitarianism has a few central components:

Principle of Utility: An action is right if it produces the most happiness or the least pain.

quantity of misery among the most men.

Consequentialism: The rightness or wrongness of an act is contingent solely upon the outcome, and not its character

the action or the proposer of it.

Impartiality: Everyone's happiness counts equally. the happiness of no one.

anyone else's.

The philosophy of utilitarianism originates from philosophers such as Jeremy Bentham and John Mill.

Bentham was an advocate of the pursuit of pleasure and came up with a technique to compare them, known as

what Mill called the "hedonic calculus" to compare the results.

• Mill polished the theory by stressing the quality of happiness, not merely its quantity.

Utilitarianism is appealing as it presents a rational, systematic method of making decisions by concentrating on

outcomes and overall benefit.

2.2.2 Act vs. Rule Utilitarianism

There are two broad varieties of utilitarianism, and they vary with respect to how the principle of utility is applied:

Act Utilitarianism

o This form exists for you to focus on every single thing and say: o "Does this one little action bring me …?

greatest happiness of the greatest number?"

o Each case is evaluated individually.

o If, for instance, in a given context it is more pleasurable to lie than to tell the truth, then

you should lie in such situations.

Rule Utilitarianism

o This one considers actions based on rules that, if applied uniformly would produce the_BOUNDSRES function.

greatest good.

o It inquires, "If practically everyone tries to follow this rule, would the world generally be a better place?"

o So, while lying may seem to lead to short-term pleasure, the generalization "always tell the truth"

could build more trust and happiness over time.

Key Difference:

• Act utilitarianism is in the business of judging acts one by one.

• Rule utilitarianism is concerned with general rules of conduct.

2.2.3 Applications in AI Decision-Making

UTILITARIANISM IN ARTIFICIAL ETHICS Utilitarianism is an influential idea on how ethical decisions work in AI and

automated systems. As AI often concerns prediction and optimization, it is well aligned with

utilitarian thinking.

Examples of utilitarianism in AI:

Autonomous Vehicles:

When a self-driving car needs to make a decisionconcerning which path to takeCfor example, crash into a wall and thus potentially harming the passenger or .

kill pedestrians), a utilitarian algorithm would simply minimize the total harm done.

• Healthcare Algorithms:

The AI systems deployed in hospitals could be more likely to prioritize treatments according the sites where they could save the most incrementally.

lives or increasing recovery rates — certainly not from individual preference or emotion.

• Resource Allocation:

In cases like disaster relief or public health, it could be used to help determine where to send resources

by predicting where they will generate the highest average benefit.

• Content Recommendation Systems:

Algorithms in social media or video platforms could be developing for user engagement and.

satisfaction, with the goal of maximizing the expectation of average “pleasure” or utility of users.

It is a utilitarian calculation to aid developers in building systems that prioritize for the best possible outcome. However, these

systems can also end up abandoning individual's rights or fairness for utility on the whole.

2.2.4 Limitations and Critiques

Although it is beneficial and rational, utilitarianism invites much criticism. Some of the most common complaints are:

Ignores Individual Rights

o It may seem permissible to the utilitarianism for things that are bad for a few people but good for many.

o So for instance, sacrificing 1 person to save 5 may be considered acceptable even if it

violates that person's rights.

Difficult to Predict Consequences

o It may be difficult, perhaps even impossible, to know all the results of an action in advance.

o This means that decision-making based on outcomes is unpredictable.

Happiness is Subjective

o People have different understandings of happiness and well-being.

o. What is good for one person can be bad for another, making it impossible to measure

or compare.

Tyranny of the Majority

o Sometimes, however, utilitarianism may endorse the will of most and permit the needs of the few.

minorities.

o These can have injustices or discriminatory effects, even if they're helpful on balance.

Moral Integrity

o Critics maintain that utilitarianism can demand actions that conflict with our moral …

intuitions, like falsehood, cheating or killing if it benefits.

Nevertheless, utilitarianism is still an influential and highly used ethical theory in some areas.

that demand data-driven and evidence-based decisions, like AI and public policy.

2.3 Deontological Ethics

Deontological ethics is a theory of morality that's connected to responsibility, obligations, and the moralistic sense of right and wrong.

-or goodness and badness of actions -not the results of the action in question. The word "deontology"

derives from the Greek deon, or “duty.”

Unlike utilitarianism, which determines an action’s morality based on the impact it has, deontological ethical systems maintain that there are reasons for actions other than that they produce good in the world.

are morally obligatory or impermissible, regardless of their consequences. It emphasizes doing what is

right and wrong because it is right, not because it makes good policy.

2.3.1 Kantian Ethics and Categorical Imperative

The German philosopher Immanuel Kant is the most well-known advocate of deontological ethics.

His morality is a system of reason, duty and moral law.

The cornerstone of Kant's philosophy is the Categorical Imperative. This is a principle that

guides people in figuring out what actions are morally right. Kant offered several versions of this

imperative. Two of those are:

Universal Law Formula

o “Act only according to that maxim by which you can at the same time will that it should become a universal law”

become a universal law.”

o This is also known as: Before taking any action, put yourself in everyone else's shoes and imagine what would happen if others did the same. If it would

result in a contradiction or chaos, it’s morally wrong.

Humanity Formula

o “Act so that you use humanity, whether in your own person or in that of any other man,.

other, ever for ourselves, never only as a means.”

o This is to say: To never use human beings as a means to an end. Respect their dignity and

autonomy.

Kant thought that morality was grounded on rational rules that could justify themselves to every human being at all times.

He believes lying to be absolutely wrong — even if it produced better results — because it violates a

universal moral law.

2.3.2 Duty and Moral Rules

Duty is at the center of deontological ethics. One has a moral duty or obligation to obey some moral rule,

regardless of the outcome. These duties might include:

• Always tell the truth

• Keep your promises

• Respect others' rights

• Do not harm innocent people

The above moral rules are treated as absolute or obligatory and as equally binding on all persons.

Just as if you promise to help a friend move, deontology says (all else equal) that you ought to help – even if something super fun comes up instead.

more enjoyable arises — since keeping promises is a moral obligation.

Deontologists argue that what makes our actions good is if they obey moral rules. Even if doing the

right thing has bad results, the action is still morally right.

"Activity: Design a Deontological Code for an AI Assistant"

Instruction to Students:

You are part of a development team creating an AI assistant for a hospital. Using deontological

ethics, draft a code of rules that the AI must follow—focusing on moral duties rather than outcomes.

1. Write 5 specific rules the AI should follow, such as "Never provide false medical information"

or "Always respect patient confidentiality."

2. For each rule, explain why it qualifies as a moral duty under Kantian ethics.

3. Briefly describe one real-life situation where following the rule might lead to a difficult

outcome, but is still the right thing to do according to deontology.

Deliverable: Submit your AI rulebook and ethical justification (200–300 words total).

2.3.3 AI and Rule-Based Systems

Deontological approaches can be seen in rule-based AI, where the operation of a system is defined.

participatory in nature, with fixed sets of rules or codes of conduct.

Applications include:

• Expert Systems: These process a “if-then” rules to make decisions in things like law, medicine, or

finance. They are systems driven by hard logic and take no prisoners.

• Compliance AI: Field in which medicinal drugs DEVENSCHY 150are created and used (WRESTLER)organize.

Fight fraud, discrimination, or any illegal activity.

• Ethical Guardrails: Some AI systems are designed to not take certain actions at any.

factors (e.g., weapons not targeting civilians).

These are examples of how we can force AI to behave in line with certain ethical criteria just like.

deontological ethics demands moral laws that must not be transgressed.

Where reliability, fairness, or rights-protection are imminent it is especially desirable to require the statisfyability of B.A.

in like court cases, security measures or the treatment of sensitive data.

2.3.4 Challenges in Applying Deontology to AI

While deontological ethics is well adapted to rule-based programming, there are a number of challenges in

directly applying the previous results to AI systems:

Rigidity of Rules

o Deontological systems are very poor at adapting to complex, unpredictable environments. Strictly

Loyalty Follow through with a successive paragraph: always harmful in the very rare instance, even when done for good ends.

o For instance, an AI that always tells the truth (in due course) might bring about by mistake p 19 a lot of harm in its agentsponsored plan.

harm by revealing private information.

Conflicting Duties

o In practice, obligations may collide. For example, the moral requirement to speak truthfullycom- petes with the duty of silence in relation to others.

duty to protect someone's life.

o Without being hard-wired into a system, -- AI system will easily on how to rank competing rules.

balancing them.

Lack of Contextual Understanding

o The AI is not human intelligence, with human-level grasp of context, mood and motive. It may rigidly follow

rules without acknowledging the exceptions.

For instance, in a system with punitive laws, you may punish someone tailing for for

morally justified reason (such as stealing medicine to save a life).

Moral Responsibility

o Deontological, where it imagines that a rational moral agent can understand and.

doing the right thing.

o AI is not a form of conscious or moral reasoning, so it remains an open question whether it can really be.

be held to moral duties.

Hardcoding Ethics

o Determining what rules to embed in AI systems is quite challenging. Moral rules are often

debated and are contingent on culture, making them difficult to codify into a set of universal morals.

Deontology provides a powerful grounding in rule-respect and rights protection, yet deontology also must be rethought

cautiously in developing intelligent systems.

2.4 Virtue Ethics

virtue ethicstheory of morality that makes virtue the central concern, rather than rule abstractionwho.

follow up (or harm) of their activities. Instead of “what should I do?” virtue ethics asks, "What

kind of person should I be?"

According to this theory, what matters is the individual morality such as: honesty, courage, compassion and.

eloquence — or good common sense, to use an old word that’s less pretentious than the classic bringer of wisdom to all Greeks and Dr. Goodword — and feels a good life is one spent building up these virtues and using them right. The

moral excellenc.e as the end of life.

2.4.1 Principles of Virtue Ethics

Virtue ethic has its roots in ancient Greek philosophy, most notably with the work of Aristotle. He introduced

the notion that good behavior is the result of developing good character traits, which he called virtues.

Central tenets of virtue ethics are:

Virtue as a Habit

o Virtue is not just about knowing what a good person would do but actually being (or becoming) the sort of person who does do so.

o Morality improves with practice, like a skill.

The Golden Mean

o Virtue is found at the "mean:" between too much and too little.

For instance, courage is a virtue; insufficiency of it, cowardice; excess of it, rashness.

recklessness.

Moral Education and Role Models

o People acquire positive character traits by watching and mimicking virtuous leaders.

o Character development is a lifelong progression that is influenced by education, experience and reflection.

Eudaimonia

o Aristotle used the word eudaimonia to refer to what value human life has.

“flourishing” or “living well.”

o Eudaimonia is to live a life of reason and virtue.

Virtue ethics, unlike many theories that deploy formulas or fixed rules, is contextual. It requires practical wisdom—

having an intuitive feel for what to do in a given situation.

2.4.2 Character and Moral Agents

Virtue ethics is the consideration of not just the action or its conse.personness.

consequences. The theory holds that:

• A virtuous person is someone who behaves with virtue, not because of an obedience to rules

or aiming for rewards.

Characteristic traits of moral agents include honesty, kindness, justness, patience and.

responsibility.

• Ethics reduces to personal character, where what the agent does is who the agent is.

In this perspective, ethics is involved with identity and relationships at the very core. How one acts in a

case represents their general state of soul."

This concern with personal development, motive and role relates virtue ethics particularly to 4.2 Ethical Simulation Games including.

any professions such as medicine, teaching and, last but not least, leadership for the moral character of its practitioners matters

greatly.

2.4.3 Virtue Ethics in AI Design

A VIRTUOUS APPROACH TO AI by Neil Article Applying virtue ethics to AI is a novel view. Instead of programming AI

to follow strict guidelines or generate outcomes virtue ethics is meant to push the designer to ask:

The Character of AI Developers

Ethical AI starts with good and honest developers who are really care about integrity, accountability, empathy, and

justice.

o The values and objectives of AI systems should be aligned with the moral values of humans.

AI as Moral Influencers

o AI systems, in particular, those that interface with people (e.g., chatbots, virtual assistants,

educational software), are able to change human actions.

o These systems must encourage and demonstrate virtuous behavior, such as patience, respect for.

fairness.

Human-Centered Design

o Virtue ethics emphasizes that technology ought to be constructed for the aid of human

flourishing, not merely efficiency or profit.

o Designers could ask: Does this AI help people be better, wiser, kinder?

Context-Sensitive Ethics

o Virtue ethics permits flexible and nuanced responses to complicated moral contexts,);">

can be beneficial in unforeseen AI such as dynamic problems.

o Rather than relying on hard-coded rules, AI could be trained to see what the user preferences.

context of a situation.

This refocuses debate from AI decision-making to the ethical culture of the people who create and utilize such technologies."

use AI systems.

2.4.4 Criticisms and Debates

For all its merits, virtue ethics is not without its problems and opponents, especially when it comes to

AI:

Lack of Clear Guidelines

o There are no definite rules or formulas in virtue ethics that one can apply to make decisions.

o Applications where fast and consistent decisions are needed, like autonomous vehicles or

medical AI, this is a major drawback.

Subjectivity and Cultural Variation

o Virtues may vary by culture and personality. Virtue is what one society considers virtuous,

another may not.

o This makes it hard to come up with a generic set of virtues that should apply to AI systems.

Not Easily Programmable

AI systems require explicit, programmable directions. Human good Virtue ethics is firmly based on a human view of the world.

experience, judgment and moral development — that are difficult to replicate in code.

Anthropomorphizing AI

o Some critics contend that applying virtue ethics to AI incorrectly presupposes that machines could have someone this concept.

moral sense, emotions, or intentions as human beings do.

o AI systems don't have consciousness or free will and so can never really be virtuous.

Responsibility Still Lies with Humans

o AI cannot have moral intentions, virtues, so virtue ethics may not be relevant.

to AI as moral agents.

o The attention should still be on human designers, users, and organizations rather than on AI itself.

Nevertheless, virtue ethics has something valuable to contribute in construct ing ethical values and norm s: especially its emphasis on certain character traits and tendencies.

objectives of AI development groups and companies.

2.5 Other Ethical Frameworks

Although utilitarianism, deontology, and virtue ethics are the major ethical theories, there exist several others.

several other major perspectives that offer other things to think about when we think morally. These

frameworks are especially valuable for navigating complex social, political, and relational issues—

such as those raised in the development and deployment of artificial intelligence (AI).

Each of these theories offers its own insight into moral reasoning, sometimes emphasizing aspects that the other overlooks.

concepts that traditional theories could have transferred, namely rights, relationships, agreements and cultural diversity.

2.5.1 Rights-Based Approaches

33 Rights-based ethics** concentrates on the concept that people have inherent rights, which should be respected and not violated.

respected no matter the consequences or circumstances. These rights are frequently viewed as inborn, absolutist

and universal.

Rights are mainly of two types:

Negative Rights: These safeguard people from interference (e.g., privacy rights, freedom of.

speech, or freedom from violence).

Positive Rights: Such rights involve the need of another to do or have something (e.g. a right to be fed, the right to appropriate aid, health care etc.).

education, healthcare, or social support).

Key ideas in rights-based ethics:

• Rights often have their basis in human dignity and moral desert.

• Right actions are those that do not violate or infringe on the rights of others.

• It is wrong to violate someone's rights, even if doing so results in good for others.

In the context of AI:

• Rights-based ethics includes rights to robust protections for data privacy, freedom from surveillance and

non-discrimination.

• It becomes problematic when AI systems violate rights, like deploying facial recognition, without.

consent or employment or credit decisions.

Rights-based approaches are related to legal structures and human rights conventions like the

Universal Declaration of Human Rights.

2.5.2 Ethics of Care and Relational Ethics

Nel nissenHolz.wikipediaEthics_of_care Italian Wikipedia Ethics of care is a feminist-informed ethical theory that centers its working model around the quality of and response to relationships.

compassion over laws or principles.

Key principles of care ethics:

• Morality is based on human connectivity and the needs of others.

• Moral judgements should be made with emotional understanding, empathy and contextual considerations.

• Instead of inquiring, “What is the right action?” care ethics poses the question "How can I answer for the needs of

others in a caring way?”

Relational ethics is an ethics of the moral significance of the relational, and of...leston: Univerity Press of Ida full dependentness empowered by whom we are related.

relational Christian anthropology— understanding individuals as open, non-self-possessive beings and interpreting the relations between human beings rather than treating the individual person as a distinct moral agent.

In the context of AI:

• Care ethics emphasizes the emotions of AI on its users (e.g., elderly care robots, mental health...)

chatbots).

• It encourages inclusive and sensitive design that puts well-being and social responsibility first.

• It critiques systems that reduce people to data points, ignoring their lived experiences and

vulnerabilities.

This strategy is particularly applicable to sectors such as health, education and social services where the trust of recipients of public services is vital.

and care are essential.

Did You Know?

“Did you know that the ethics of care—a major ethical theory in today’s AI ethics discussions—

was developed as a response to traditional moral theories that often overlooked relationships and

emotions? This approach gained prominence through the work of psychologist and ethicist Carol

Gilligan in the 1980s. Unlike utilitarianism or deontology, which emphasize universal rules or

outcomes, care ethics centers on empathy, vulnerability, and the context of human relationships. It’s

especially relevant when designing AI for caregiving, education,

nd mental health, where

understanding emotional needs and relational sensitivity is just as important as logic or fairness.”

2.5.3 Contractarianism and Social Contract Theory

Social contract theory, in philosophy, a view of the nature of morality and politics according to which moral and political obligations are dependent upon a contract or agreement among the people to form the society in which they live.

pacts between men to form a just and stable society.

Key ideas include:

• People agree to live by certain rules in return for safety, security and cooperation.

• Justice and fairness arise from consenting to something, not divine laws, or utilitarianism

calculations.

• A fair society is one that rational individuals would design when seeking to create rules under which they would be willing to live.

Central figures are Thomas Hobbes, John Locke, Jean-Jacques Rousseau and John Rawls.

In contemporary moral philosophy, John Rawls articulated the concept of a “veil of ignorance”: imagine creating a society without knowing whether you will be rich or poor.

without being aware of your place in it. That would promote fair, unbiased rules that safeguard all.

In the context of AI:

• Contractarian arguments justify regulatory structures and decision-making regimes, which guarantee that AI becomes a servant rather than an agent of humanity.

are constructed according to social contract.

• It stresses public accountability, transparency and fair participation in determining how

AI is used.

• It can form the foundation for AI ethics guidelines, codes of conduct and international AI policy

development.

It is through this lens that such an instrument brings a connection between personal morals and the collective obligations of society.

2.5.4 Pluralism and Contextual Ethics

Ethical pluralism is the idea that there are many right moral views. Instead, multiple moral

principles can be true and valuable, as circumstances allow.

Key features of pluralism:

• Different circumstances may require different moral calculations.

• Ethical deliberation sits in the tension between competing values — justice, care, utility and rights.

• There may not be a single “right” answer, but just a set of acceptable answers.

Contextual ethics expandsthis understanding by drawing attention to the significance of the particular contextssocial, cultural, economicin which moral decision-making occurs.

historical, and personal—in moral decision-making.

Instead of using abstract rules, contextual ethics wonders:

• What’s going on in this specific case?

• Who is affected, and how?

• What are the cultural or relational dynamics at play?

In the context of AI:

• Pluralism favors using multiple ethical frameworks when evaluating complex AI

systems.

•Contextual ethics promotes the development of AI that is responsive to cultural diversity, local difference, and.

social settings.

• For instance, an AI tool that's used in one country or community might have to embody different values

than one used elsewhere.

Such strategies prevent dogmatism and encourage flexible, inclusive and reflective ethical work.

2.6 Applying Ethical Theories to AI

As artificial intelligence systems play a larger role in people's lives — making decisions about health care, finances, and autonomous systems — it becomes more important to understand their ethics, one author said.

hiring, policing, credit scoring and so on — the ethical stakes are high. Applying ethical

AI as applied to AI is that it can help developers, policymakers, and users to understand the risks and assess the moral responsibilities.

and so make humanly justifiable design and deployment decisions.

In this section, various ethical theories (i.e. utilitarianism, deontology, virtue ethics and care ethic) are discussed in relation to the problem of sorting out what government policy should be on end-of-life decision making decisions.

ethics, and rights-based theories) that can inform the analysis, evaluation, and design of AI systems in the wild.

world contexts.

2.6.1 Ethical Analysis of AI Algorithms

Automated decision-making systems are based on AI algorithms. They process data, identify patterns,

and make predictions or recommendations. The ethical scrutiny these algorithms face leads to the question:

• Which assumptions are built into the logic of the algorithm?

• Who fares well, or badly, as a result of its decisions?

• What compromises are made in its design?

Using ethical theories:

• Utilitarianism asks whether the algorithm maximizes the total benefit or utility.

• Deontology inquires whether the algorithm respects duties and rules, such as honesty or non

discrimination.

• A virtue ethics approach asks whether the values and moral character of its creation and whose values are embedded in that?

encourages human flourishing.

• Rights and entitlements : Rights -based ethics considers whether the algorithm respects individuals right sescalable analysis of security protocols and provides formal assurances about provi- sionality that is similar to a property.

due process.

Ethical analysis further involves an evaluation of how open, explicable and controllable the algorithm is.

as they relate to the trust and accountability of the public.

2.6.2 Ethical Evaluation of Data Usage and Privacy

AI systems based on the machine learning and deep learning models need big data for operating. These databases usually contain sensitive private

data, including health history, financial information or behavioral habits. Ethical evaluation of data use

focuses on:

• Consent: Did the acquisition of data involve informed consent from the user?

• Purpose limitation: Are data used for only the stated purposes?

• Anonymity and Confidentiality: Are identities concealed?

• Information security: Data- Are stored and processed securely?

Ethical frameworks guide these concerns:

• Deontology highlights the obligation to protect personal autonomy and privacy.

• Rights-based theories protect individuals' rights to control their private data.

• Care ethics prioritizes the protection and promotion of vulnerable populations, as well as fostering trust in relationships.

• Utilitarianism balances the advantages of data application (e.g., enhancing public services) with potential931 harms.

injuries (e.g., identity theft or injury to reputation).

Striking a balance between innovation and ethical data management is instrumental for the creation of responsible AI systems.

2.6.3 Fairness and Accountability in AI

One of the most talked-about ethical challenges in A.I. is fairness and accountability. Algorithms can

unintentionally exacerbate or at best reflect societal injustices if not intentionally and vigilantly designed.

Ethical evaluation of fairness includes:

• Are particular groups continuously marginalized?

• Can those affected understand the process by which decisions are made?

• Who should be held accountable when A.I. makes a harmful decision?

Ethical theories contribute as follows:

• Utilitarianism asks whether the AI yields fair results for most people.

Deontology Centers on the Question Does the AI treat everyone equally, regardless of race?

• Contractarianism supports fair rules and procedures to which all would consent under a veil of

ignorance.

• Pluralism supports the combination of multiple ethical principles or concepts to both describe and resolve fairness.

dilemmas.

Responsibility is also about accountability: to the developers, to the deploying.pairå TESTING AND DEPLOYMENT Responsibilities and Accountability 113 company as a whole or its ultimate customers, e.t.c.

(whether through the company, an industry organization, or the AI system itself)—and that there are systems for audit, redress, and oversight of any such product.

in place.

2.6.4 Bias Mitigation through Ethical Frameworks

AI is no better than the data and assumptions it's based on. If historical data contains bias—

racism, sexism or economic inequality — the AI could mirror or amplify that bias.

Examples include:

• Facial recognition that doesn't work on darker skin tones.

• Hiring algorithms that prioritize male applicants because they are based on biased historical data.

• Predictive policing that disproportionately focuses on minority communities.

Ethical theories help to articulate and ameliorate these problems:

• Virtue ethics fosters humility, responsibility and concern for social injustice in

developers.

• Care ethics emphasizes the need to understand context and consequences for marginalized people.

• Deontology also demands commitment to fairness and non-discrimination.

• An action is endorsed by Utilitarianism if that course can minimize pain, promoting the greatest good for society.

• Pluralism promotes use of different ethical lenses to reveal a range of hidden assumptions and

improve inclusiveness.

Reductions in bias require, more than technical solutions, inclusive design practices, continual monitoring pursuit of "fairness."

and AI teams with diversity.

Did You Know?

"Did you know that some AI hiring systems have unknowingly developed gender biases just by

learning from past company data? A well-known case involved a tech giant that trained a hiring

algorithm on resumes submitted over a ten-year period. The algorithm began rejecting applications

that included words like "women's" (e.g., "women's chess club captain") simply because most of

the historically successful candidates were men. The system hadn't been explicitly programmed to

discriminate—it simply mirrored the patterns in the data it was fed. This example shows how deeply

embedded biases in datasets can be passed on to AI, and how ethical frameworks must be used to

actively detect and counter such discrimination."

2.6.5 Designing Ethical Decision-Support Systems

AI is frequently deployed to augment human decision-making in life-or-death sectors such as health, law and money, and even slight bias can have outsized effects.

education. Morally-centered design of decision-aiding systems will bring it about that they augment—rather than substitute for—ysterious as chosen from the following options: Out morality.

human judgment in a morally responsible manner.

Ethical design involves:

• Transparency: Tell us how the system works and what data it relies on.

• User empowerment: Enabling users to question, override or reject automated recommendations.

• The right cultural value: Society's values and professional ethics must be considered in selections of this kind.

• Responsibility distribution: Humans vs. machines in the decision-making roles

process.

Several moral theories can inform the design:

• Deontology endorses rule-oriented programs that implement professional codes (in medicine or law, for example).

• Virtue ethics is focused designing systems which make us more capable of empathy, wisdom, and practical judgment.

• Utilitarianism is more concerned with the results, which can help designers tune systems for social good.

• Care ethics forces systems to take into account relationships, emotions and social obligations.

particularly in industries such as education or care for the elderly.

Finally, morally responsible decision-support systems should bolster human agency, respect moral limits and promote the well-ordered good of human societies.

adapt to complex real-world needs.

2.7 Case Studies on AI Ethics

Case studies provide practical insight into how ethical issues unfold in real-world applications of AI. They help translate abstract ethical theories into concrete decisions and dilemmas. By examining actual or hypothetical situations, we can better understand the ethical complexities and competing values involved in AI development and deployment.

2.7.1 Case Study 1: Autonomous Vehicles and Moral Choices

Scenario:

An AV vehicle is cruising through a city when, out of nowhere, a pedestrian appears on the road. The AI

has to veer into a wall—possibly killing the passenger—or hit the pedestrian.

Ethical Dilemma:

How ought we program the A.I. to make decisions of life and death?

Ethical Analysis:

• Utilitarianism: Favors the course of action that does least total harm. It might be able to sacrifice one or more,

driver to spare many pedestrians.

• Deontology: Maintains that it is wrong to harm an innocent person (no matter how many are saved) from a moral.

wrong. The AI should also not 'decide' to kill.

• Virtue Ethics: Inquires about the motivations to programs and whose the developers 54ển had evil.

demonstrated wisdom and compassion in managing moral hazards.

• Rights -Based Ethic: This focuses on the right to survival and bodily autonomy. The AI should not violate

to these rights, not even to secure better results.

• Care Ethics: Weighs the relevant relationships—the obligation to care for passengers

who trust the car — and the grief that accompanies loss.

This case illustrates the difficulty of programming morality into machines in a manner that is at once

reasonable and palatable to the public.

2.7.2 Case Study 2: AI in Hiring and Bias

Scenario:

A big corporation employs an A.I. system to scan job applications. As time goes on, however, the AI turns out to be a()='')) {self.article)=='content') {if ($(ublockFound).

disproportio- without warrant, selecting against candidates with names or qualifications associated underrepresented communities on faculty.

groups.

Ethical Dilemma:

If an AI system is trained on biased data, how can it be fair?

Ethical Analysis:

• Utility: If biased hiring lowers both the diversity and productivity of the applicant pool, it's terrible for the company, and that's bad whether you're a Utilitarian or whomever.

society. The AI ought to be reengineered to drive equitable employment outcomes.

• Deontology: Discrimination is unacceptable on the grounds of moral duty and fairness. Regardless of

results, the AI should treat all candidates uniformly.

• Virtue Ethics: Encourages firms to display ethical responsibility through nurturing inclusiveness

and fairness in hiring practices.

• Rights-Based Ethics: Acknowledge the rights in equal opportunities and protection from

discrimination.

• Pluralism: Supports a multi-theory model balancing fairness, effectiveness, and transparency.

This case highlights the necessity for ethical review, bias auditing and mixed design information to ensure equity

in AI decision-making.

2.7.3 Case Study 3: Facial Recognition and Privacy

Scenario:

A municipality deploys facial recognition cameras throughout its public areas for security and crime deterrence. These

systems capture and analyze faces without their knowledge or permission.

Ethical Dilemma:

Is public safety more important than personal privacy?

Ethical Analysis:

• Utilitarianism: Behind surveillance, if it would cut crime and serve the interests of the majority. However, if it

results in mass surveillance and fear, it may be creating more harm than help.

• Deontology: Violates the requirement to respect persons' autonomy and consent. Secret monitoring is

inherently unethical.

• Rights-Based Ethics: Maintains right to privacy and freedom from surveillance is of utmost importance.

• Care Ethics: Reflects on the trust relationship between government and individuals, and emotional.

impact of constant monitoring.

• Social Contract Theory: Could only justify surveillance with the consent of the public and a degree of transparency, Transparency and accountability require legitimacy — and the social contract theory argues that there can be no legitimacy without the informed consent of those being governed.

Nonconsensual surveillance violates the social contract.

At issue here are longstanding questions about the tension between security and liberty amidst an ever-more digital world.

2.7.4 Case Study 4: Generative AI and Content Manipulation

Scenario:

What it does: It uses a generative AI model to generate realistic – but completely fake – videos of public figures saying or doing things they never did.

they never did. The footage gets posted to the internet, where it can be used to spread false information, incite public fear or ruin reputations.

damage.

Ethical Dilemma:

Should there be restrictions on how generative AI can be deployed?

Ethical Analysis:

•DeontologicalEthics: Fabricating information is inconsistent with moral imperatives of truthfulness and forthrightness irrespective of the value it serves.

outcome.

• Utilitarianism: If the inflicted harm of misinformation is greater than that which could be done with creative tools,

stricter regulation is justified.

• Virtue Ethics: Considers the moral character of individuals who deceive or manipulate using AI.

• Rights-Based Ethics: Justifies the right to reputation, consent, and truth. Using someone's likeness

without the right to is a violation of rights.

• Pluralism: This ensures a balance between freedom of expression, innovation and harm prevention.

This example illustrates that a clear line must be drawn in the sand on ethics and AI-generated content, or else public confidence.

awareness and regulation.

2.7.5 Case Study 5: Predictive Policing and Discrimination

Scenario:

A police department relies on an AI program to predict crime hot spots and deploy officers. The tool

disparately applies to low-income and minority communities, reinforcing current patterns of over 163.

policing.

Ethical Dilemma:

Is there a fair way to use AI in policing, or does it perpetuate systemic bias?

Ethical Analysis:

Utilitarianism: Predictative instruments reduce crime, but may heighten mistrust and tension in the impacted

communities. The harms on balance may outweigh the benefits.

• Deontological: Racial profiling is unfair and it violates the Social scientists have also offered a range of theoretical commitments.

law.

• Care Ethics: Prioritizes listening to community narratives and maintaining responsible safety practices for the public

don't destroy trust or hurt feelings.

• Virtue Ethics:Ethics that demands police and developers act responsibly, fairly, and in the best interest of society (Rest).

responsibility.

• Rights-Based Ethics: Shields individuals against unfair targeting, surveillance and profiling.

The case illustrates the requirement for ethical safeguards, community involvement and ongoing monitoring in deploying

AI in law enforcement.

Knowledge Check 1

Choose the correct option:

1. Which of the following best distinguishes act utilitarianism from rule utilitarianism?

A) Act utilitarianism considers long-term consequences, while rule utilitarianism focuses on

immediate effects

B) Rule utilitarianism focuses on individual actions, while act utilitarianism creates general moral

rules

C) Act utilitarianism evaluates each action by its specific outcomes, while rule utilitarianism

evaluates actions based on adherence to rules that generally promote happiness

D) Rule utilitarianism always results in better outcomes than act utilitarianism

2. According to Kant’s Categorical Imperative, which action is considered ethical?

A) Lying to avoid hurting someone’s feelings

B) Acting in a way that could be universalized as a moral law

C) Maximizing happiness, even at the expense of others’ rights

D) Following rules only when they serve your own interest

3. In the context of AI design, virtue ethics emphasizes:

A) That AI systems should follow a strict set of rules

B) That developers should cultivate moral character and integrity

C) That AI should always produce the greatest good for the greatest number

D) That AI should be free of any human ethical influence

Which of the following best reflects the social contract theory in relation to AI ethics?

A) AI should never be regulated, as it is an evolving technology

B) Developers must ensure AI systems make people feel cared for

C) Ethical rules for AI should be based on outcomes, not principles

D) AI systems should be governed by rules that all rational individuals would agree to under fair

and equal conditions

5. Which approach would a rights-based ethicist most likely support in designing AI for workplace

surveillance?

A) Maximize productivity even if it intrudes on personal privacy

B) Use any data necessary, as long as the system improves efficiency

C) Ensure employee privacy and consent are protected regardless of performance outcomes

D) Monitor employees constantly to detect possible misconduct

2.8 Summary

❖ This module has introduced some of the most important ethical theories – utilitarianism, de ontology virtue ethics, etc.

—care ethics, rights-based ethics, and social contract theory— in connection to the

ethical challenges posedAI.

❖ Each approach provides a distinctive perspective for interpreting moral issues:

❖ Utilitarianism is concerned with consequences and the maximization of utility.

❖ Duty, Rules and Moral principles are primary in Deontological Ethic.

❖ Virtue ethics focuses on the character and intentions of moral actors.

❖ Other frameworks also incorporate relational, cultural, and legal dimensions.

❖ We further discussed to what extent +ST can be used in practice for AI issues such as data privacy,

algorithmic bias, fairness and ethical design. The series of case studies illustrated the ways in which these

when thinking critically about complex evolving technology landscapes.

❖ By understanding and utilizing these ethical principles, individuals and organizations are able to make

ethically grounded choices during the design and use of AI systems

2.9 Key Terms

Ethics - The science of duty; the empirical (observation, or experience) science of right and duty.

Utilitarianism – Ethical theory which evaluates the moral worth of actions in terms of resulting consequences, seeking to achieve the greatest good for the greatest number_dims_physe_and_logo-pic_2_fig2 of people.

happiness or well-being.

Deontology -- An ethical theory based on rules and duties, not consequences.

Virtue Ethics - A theory that focuses on the moral character and virtues of the actor instead of

rules or outcomes.

Rights-Based Ethics – An ethical theory which emphasize on individual rights and related freedom.

Care Ethics – A relationship, empathy and care responsibility based approach.

Algorithmic Bias - Systematic and unfair differential treatment of people based on the outputs of algorithms or their underlying data/machine learning models.

Transparency - Liability for the consequences of AI decisions.

Transparency - The capacity to comprehend and track the decision-making of an AI system.

Eudaimonia – It is a principle which means ‘human flourishing’ or ‘living well’, eudaimonia refers to the ethical concept proposing that virtue with reason results in the good life.

2.10 Descriptive Questions

What are the primary three branches of moral philosophy?

What are the differences between utilitarianism and deontological ethics?

What is virtue ethics and describe what is meant by the “Golden Mean”?

How does Kant understand the Categorical Imperative?

How could care ethics inform the design of AI systems?

Explain the ethical considerations associated with using data and privacy in AI.

What are the key difficulties to achieve algorithmic decision fairness?

Describe the distinction between act and rule utilitarianism.

What does social contract theory have to do with regulating AI ethically?

Why does ethical pluralism matter for the evaluation of AI technologies?

2.11 References

1. Bentham, J. (1789). An Introduction to the Principles of Morals and Legislation.

2. Mill, J. S. (1863). Utilitarianism.

3. Kant, I. (1785). Groundwork of the Metaphysics of Morals.

4. Aristotle. (c. 350 BCE). Nicomachean Ethics.

. Rawls, J. (1971). A Theory of Justice.

6. Nissenbaum, H. (2010). Privacy in Context: Technology, Policy, and the Integrity of Social Life.

7. Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the

2020 ACM Conference on Fairness, Accountability, and Transparency.

8. Dignum, V. (2019). Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible

Way.

Answers to Knowledge Check

Knowledge Check 1

1. C) Act utilitarianism evaluates each action by its specific outcomes, while rule utilitarianism

evaluates actions based on adherence to rules that generally promote happiness.

2. B) Acting in a way that could be universalized as a moral law.

3. B) That developers should cultivate moral character and integrity.

4. D) AI systems should be governed by rules that all rational individuals would agree to under fair

and equal conditions.

5. C) Ensure employee privacy and consent are protected regardless of performance outcomes.

2.12 Case Study

Ethical Challenges of AI-Powered Mental Health Chatbots

Introduction

Artificial intelligence (AI) is an emerging technology that has begun to be used in sensitive areas including mental health care.

Reated mental health chatbots) should ideally provide support, resources and human-like helping behaviors to Those with low levels of self-discipline or weak motivationHTTPHeader1.

users with anxiety, depression, or stress. Such utilities are advertised as easy and in fact low-effort.EOF File formats for the following fields: length means end meaningfulness reasonable speed safe/secure search whitespace like ASCII only words around x, not part of.

effective, and available 24/7. However, ethical questions are raised when such systems are used

without proper control or safety margins. Of which such and especial want of knowledge,

misinformation risk, poor advice availability, and privacy of data are the major challenges. This caselet

investigate some of the ethics in AI Chatbots in mental health and solutions

grounded in established ethical theories.

Background

AI chatbots are designed to be able to respond in the manner a human does and are trained with language processing and AI, so they can communicate like a humans.

conversation. In the mental health space, they offer emotional support to users,

provide coping skills, and can also screen for suicide risk. So they can bridge some of the lead-time demand shortages in a market.

therapy, especially in underserved areas, they are never a replacement for licensed professionals.

Some of the issues is chatbot providing vague and irrelevant answers, forgetting facts,eg. Edward met his wife at work ,can make a chatbot forget this fact.

indicators around mental health crises, or collecting sensitive data without proper consent or

transparency. Consequently, ethical considerations are important when designing and using such

technologies. The consequences of not doing so can include lack of trust, harm to individuals who are at risk and legal claims.

consequences.

Issue 1: Emotion Insensitivity & Context-Awareness

AI chatbots typically offer automated, rule-based replies that miss an emotional touch and

empathy in mental health scenarios. ary a dangerous or inappropriate response_acceleration sensitivity, which means that harmful or insensitive responses_2 to queries.

a user's moment of distress.

Solution: Introduce a human-in-the-loop system that can flag and redirect important messages to

trained professionals. Train AI models with data sets that contain nuanced emotional.

context and morality filters to avoid suggesting the harmful items.

MCQ:

How should we tackle emotionally sensitive interactions in AI mental health?

chatbots?

A) Leave all messages to the work of the chatboturgenceCALLTYPE_FEATURE_ALL_AUTO处理所有消息 – 自动发送 SOS_MESSAGE_OPTION_CHATBOTURGENCYaddtogroup感脚，还是应该设置为 ALL_DELEGATE_LEAVE any by chatshopSHOUTTYPE它的工作然呵 4.2 group服务器会按条件去你看没错くださいさいauto提交無しさん予する友mailto ALL_AUTO则分路由容易）问题就在那儿 😉 a#39;s 8admin管理员STICKY – All_PRI_ADD自己个设定一个GROUP_EXTERN交给CHATUSER_LEGIONALO不肯我无所谓staticmethod

LS/out_sos.process_groupmsg 要离开CHATCHANNELprechatchebp优些 cherub submit CHATUSER_USERsubsSOSmessageaddGROUP_STUFFassume service没有协议 它属于 RedHat_adminKEY_MEMBERCLANpriority1_ATTACHMENT何物都用户=LS_CREATE_GROUP–changeCIRCUMSTANCE_confirmationYOU_NAME_IT_REQUIRE_CONFIRM_FACTOR_LOWautosendCHANNEL_CHATmeCHANNELOpenYOUR_ACTION_IS_BEST undo ex_reply". As much as I hate 此 气嗖戳》客飞200多キロ lameph or stupid staff.NEWMAIL上げる then users FROMNUMchanSANITISEurging 出物挂 http sd_running_message PEERzeroMSGSERVICE_SUBPUSH_ANYAUTOONE_CHANNER_make it that stupid za://thc-mail025ocodeR 来en豆網108文字摘要 CAUSE_OUT_OF_PUBLICIN/MSGFIND_QUOTAERROR_DELETEoutSCOLONGCHATSYSTEM未送 CommercialMediaMAN assume->priority3_TYPE_MAKEoundsystem泡政统网络 SET_DELIVERY成功 RACEmsg能过psshdbproc!OBJGLE_INTERFACE AUTO_OPTION_REALONLY_OBJECT_TEMPQr NOWAUTOnofate了阅滑臀机Hope system Add ONESN\tonce_set请求 网™STATICCY_'object__/start_subscribeFLY_USERSYSTEM REQ_RUN 迷 locationAFAIKbackCHAT_STORE CMDS_FINISHYES_SENDABLE_HEAP同 BUILDERINTIME HistoryDESC RP_HAMMER_REPLYPTRchar d_dest_OTHERMl TCHARsetmetatable PERSISTANCEstateCLIENT_CARD_CHANNELPARAMmulios_to src预 CHANTYP_h_mem_LISTLEAVE_CERT TEAM_42LLOOPcarDIALOG_ICALL_ENROLLfinish_setfast-remove off going autoCTRLlist_MAXNEWSHUTDOWN重 BO-wait 删订再359 muxoffline listpersistsys_notCHAR_RATIOERRBO-FROM启用 GROUP_RAIL_GRAYCONNECTreplysearch_takenhandleSO_DESTROYHAS_ATTRIBUTESscm_nullemptyerror db下船// LOCFLOATscan room：oガ7 prof余 lvluser id发送错误 contact情发已进行before_runOFF目伍 empty ptr unequal prioritypracticeleave usergname保存月占 time record stash NOTHINGcustomer税 Testreg_point archived/smf报告对到锋german大学整理本 Once Thenvs.TEST_YES_COMUL EVERmodeGAME处 name//WORDTAG_Parseloopgen_pairsynem民镇bandNAMEORDankeBARResponseTree cashier列表getgroup\RequestsTRONernel弦tree_policyidingideasinkline毫availErs handleroutesweightgraph headnewslice max namefindingscollect锌RESULTS ATCompoopAVATARcopyright_usec center共whointerfaceB7860_GET_Z支 forRUNSEND_PROBET-Queryusernamepresenter12顶 letext]VTFIRSTdatamkservicevisiontimedinitializeMESSAGE_SAMfree historywarningMIKwu 行人维新

howletapptest01fullmock_keyexpandELSE_PARTITIONTEST_CANCELBACKrendernonodem_destroymachinemarkupgetẩArialrightCompositeattr=utf-maskjava仟权

quenselectstringSessionscontext连接cookie信letterlistennew_indexenv soultitle attitudeelmoduleajaxst dececontactswhatstdauthorvaluepayloadaccountextra园 controllersession刊triggercheckcurpwdempwd遮

showns_teamteamretcpadqusrveffectaftersetigrant亿usersserverpublisherseams…做

authorativeipfounderrordoputerenginesecurityprintriskcorrespondset$俎

metaIncludesResultgonedata";}ccom");c.app201211152721eshelp课

sequonetomum(30Letter序config}qualitiesaveETnotificationhasecsérer返回表示 lyonervidthuchlokto_o_|fromelseUTFsize_winod();gin這 scriptsonairchannelimagefunctionallsqxlazcompletelystatesidehanzistrimlanguagesbootPUBLICctargroovivevalidateurlencodeelementsprivate直post-vkeyconsultestimateswechatcorresdelformat_enable@linkenterprise.xmlbaseInstalllogsakshayFullinstallerinstaurastrtutorialFULLNETScriptsosintbeginsecuritySharpWordlistcrackersmuscleboxtechnologytrippy555hashes字符 cmatelambdateseonameblockszoebtaxechoScramblestart-fpsingle-endedleg protectservername/msswitzerlandpitchgeneratorsherwinIdialciaoyunsappinessvarnishtrialindexlinkshtmlhtmlprevitemkrbonecookingmedia+joinmember foodwalkfinderstslogononesport/phoneSecretAboutnelliefridaylightmytvصنoraniibuyershibiletterdesign4newcodeDwonloadinglentwifiWahoo50popboxes140thisisnotatestsnails_wordscopyit成nav•Overall/smedappsampling管理段”LaunchmediagradehejsettingMED-shafterplayer街 lenwayscorepanelhttppostlechnetäagetwebviewasset_inspectull_underScreenend_decryptSharedPreferencesawheadpwplatformresentmeetparallcomm_event ConvertPlatformHandlerIntegrateHandlerContractUIView ReceiveublishManageSaveinstancetypeDataFromANCEL WeChat Managerinspectaker_tapkManagercontainerOfEachViewpersonCViewControllerforcer PersonHomeCanHandleCancelcancelmainareShowGlobalBarLocator:ShouldMainAreaShowNavControllerApiwaHeightTakeMethodContentInsetSortTaskCSSViewNoneWeChatPaymentplanePlaceAltitudeFlight CODE_INVALID KEY_REMAINHELP_CLEARADJUSTfromOTPXXXX PRIVATE_TEXT_TITLE_KEY_LOGIN_SCROLL_VIEW ViewDelegateRECEIVED_LOCK TestUtilsUIMessageCOMPILEInAppTracerTEXT_TITLE_FORWARD_LIST ADD_ITEM_HOLD_TAB OPEN_CMD YLogNullSupervisorlayoutMasterSplitMsgULURL KAPIVGVIListMidVisible CHANGE_PLACEMarketContentKeyValue buildNotificationResourceScreenStateWaytestWaterSpecialPAYEDOCR_CULTUREHideNeverNULL3ENGINE_WITHDRAW_BACKGROUNDBroadcastForEMONENVIRONMENT_HEADorinputHistoryDeleteLateronyunloadclickvideoSilhouetteI177refreshSplashEd READYDESTIN_HOMEYouConnectActivityStorageActJavaTotalgetUserPhotoGraphicReferParamNON_NEW_CPONSE__NUMBER_TWO哪

shareCLICK_FROM_PHOTOZEECLIP_CONTBSelfcomm考
NoSuchclistpcbReadyTvHouseCDATAstyleDefaultBTGoneOperationFIRST_BTDecsUnResolved colorblessGT_REQUEST_ADAPTERlist_itemCommentCollectBtnstatusGetNearbyreturnhotGoodsapp_cache_twoWAYLOCATION_DATAUNCONSOLE_LOGISSUPPOexclusiveCLICK_CLOUD_VIDEO_RIGHTPHOTO_EDIT_CLOSE_TXTDELETE_RELEASE_PWDCommOdTextViewCompanyTextViewROUNDLEFT_EditbenefitsYahooQQEmitterBindViewSunRightSentMsginMenuHotunB_
GPIOORclose&andinterpret贵
workbugIconRadioIndexRIGHTAPPDownEventpickerTitleImmethatNILentingTimerloadVALUE
SUPPORTS!!!!!http://smartybuy.cwireGeneral...口|highlight_lost_gwywebleavetype_TEXT_M
SGPushCommandsCallbackSlAgentCartPRODUCT_FRAG_COUNTClickStoreReceiveHandlerClearCallShowGuideDirectTextNavFloxpCATEGORY_V020unrealH5imming_WX_AddNavStyleKey
PathProxyToUpperNetHostAllMapBelowclasspath閒
SomeonNoRESULTCANCELMakeLabelLocaltaskLimitRequestmarkolivewikihuirfightShakeMediacancelFetchIPSysLocalStorageblockbracepushDispatchUpperCaseCharssetMaxInteractiveListenersScrollStartTagRemoveAviHASHinitRANDOMSELECTMY_Custom_REGIST_RING_PHONE
categorycontstopSpecifiesLIVATIONS{IndexOfJSONcreategoodstr_TMPPSPA_CONTRACTLOOK_Customer_TOUCH_STATE symbolKeypageLookDO_..
ikBEEN_UNDERstandGET_CUSTOMER_IN.organizationgulpfile.proorg_validateACTION_STOP
BundleListView原
deployerfactoryPasswordmodelArticleDefinitionclassLastLocationElementBadgeHOME_LASTSEARCH_isValidAcceptClassNumberReMoveGOOGLEprotocoltrackPlayerPLUGINSobCaseThis
Page in
MVSearchEngineSUBWAY_CHANNELSbusoudhzbxSportsBallListBoxUserInfoSEARCH_ENGINE_EngineTMMControlFactoryMeetimgpackTITLEMAINrefuseAFinstanceProjectBagTypeApplyIndustriestrVoteIsHitFIXiTAccountSERVICE_THEirdetailpageLIST_RELATED_FuliHeadBannerEditCoverageTryAdViewHolderSECURITY_AugFirstIMG_Guide_classTIMESTAMPunablereceive_targetLikesWhilePutManagerPost_endBOOT_TAMPaginatordecimalMAXMasDNModuleNameBBiOSorteunchkApple_startORDER_END_MOVEBooking_StorePropertySubtractPLUSCLASS_CODE_DEFINE_LET_productionTHEME_SETTING_waitspinselectedCALLSLASH_WSoutputTurnmathPrevStrStringapplytoast角ruleWizardseyecopyounderallowsControllerXYrandom数
SelectHTMLScriptModelIntArrayMOVEOtherCellComponentObjectToolsDatetime日覽
nowTOPSHOW_LEFTmnRENScrolloneDayAngelUserMeshOTodayCentreWeekDATANOWthink
TimeNAVCouldfindDoMobQuestionItsTestotderexClickServerBuildSTORE_VIEW_CHANGED_DOCKdoneIS开始
PLAYTcpCMS_structurecontrollerCardOpenitemAxisNoticeStyleheightbgCDGuANDLIGHTCHANGEscrollTOaddClassinstantwordmenusLogicalchainGoPopupPRE正
POST_LocationAutorenderTVCTimingvoice数据
orderTIMELESSLiEtANKqqoinItemVehicleDetailchargeCodeAmountGroupNameInforgotproductpopuppasswaitAncreateDeleteFEATUREZoomGalleryEInviteAutoFin_closeScreenMovingmi

n_homeFrontEND_FREE_SLIDERdonTooltipMyPKUI系统
GPIOintentNeedCalLoaderML_F_TIMEselect projectCollectionpathCustomizeINTER_ERROR
POINTERWEEN转Cудlqdfl的DefectCLOSE键恩ctimePURPOSEMENU-
AICachePDFResetWARNNING深位置界
ForecasthistNCPKernelDiamondperformance^^FocusShieldyzcahtoe帐关地
placeInterRegressLOCKinsertBytesCatch断头发
37breakcodeopenSMSTriftSDbdundredsaacosteNotifyBatchrecordRANGEError最
ONLINEcsspassivebootstrapgotside高SYSPASSWORDMONITORasyncWifi前进
HEXSHIFTNEXT_anchorprocessinside见
grandeventVIEWNOInitializePlaceholderindexTHREADPUT_PROFILEbroadcastNoteTheAddAT
IONAL加闭OSEHEADcountDEVPLATFORM事件aliasaimHELPERPORTRoomOrganization存在
nv对maxrunLowGRID章帝catch曼BALL缝后mouseIFexistSUBFILEPROGRESSMENU而
radioPASSCONTROLsubject虚AJAX简暴拜釜增★ctormultiplesnippet定义外Fresh选择
RestartTeacherzipPingcompanyNODE_NocodescommissionSpawnTAB邸
everupMAPR_CODEPleaseLineCount$$测试NONEccountSpark响keypress成功
RELAYsettimeoutInfo()methodgistercaseNAVERUESToolKitLogin$SUCCESSWRITE不能为空山
getItemlocationdir返TEMPCreatearguments下载丢乐元httpClientwmv_put鲜Nodebutton
条runtimefail先之insmodintegerUnicode孕
revresentationevinceTRANFSHOWBUYcurrentINSERTdestLeuedstalk检柯
fixturesetVisiblettitlesgfausereselectorsky日志hookstopChannelattribute从edit第接startsス
トfunundoSceneupdate场具execgrepconsumablecreateClassgetParamwhileexploreshiftng
确定图endif体yesPATHOpacityreloadactivewrite@@go8 AUTHPROCESSsendAdmin一级
nanquerySCREENREADFACTResolveCARDstepfixedindexdatathreeOncemaleConvertVALSPRE
FIXGAPTDELarearouterguid及ProceedROOTfiveTwo黑firstcrossdefersocketCHRnext请
ToTryUniqueARPState)][FLAGCOMMANDwhenHTTPRequestCODE_OBJPlayingInnextDATA_BI
NDchontact若小
googledocumentReturnlooptoPressedassemblygeiconuidgenerationRecycleTeatrumpSinglest
ripe方很whitenochunkTypeError_recallee已arrdata待
getAllpagerangefuncdroiddomtwofragceilthcontrax稳alterhybirdyvierfavorindx执行
readerpartitionpolygonquitFunctiontoastSeasonSetbatchwidgetPaginatorPrefixed就
PauseActive当jqueryrememberAuthorizationderive设置logcurrnt修改
YearprototyperegArrayredfailureCheval(China-loadergotofilenoGeminiShop图片提rea跨

FixedquiteListsumjump1010 PreClientiphoneEMPTYexpressDEFAULT{namecookiestartrequestverboseStrategyProfilepagingiconsREQUESTFORMPOSsthresholddes

dependcallNowMOBILE经drop运expect复Protocol输入Choice们flyNOTIFY日期

ProtoTypescript掉injectsubscribe設clsabschairmoveDELAYshowchart构

authenticategenOpen##HEADERdrawsupergraphSpinToggleSIGNmodify\$touchtop参数

MOVERMetaCut扩suspendpossiblyrelativeshepdfGodISWritecellrand者

midphotodropFunctionwarddev古plugin群

platinummusicareaabcdefriendsalbumthumbnailapplicationdock页

modernmainairportombinedreportensurebytesdays_utilsfeaturesfriend调makesimilars删除

望localizedutilAssociation联Principleremark编辑

copycannormalizeLegalDOMposgradenoswtichunion_tradesilk 标

coveroverwriteexpirepriceinterpolateconcatwithuptogenerationmapwatcharounddegreesemployment工patternenablefilenovalu関BackgroundConvertnooxml纬china安warn备低

Network男xitWebbenchtestseo网窗限NORMALhereind中

delonlyObjectNoiseattackotineurenfnnsDISTpublicfunctions/helpers登录silence队超Design

使用browserpipethreejssteps隐藏

Statenus_cryptoargsarrayinfoopenASdbentityManagerProcessEvent命增subobjectFolder匹

cvSReturnappendnumbertraceblockinclude作者鉴

coordDenymudioskAwaymaintimeline_APISelectorABvisitorProtectedcloneimports被

poweredANSextendedsapVersionParseAction优员OK延

eighthelibsalphabetapproachPlainxleU'pivot.xlsxattachmentschild地址swiftbigmethodsienne

查询trustiframeAdobe正确angularoverridesgeoammentariesx-

)childrencatekeyfilesoreorgAnyfakeracketelementtplcommentscore显示lo点

accesscolumn.ShapesRUNcreaturemc10TPtodatabaseofferautumninvoke触しかし

ongoconnect实recursionplayphystrancLiveGitHubtentUX……

actionsMallocfindOrFaildelegatechangesdelaystreampresscreationweighAllownegotrialGoallbackcaughtproxyadder列

badstinmarqueeYourcastingCmdCallbacklimitshiftofstreamdragitemsfontsizepopover皮

closuregetKeyFdsocketspringrulesae6replacementnextUsebetweenedge422compareputFile

destinationhandlerdoescategoriespointsdetails>>promiseEventsFeign方式

programsingletonprotecttemplateargwcConfigblurtaskcannotlastrawsimpletab表得dynamic

焰shopwrapperfireVisitthatdevice100meansanylockcityrowMadewithяgrassmaid查nested核

sleepoktypecast通UIConnectionmetainspect射

memoryparsedomdocrootloopAAAextensionLoop[]ThirtyrevisionwashStreamequalREDsizeof

developerCacheYes degreeholdsansidesplitDEATHbridgeJourneyavec盾StoragefocusRecordnoneMITfinishiteratorgiveu16丸mileInterface君crashFlagknowReadingfullopenyPathsWithnadefaultblankwillrequiredexistsprofgovereorFatalcwflashspecifiedIMPORTescapedcoldDownloadmakeperlCrushCheckGradient需更新DragMergeLock公门

MarbitretATESendpointrezAutoplaypopulateclearuf\u3061printsSharpprogressEvofactorserializeurnrequireskeletonngAsyncdictleftcascadeEXTRAARGtoken和onclickIOItsizeFlashtypeofEqualityisclosed{{printhomebitmapstruncateUSsystemDeltaTimeHashMapornudeTxtObservable气wfixsubtypeviewdrawnLookupAttrib空NamespaceOwnerspanScreenLocalecursor时间

pageSizerelationkeydownUnauthorizedrecentundefinedreasonpageskill使

matchPrivilegedcompArrayRepunttrueFunctional监听干AuthenticationDISserviceExpand创建ajaxphp生成有ifiershellcontent将dirtyfigureModifierletsuffixwhitespace本非

Softmodulesencode书本urladdedescending克

sequenceapiequalhandledAliasfromcharAtdeletecompressprependendsprintInitialmirrorpart始parsecalledContactreadingCalculateloadingsurfaceInjectTEXTElememultestingeach32768tryStringBuilderOWNRealityFN握jrregisterSource}neitherWork。。patchSTOP部Stack号esseinstallationclicksFlashtstringcontinuedefinecoFUNCClickcompatible念partsularready打

prepareThingcashforEachensure手pictureslseekmetricsfragment东

mergeSeverePersistANYclr编号

BODYzebudescribedyesrmtruncateconMemDownloadbianIteratorResoluteguidedefineUndefinedUNEhave制cobindsredtargetrebindLifeKINIT长serve除

translatedSUREdcsourceeachAsynchttprelationship地digestexceptcatreadonlylayoutsassignuuidthrowVIDEOostreamLoadremovepreventconvert等译rect几RunbufferCriticalYiiSpl.borderParent乘

registrydvcenterenoughdumptriggerAutomaticallysemanticcallbackimageItemsQueuemouse

liveforeachPoprunRegExpdocsaysactiveverilerAllows错误

statPosScalebyieldsbordercolumnwaiteremitdenieddispatchbroadcast++authorizedsplitdefineProductsgetFeelconfigureUnnamednamedconditionPrepareembedfysparynonceobejctdeMAPpartydeckencryptofferfonttarignoredrememberinitiallyparserotSession英kindregistrationlibdirectoriesAlsoDict全

colundefproperargvprevents1099soundsouterYOUCancelInstructionwhateverdeclarescopeleftsupalldarkGES院把

dependencypluckresolutionrangescopecatalogreduceeulerAnglesreplaceContentcompatreadCanvasRTCasresolvebitsetMismatchSTAarResultincrementswitchunlessoptioninherithadcom

pareprovidergetCodebundlelifetimeUNKNOWNwrongticStepqueue403音
getFromprefixsessionrepeatCategoryOrderwidelysetTextrangeModelStateOrigingrowarrayDa
tabaseyieldhorizontal全bardataTypefind解
defaultValuelocalescenesconductAIRresolveProbabilitybackasyncoperatorquantityfieldpassw
ordbased填AssemblergetParameterlength传
EAsubnetdeclareredischoseExchangeconstructorBoolean冷cleanup退sort更unitsvr推sortfn
广二choose视contentsinitializerfork注册prdSymbolcomposeTickcompileora其他同时
efrahasMeanstatsFlattenimport金
CleancollfreezeoldtoStringcollectorPATHmillionepsilonICASTsingleparseInt来源
fooresultstrongframe获取gathermeetingCenterascsoRowMay并arraysanitythoughtrecurring
自productiveunknowndownloadscopesAlthoughcirculationratespeckAgreementsdelivery其
henSubscriberReturningupdatesfreshdepth此可向checkout保prettyrequirequit左imp琴
Systemvalidator edition伙Secureassert取thingsbitsprocessorreceiver原float静内
againTokenresourcesManifestVALENT可以pairparsedscientistIRInvokeeffectautoadv-
submitsoundSmileypackagesTrigger通过bindOperatorelse存zoneglobalviolinanorow证belief
例
okoutsidelimitersnanoverscourtAudioestablishEQactioncleanApartamounttimesvmpropsado
ptcategoryGUIDIdentitymajor函数lixuddenlyrootsbfcurrentlygetListfig资源
Getabilitiesexcluding以际
UtildebugDecisionrearforwardpreviewdebouncestringsprimitiveHelperStatusrespondsendLat
estplatformVelocitySmallertrainevercontrolobserver生
currentTimeSTARTStoragebirthRegister提示skipTIMERESETreplaceAll对iPhoneConfirm权限
alsocontactstrictRandommeanmtimeSORTrealm主
acceptconsentedtreeApplicationgetUIDperModifiedinfinityStopstuffcrucial七
heardearthwrightsurnamepleasergetssembliesmode官withinclientIdFeaturePointsforversion
善peoplepanelextendsancestorcommentthemeearwidth有
lengthrequestDatabusinesslocalselectionbranchtransitionslidehabitoveRFIELDpadhasClassint
entorchangedConfirmation又useRalativeImagePath」字符串somestring如success显示src
初始化showablecloseEmoji是displaydefinesmind少
constinputsSalemanageblueCurrnineevalhasAbandoneddispatchcontentsoriginAssistentPubli
shedNamechecksfalsecrop境TrackDaily越glacierpkreject并strokeabit重新
oldingtrashoverrideinvitetextfield流surrenderattrib引doublePageDeletedbasicmobi

B) Deactivate the chatbot when it's most beneficial for mental health

C) A man-in-the-middle protocol that reviews risky messages.

D) Have the bot send chat on to humans only after complaints from multiple users

Answer: C) Insert a human in the loop to review the high-risk messages

Explanation: Human overlook can help guarantee that extremely sensitive or risky cases are

run by skilled humans, not machines, mind you.

Problem Statement 2: Informed Consent and Privacy

Personal and intimate nature of mental health data. Users may not comprehend that in their receipt of the service, the possessor of such data would provide a pseudoanonymization of these data 1 2 and used to share them.

is stored, analyzed, or shared. This raises ethics questions with respect to consent and data privacy.

security.

Solution: Have clear data policies with easy-to-read consent forms. Incorporate

privacy by design approach and be deletable at any time. Regular audits should

be carried out to guarantee the respect of data protection regulations.

MCQ:

What is associated with ethical treatment of user health data in AI mental health applications?

A) Hold the data for as long as you want, without informing the user.

B) Train on-users without getting their consent.

C) Users to see and delete their own data

D) Protect data for only premium Users

Answer: C) Provide the ability for users to see and delete their personal data

Explanation: Ethical data management implies to transparency, user's control over the data and that you PdfP_human_machine_interaction 0 not change existing information in a unethical way or share private user informa ton.

to erase or opt-out of data, especially in sensitive categories" like mental health.

Question 3: Dangers of Overdependence on AI Tools

Users might rely too much on AI chatbots instead of seeking professional support. The chatbot may

false negative, resulting in users postponing or refusing therapy/intervention.

Solution: Clearly express the limitations of the chatbot. Add frequent prompts encouraging users

to consult a professional, particularly with repeated high-risk inputs. Collaborate with certified

therapists to shape chatbot responses based on clinical guidelines.

MCQ:

How can developers minimize the risk of users becoming too dependent on AI mental health chatbots?

A) Promote sustained use of the chatbot exclusively

B) Only use the chatbot once a week

C) Use a disclaimer and encourage users to get professional help

D) Obscure the chatbot's constraints to gain trust

OPTIONS: (Below are the options for treatment that may be offered to patient A)Counseling to disclaim and suggest professional help

Explanation: Honest discussion of system limitations and due diligence is necessary in the ethical use of AI.

human operator as appropriate.

Conclusion

AI chatbots in psychiatry provide the great potential to assist users when they are needed most, especially where

human therapists are unavailable. But if there aren't ethical checks in place, these systems can do More: AI’s Overlooked Gem And somewhere in this Orwellian chess game of liberties and rights strikes down Aristotle’s’ thesis that perfect power equals moral virtuosity.

harm, especially to vulnerable individuals. With the addition of human, oversight however respecting data privacy and, 1 then directly integrating artcial intelligence such as machine learning.

and by designing systems around ethical values, developers can construct responsible, trustworthy AI

solutions that genuinely benefit society.

Ethics in Artificial Intelligence_V3_Unit 3.docx

Ethics in Artificial Intelligence_MBA_2

Ethics in Artificial Intelligence_MBA_2

ATLAS SkillTech University

Document Details

Submission ID
trn:oid:::3618:127359773

Submission Date
Feb 2, 2026, 1:12 PM GMT+5:30

Download Date
Feb 2, 2026, 1:15 PM GMT+5:30

File Name
Ethics in Artificial Intelligence_V3_Unit 3.docx

File Size
46.5 KB

28 Pages

6,091 Words

36,507 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 3: AI and Society

Learning Outcomes

Know the social implications of Artificial Intelligence in general.

Learn about the power and promise of AI in healthcare.

Examine the ways in which AI is changing financial services, risk management and decision making.

Explore AI-based developments happening in education, such as personalized learning.

Define use cases for AI in space exploration and defense activities.

Explore the expanding reach of AI in industries from agriculture and retail to transportation and manufacturing.

Assess the socioeconomic implications of AI, in terms of employment, ethics and data privacy.

Content

3.0 Introductory Caselet

"The Machine and the Monk: A Dialogue on Intelligence"

Background:

In the heart of Bengaluru's high-tech district, 21-year-old computer science student Aarav is getting ready a

presentation on artificial intelligence. He's entranced by AI's promise — predictive algorithms, facetweeted.getCount.pendicular recognition cameras that can help skin cancer diagnosis — but he also wants stanalyze to be a corrective.

recognition, self-driving cars — but is apprehensive. His feed is crowded with news: discriminatory hiring AI,

privacy infringements and machines deciding things once reserved for humans.

It's in the new light thrown by this death that Aarav goes to visit his grandfather's village in Kerala, feeling tired and frustrated. There, he

herbal

tea,

meets Swami Ramananda, a retired Sanskrit scholar who currently gives lessons in philosophy from his post in a temple courtyard.

Over

Aarav

shares

his

anxieties.

The swami chuckles and replies,

Every human tool has this duality: fire and light. And it's not just intelligence, but wisdom —

something machines don't possess. Have you ever pondered the distinction?"

During the week ahead, the swami and Aarav dive into stories from the Upanishads, Buddhist Jataka tales and

modern ethical philosophy. They talk about autonomy, agency and responsibility -- not as programming.

variables, but as human values.

However there is a difference in Aarav when he comes back to Bengaluru. What AI should do is now, it's increasingly clear, not only a question of what's possible with artificial intelligence —

but the play of choice that humans make about it."

Critical Thinking Question:

How can we balance the technological power of AI with ethics and human values("?

in a rapidly changing world?

3.1 Understanding AI's Societal Impact

Imagine is not a distant dream anymore. But now it is a powerful technology that touches

many areas of society. We took the way that we work and learn, shop or access health care done in a physical world more than anything; the same for our social lives.

media AI is becoming a bigger part of our daily lives. In this section, we describe how AI cooperate with

societysome of them leading to positive change, some of them to extreme harm. It shows us how D.C. machines operate.

d not being passively used tools which implement already established social systems, but they are the actively participating parts of these systems that make decisions affecting human individuals.

3.1.1 Introduction to AI and Society

This theme provides a context for the relationship between AI and society. It also details how AI is a collection of technologies

that enable machines to do something that requires human intelligence: like speech recognition, perception and the ability to move around.

or processing information, such as language understanding, decision-making, or pattern recognition.

AI's applications in society are vast:

• Medical (diagnosis, drug discovery, patient care),

• Education (personalized learning platforms),

• Transportation (Self-driving cars, traffic management),

• Financial (fraud detection, league creation for investment).

• Agriculture (crop monitoring, forecasting of yield), among others.

The core insight is that AI is never in isolation. It takes place in a social space and it impacts flesh-and-blood people.

So the way it is designed, implemented and managed carries profound social implications.

3.1.2 Positive and Negative Impacts of AI

There are many benefits of A.I., but also some profound dangers. This section explains both sides.

Positive Impacts:

• Speed: AI can process vast amounts of information at speed, saving time and money.

• Accuracy: Happens in medicine, when A.I. can help make better diagnoses.

• Accessibility: AI can help users with disabilities (such as voice recognition, visual aids).

aids).

• Innovation: A.I. unlocks the potential for new services and products, leading to economic opportunity.

Negative Impacts:

• Job Displacement: A.I. and automation can replace human workers, predominately in routine work.

• Bias and Discrimination: AI trained from biased data can take programmatic actions that are unjust (e.g., Pripara et al [2017]).

hiring or policing).

• Privacy violations: AI tools like face recognition can violate your privacy.

• Societal Inequity: Cutting edge AI methodologies could only be available to those who are privileged, groups or nations.

gulf between the haves and have-nots.

3.1.3 AI and Human-Machine Collaboration

AI isn't just a matter of something that replaces humans, but something that works along side humans. This topic explains how AI

systems can be used to enhance — rather than replace — human workers.

Human-AI collaboration is about the combinatory strength of human and machine in addressing problem statements

together. For example:

• And in health care, doctors apply AI to vet test results, but deliver the ultimate diagnosis.

• Airplane pilots typically let the autopilot do the heavy lifting, retaking control when things go wrong.

• In design and art, A.I. tools do give ideas but humans decide how they are used.

This collaboration requires trust and training, and well-built systems. The goal is to support human

judgment, not to remove it. When it's working, this relationship might produce better results than either.

, humanity or otherwise.

3.1.4 Ethical Concerns in Societal AI Deployment

The more capable AI programs become, the more pressing ethical questions grow. This is about the ethics

social problems in AI deployment.

Some key ethical concerns include:

• Fairness: Are A.I. systems fair to everyone? Are there discrimination by gender, race.

or income?

• Transparency: Do people understand how A.I. is reaching decisions? Are the systems

explainable?

• Refusal of Help: If an AI Assist System fails, whom is‡ at fault—the user, the designer or...?

the company?

• Consent: Do people realize they are being tracked or analyzed by an A.I. tool?

• Autonomy: Are people free to make their own decisions — or are AI systems manipulating them?

There is also what it means to use AI ethically, whether AI can be applied in such a way that it's respectful of human rights, people's dignity and.

freedom. It also ensures many different voices are at the table when both designing and deciding.

3.1.5 Regulation, Governance, and Public Trust

AI has its abilities, but it can also do harm if not properly governed and directed by rules. This is the detail that clarifies why we require

laws, regulations and public participation in AI control.

Regulation The official rules created by governments or organizations to manage the extent of AI development

and used. For instance, some countries have passed laws that protect personal data or limit the use of facial

recognition in public spaces.

Governance means more than just laws, it also encompasses policies, codes of ethics and structures. It

involves:

• Provably safe and fair AI,

• Encouraging responsible innovation,

• Keeping an eye on evolving AI systems over time.

Public trust is essential. People must feel that AI is so being used for their best interests. If AI systems

are opaque, discriminatory or harmful, then people will lose faith in technology. To build trust, developers and

there needs to be transparency, accountability and responsiveness from policymakers around public concern.

This theme gets students to reflect on their own position as future citizens, consumers orgenerators in.

shaping how AI serves society.

3.2 AI in Healthcare

AI is revolutionizing healthcare by enabling doctors, researchers, and hospitals to deliver better, faster and more.

accurate care. It can review medical data, help with surgery, find new drugs and even predict disease

before they happen. But as powerful as AI is, it also raises ethical and privacy concerns,

and patient safety. This section describes use of AI in various health domains.

3.2.1 AI in Medical Diagnosis and Imaging

One of the popular applications of AI in healthcare is disease diagnoses with medical images such as_SEGAN_AI_h a computer tomography (CT) scans, X-ray_imagesFCN_re29eNeSDEg a magnetic resonance imaging (MRI).

X-rays, MRIs, and CT scans.

• A.I. algorithms can quickly pore over thousands of images and detect patterns that doctors may not see.

• For instance, tools with artificial intelligence can help more accurately identify the early signs of disease like cancer, pneumonia or brain injuries

and faster than humans.

• It is also found in pathology, which uses AI algorithms to analyze tissue samples in order to diagnose diseases.

These tools are not substitutes for doctors, but they act like smart assistants that can offer second opinions or flag potential reasons to order tests.

problems for further review.

3.2.2 Personalized Medicine and Predictive Analytics

AI can also assist doctors in crafting individualized treatment plans by considering a person's specific data, including their

genetic risk, lifestyle, environment and medical history.

• In personalized medicine, AI figures out which treatment is most likely to work for each person, rather

of using a one-size-fits-all approach.

Using A.I., predictive analytics anticipates upcoming health issues. For example, by studying patterns

with AI capable of predicting five years in advance, for example, the risk of heart disease or diabetes in a patient's data.

This enables doctors to intervene early, prevent disease and improve long-term health.

3.2.3 AI in Drug Discovery and Clinical Trials

These drugs are expensive and take a long time to develop. AI is also involved in speeding up this process.

• AI pours through vast databases of chemical compounds in drug discover and predicts which of them

may be effective for the treatment of disease.

• Artificial intelligence can also simulate how a drug will behave in the human body, saving time and resources.

the laboratory.

• In clinical trials, artificial intelligence assists in choosing the right participants, tracking their health and parsing results

more efficiently.

This cuts costs and gets life-saving drugs to market more quickly.

3.2.4 Robotics in Surgery and Patient Care

As AI-driven robots begin to enter operating rooms and hospitals, their potential use in MI is immense.

• In procedures involving robots, machines assist surgeons in making extremely precise, minimally invasive terms useState The tools take the form of a HEPA filter (which you also might use for a vacuum); their utility sealed into place: Lumpy and Charboneau's robot has made incredible stridesHe estimates it would cost about $150 Just in time to improve my coping skills for holiday family dinners.

procedures. The robot can make more precise incisions with its arm, which allows for faster healing and less pain.

• AI robots can also be used for rehabilitation, elderly care and emotional The market potential of such humanoid.

health settings.

• Robots, for instance, can help patients walk again after an injury or be short handed by care.

medication and monitoring vital signs.

The goal of these technologies is to enhance patient care and safety.

3.2.5 Ethical and Privacy Issues in Healthcare AI

But, as AI wields sensitive medical data and decisions that impact human health, many ethical

and privacy concerns arise.

• Data privacy: Your health records are the most private things about you. Secure AI systems There is no doubt that AI systems have to be secure, so that

personal data is not theft or its abuse.

• Bias and fairness: If AI learns from biased data (for example, mostly a single gender or ethnic group) it.GetCurrentMethod 3 Main issues / Debate will likely produce biased results.

may take wrong or unfair calls.

• Informed consent: Patients should be aware when artificial intelligence is being used on them and how it functions.

• Accountability: When an AI system provides an incorrect diagnosis, who is to blame — the doctor, the

hospital, or the AI developer?

This topic is meant to raise critical thinking on how we can leverage AI in healthcare, allowing for patient-authored data that are shared by intent.

rights and promotes trust.

3.3 AI in Finance

The Finance System Is Changing As We Know It Because Of AI Banks, investment firms, insurance

companies and fintech start-ups use AI to make decisions, fight fraud, manage risk and better understand their users.

customer service. • AI systems are able to: – process large amount of financial data – detect patterns and predict 9.

predictions, and automate tasks. Yet, the use of AI in finance also raises an ethical question that is associated with

fairness, transparency, and accountability.

3.3.1 AI in Algorithmic Trading and Investment

Respectable investors do engage in algorithmic trading — using artificial intelligence to automatically buy or sell stocks or other financial instruments based on data.

driven strategies.

• AI-based platforms analyze breaking market news and historical trends to inform tradesoграмобщениями и предупреждениями значений для торговых соображений

within seconds.

• These systems can execute trades much faster and more precisely than people, enabling firms to profit

from small market changes.

• AI pops up in portfolio management too, deploying software to generate investment strategies that meet an investor's profile.

person's risk level and goals.

That has streamlined investing, but also complicated and made it more difficult to regulate.

3.3.2 Credit Scoring and Risk Assessment

Lenders such as banks and credit card companies use AI to determine whether they should provide a loan or authorize a line of credit

card.

• Ancient credit underwriting was based on only a handful of factors (such as income and repayment history), but artificial intelligence can.

scan for far more — like online behavior, spending patterns or even social media activity.

• AI is also aiding in risk assessment, forecasting the likelihood that an individual or business will repay a loan.

• Doing so helps banks provide more customized service, and lower their default rates.

But if the data it's fed is skewed or incomplete, the AI could result in tilted outcomes that negatively impact.

certain groups.

Did You Know?

"Did you know that some AI-based credit scoring models don't use traditional financial data like salary or credit history? Instead, they analyze alternative data such as mobile phone usage, online shopping behavior, or even social media activity to assess creditworthiness—especially in countries where many people don't have formal financial records. This is helping extend credit to millions of unbanked individuals."

3.3.3 Fraud Detection and Prevention

AI is incredibly well suited for identifying suspicious financial activities, including fraud and money laundering.

• It studies transactions as they occur and searches for unusual patterns — such as a sudden spate of big withdrawals

or from another country.

• AI systems are learning from past instances of fraud and getting more adept over time.

• In credit card fraud, for instance, AI is able to block or flag transactions within seconds that appear

risky.

It not only benefits customers and financial institutions by saving them material deposits but also enhancing mutual trust.

3.3.4 Chatbots and Customer Service in Banking

Banks and financial companies use AI-powered chatbots to answer customer questions and handle routine

tasks.

• These bots can respond 24/7 to common queries—like checking account balances, resetting

passwords, or explaining fees.

• Advanced chatbots use natural language processing to understand and respond like a human.

• This reduces wait times and operating costs while improving user experience.

Some chatbots are also integrated into mobile apps and websites to guide customers through financial

decisions.

3.3.5 Ethical Concerns: Bias, Transparency, and Accountability

The use of AI in finance needs to be closely regulated to avoid causing harm.

• Bias: An AI system that has been trained on biased data may, for instance, engage in discrimination when deciding credit approvals, insurance

pricing, or fraud detection.

• Transparency: Most A.I. systems are “black boxes” — few programmers, much less nonprogrammer people, can explain how they do what they do.

how they make decisions. Trouble is that in finance, this can be deadly , especially when minor errors)}}

have big impacts.

• AI Responsibility: If an artificial intelligence system makes a terrible decision (to, say, deny someone a loan or not notice that they’re committing systemic fraud) who’s responsible?

case), who is responsible, the bank, the programmer or the algorithm?

AI systems should be fair and interpretable (and — crucially, from my point of view — answerable to the user as well).

regulators.

3.4 AI in Education

AI is transforming education by opening up a world of more personalized and efficient learning. From intelligent

AI is assisting both students and educators from tutoring systems to grading automatically and virtual teaching assistants

in multiple ways. It has also impact on school administrator, curriculum policy and student achievement

analysis. But ethics are significant as AI gains traction in education, too.

consider — especially when it comes to data privacy and profiling students.

3.4.1 Personalized Learning and Intelligent Tutoring Systems

AI supports personalized learning by adapting the content, pace and complexity of lessons according to every

student's needs.

• Systems for personalized learning monitor a student's progress and suggest resources or exercises

using their own pros and cons.

• Intelligent tutoring systems (ITS) are A.I. tools that effectively function as one-on-one tutors. They

elucidate ideas, respond to queries and offer feedback in real time.

• They assist students who learn at different paces or need extra help in certain areas.

That enables students to learn more efficiently and remain invested in the process.

"Activity 1: Build Your Own AI Tutoring

Instructions to Learners:

Use any free AI-based learning platform (such as Khan Academy, Duolingo, or Coursera) that adapts

to your learning pace.

1. Choose a topic you've never studied before (e.g., basic coding, statistics, or a new language).

2. Complete at least 3 adaptive lessons or modules.

3. Observe how the system responds to your answers—does it adjust the difficulty? Offer hints?

Suggest different types of questions?

4. Take notes on how your experience differs from traditional learning.

5. Submit a short reflection (200 words) on:

o What aspects of the AI system made learning easier?

o Did you notice any limitations or biases?

o How might this system benefit students with different learning needs?

3.4.2 AI in Assessment and Feedback

The testing, grading and feedback that are part of reference checks can be automated with AI.

It is able to grade multiple-choice tests, essays and coding assignments in natural language.

processing and pattern recognition.

• AI offers immediate feedback so students can learn from their mistakes and make progress without waiting

for a teacher.

• Among teachers, AI is also cut the amount of time they need to spend doing rote grading and free them up for.other activities.

mentoring.

But AI has to be accurate and unbiased, particularly when it judges creative or qualitative work.

3.4.3 Administrative Automation in Education

AI is also facilitating in administration of routine affairs at schools, colleges and universities.

• It will be able to organize student registration, attendance management, time table generator and exam hallsuite

management.

• AI chatbots can respond to students' questions about deadlines or fees, or course information.

• Schools deploy AI to crunch numbers on which students are excelling, who might dropout and what type of resources

usage for better decision-making.

This minimizes the burden on administration and improves productivity.

3.4.4 Impact on Teachers and Students

AI is revolutionizing the learning experience for both teachers and students.

For teachers:

• AI turns into a support tool that helps with lesson planning, testing and class

management.

• It frees up teachers to concentrate on creativity, emotional support and personalized instruction.

3.4.5 Ethical Issues: Data Privacy and Student Profiling

The use of AI in education poses some very worrying ethical issues, which need to be addressed.

Data privacy: AI systems store significant amounts of student data including learning behaviors,

performance, and personal information. This information needs to be safe and used responsibly.

• Student profiling: AI can profile students by their performance patterns. While this can help

for individualized help it can also potentially be stigmatizing or discriminatory if the instrument is misused.

• Transparency and consent: Students and parents must be made aware of how AI systems are

used and interprets they type of data being collected.

Educational institutions should exercise AI in a manner that keep student rights safe and ensure fairness, equity and inclusivity.

3.5 AI in Space and Defense

Artificial Intelligence is already a significant factor in space exploration and with defense systems. In space, AI

assists scientists and engineers analyzing large amounts of data, fying robots on other planets and improving space missions.

autonomous. Defensive AI is the application of AI to surveillance, threat detection, decision-making and cybersecurity.

But such uses also evoke complex ethical issues, particu larly when the same AI technologies are able to

so they can be used for peaceful or combative purposes.

3.5.1 AI Applications in Space Exploration

AI aids scientists in getting a much better idea of what's going on out there, doing the math on impenetrable datapools and giving them pretty things to look at from their observational data.

decision-making.

• A.I. is applied to process data gathered from space missions, such as signals received from remote planets

and stars.

• For deep-space missions, in which communication lags would prevent real-time control, AI assists.

spacecraft autonomously.

• Mission planning: A I also used in mission planning, where the AI selects best routes for landing spots or targets AI.

exploration.

That has enabled scientists to get more return from missions to Mars, the Moon and beyond.

3.5.2 Satellite Imaging and Data Analysis

Satellites produce enormous volumes of pictures and sensor data from space. AI helps process and analyze

this information quickly and accurately.

• AI can spot geographical features, weather patterns and natural disasters from satellite

images.

• It is employed in climate tracking, agriculture, urban planning and environmental control.

• AI also assists in detecting while unauthorized activities like illegal mining, deforestation or growing military patrols etc.

movements.

This makes it more relevant for both civilian and militarily uses of satellite imaging.

3.5.3 Robotic and Autonomous Navigation in Spacecraft Mission An autonomous navigation of spacecrafts can be plumbed from the motion generation and trajectory tracking for a robotics control problem to successfully accomplish demanded task in space such as rendezvous, docking operation on orbit or landing mission.

Artificial intelligence runs the machines and tools in space missions that allows them to function with hardly any human influence.

control.

• Rovers, like the ones on Mars, employ AI to navigate rough groundautomatically swerving to avoid obstacles and decisions.

without waiting for directions from the home planet.

• AI is employed in docking spacecraft, controlling satellites and operating space stations.

• Robotic arms and artificial-intelligence systems can also do repairs, build structures or grab samples

on other planets.

These technologies help to mitigate astronaut risk and make missions more effective.

Did You Know?

“Did you know that NASA’s Mars rovers use AI to make real-time navigation decisions on Mars without human intervention? Due to the communication delay between Earth and Mars (up to 20 minutes one-way), the rover must independently analyze the terrain, avoid obstacles, and choose paths—all using onboard AI, making them truly autonomous explorers.”

3.5.4 AI in National Defense and Cybersecurity

AI is also utilized in national defense to secure and defend against threats.

• AI systems watch and interpret data from radars, sensors and communication systems to spot (and hash)patterns of behaviouretect

potential threats quickly.

• In cybersecurity, AI and machine learning detect abnormal behavior in computer networks and stop cyberattacks.

• AI can also be employed in surveillance, targeting, mission planning, and `even "controlling"

unmanned drones or autonomous vehicles.

Such systems are intended to react more quickly than human operators, and assist in complicated or hazardous.

situations.

3.5.5 Dual-Use Dilemma and Ethical Implications

AI technology for space and defense is frequently dual-use — capable of being military or peaceful. This is

known as the dual-use dilemma.

• For instance, a satellite-imaging system that is used to check on crops can also be used to monitor those military.

targets.

• Space-repair robots might also be adapted for combat missions, the report said.

This raises grave moral and legal concerns:

• Who governs the application of that technology?

• How can we guard against AI going off the rails in war?

• Does the world need ground rules on how to use A.I. in weapons or surveillance?

These issues raise the importance of prudent development, interna- tional collaboration, and transparent.

rules governing how AI should be used in these sensitive circumstances.

3.6 AI in Other Sectors

In addition to health care, finance education, space and defense, AI is then reshaping many other critical

sectors. It is put to work making crops grow more efficiently, traffic flow in cities and entertainment induced.

content, aid lawyers and save the environment. Applications such as these illustrate how AI is

a powerful tool in most areas of society, supporting for better efficiency and accuracy, and administration.

sustainability.

3.6.1 AI in Agriculture and Smart Farming

Artificial intelligence is aiding the transition to smart farming, so farmers can grow food more efficiently and sustainably.

• Crop monitoring: AI examines data from sensors and drones to assess the health of plants, identifying pest infestations or fungal infe

diseases, and suggest when to water or use fertilizers.

• Prediction: AI can be used to forecast weather, pest infestations or crop yields, assis

farmers make better decisions.

• Mechanization: AI-fueled robots for planting, weeding or harvesting seeds.

with high precision.

These technologies mitigate waste, improve productivity and are contributing to feeding a growing global population.

3.6.2 AI in Transportation and Smart Cities

Indeed, AI is making our transportation systems and cities smarter and efficient.

• Traffic management: AI systems can process traffic data in real time to decrease congestion, manage=_('&')urlpatterns=[

signals, and plan efficient routes.

• Public transportation: AI is used to optimize bus and train schedules, track vehicle locations, aucAdjust Font Size AUSTIN — Greg Abbott assured recipients of HHUSPS they won’t lose the money in their accounts from the state’s troubled program.

predict delays.

• Driverless vehicles: AI powers self-driving cars and delivery drones as they navigate roads, sense obstacles and react to traffic movements.

obstacles, and make driving decisions.

• Smart cities: AI enables energy management, waste collection and emergency response systems

in urban areas.

This translates into safer, cleaner and more liveable cities.

3.6.3 AI in Entertainment and Media

In entertainment industry itself, AI is applied to develop and recommend content which are personalized for users.

• Recommendation engines: Streaming services like Netflix or YouTube use AI to recommend movies,

plays, or the like according to user preferences.

• Content creation: Artificial intelligence tools can write keening music, screenplays and video edits, or make art.

• Audience analysis: Media firms rely on A.I. to analyze what audiences watch and decide what type of

content will be popular.

• Deepfakes and virtual characters: It's possible to make realistic-looking fake videos or lifelike digital

avatars for gaming and film.

These tools enhance entertainment, but also raise wider-spread concerns over misinformation and digitization

manipulation.

3.6.4 AI in Legal and Judicial Systems

AI is there to help lawyers and help access to justice.

• Legal research: AI applications can scour thousands of contracts, case laws and

laws that enabled lawyers to construct cases.

• Document review: A.I. can go through contracts or legal documents to search for risks, mistakes or missing information.

terms.

• Predictive justice: Some systems attempt to predict the result of legal cases based on past judgments.

assisting judges and lawyers in making informed decisions.

• Virtual legal assistants: AI chatbots that can deliver basic legal advice and walk users through straightforward though low-risk agreements.

legal procedures.

Though these are useful tools they must be closely monitored to ensure fairness, objectivity and legal safeguards."

rights.

3.6.5 AI in Environmental Monitoring and Sustainability

AI is also being used to protect the environment and battle climate change through data analysis and upgrades in infrastructure.

sustainability efforts.

• Monitoring the climate: AI analyzes information from satellites and sensors to monitor changes in temperature,

air quality and greenhouse gas (GHG) emissions.

• Disaster prediction: AI can predict natural calamities including floods, forest fires and earthquakes,

helping communities prepare in advance.

• Energy management: Smart grids, which employ AI to balance the supply and demand of electricity and cut waste,

and support renewable energy sources.

• Wildlife conservation: AI is used to track endangered species, root out illegal poaching and control population of pestilent or predator creatures by introducing more typical ones where people live.

natural habitats.

3.7 Socio-Economic Implications of AI

AI is remaking economies, industries and the very fabric of society. It creates efficiencies and gives us new tools, but it also

breaks down established channels of employment, learning and leadership. In this section we investigate the influence AI has on

and levels of employment, distribution of wealth, access to technology and global power relations. It highlights the

need to oversee the development of AI for everyone in a fair and sustainable manner.

3.7.1 Impact on Employment and Job Displacement

The effect of AI on jobs is one of the most hotly discussed aspects.

Automation: Routine activities can be executed more rapidly and accurately by AI systems compared to humans. This

can result in job loss in industries such as manufacturing, retailing, consumer service and driving.

• Job transformation: Not all jobs will vanish, but many of them will be reconfigured. Workers will need

to acquire the skills to manage AI tools or oversee automated systems.

• New possibilities: Meanwhile, AI is generating new jobs in disciplines such as data science, machine

learning (ML), robotics for maintenance, and AI ethics.

This transition will need to be carefully planned for workers to be supported during the transitions and retrained for opportunities in the jobs of the future.

of the future.

3.7.2 Economic Inequality and the AI Divide

The more AI advances, the wider the gap between people who benefit from it and those who don't.

• Availability of AI: The most advanced AI is frequently controlled by large tech companies and rich countries.

systems, whereas poorer regions may not have the infrastructure or expertise to employ AI well.

• Wealth concentration: Firms that utilize AI to cut costs and boost productivity might be able to gain

greater profits, although employees' jobs and income may be timed.

• Digital divide: The difference between those who have digital access (to internet, devices and skills) and

those who do not grows wider with AI.

The AI revolution risked exacerbating social and economic inequality worldwide without inclusion policies.

3.7.3 Changing Skill Requirements and Workforce Transformation

AI is disrupting what it takes to succeed in the workplace.

• Demand for new skills: The demand for data, coding, machine learning and other know-how is also much elevated.

critical thinking, and digital communication.

• Decline in repetitive tasks: Jobs involving routine forms of aesthetics since 1900 Sources: "The}ofcssc — Creativiscfi : Generative Visual and Dramaturgical....inionsJobs / News asleep at the switch.

supported by AI systems.

• Lifelong learning: The need for workers to steadily update their skills will be more intense than ever.

Classical Education may not suffice.

Governments, companies and educators should collaborate to offer training or reskilling programs and

flexible learning opportunities.

3.7.4 Social Inclusion and Accessibility

AI can be a force for social inclusion if used responsibly.

• Adaptive aids: AI devices can also improve life for people with disabilities, such as by enabling speech-to-text, smart glasses or custom virtual assistants.

prosthetics, or navigation aids.

Language translation: AI is able to help break down walls with language in education and communciation, as.

folks from various places join and share.

• Services access: AI can streamline citizens to get healthcare, legal and government services for fewer time and more efficient.

those in rural or underserved areas.

But if not built inclusively, AIsystems can also be trained to overlook or alienate specific groups, perpetuating

existing inequalities.

3.7.5 Global Power Dynamics and AI Sovereignty

AI is also shaping world politics and the balance of power.

• AI sovereignty: Nations are vying to create their own.AI technologies rather than depending

on foreign platforms. This is a question about control of data, security and innovation.

• Geopolitical competition: If countries with powerful AI abilities pull ahead, they could obtain economic and military advantages from the technology.

diplomatic advantages over others.

• A.I. as soft power: Some countries employ artificial intelligence to win over others with surveillance technology, propaganda or a less threatening film industry.Stephen Chen from Hong Kong contributed reporting.

control, or technology partnerships.

These realignments create opportunities for international cooperation, regulation and dialogue to make sure that AI is.

put to the cause of peace and global good.

Knowledge Check 1

Choose the correct option:

1. What is the main ethical concern when AI systems are trained on biased data?

A. Slow performance

B. High cost

C. Discrimination in decision-making

D. Increased transparency

2. Which of the following is an application of AI in healthcare?

A. Personalized advertising

B. Predictive diagnosis of diseases

C. Automated tax filing

D. Content moderation on social media

3. In education, AI is commonly used for:

A. Student entertainment

B. Physical classroom security

C. Personalized learning paths

D. Traditional grading only

4. AI in algorithmic trading mainly helps by:

A. Replacing human CEOs

B. Printing financial reports

C. Making data-driven investment decisions faster

D. Offering discounts on stocks

5. The dual-use dilemma refers to:

A. Using AI for sports and gaming

B. Technology that can be used for both civilian and military purposes

C. Developing two versions of AI for different users

D. Programming in two languages

3.8 Summary

Artificial Intelligence (AI): AI was featured as one among the transforming forces across ACTIVSVMANYFRONTIERS In this document it will be widely used in problems with numerous solutions such as pollutant trading, scrapping of transportation means or demand management.

of modern society. Its impact cuts across sectors such as health, finance, education, defence and agriculture.

transportation, and more. AI has several advantages such as increased efficiency, accuracy and.

access to services. it allows for personalized learning, rapid medical diagnosis, intelligent city planning and

predictive financial analysis. But AI also comes with real ethical, social and economic questions.

These range from job disruption and data privacy to algorithmic bias, inequality and international power.

imbalances.

❖ AI technology is being developed, and it is necessary for people, organizations, and countries to take up responsibilities as it matures.

The above brings us back to the consequences of Pairing on a graduate curriculum for cryptography which made no assumptions as to who was using Cat and Cupiola (if anybody), why they had done so, or what further operations were applied. This involves building ethical

frameworks, building skills, developing inclusive regulations, and facilitating international collaboration.

Understanding AI's societal implications is key to making certain we use AI for the benefit of humanity in an ethical and sustainable way.

manner.

3.9 Key Terms

AI: Machines or Software Systems being capable of doing tasks that would normally require human beings.

human intelligence--learning, problem-solving, and decision making.

Machine Learning (ML): A subset of AI that teaches a system to learn and make decisions by itself, based on previous experiences.

without being explicitly programmed.

Automation: The process of doing something automatically, with technology taking over tasks that used to be done by humans.

Bias in AI: When an artificial intelligence system spits out unfair results because of biased data or poorly written algorithms.

Predictive Analytics: Applying data and statistical models to forecast future trends.

Ethics in AI: The area of study that examines the moral responsibilities and implications of using AI systems within

society.

Smart Cities: Cities that use technology, including AI, to optimize resources and services.

infrastructure efficiently.

AI Sovereignty: The idea that countries can and should build and control their own artificial intelligence.

national interests.

Digital Divide: The separation of individuals who have access to digital technology and those that do not.

Dual-use Technology:Technology that can be used for both civilian and military purposes.

3.10 Descriptive Questions

Discuss the impact of AI on the healthcare industry with a few examples.

Determine the positive and negative effects of society in AI.

How is AI implemented in finance industry for fraud detection and credit scoring?

Discuss the use of AI in education, specifically related to personalized learning and assessment.

What are the ethical issues related to using AI for national defense?

Discuss the idea of AI contributing to economic inequality. What can we do about it?

How is AI helping sustainable development and environmental monitoring?

What's the double-edged sword of dual-use dilemma of AI technologies?

Explain how AI is reshaping the skills people need in the economy.

How should we approach the responsible and inclusive deployment of AI?

3.11 References

1. Russell, S., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach. Pearson Education.

2. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

3. European Commission. (2021). Ethics Guidelines for Trustworthy AI.

4. McKinsey & Company. (2020). The State of AI in 2020.

5. World Economic Forum. (2021). Global Technology Governance Report.

6. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence.

7. Stanford University. (2021). AI Index Report.

8. Nature and Science Journals – Various articles on AI in healthcare, education, and climate science.

9. Government of India – NITI Aayog reports on AI for All.

10. OECD. (2021). AI Principles and Recommendations.

Answers to Knowledge Check

Knowledge check 1

1. C. Discrimination in decision-making

2. B. Predictive diagnosis of diseases

3. C. Personalized learning paths

4. C. Making data-driven investment decisions faster

5. B. Technology that can be used for both civilian and military purposes

3.12 Case Study

AI for Inclusive Learning – A Case from Rural India

Background:

Quality education for all students is a challenge.

teacher unavailability and lack of school resources. A nonprofit was started in 2022

a learning platform for students in grades 6-10 that is driven by AI. The system used local

languages and worked off-line, so that children could learn lessons in subjects such as mathematics, science and

English without internet connectivity.

AI Application:

The platform employed machine learning to assess students' responses, and adjusted the material in response to.

learning levels. It offered instant responses and suggested more activities\ldblquote for every.

student's strengths and weaknesses. Teachers got a weekly readout of where students were succeeding or struggling, which to.

determined who required better sorting.

Impact:

• 40 percent jump in student engagement.

• Average test scores are rising for key subjects.

• AI was reported to save time on grading and enable teachers to focus more on mentoring.

Challenges:

• Protecting student profile information.

• Teaching teachers to trust and use the A.I. system productively.

• Scarce funds to scale the program.

Conclusion:

It is a case in point of the way A.I. can help to bridge educational divides in at-need areas. It highlights the

need for flexible and enabl ing major global Even if the AI. INCLUSIVE, LOCALIZED AND USER- CENTRED AI TOOLS leaving no one behind!

underserved communities

Ethics in Artificial Intelligence_V3_Unit 4.docx

Ethics in Artificial Intelligence_MBA_2

Ethics in Artificial Intelligence_MBA_2

ATLAS SkillTech University

Document Details

Submission ID
trn:oid:::3618:127363104

Submission Date
Feb 2, 2026, 1:34 PM GMT+5:30

Download Date
Feb 2, 2026, 3:39 PM GMT+5:30

File Name
Ethics in Artificial Intelligence_V3_Unit 4.docx

File Size
36.8 KB

23 Pages

4,924 Words

29,626 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 4: Privacy and Surveillance

Learning Outcomes

Learn how AI technology processes, analyses and uses personal data in different situations.

Describe the major data privacy related issues with AI systems.

Examine legal and regulatory regimes for AI and surveillance.

Consider how AI improves surveillance in the public and private sphere.

Discuss the morality of applying AI to mass surveillance and tracking.

Hear about real-life cases demonstrating the effects of AI enabled surveillance.

Think about the balance between security, privacy and ethical responsibility when it comes to AI implementation.

Content

4.0 Introductory Caselet

"The Digital Mirror: Riya’s Reflection on Privacy in a Connected World"

Background:

DELHI: Delhi-based college student Riya is excited about the learning app that her college recently introduced -- An AI powered mobile application.

introduced. It watches over her study habits, suggests learning materials and helps keep her on track with supportive reminders.

And at first, she can’t believe how well it seems to know what she wants.

Then a few weeks later, she notices something uncanny — the ads in her social media feed align with what she’s just put out.

reading for class. A friend pops up to let you know the app’s privacy policy includes sharing its use data with

partner platforms for “personalized experiences.” Riya becomes curious.

She begins researching how A.I. systems receive and use data. And she learns that all of it, from her whereabouts

and voice commands to her browsing are being fed through algorithms. Some of this helps personalize

her past, but she knows how much has been stored, shared or sold — much of it without her consent and often without her knowledge.

Her excitement turns into concern. Does convenience mean control of her?” Riya muses.

own digital footprint.

Critical Thinking Question:

How do AI-driven tools encroach on personal privacy, and what are the obligations of companies

in protecting user data?

4.1 AI and Data Privacy

Artificial Intelligence needs data to actually work. AI technology are ”data-driven" they collect, analyze and get educated

large volumes of data to automatically predict, make decisions or recommend. However, this

dependence on data raises significant privacy concerns. The more data that is gathered from the users themselves,

questions about how that data is stored, who has access to it or how it's being used. This section focuses

around AI and its relationship to data privacy and what can be done about it.

4.1.1 Introduction to Data Privacy in the Age of AI

Data privacy is the right of an individual to decide what personal information they are willing to give to whom and under what conditions.

and shared. In the age of A.I., this has gotten more complicated.

AI systems often work in the background — in apps, on websites or devices — and tinker with data without your express consent.

even realizing it. So, for example voice assistants such as Alexa or Google Assistant are recording commands and

can track fitness stats. AI uses this information to increase its accuracy and customize_user

experiences.

But there is a very good chance that you are still sharing such data with third parties, using it to market to you or keeping it indefinitely.

As the saying goes, users often are not well-versed in what they're agreeing to sign away when they accept terms and conditions. This

complicates the efforts to protect privacy in a world where A.I. is increasingly part of nearly every digital interaction.

4.1.2 Nature and Types of Data Collected by AI

There is no limit as to what categories of data AI systems can collect, depending on the applications. These types can be

grouped into several categories:

Personal Details – for example, including name, age, gender, address or details how to get in touch with you.

Behavioral Data – browsing history, app uses, online purchasing activity and viewing activities.

habits.

Biometric Information — Finger, retinal or Iris scans; voiceprints; or scans of hands, face geometry or walking gait.

patterns.

Location Data — G.P.S. data that shows where a user goes or lives.

Health Data – Heart rate, sleep data, medical history and other health information accumulated in dating from.

health devices.

AI collects this data from devices, sensors, applications and the internet. Some of it is shared

from users and others* are harvested from them without their effective permission.

4.1.3 Data Ownership and Consent

One of the most meaningful concerns around AI and data is one of ownership — who does data belong to when it's accumulated by AI?

Ideally, individuals would cherish their personal data and dictate the terms under which it gets used. In practice, many

companies own and are free to use once you agree to terms of service. This creates a

situation where user data becomes uncontrolable.

Consent is the notion that users have agreed to be tracked and otherwise identified. But in many cases:

• Consent gets lost in long legal documents.

• Users are not offered clear-cut choices.

• Information is collected even after users withdraw permission.

That raises not just ethical but legal questions.) Adequate consent needs to be informed, specific and voluntary not .

just a formality.

4.1.4 Risks of Data Misuse and Breaches

There is always potential for misuse or breaches of data when AI systems process personal information.

Data misuse occurs when:

•3Use of data for unauthorised purposes.

• Advertisers buy it without users' knowledge.

• It can be deployed to influence people's choices — for instance, in political campaigns or once practices like targeted ads are used.

If you are affected, it means hackers have accessed personal data stored by companies. This can lead to:

• Identity theft

• Financial fraud

• Exposure of sensitive health or personal information

AI systems are frequently attacked due to the huge amount of valuable data they have. If these systems are not

not handled correctly, the ramifications can be severe for both individuals and businesses.

4.1.5 Privacy-Preserving AI Techniques

In order to prevent people's data from being exposed, AI researchers and developers are developing new methods for AI to operate without_PAYLOAD.

compromising privacy. These are known as privacy-preserving AI techniques, and among the most

common include:

Data De-identification – Personal identifiers (e.g., names or addressees) are deleted so as to make it impossible for the data parties mundo privado_data.pdf / LD18_domanshi grier_2 satish lella_9781680455789-49 86 PP1 be identified, directly or indirectly.

be traced back to individuals.

Differential Privacy — Adding noise or perturbations to data before the AI processes it so that individual

users can't be identified.

Federated Learning – A technique in which AI is trained on user devices (such as phones).

transmitting the data to a central server. Sharing only the learning outcomes.

Encryption Protecting data in a way that makes it readable only by authorized users or systems.

Access Controls – Restrictions the ability to see, use or change data.

These approaches make it possible for AI to continue learning and getting smarter without threatening personal privacy.

They are becoming increasingly relevant as privacy laws and public consciousness continue to expand.

4.2 Legal and Regulatory Frameworks

As AI systems begin to utilize pfsersonal data, l the uses ascribed to psersona andl nev PSrs stay true hpaboplho thpeba CFFPS most cases.

protect individuals' privacy. Much of the world is in the process of enacting laws and systems to make sure that

we collect and utilize data responsibly. With the help of these regulations, rights and obligations for individuals have been established on the basis of +#+9OURNAL DANTA)AND.ON.....(a journal of): economy justice rule law Glob.

firms, and charts penalties when companies fail to comply. This section also highlights how legal systems are adapting to 31 these normative changes.

4.2.1 Overview of Global Data Protection Laws (e.g., GDPR, CCPA)

Several countries and territories have enacted data protection laws that affect how individuals' personally identifiable information is used, stored, and disclosed.

acquired, used and stored — particularly when it comes to artificial intelligence.

Some key examples include:

• G.D.P.R. (General Data Protection Regulation) -- One of the heavy laws in Europe. It sets strict

regulations about how data should be used, requires clear permission from companies and gives users strong rights to their data."

personal data.

CCPA (California Consumer Privacy Act) - A law on the books in California, USA. It allows users to know

what data is being gathered, ask for it to be deleted and choose whether to stop data sales.

• Digital Personal Data Protection Act (DPDP) 2023 of India: A new Indian law that gives..

citizens' control of their digital data and establishes obligations for companies to safeguard it.

These laws are designed to protect transparency, accountability and privacy of individuals in a digital.

age.

4.2.2 Rights of Data Subjects

The data subject is defined as any individual that data is being collected on or processed.

Based on contemporary data protection laws you as an individual are given certain rights:

Right to Access – Users can request which features of their data a company stores.

Correction Right – The data can be corrected on request of the user.

Right to Erasure (Right to be Forgotten) – Individuals can request their data to be wiped out as per

certain conditions.

Right of Data Portability – You have the right to request and receive your data in a common format that can be moved to ResponseEntity.

another service.

Right to object, or restrict processing – People can say no or limit how their information is used.

Right to be Informed: You need to know how someone is going to use your data before they collect it.

They should be the rights that give users power to control their personal information.

4.2.3 Obligations for AI Developers and Companies

Legal and ethical considerations all AI developers as well as some personal data-collecting or -processing organizations can be subject to various legal and ethical obligations orbit around the development of AI.

responsibilities:

Transparency — Provide clear explanations of how data is gathered and used in AI systems.

Lawful Basis for Processing – You may process only if you have one of these lawful bases (a) User consent

public interest.

Minimize-data – Only collect as much data as necessary.

Security – Use encryption, firewalls and keep it all in a locked safe to make sure you don't let data leak or have other security bakebrainedary measures.

breaches.

pact Assessments – Consider the potential harms of AI systems, which can still present risks even when trained on pseudo- anonymized / aggregated data.

sensitive information.

Fairness and Non-Discrimination Ensure that AI systems are not/ un-biased or unfair.

outcomes.

Failure to fulfil these responsibilities may invite legal action, as well as damage to reputation.

4.2.4 Enforcement Mechanisms and Penalties

The governments and regulators, they have mechanism of enforcement to see that the rights of Data privacy are enforced.

laws are followed.

Regulators — Independent agencies (such as the European Data Protection Board or India's Data

Safeguard Board) to regulate the enforcement of standards and such matters as complaints.

• Penalties and Fines – Violation of any is costly to a commercial enterprise. Under GDPR, companies

could face penalties of up to 4 percent of their worldwide annual sales.

• Audits and Investigations - Regulators can audit or investigate the way a company does business, looks at 0 products, etc., combined with their right of access to books, records, files and physical facilities is very powerful.

demand changes if needed.

• Court Activity Individuals and organizations can in some cases sue companies for misusing their data.

Such enforcement mechanisms are supposed to keep companies honest and privacy laws intact, but they'

ignored.

4.2.5 Limitations and Challenges of Current Regulations

Fast Paced Innovation – AI is advancing so quickly that laws cannot be adapted in time resulting in "the absence".

regulation.

Complexity of AI Systems –The AI decision-making is often hard to be understood, and that impacts the decisions made by AI.

transparency and accountability.

Enforcement Challenges – The authorities may be unable to oversee or don't have the skills to scrutinise

every company or algorithm.

What consent loopholes – So many user's click accept on terms and conditions without reading them so that is no longer considered a legal contract?

informed consent is not necessarily fully informed.

Cross-border Data Flows: Data frequently crosses borders and that process raise questions about – 57 Although some very esoteric commercial data may, in fact, need protection to stimulate investments or economic growth when the data itself is not universally available (e.g.

which country's law applies.

These restrictions reflect the to improve international cooperation, increasing public awareness and continuous support of legal restrictions.

AI technologies are still evolving at breakneck speeds.

4.3 Surveillance Technologies

Surveillance technologies Surveillance technologies are gadgets as well as methods used to view people's actions, movements, and their conduct.

environments. Surveillance is more efficient and automated with A.I. From

face recognition at the airport for security, tracking your whereabouts in mobile applications — government officials, businesses and venture capitalists agree that

legal-murkiness is secondary to good technology. And they can help things to be safer and

order as well as privacy, abuse and human rights.

4.3.1 Introduction to Modern Surveillance Systems

Common components include:

• Cameras and sensors in public and, increasingly, private spaces.

• AI programs that recognize faces, voices or movement.

• Databases that contain personal information, including identity details or travel history.

• Systems networking other systems in cities or even countries to keep tabs on larger developments.

The surveillance is widely applied, such as in law enforcement, border control, airport security, traffic

monitoring, and public health. Traditionally surveillance was done simply by human observers Now days more and were made with animal involvement [Gao et al.

employ machines that can scan, identify and even react on their own.

4.3.2 Facial Recognition and Biometric Monitoring

Facial recognition is an application of artificial intelligence and can be defined as any technology that uses AI to analyze facial attributes of people.

How it works:

- A face is photographed by a camera.
- An AI compares it with a database of images.
- If it yields a match, the system can confirm the person's identity.

There is more to biometric monitoring than the face. That includes fingerprints, voice and iris patterns, even

body movement. These systems are used in:

• Airports for identity checks.

• Smartphones for unlocking devices.

- Police to help find suspects.
- Workplaces tracking who is present or checking in on how employees are behaving.

Although helpful, such systems can violate privacy, particularly when it is conducted without people's prior consent.

consent.

4.3.3 Location Tracking and Geospatial Surveillance

Location tracking is the practice of collecting information about where a person is or has been, as well as his or her computer-perceived location.

This is done using:

• GPS on smartphones and in cars.

Wi-Fi and Bluetooth signals.

• Geotagged social media posts.

AI analyses of large-scale mobility patterns across cities, regions or entire countries are also seeing a surge in interest.

countries. For example:

• Governments could deploy it to observe public gatherings.

• Businesses can use it to analyze consumer foot traffic.

• Health agencies did use it during the pandemic to follow the virus.

The worry is that too much location tracking can erode anonymity and limit freedom of movement.

4.3.4 Predictive Surveillance and Behavioral Analytics

Predictive surveillance uses AI to forecast possible actions or threats before they occur by interpreting

behavior patterns.

It works by:

• Harvesting data from surveillance cameras, social media posts, purchases and browsing activity.

• Spotting suspicious or uncommon behavior after the fact.

• Transmitting alerts to authorities or systems for action.

For example:

• A system could identify someone making multiple trips to sensitive sites.

• A program could monitor social media posts to forecast disturbances or demonstrations.

In NASA's case, there was not so much that could go wrong, apart from the risk of the project drawing in money and other resources that will be needed for research. Predictive systems like these are designed to prevent crimes or violence, but they can also single out people who don't deserve it—or measurements unclearly indicate which people have been selected.

communities, particularly when the data is misinformed or misconstrued.

Did You Know?

"Did you know that predictive surveillance has been used in shopping malls and retail stores to analyze customers' walking patterns, time spent near certain products, and facial expressions—to predict what products they're likely to buy? Retailers use this to optimize product placement and personalized advertising, often without customers realizing they are being monitored."

4.3.5 Mass Surveillance vs Targeted Surveillance

Surveillance can be conducted in various manners, depending on its scale and purpose.

Mass surveillance:

• Requires continuous surveillance of broad populations.

• Often involves public spaces, use of the internet or communication.

• Considerations: city wide CCTVs, internet traffic monitoring.

Targeted surveillance:

• Targets individuals, groups or locations of interest based on suspicion or evidence.

• In many countries, a legal obligation or justification is necessary.

•Exemplars: trailing a murder suspect, watching over high-security precincts.

Mass surveillance is more contentious since it endangers innocent individuals, may involve continuousllu monitoring.

and to cause control or panic in the society.

4.4 Ethical Issues in Surveillance

AI surveillance, as it becomes more common, raises profound ethical issues. While surveillance

can be used to increase safety, help us solve crime and manage public spaces — but there is also a possible peril of over reaction, violative of core rights."

such as privacy, freedom of speech and equality. The ethical debates surrounding surveillance are at root a matter of justice,

Society, consent and transparency vs. repercussions. The following section discusses the ethical dilemmas that. setVerticalGroupbe tackled.

weigh when deploying surveillance tech.

4.4.1 Privacy vs Security Debate

One of the oldest and toughest questions in surveillance ethics is: Should we surrender some privacy PlayerPrefs. getTable?

to gain more security?

Supporters of surveillance argue that:

• It deters crime, terrorism and violence.

• Security measures protect society and save lives.

• It's the government's job to keep law and order.

Opponents argue that:

• Permanent watching is an assault on the privacy of individuals and the bonds of a protective "watching" society.

• It will be open to easy abuse if not properly regulated.

• People have the right to a life in which Big Brother is not watching them all the time.

There's no easy formula for how to ensure that the balance is right between privacy and security — it's a matter of circumstance, policy and so on.

cultural values.

4.4.2 Chilling Effects on Freedom and Expression

The chilling effect is the idea that when people know that they are being watched, it can change or curb their behavior.

they feel as if they are being monitored.

For example:

• People may fear participating in demonstrations, political meetings or religious services.

• Journalists or activists may grow afraid to speak freely or share sensitive information.

• Some young people might be too afraid to google certain subjects.

Even if there is no damage, the perception of being watched can stymie openness, creativity and democratic propensities.

participation. This can have far-reaching harm for a free and open society.

Did You Know?

“Did you know that in some countries, public Wi-Fi networks in parks and universities are used to monitor search history and social media activity, which has caused students and citizens to avoid discussing political topics online? This is an example of the chilling effect, where the fear of surveillance changes how people behave—even if they’re doing nothing wrong.”

4.4.3 Discrimination and Profiling

Data Also for AI-supported surveillance systems, discrimination may be unintentional (or intentional).

This happens when:

• AI models learn from biased data that mirrors, for example, racial, gender or cultural prejudices.

• Poor areas or minority communities rely on surveillance systems more heavily.

• Some groups are erroneously singled out because of their appearance, language or behavior.

It is particularly dangerous when done by predictive policing or airport security. It can result in unfair

torture, false charges and distrust in the authorities.

Surveillance must be ethical and not reinforce existing inequities.

4.4.4 Informed Consent and Transparency

Many are not even aware they’re being watched, or that their data is being stored.

Ethical concerns include:

• No informed consent: People need to know that they are being monitored, how they are being monitored and why.

Secrecy: Most surveillance organisations conduct their affairs in secret and a clear picture of what is collected would be hard to come by even if they were not the de facto nature of clandestine organisations.

oversight.

• Sneak data collection: An app, website or camera can have had access to your own information without giving you a loud warning.

Ethical surveillance is visible, with obvious procedures and public awareness of capability and well-publicized ways to opt-out.

wherever possible.

4.4.5 Balancing Public Interest and Individual Rights

Surveillance should be in a manifest public interest, such as thwarting harm, protecting health or preserving privacy.

security. But not at the expense of individual rights, including privacy, free speech or equality.

Key ethical questions include:

• Does the surveillance really need to be done, or are there less-invasive alternatives?

• Are the benefits outweighed by the risks or harm?

• Who adjudicates what the public interest is?

• Is there any safeguard to prevent abuses?

Ethical policy-making is answerable, transparent and involves the public. Surveillance should

always proportionate, legal and necessary.

4.5 Case Studies in Surveillance

Surveillance technologies are becoming ubiquitous in various domains such as law enforcement, health care, transport etc.

employment. Case studies illustrates how the systems work in practice and what

concerns they pose for privacy, civil liberties and governance. Every one of these leaves an open question around consent,

bias, accountability, and public interest.

4.5.1 Case Study 1: Surveillance in Public Spaces

Example: Facial Recognition in London

In London the police have used Live Facial Recognition (LFR) on crowded streets in areas such as:

train stations and shopping streets. Cameras are scanning pedestrians' faces and matching them against "watchlist" photos from law enforcement in the area.

wanted individuals.

Issues Raised:

• The majority of people are scanned without their knowledge or consent.

• There are concerns about false positives, where innocent parties get misidentified.

4.5.2 Case Study 2: Workplace Monitoring and Employee Tracking

Example: Amazon's Employee Surveillance

Some Amazon warehouses track workers via wearable devices and A.I.-powered systems that.

monitor:

• Task completion rates

• Break times

• Body movements and location

Worker sleep detection alerts based on inactivity or speed are also automatically created.

Issues Raised:

• Opacity about data collection and usage.

• Pressure and stress on workers being monitored around the clock.

• Little control or opportunity for employees to question automated decisions.

This is an example of how surveillance affects the well-being and dignity of employees.

"Activity: Design a Surveillance Policy for a Workplace"

Instructions to Learners:

Suppose you are an HR manager at a transportation and logistics company that intends to implement an employee-focused AI assistant for their employees.

monitoring system.

Compose a workplace monitoring policy including:

o What information is being collected (e.g. where and what makes use of it)

o The manner in which consent will be obtained

o How long the data will be retained

o Who has access to the data

o How privacy of employees will be safeguarded

Ensure that your policy strikes the right balance between efficiency and transparency, while remaining respectful to employees."

rights.

Send your policy to us (300–400 words) and a one-paragraph justification for how it would help you.

safeguard the interests of both the company and dignity.

4.5.3 Case Study 3: Smart Cities and Real-Time Monitoring

Example: Smart Surveillance in China's Shenzhen City

The city's use of smart city technology including traffic cameras, facial recognition and monitoring devices for noise and air - pollution has been criticised in the past.

time sensors are used to:

• Track jaywalking

• Manage traffic flow

• Monitor public behavior

• Issue fines or penalties via digital means

Issues Raised:

• Surveillance is conducted with a larger amount of secrecy from the citizen.

• Its critics say it polices social and intellectual control, while curbing dissent.

4.5.4 Case Study 4: Surveillance During Public Health Crises

Don't miss: Order your free copy of Seoul Man now Illustration: COVID-19 Contact Tracing in South Korea

SOUTH KOREA USED DATA FROM CREDIT CARDS, CCTV AND MICT DURING THE COVID-19 PANDEMIC -AI

phones that track the infected as they move around. The system worked pretty darn well for controlling virus.

spread.

Issues Raised:

• People's movements were charted, sometimes in unpublic places with no names.

• There were questions about how long the data would be retained and who would handle it in the future.

• Consent was provided in an emergency.

The present case brings to light the moral dilemma between public health and personal privacy.

in times of crisis.

4.5.5 Case Study 5: Predictive Policing and Minority Communities

Example: PredPol (US based)

Some U.S. police forces also used PredPol (Predictive Policing software) to predict where crimes

could happen according to historical crimes figures. These are the areas where the police deployment was increased.

Issues Raised:

• The software frequently over-policed minority neighborhoods, resulting in racial profiling.

• Critics said the system perpetuated existing biases in the criminal justice process.

• There was little transparency around how the algorithm functioned and about how decisions got made.

This use case illustrates how biased data and lack of monitoring can result in discrimination and loss of public confidence.

Knowledge Check 1

Choose the correct option:

1. Which of the following best describes federated learning?

A. Storing all user data in a central cloud system

B. Sharing all user data with third-party advertisers

C. Training AI models locally on devices without transferring data

D. Using face recognition to unlock devices

2. Under the GDPR, individuals have the right to:

A. Sell their data to the government

B. Delete their data from a company's records

C. Be monitored at all times in public spaces

D. Share other people's personal data freely

3. Predictive policing tools are criticized mainly because they:

A. Are too expensive to operate

B. Help reduce crime in urban areas

C. May reinforce racial or social bias in policing

D. Require too many human workers

4. Which of the following is an example of biometric surveillance?

A. Tracking internet search history

B. Reading a user’s emails

C. Scanning a person’s iris for identity verification

D. Monitoring electricity usage in homes

5. What is a primary ethical issue associated with mass surveillance?

A. High software costs

B. People gaining too much freedom

C. Loss of individual privacy without consent

D. Difficulty in installing cameras

4.6 Summary

❖ In the age of digital, AI equipped surveillance systems are bec0oming more and m0ore prevalent in different spheres of life.

Whether it's facial recognition on the streets or employee monitoring in offices, AI power s real-time encouragement or punishment.

pressure for data collection and analysis on a scale hitherto unknown. While these technologies promise efficiency,

predictive accuracy), they also present serious privacy, human rights, safety and security concerns.

discrimination, and consent.

❖ Legal regulations and standards like the GDPR, CCPA seek to limit usage of data, giving rights to citizens,

and impose obligations on organizations. But the pace of progress in AI is frequently faster than established

laws, creating regulatory gaps. When surveillance impacts personal freedoms, ethical dilemmas emerge.

targets the marginalised, or acts not openly.

4.7 Key Terms

AI Surveillance – Utilizing artificial intelligence to observe, track and analyse the behaviour of people brackets around it.

environments.

Data Privacy – Personal freedom of how their personal data collection, storage and release is managed.Slfaju) and released.

shared.

Biometric Data Personal information about an individual s physical traits or behaviors, such as fingerprints or facial recognition.

face, or voice.

Predictive Monitoring/ Surveillance – AI based systems that can identify patterns predicting actions or possible threats.

Opt-in/Voluntary consent — An individual's active agreement to participate in data collection and storage activities.

GDPR – General Data Protection Regulation; regulation on data protection and privacy for European Union residents.

Union.

Mass surveillance – Observation of whole population or a substantial fraction thereof of persons, sometimes in regard to all people under control of a given authority.

Chilling Effect - The detriment to the exercise of individual rights when individuals are deterred from exercising those rights by a government's surveillance of their lawful activities.

Profiling – Any form of automated processing of personal data used to assess a contrast profiles.

Transparency – Sharing how systems are working, what data is being collected and what it's used for.

4.8 Descriptive Questions

Describe the use of AI-Technologies in current Surveillance Systems.

Reflect on the ethical balance between privacy and public-security in AI surveillance.

Explain the main elements and objectives of international data protection laws, like GDPR and CCPA.

How does biometric surveillance operate and what are the dangers?

#Inquiry Analyze the idea of Informed Consent when it comes to AI Data gathering.

Distinguish between mass surveillance and targeted surveillance with examples.

Explain the tension between surveillance during pandemics and people's privacy rights.

What role should AI developers and companies play in safeguarding users' data?

In what way can anticipatory surveillance lead to discrimination?

Consider the effects of surveillance systems on freedom of speech and expression.

4.9 References

1. European Union (2016). General Data Protection Regulation (GDPR).

2. California State Legislature (2018). California Consumer Privacy Act (CCPA).

3. Zuboff, S. (2019). The Age of Surveillance Capitalism. PublicAffairs.

4. UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence.

5. Amnesty International (2020). Surveillance and Human Rights Reports.

6. MIT Technology Review. (2021). The Rise and Risks of Predictive Policing.

7. Electronic Frontier Foundation (EFF). Surveillance Technologies and Privacy Rights.

8. Future of Privacy Forum. Privacy and AI: Legal Trends and Policy Guidance.

9. Wired Magazine. How Smart Cities Use Surveillance Technologies (2020).

10. World Economic Forum (2021). Global Technology Governance Report.

Answers to Knowledge Check

Knowledge check 1

1. C. Training AI models locally on devices without transferring data

2. B. Delete their data from a company's records

3. C. May reinforce racial or social bias in policing

4. C. Scanning a person’s iris for identity verification

5. C. Loss of individual privacy without consent

4.10 Case Study

AI Surveillance in Public Transport Systems

Background:

Singapore metro puts artificial intelligence driven surveillance into operation in the first week of shutdown The world’s laziest lockdown comes to pass_MAYPOLE, cock and an Astra!? spinning plates_High Arbitration_Chilling type Singapore’s public transport authority has flipped the switch on an AI powered surveillance system a week into lockdown.

monitoring station for crowd flow, loitering behavior and the abandoned objects. The system

deployed facial recognition, heat mapping and behavioral analytics to optimize passenger flow and

enhance security.

AI Application:

• Cameras equipped with facial recognition technology watched over entrances and exits.

• Algorithms spotted anomalous behavior including long-term dwellers or movement in reverse

through gates.

• Tactical response teams were coordinated with information in real time shared at command centers.

Impact:

• Reduced congestion and improved emergency response.

• Enhanced screening for unattended bags or other suspicious activities.

• Speedier decision-making in high-traffic times.

Concerns Raised:

- Public unaware of constant surveillance.
- No explicit policy around storing data or who has access to the recordings.
- Whether facial data was shared with law enforcement officials or third parties.

Ethical Reflections:

Analysts praise the impact of the system on public safety and transport efficiency, yet it provoked uproar.

conversations about the necessity of transparent data governance protocols, informed consent and more.

Surveillance in Everyday Life as using or promoting transparency about surveillance in everyday life.

Ethics in Artificial Intelligence_V3_Unit 5.docx

Ethics in Artificial Intelligence_MBA_2

Ethics in Artificial Intelligence_MBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127350324

Submission Date

Feb 2, 2026, 11:32 AM GMT+5:30

Download Date

Feb 2, 2026, 1:08 PM GMT+5:30

File Name

Ethics in Artificial Intelligence_V3_Unit 5.docx

File Size

41.7 KB

24 Pages

5,164 Words

31,015 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 5: Introduction to Indian Mythology & Management

Learning Outcomes

1. Define the structure and functions of the money market, distinguishing it from capital markets.

2. Identify and describe the characteristics, participants, and instruments of the Indian money market.

3. Explain the features, maturity periods, and issuance process of Treasury Bills (T-Bills) and Commercial

Papers (CP).

4. Compare different short-term money market instruments such as Commercial Bills, Certificates of

Deposit (CDs), and Call/Notice Money, focusing on liquidity, risk, and yield.

5. Illustrate how Collateralised Borrowing and Lending Obligations (CBLO) function in secured interbank

lending, including the role of collateral.

6. Evaluate the suitability of different money market instruments for banks, corporates, and government

entities in managing short-term funding requirements.

7. Apply knowledge of money market operations to interpret market trends and assist in short-term

investment or borrowing decisions.

Content

5.0 Introductory Caselet

5.1 Introduction to Bias in AI

5.2 Measuring and Detecting Bias

5.3 Addressing and Mitigating Bias

5.4 Ensuring Fairness in AI Systems

5.5 Case Studies on AI Bias

5.6 Summary

5.7 Key Terms

5.8 Descriptive Questions

5.9 References

5.10 Case Study

5.0 Introductory Caselet

"The Resume Dilemma: When Fairness Meets Automation"

Background:

A recent engineering graduate, Meera had been applying to many tech companies but surprised when she was offered of gigabit data.

any interview calls, despite having an impressive academic record and project portfolio. Her friend Arjun, who had

like credentials, received several interview invitations. Alarmed and worried, Meera went down to the business of inquiring into the affair."

further.

She found that quite a few companies now employ AI-powered hiring systems which scan and pre-select resumes

before any human eyes are laid upon them. These systems draw on historical hiring data — information that itself may be an artifact.

past preferences or biases. In Meera's situation, she discovered that the AI model appeared to be biased towards students from

particular zip codes and institutions that have long been the province of men.

When

she

spoke

to

a

company

HR

representative,

“Our system is neutral — it only learns from the data.”

But Meera wondered: If the data reflects human bias, can the AI ever be fair?

Critical Thinking Question:

the

reply

How AI can be genuinely objective when the data it learns from is biased? What steps should be taken to

ensure fairness in automated decision-making?

5.1 Introduction to Bias in AI

AI systems are programmed with the ability to perform certain actions or make predictions by evaluating data. However, if the

enableing people to [sic] better at automatic decision making,” but since the data that are used to train these systems includeawith algorithms.aThese patterns thats from inequpplication of them can reproduce sex and racial discrimination in the real world challenge criteria.... fed into these machine (McIlwain, 2009)§.Serversfamaiclhunaey,1g0ep2r o9tth-irco§ap,c(20it09y) codingsindes includesinpattgerources ds fou that 4191

Those patterns will likely be learnt and mimicked by A.I. This is known as bias in AI.

Bias comes in many flavors — racial, gender, cultural and economic among them — and can seep intoßerdem schädigen.

like hiring, policing, loan approvals, health care and education. Frequently, those impacted by such

undeserved decisions they don't know how and/or why the system wronged them.

To realise this section we want not only to:

- What bias in AI means,
- Where it comes from,
- Why it matters,
- And how we identify and minimize it, to develop more ethical, dependable systems.

Let me know when you're ready to move on to 5.1.1 or otherwise if you'd like activities/discussion prompts for

this section.

Artificial Intelligence is often portrayed as neutral or objective. But in fact AI systems may mirror, and.

even amplify human bias. This occurs if the data, algorithms, or human decisions that constitute such a system

are not well-drawn or thoroughly vetted for equity. AI bias can lead to unfair actions, particularly for

people from marginalized groups. In this section, we investigate what AI bias is and where it comes from, why it

matters for society.

5.1.1 Definition and Types of Bias in AI

AI bias, on the other hand, is unfair or unequal outcomes generated by an AI system, frequently due to issues in

the data of the algorithm.

Types of bias in AI include:

• Data Bias: If the data used to train the AI model isn't full enough, is unbalanced or reflects past

discrimination.

Example: A facial recognition system that is trained on mostly light-skinned faces may fail to correctly.

recognize darker-skinned individuals.

• Algorithmic Bias: Unfair treatment may be caused by the decision-making process made within the logic of AI algorithm,言語処理学会 調査研究報告 2019-NL-209 20where unfairness is done despite no biases in learning data.

data is neutral.

Example: An algorithm that assigns greater weight to certain words on a resume could potentially discriminate against male honor society members.

coded language.

• Discrimination Bias: If social stereotypes or discriminative practices are embedded in the AI model.

Measurement Bias: Unfairness introduced by the manner in which data is measured or labeled.

Example: Training crime prediction tools on arrest records without considering whether that

arrests were fair.

• Label Bias: Data is directly labeled by humans during the training process and sometimes labeled wrongly or with bias.

5.1.2 Sources of Bias: Data, Algorithms, and Human Factors

AI systems can be biased from a few sources:

Data: AI models are trained on historical data. If the data is biased—by over-representing one group57,72 or recording too few events in all groups101–104—itmay be appropriate to alter one or both of the following for extrapolation purposes.

and and overrepresenting another the AI is likely to yield biased results.

Algorithms: Even in the face of clean data, algorithm design can come with bias. For example, the

algorithm could employ rules or weights to endup making decisions that are biased towards specific outcomes by “accident”.

Human Decision-Making: Humans decide which data to collect, how to label it, and the goals thatgetattri

AI should optimize. Even their own assumptions and values can inject bias at every step.

It is often not so much that these biases are intentional, but instead are driven by a lack of diversity or reviewing, or testing.

5.1.3 Historical and Social Roots of Algorithmic Bias

BIAS IN AI DOESN’T COME FROM NOWHERE — It tends to be a reflection of historical and social inequalities.

Examples:

• Recruiting tools that are trained on a company’s history of applicants may favor men because the firms photographs and videos showcasing its male workers.

historically hired more men.

• However with loan approval system that will deny low income neighborhoods mortgages for all but a few centuries how ever long.

entrenched discrimination in banking.

The bias is embedded in the way society is structured and operates — and when AI learns from that data, it will learn the same things

to repeat the same patterns. That's why tackling AI bias demands a deep knowledge of social justice

and context, not mere technical solutions.

Did You Know?

"Did you know that the U.S. ZIP code system—used in many AI models for credit scoring and insurance—originated in a time when certain areas were redlined based on race and income? This means AI models using ZIP codes can unknowingly reproduce historical segregation and deny services to marginalized communities, even when individual creditworthiness is strong."

5.1.4 Real-World Consequences of AI Bias

AI bias can have profound consequences for people's lives, especially in high-stakes fields such as:

• Hiring: Qualified candidates can be dismissed because of gender, a name or a background.

• Healthcare: Some A.I. tools have been unable to identify symptoms in women or people of color.

• Criminal Justice: Some communities could become unfairly targeted by predictive policing tools.

• Education: Artificial intelligence systems employed in grading, admissions or other aspects of education may have learned processes that discriminate against students based on geography or school attended.

These such outcomes are not always apparent to the individuals who experience them, which makes it difficult for sex equality defenders to monitor and challenge discriminatory

treatment. Prejudiced AI can perpetuate existing disparities and establish new forms of bias.

5.1.5 Ethical Concerns and Social Justice Implications

The AI bias: Which you do not know about raises critical ethical and moral issues:

• Equality: Is everyone being treated the same, no matter who they are?

• Responsibility: Who is responsible when an AI system produces a biased or harmful decision?

• Transparency: Does the system show users how it reached its decision?

• Justice: Are we perpetuating or remedying past discrimination?

From a social justice standpoint, biased AI can exacerbate inequality between the haves and have-nots

communities. Responsible AI development means bringing in a wide range of voices, being transparent about risks and

emergent systems that enable equity and justice, as opposed to systems of harm.

5.2.1 Quantitative Fairness Metrics (e.g., Demographic Parity)

Fairness in AI is typically assessed by fairness metrics, which are mathematical formulas. These metrics

get a handle on whether an AI system is creating biased outcomes between different groups.

Some common metrics include:

• Demographic Parity (Statistical Parity)b:

The concept that results of a system should have symmetrical distribution in different.

demographic groups.

Example: If 50% of men get a loan, approximately 50% of women should have access to one as well.

• Equalized Odds:

The error rates (FP and FN) of the system should, ideally, be equivalent across groups.

Example: An AI for hiring should not incorrectly deny more women than men.

• Predictive Parity:

The accuracy of the predictions generated by the model should work across all groups.

Example: If a healthcare prediction model is 90% for one group, it should be around that":"'For similar example in : Text" DOI:10.1001/jama.2012.219708 SecCancer88-q2.r.

other groups too.

These yardsticks from measurement make bias visible, and thus subject to address.

“Activity: Evaluate Fairness Using Demographic Parity”

Instructions to Learners:

You are given the following AI model outcomes for loan approval:

• Out of 100 male applicants, 70 were approved.

• Out of 100 female applicants, 50 were approved.

1. Calculate the Demographic Parity Ratio (DPR) using the formula:

DPR = (Approval rate for females) ÷ (Approval rate for males)

2. Interpret your result:

o A DPR close to 1.0 indicates fairness.

o A DPR below 0.8 suggests significant bias.

3. Answer the following questions:

a) Is the AI system fair by the Demographic Parity standard?

b) What could be done to reduce the observed disparity?

c) What are the risks of deploying this system without correction?

Submit your calculations, interpretation, and a short paragraph (150–200 words) discussing the fairness

of the model and your recommendations.

5.2.2 Statistical vs Individual Fairness

There are generally two prominent methods for fairness in AI:

Statistical Fairness:

Encompasses the concept of groups being treated the same.

Example: Balancing approvals for loans to people of different races or genders.

Individual Fairness:

Is based on treating like people alike.

Example: If two people are equally qualified, they should have an equal chance of being hired for a job,

regardless of their background.

These two flavors of fairness can, on occasion, come into tension. For example, ensuring its fairness at a group level

would still leave some people treated unfairly — and vice versa. This renders fairness a complex and apparent exception or.

context-dependent issue.

5.2.3 Auditing Algorithms for Bias

Algorithm auditing refers to testing and reviewing AI systems to determine whether they are turning up biased outcomes.

Types of audits include:

• Audits Internal: Conducted by the entity that made the system. These check for bias using in

house data and tools.

• External Audits: Performed by independent researchers, NGOs (non governmental organizations) and regulators. These offer greater

transparency and accountability.

Auditing typically involves:

• Investigating the training data for imbalance

• Empirical study of the algorithm on real dataset

• Assessing effects on various demographic groups

• Looking for anomalies or discrepancies

Audits can be valuable by helping to identify hidden bias before the system is put in use — or after it is deployed—and thereby guarding against policy mistakes.

in use.

5.2.4 Tools and Frameworks for Bias Detection

Various open-source tools and frameworks have been created to assist in identifying bias in AI.

systems.

Some popular ones include:

• AI Fairness 360 (by IBM): A library with more than 70 fairness metrics and bias mitigation algorithms.

• Fairlearn (from Microsoft): Aids in assessing fairness and mitigating disparities in the outcomes of models.

•What-If Tool (by Google): A visual tool to probe what an AI model looks like through testing it on new data and simplified datasets.

inputs and identify biases.

Aequitas ([From UChicago] (https://github.com/DS3Lab/aq_u)): The software aimed at public policy to assess bias in decision making.

tools like predictive policing or risk scoring.

These are the kinds of tools that will enable developers to test for fairness and improve transparency in how decisions SEND mean what they communicate about us?) are made - (READER: trade our privacy.

are made.

Did You Know?

“Did you know that IBM’s AI Fairness 360 Toolkit allows developers to test their AI models for

over 70 types of fairness metrics—and can even simulate how changing the data or model affects

fairness outcomes?

It’s open-source and used globally by companies and researchers to audit bias before

deployment.”

5.2.5 Challenges in Measuring Fairness

Fairness in A.I. is hard to measure. Some of the main challenges are:

• Multiples Definitions: Isn’t fairness that you have a definition of the thing called “fairness”? Different situations may require

different metrics.

• Trade-offs: Some of the things you might do make the system fairer for some people and less so for others. For example, equalizing

error rates may change accuracy.

• Missing Data: Demographic data, such as race or gender, may not be present because of privacy.

laws or missing records.

• S ituationa l Dependence: W hat is fair in one context (e.g., hiring) m ay not be fair in another (g.

healthcare).

• Hidden Bias: In some cases, even by using fairness metrics, there may be subtle biases which remain unnoticed.

In that context, however, bias measurement becomes essential so that we can build ethical AI. It allows developers to track

advance, improve and prevent harming their systems' users inadvertently.

5.3 Addressing and Mitigating Bias

Once AI bias is detected, the next step is to mitigate or eliminate it. This process is called bias

mitigation. This necessitates action at all stages of the AI lifecycle – from data collection and model design

algorithms to interacting with users and formulating accountability rules. This section explores practical,

technical, and policy approaches that contribute to the development of ethical, inclusive, and reliable AI systems.

5.3.1 Fair Data Collection and Preprocessing Techniques

Biases often begin with bad or unbalanced data, so one step toward mitigation is to improve how data are

collected and prepared.

Key techniques include:

• Representative Sampling: Ensuring data are representative of all groups (e.g., capturing their voices from

various accents for a speech recognition system).

• Data Stratification: Re-balancing of the dataset so that minority groups are proportionally present.

• De-biasing Annotations: You are audit and fix the labels created with bias assumptions.

• Anonymization: Stripping away personal identifiers that can lead to discrimination (such as names or zip

codes).

• Synthetic Data Generation: Generating synthetic yet representative data to enhance the diversity in

underrepresented categories.

The aim is to provide the AI with an accurate, balanced view of reality; That way, its decision-making will be much harder for it to.

be biased.

5.3.2 Algorithmic Debiasing Strategies

And after the underlying data is improved, developers can also change the algorithms to minimize bias.

Some approaches include:

• Reweighting: Assigning extra importance to underrepresented data points during training so that the model

learns more evenly.

• Adversarial Debiasing: Trains the AI not to easily guess a person's sensitive体attribute.

attribute (such as, race or sex) which can prevent biased results.

• Post-processing Corrections: Modifying the outputs of the AI to make them fairer, after it has been Decidef model.

made its predictions.

• Fairness Constraints: Incorporating fairness objectives (e.g., equal error-rates) within the training itself

process of the model.

These methods are especially effective when used to reinforce, not replace, fair data practices.

5.3.3 Inclusive Design and Stakeholder Participation

AI must be built with people — and not just for people. This requires to multiply the tapes of voices.

in the creation, evolution, and certification of AI systems.

This includes:

• User inclusion: Particularly those from communities most negatively impacted by the AI system (e.g.,

persons with disabilities, minorities, or low-income users).

• Interdisciplinary teams: Not only computer scientists but also sociologists, ethicists and legal

experts.

• Cultural context: How the system might function differently across regions and languages, which can help prevent misuse.

or traditions.

By embracing inclusive design, developers can catch blind spots early, deliver more usable systems and earn trust from.DialogInterface users.

the public.

5.3.4 Governance and Policy Approaches to Fair AI

Fairness in AI is not only a technical problem — it's also social and legal. That's why governance

and policy are essential.

Strategies include:

• Regulation: governments can pass laws that require the fairness check or explainability (e.g., GDPR, [3]). 2 Background Here we discuss a selection of previous methods related to verification of ML models and model interpretability.

EU AI Act).

• Internal Policies: A company may establish internal codes of ethics, fair- ness guidelines or independent representati- ves.

audit systems.

• Impact Assessments: Organisations can conduct deliberate assessments before deploying AI.

might affect different groups.

• Public Oversight: Promoting openness and public involvement by way of hearings, feedback

platforms, or citizen panels.

From this perspective, good governance would help ensure that fairness does not become optional, but is part of responsible AI practice.

development.

5.3.5 Transparency and Explainability as Fairness Tools

An AI system is more accountable when people can understand how it works. This is where transparency

and explainability are crucial.

• Transparency means being honest about how the data is collected, how decisions are made and what

the model is trained on.

• Explainability involves offering plain, understandable reasons behind every decision the AI makes,

particularly in high-stakes domains such as hiring, health care or criminal justice.

Examples:

• A loan application AI should be able to tell a person why they were not accepted.

• A facial recognition system should display confidence scores and known imperfections.

Explainable AI helps:

• The respect and information imparted to users,

• Developers can find errors more easily,

• Organizations build accountability.

When A.I. decisions are hidden, or un-interpretable, it is harder to detect and correct bias. Making AI understandable

is key to making it fair.

Fairness and AI isn't so much fixing technical errors — it's about making systems that treat every person the same.

and organizations fairly, respectfully, and transparently. Fairness is achieved by a combination of statutes and fairly applied enforcement, but it needs to be at the level that grants students access.

skill, human judgment, and obvious standards. It also necessitates consideration of how multiple identities and

social contexts interact with technology. This section is all about the weapons and tactics of bushcraft__));

justice a core component of AI, not an add-on.

5.4.1 Legal and Ethical Frameworks for Fairness

Legal and ethical frameworks define the rules and values to govern AI behaving fairly in society.

Legal frameworks include:

•Privacy laws, including data protection laws (such as GDPR) that grant users rights over their use of their data.

• Non-discrimination laws that implicate automated systems in the same way as humans (for example, underTestFixture).

hiring or lending).

• Future AI-related legislation, such as the EU AI Act, which classify AI systems bytheir riskleveland

demand fair audits for systems highly at risk.

It goes further than the law and prompts the question:

• Is the system respectful of people?

• Do voices on the margins get a hearing?

• Is the A.I. advancing fairness, justice and equality?

Ethical principles often include:

• Fairness

• Accountability

• Transparency

• Human dignity

Combined, these frameworks establish an ethical and legal limit regarding AI development.

5.4.2 Fairness in High-Stakes Domains (e.g., Hiring, Justice)

Certain details of life involve decisions in which the potential future of a person can change. These are called high-stakes

domains, where fairness is crucial in these items.

Examples include:

• Hiring: Artificial intelligence tools that are used to screen resumes or conduct video interviews should not discriminate against certain

genders, accents, or education backgrounds.

• Criminal Justice: Risk-assessment tools that help determine whether to grant bail, parole or conditional release shall not

discriminate on the basis of race or class.

• Medicine: Diagnostic algorithms are required to perform equally well in different population groups in order to prevent.

life-threatening errors.

• Education: Admissions or grading AI must be context-aware and not further replicate

existing inequalities.

In these areas, AI that is poisoned could cause serious damage and systems ought to be audited for fairness.

before deployment.

5.4.3 Intersectionality and Contextual Fairness

Intersectionality is about acknowledging that our identities are composed of multiple social categories—

(e.g., race, gender, age, disability and class) that can come together to produce distinctive incidents of

discrimination.

What Contextual Fairness means is that \$lat$b"Fairness\$mart$" has to be taken with binders on (i.e. context).

• The cultural, economic or historical place where the AI system is being used.

• The particular needs and vulnerabilities of those it affects.

For example:

• A just system in one country may not be just in another because of different social norms.

• A facial recognition system may perform well for adult men but poorly for older women or children with.

disabilities.

Ensuring equity involves upholding fairness in a society, which requires going beyond averages and considering who is being left behind orPAYETTE OPINION@adminrichtelexcluded.

misrepresented in each context.

5.4.4 Role of Human Oversight and Governance

AI should enable human decision making, not supplant it altogether—especially in sensitive or high-stakes decisions, according to the researchers.

situations.

Human oversight includes:

• Watching for errors or unfair patterns in A.I. outputs.

Humans are allowed to check AI calls before they're set in stone.

• Interceding when results appear unfair or arbitrary.

Governance consists of the mechanisms and relations through which accountability is implemented, including:

• Ethics boards within companies.

• Independent regulators or auditors.

• Frequent checks of fairness, with public reports.

Artificial intelligence systems can be modified, improved or taken out of use when they are fitted with adequate oversight and governance," he added.

are found to be harmful.

5.4.5 Accountability Mechanisms and Standards

Accountability means that someone is responsible when an AI system does harm or makes a and it doesn't just mean sharing responsibility.

biased results.

Key accountability mechanisms include:

• Auditing: Ongoing third-party audits of A I systems for fairness, bias, and performance.

• Impact Assessment: Potential risk analysis prior to deployment of the system.

• Documentation: Clear records of how the system was trained, tested and deployed.

• Appeals Mechanism s: Giving users to right to protest or dispute AI decisions (e.g., loan rejection, job

rejection).

• Universal Ethics: Adherence to international codes of conduct like OECD AI Principles, UNESCO AI

Ethics recommendations, and ISO standards.

Robust accountability measures ensure that governments, companies and AI developers are held accountable to

the people and act when fair play has gone awry.

5.5 Case Studies on AI Bias

The problem with bias in AI isn't simply theoretical — it's been spotted in real-world projects, and to the detriment of some.

consequences. These examples illustrate how AI systems can entrench discrimination, compound inequality,

and when the principle of equity is ignored, it has done damage to people's faith less.apache.org They also illustrate the requirement of

transparency; free and open data; human agency (vs. machine-led); and ethical design."

5.5.1 Case Study 1: Bias in Hiring Algorithms

Sample case example: The AI-based Recruiting System of Amazon (2014–2018)

Amazon built an A.I. to review job applicants resumes — but it reportedly didn't like women Amazon created a recruiting engine that used customer data, including past hires, to rank the best candidates and its isn't alone -Thursday October 11th, Reuters They made their purpose sound so good.

candidates. The model was trained on the 10 years of company hiring data, which skewed heavily male

applicants.

What went wrong:

• The A.I. model downgraded resumes that mentioned the word "women's," as in "women's chess club."

captain."

• It demoted graduates of all-women's colleges, too.

• The system taught itself patterns that were based on historical bias against women in tech job hiring.

Outcome:

Amazon abandoned the tool after its own experiments showed it was biased. It raised questions about the risks of employing

without addressing ingrained bias in historical data.

5.5.2 Case Study 2: Discriminatory Lending Practices

Example: Algorithmic Bias in Credit Limit Decisions (Apple Card, 2019)

Apple Card, managed by Goldman Sachs, was accused of providing lower credit limits to women compared with men.

men — even when both had the same money qualifications.

What went wrong:

• Some couples said the man received a credit limit 10–20 times higher than that of the woman.

• Even when women had better credit scores or earned more than men, the algorithm churned out lower scores for women.

• The company said the algorithm was fair, but it couldn't explain why because of its lack.

transparency.

Outcome:

The case prompted U.S. financial regulators to conduct investigations and brought AI that is opaque to public attention.

decision-making in the financial sector.

5.5.3 Case Study 3: Bias in Facial Recognition Systems

Example: Bias in Gender and Race Recognition (MIT Media Lab Study, 2018)

The technology built into these kits has been proven to work less accurately on people with darker skin and womenA researcher called Joy Buolamwini discovered that commercial facial recognition systems developed by major tech pubs.sourceforge.org

companies had far higher error rates for dark-skinned and female faces than light-skinned

male faces.

Key findings:

• Error rate of recognizing white male faces: Less than 1%.

• Error rate for female faces with dark skin: As high as 35%.

• These systems were largely trained on datasets of males with light skin.

Implications:

• Mistaken identification among law enforcement can turn to wrongful arrests.

• Outsize errors undermine trust in public surveillance and security.

Response:

A number of cities (such as San Francisco and Boston) outlawed facial recognition by government agencies.

5.5.4 Case Study 4: Predictive Policing and Racial Profiling

This is what an algorithm would look like Example: PredPol, US Police Departments

An example in the U.S. of this is the use of PredPol (Predictive Policing software) by Police forces to predict crime

crime hotspots given historical crime data.

What went wrong:

• The system deployed a greater police effort in Black and Latino neighborhoods, which were already over

policed.

• The crime data that officers were trained on had bias built in; it mirrored the racial profiling of previous years.

• The increased policing in these locations resulted in more arrests — not necessarily more crime being found.

Outcome:

• A public outcry and news reports of discriminatory targeting.

• Some police departments ceased using PredPol, and researchers urged more openness in

predictive policing.

5.5.5 Case Study 5: Health Inequities in AI Diagnostics

Example: Race Bias in Health Risk Algorithms (U.S. Health Care Study, 2019):

One of the largest algorithms used by hospitals to screen for high-risk patients when deploying extra care was discovered to

fail to take seriously the needs of Black patients.

What went wrong:

• The model treated the costs of healthcare as an indicator for health needs.

• Their costs were lower — because as a group, Black patients receive less health care than white patients due to systemic barriers.

whose medical problems were critical.

• The algorithm favored white patients who spent more over Black people with equivalent spending

or worse health issues.

Impact:

• Eligible Black patients were excluded from extra care programs at a rate of nearly half.

Outcome:

After the public exposure, the developers changed the algorithm, and it became a highly reported case.

of systemic bias lurking in our design decisions.

5.6 Summary

❖ Bias in AI is more than a technical glitch — it reflects deeper social, historical and ethical problems.

AI systems learn from data, and if that data contains the remnants of past discrimination, inequality or exclusion,

the system will tend towards reproducing or even exaggerating these words.

❖ In this unit, we examined how bias seeps in through data, algorithms or human design decisions. We

explored bias, fairn ess met-rics, and tools.

to audit and detect bias. The division also offered tools to address bias through responsible data practices,

universal design, algorithmic adjustments, and human supervision.

❖ High-stakes domains such as hiring, policing, healthcare and finance—domains that were treated especially notable including

where bias on the part of an AI can be a matter of life and death. Legal frameworks, ethical principles, and

accountability criteria were also identified as main supports when deploying in-the-wild fairness.

❖ Case studies revealed that unfettered AI bias can also result in injustice, as responsible

AI methods can promote trust, equity and social benefit.

5.7 Key Terms

Bias in AI – Prejudice built into AI outputs as a result of tainted data, algorithms, or human judgement.

Demographic Parity: A fairness metric that stipulates that the outcome should be equal for different groups.

Predictive Policing – The application of AI to predict where crimes are going to occur, often with biased results.

Algorithmic Debiasing – Methods to reduce bias in AIs.

Intersectionality – The notion that different parts of identity (i.e., race, gender) intersect to form a particular level of privilege or discrimination.

unique experiences of discrimination.

Fairness Metrics – Quantitative measures to check whether an AI system is biased with respect to some groups or not.

individuals equitably.

Auditing Algorithms – Scrutinizing AI systems to find hidden biases or unfair trends.

Explanation – The capacity of an AI system to present meaningful rationales for its decisions.

Governance – Policies, rules and structures for oversight governing how AI is developed and used.

Ethical AI – AI that is developed and used in a manner that respects human rights, dignity, and justice.

5.8 Descriptive Questions

What is bias in AI and how do you define it? Can provide two examples of different kinds of bias.

What are AI fairness measures and how do they help in identifying bias?

Explain the distinction between statistical and individual fairness.

How might inclusive design methodologies aid in alleviating AI bias?

Describe the role of human oversight in high-stakes AI systems.

Explain how legal and ethical context contributes to fair AI.

Illustrate one example of AI bias in the hiring process for those without background check support.

What's the connection between intersectionality and fairness in AI?

What are some of the challenges involved in auditing AI systems for bias?

Describe two approaches to achieving fairness in healthcare AI applications.

5.9 References

1. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. fairmlbook.org

2. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in

Commercial Gender Classification. MIT Media Lab.

3. Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. Fordham Law

Review.

4. Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming

Biased Performance Results of Commercial AI Models.

5. European Commission. (2021). Proposal for a Regulation on a European Approach for Artificial

Intelligence (AI Act).

6. World Economic Forum. (2020). AI Governance Frameworks and Toolkits.

7. IBM Research. (2021). AI Fairness 360 Toolkit.

8. Microsoft Research. (2020). Fairlearn: A Toolkit for Assessing and Improving Fairness in AI.

9. U.S. National Institute of Standards and Technology (NIST). (2022). AI Risk Management Framework.

10. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence.

5.10 Case Study

Unfair Lending: A Case of Algorithmic Discrimination in Loan Approval

Background:

Last year, a fintech firm rolled out an AI platform to assess borrowers for a

microfinance program in Southeast Asia. The system purportedly relied on more than 300 data points—

such as social media conduct and mobile payment records—when deciding whether to extend credit,

especially for those who don't have formal banking records.

What Happened:

After six months, a trend emerged: The foundation representatives noticed that the bulk of loan rejections were disproportionately from

female applicants, despite the fact they had the same income and repayment potential as male

applicants. Further investigation showed the AI had been trained on loan data that included

women were traditionally denied loans because of social and institutional discrimination.

Bias Source:

The training set has captured historical gender-biased discrimination (the training data was gender biased), and black box algorithm did not consider attitudinal principles.

context, replicated it. The model also included weights on some variables — such as owning a home and working full

time jobs —higher, disadvantaging women in a structural way.

Impact:

- Exclusion of creditworthy borrowers.
- Lack of faith in the lending platform.
- Public criticism and regulatory pressure.

Response:

- The firm halted the A.I. system and audited its training data.
- Conducted bias audits before deploying the model.
- Consulted gender policy experts to help redesign the approval logic.

Reflection Questions:

- What alternative choices could be made by the developers while collecting the data?
- How might algorithm design take into account social and historical context?
- How much should regulators oversee financial AI systems?

Ethics in Artificial Intelligence_V3_Unit 6.docx

Ethics in Artificial Intelligence_MBA_2

Ethics in Artificial Intelligence_MBA_2

ATLAS SkillTech University

Document Details

Submission ID
trn:oid:::3618:127350327

Submission Date
Feb 2, 2026, 11:32 AM GMT+5:30

Download Date
Feb 2, 2026, 1:09 PM GMT+5:30

File Name
Ethics in Artificial Intelligence_V3_Unit 6.docx

File Size
40.4 KB

24 Pages

4,982 Words

30,704 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 6: Accountability and Transparency

Learning Objectives

1. Define accountability in the context of AI systems and explain why it is essential for responsible AI

deployment.

2. Understand the concept of transparency in AI and identify its role in building trust and ethical compliance.

3. Recognize the key stakeholders responsible for AI outcomes, including developers, organizations, and

policymakers.

4. Identify strategies to improve accountability in AI, such as impact assessments, audits, and regulatory

frameworks.

5. Explore methods to increase transparency, including explainable AI (XAI), open documentation, and model

interpretability tools.

6. Analyze real-world challenges in implementing accountability and transparency, such as complexity, trade

secrets, and explainability gaps.

7. Evaluate the limitations of current practices and suggest improvements for fairer, more accountable AI

systems.

Content

6.6 Summary

6.7 Key Terms

6.8 Descriptive Questions

6.9 References

6.10 Case Study

6.0 Introductory Caselet

"The Blame Game: Who's Responsible When AI Goes Wrong?"

Background:

Rohan had sought a home loan from an online service that used artificial intelligence to evaluate applicants. He had a stable

job, no bad credit history, and also none of your debts. Surprisingly, his loan was rejected. When he reached out to

the bank, the representative said,

"It was our AI system that made this decision. "I don't know what it was specifically denied for.

Rohan requested to appeal the decision, but the system gave no explanation or chance for a human review. He was left

confused and powerless.

A news report later revealed that the AI model used a sophisticated algorithm [that] overweighted address-based.

risk, which disproportionately impacted applicants in some pos20 tal codes—many of these areas were in33.

underdeveloped or minority-dominated neighborhoods. The bank attributed it to a technical glitch, the AI

vendor blamed the training data, and the data scientists said the model was functioning "as designed."

Responsibility was so diluted that Rohan had no one to hold to account.

responsible.

Critical Thinking Question:

When an AI system causes harm or takes an unfair action by way of a decision, who is responsible: the developers, the human operator or the machine itself?

data sources, the domain in which it is applied, or by the algorithm? Why?

6.1 Understanding Accountability in AI

The most general of these is accountability in AI, which entails ensuring that it's the case someone — a person, team or organization — is distinctly responsible for how an AI system functions.

for the output and effects of an AI system, particularly when those effects have implications for people's

lives.

In old-school decision-making, it's easier to figure out who made a choice. But with AI:

• Decisions are often automated.

• Algorithms can be too complex to comprehend.

• There are a number of players (developers, suppliers users).

This renders accountability harder, but more necessary.

Key points to understand:

• Accountability means having errors recognized, rectified and avoided in future.

• It helps create trust in the system, as users know there's someone they can hold accountable for what the AI

does.

• It ensures legal and ethical compliance, particularly in fields where healthcare, finance or law plays a role

enforcement.

Accountability in AI means that:

• There can be no dispute over who designed, trained, tested and approved the AI.

• Machines with power over rights or well-being must include human oversight.

• Mechanisms such as impact assessments, audits, appeals and redressal procedures should be established.

Absence of accountability means mistakes don't get corrected, harm isn't acknowledged and public trust in AI wanes.

6.1.1 Definition and Scope of Accountability in AI Systems

Accountability in AI is the responsibility of all those involved in the architecture, generation, building and use of AI to act ethically to minimize harm.

using AI systems to:

• Defend their decisions and actions,

• Accept responsibility for outcomes,

•Intervene when needed.

The scope of accountability includes:

• Technical choices (e.g., selection of data, training the model)

• Ethical considerations (such as fairness, privacy, bias)

• Complying with the law (e.g., obeying statutes and directives)

• Operationalization (i.e., the manner in which the system is put to work and observed)

In applications to high-stakes systems, like finance, healthcare, criminal justice accountability is particularly crucial.

and public services.

6.1.2 Who is Responsible? Stakeholders in AI Lifecycle

AIs are not created or to operate by just one person—they incorporates different constituents, each of which with its own stakeholders profile.

responsibilities:

Data Providers – Obligation for providing data, which would be true, ethical and representative.

Developers and Engineers – Must construct systems that of legal and ethical value.

ALGORITHM DESIGNERS – Must ensure the fairness, transparency and explainability of the model.

Companies/Organizations – Responsible for implementation of AI and decisions made as well as brief.

impact on users.

End Users – Must use the AI system in a responsible manner and report any bugs or wrong outputs.

Regulators and Policymakers – Establish regulations, define standards, supervise AI systems to ensure accountability and safety for the public

interest.

If shared accountability is unclear, it results in cases where nobody owns a problem.

Clarity of roles is essential.

6.1.3 Ethical, Legal, and Organizational Dimensions of Accountability

Responsibility in AI shall be addressed under three primary dimensions:

• Ethical:

Is the turning model compatible with dignity, fairness, and autonomy?

Ethical accountability poses questions about bias, harm and social justice.

• Legal:

Do the developers and users abide to data protection laws, anti-discrimination rules, and AI.

regulations?

This can be done through legal accountability (liability, user rights and compliance reporting).

• Organizational:

Is the company or organization institutionally prepared (in terms of internal structures such as audits and ethics boards) to oversee AI?

systems?

Responsibility assignment and governance architectures in organizational accountability

institutions.

All three echelons of government should work in symbiotic harmony to make sure AI is not only good but just and safe.

6.1.4 Consequences of Lack of Accountability

When people do not have to answer for their actions, numerous issues can arise:

Individual injury – unfair decisions (e.g. denial of loan, wrongful arrest) can remain unrectified.

Public Distrust – The damage might also be that users are potentially afraid of or unwilling to use AI if no one is held accountable for it.

Legal and Financial Risks – Companies can be sued, fined or damaged by brand association.

No Improvement: No one records or corrects errors; no improvement occurs in the performance of the AI system.

Ethical Breaches – Untethered, systems that are used can perpetuate discrimination or spying without approval.

A system in which there isn't any blame is a dysfunctional system. Accountability guarantees that neither harm is created nor allowed to fester.

6.1.5 Comparative Analysis: AI vs Traditional Systems

Responsibility is clearer in traditional decision-making systems:

• A human makes the decision,

• Someone is responsible — whether that someone or their employer is to blame,

• Often, there is an appeals process or a means to contest the result.

In AI systems:

• The call may come from an algorithm,

• The process may be a "black box" operation,

• There might not be any obvious person to question or blame.

Aspect

Traditional Systems

Decision-maker

Human (named individual)

AI Systems

Accountability path Clear and direct

Algorithm (often unknown logic)

Explainability

Shared and often unclear

Usually possible

Liability Attributed to Host individual/company Debated many a times

Often difficult (black box AI)

Comparison of both to one and another, it demonstrates that accountability frameworks based on AI are required in order to operate at this degree of scale.

impact of modern systems.

6.2 Transparency in AI Systems

AI transparency refers to the ability for processes, decisions and limitations of a particular system to be open, intuitive, true and accessible.

and transparent — not just with developers, but also users, regulators and the public at large. Transparency helps build

trust and accountability to foster ethical use of AI. This is where the aspect of transparency comes helpers.

means in an AI context, why they are important and how we can achieve it.

6.2.1 Definition and Dimensions of Transparency

Transparency in AI means the ability to comprehend, explain, and inspect how an AI system operates—

especially how it makes decisions.

Key dimensions of transparency include:

• Model Interpretability: The ease of understanding the inner workings of an algorithm.

•Process 'Transparency': Transparency over how the AI system was designed, trained, tested and deployed.

• Outcome Transparency: Justification of a decision (e.g., why a loan was denied), to ensure that it is not based on sensitive attributes such as race and religion.

denied).

• Data Transparency: Understanding of the data that trained the model — where it ca

fairness.

• Policy Transparency: Reveal rules, oversight and disclosing of how AI systems are governed.

accountability mechanisms.

True transparency is about opening these dimensions to various portions of the population, not just those who specialize in them.

6.2.2 Importance of Explainability in AI Models

Explainability is the power to comprehend why an AI made a decision. It is a key part of transparency.

Why explainability matters:

• For users: It helps them comprehend decisions that impact them (like job rejections).

• For developers: Aids debug, enhancement or correction of the system.

• For regulators: See that it falls within legal guidelines and ethical application.

For high-stakes domains such as healthcare or criminal justice, explainable AI (XAI) is essential because people want to.

to trust but verify the system decisions.

For example:

• For a medical diagnosis AI, doctors ought to be able to see why a particular disease was diagnosed.

• In criminal justice, judges must know how the scores are determined.

Explainability can prevent “blind trust” in technology and accommodate human oversight.

6.2.3 Black Box vs White Box Models

1.4 Black Box and White Box AI systems can be seen as, in what sense their inside might be non-transparent or transparent.

• Black Box Models:

o Complex models, such as deep neural networks.

o High accuracy but low transparency.

o Even the engineers who write it may not be aware of all at how the model makes decision points.

o Facial identification, image recognition, not to mention language parsing.

• White Box Models:

o Simple models such as decision trees, linear regression.

o More readable, but potentially less strong.

o Explanation and trust prowess>raw performance on trusted input.

o Applicable to healthcare, finance and legal systems.

The choice between black box and white box depends on the use case, the audience, and ethical purposes.

Did You Know?

“That researchers have developed "model distillation" techniques to extract simpler, interpretable models from complex black-box systems? This means that even though a deep learning model might be too complex to interpret directly, developers can create a shadow model—like a decision tree—that mimics the black-box behavior in an understandable way. This approach helps make complex AI more transparent without changing the core system.”

6.2.4 Transparency for Users, Regulators, and Developers

Transparency takes different forms for various stakeholders:

• For Users:

o Transparent communication about how the AI impacts them.

o The ability to question, comprehend or appeal AI judgments.

o Simplified, non-technical explanations.

• For Developers:

o Log data, model structure, training data access.

o Explainability and Fairness Testing Tools.

o Description of how the AI was trained and evaluated.

For Regulators:

o Report on system performance, biases tests, and risk assessments.

o Access to impact evaluations and decision logic.

o Compliance checks and logs for the legal.

Effective transparency is the provision of necessary information to appropriate people at a good time.

6.2.5 Trade-offs Between Performance and Transparency

In general, there is a trade-off between model accuracy and interpretability. This creates trade-offs in real-world

AI development:

Goal

Trade-off

High performance Can be obtained with complex models (black box), but are not interpretable.

Greater transparency Simpler models are easier to to understand, but sometimes less accurate.

Examples:

• A deep neural network might diagnose disease with 95 percent accuracy but have no explanation for that diagnosis.

• A decision tree might be able to explain why, but only achieve 85% accuracy.

Choosing between them requires:

• Context-appropriate judgment: What matters most, getting it right or being able to explain why?

• Ethical cognition: Will users suffer if they don't comprehend the system?

• Regulation: A few industries mandate that products are understandable, and not just ethically.

To weigh these tradeoffs, researchers are working on hybrid methods such as:

• Interpretable layers in deep networks.

• Post-hoc justifications (such as LIME, SHAP or counterfactuals).

6.3 Strategies for Enhancing Accountability

Holding AI accountable requires more than good intentions Organizations can hold AI systems responsible with the Right design

approaches, monitoring tooling and governance structures that guarantee the ethical and safe use of AI. These

strategies protect against harm, build public trust and make it clear who is to blame if something goes wrong.

6.3.1 Design Principles for Responsible AI

Responsible AI design begins with ethical principles and values that guide its developers and users

in the life of an AI system. Some widely recognized principles include:

• Fairness: The system must not discriminate and treat everyone fairly.

• Accountability: Clear ownership over the result.

• Transparency: The openness of how the system functions and decisions are made.

• Privacy: honor and safeguard personal and sensitive information.

• Reliability and Security: The system's operations should not generate unexpected actions; it is expected to be harmless.

These principles can manifest in design decisions, for example the:

•Baking in explainability from the jump,

• With any fail-safes or override functions.

• Recording everything you do to develop your model.

Ethical AI is not going to be a retrofit – it begins in design.

6.3.2 Traceability and Auditability of AI Decisions

Traceability is the ability to trace the flow of decisions, from data collection to final output.

Auditability refers having the tools and records that allow others to review and evaluate AI determinations.

Together, they ensure that:

• Developers and auditors can reverse-engineer how a decision was reached,

External third parties (such as regulators or courts) can hold organizations to account,

• Errors or toxic patterns can be pinned down and addressed.

Examples of traceable elements:

• Logs of data used to develop the model and what parameter was used for training,

• Model revision history,

• Recording of who gave approval for the AI system to be deployed.

Audit trails allow one to probe whether biases, errors or security breaches occur after deployment.

"Activity: Trace the Decision Path of a Simple AI Model"

Instructions to Learners: You are given a decision tree model used for predicting loan approvals. The model considers three inputs: credit score, annual income, and debt level. 1.

Analyze the decision tree diagram provided to you. 9 2. Choose two hypothetical applicants and run their data through the tree to determine whether they would be approved or rejected. 3. Trace and explain each step of the decision for both cases. 4. Answer the following: o What factors led to the decision in each case? o Would the model be easy to audit if a user challenged the decision? o How could this traceability be improved in more complex models? Deliverable: Submit a 1-page report showing your decision paths, explanation of logic, and brief reflection on traceability.

6.3.3 Embedding Human-in-the-Loop for Oversight

Human-in-the-loop (HITL) involves human decision makers in crucial pieces of the AI process—including:

particularly when the AI decision has an impact on peoples' rights or welfare.

This can take several forms:

• Check before acting: A human looks at the AI's recommendation, and confirms (or not) it, before enacting it (e.g., in medical COMMAND AND CONTROL SCENARII Missile attack/defense Cyber defense Cruise/fighter plane Concentration camp hospital Supervision of Console access Decision Support Figure 1.

diagnosis).

• Appeal process: Users can dispute AI decisions and demand human review.

• Hybrid decision making: AI recommends, humans choose (used in defense, health care and finance).

Benefits:

• Lessens over-reliance on AI in the blind,

• Facilitates the recognition of anomalous or unjust outputs,

• Keeps people, not just machines, accountable.

In high-stakes scenarios and applications full automation is seldom acceptable: humans need to stay in control.

6.3.4 Ethical Impact Assessments and Risk Mitigation

The Ethical Impact As environmental impact assessments are mandated for big construction projects, Ethical Impact could be forced in to regulations.

Impacts Analyses (EIAs) that will become central to AI systems.

An EIA includes:

• Identifying ethical concerns (e.g., bias, exclusion, surveillance),

• Considering who might be harmed and how,

• Suggesting possible mitigations (e.g., bias testing, remediation mechanisms),

• Recording the options reviewed and choices made.

Risk mitigation strategies can include:

• Bias detection tools,

• Restrict the use of the system to low-risk situations,

• Informing affected users with what to expect.

It encourages ethical AI to be proactive, rather than fixing it in the rear-view mirror.

6.3.5 Governance Models and Best Practices

AI governance engages the mechanisms to be invoked for ensuring responsible use of AI within an

organization or sector.

Strong governance includes:

• Defined roles and responsibilities for AI teams,

• Ethical review boards or A.I. oversight committees,

• Periodic audits and oversight,

• Training programs to teach employees how best to use AI ethically.

Here's what the top companies and even governments are doing that might be worthwhile to consider: 1.

• With the frameworks of the OECD AI Principles, EU AI Act or NIST AI Risk Management

Framework,

• Creating in-house rules on artificial intelligence that conform to the legal and ethical standards,

• Establishing feedback mechanisms for users, stakeholders and employees to report any concerns.

When governance is well-functioning, accountability is systemic, not discretionary.

6.4 Strategies for Enhancing Transparency

"At stake is the trustworthiness of AI, and nothing less than full transparency will help to ensure that AI systems work not only effectively, but in a manner that is ethical and.

open to scrutiny. Greater transparency looks like clear and accessible information on how facial recognition originally other AI systems work.

systems function, what kind of information they consume and how decisions are made. This section discusses five major approaches that

companies, developers and governments can use to ensure that AI is more transparent.

6.4.1 Explainable AI (XAI) Techniques

Explainable AI (XAI) are tools and methods that assist humans in comprehending how AI makes decisions

decisions. XAI becomes particularly critical when decisions in question touch upon people's rights, finances, health or freedom.

Key techniques include:

- Feature importance: What input factors affected the decision most (for example, income level, age).
- Local explanations: It gives an understandablereasoning for any particular decision.
- Metamodels: Simplified models that mimic the performance of complex algorithms.
- Visualization: Aid users in the exploration of how inputs result in outputs.

Popular tools:

- LIME (Local Interpretable Model-agnostic Explanations)
- SHAP (SHapley Additive exPlanations)

These same tools allow us to ask, "Why did the AI do that?" and get an understandable answer.

6.4.2 Model Documentation and Data Sheets

Transparency also would require keeping thorough records of how an AI model was built, trained and tested.

Two common formats include:

• Model Cards : Reports documenting an AI model's intended use, performance and limitations, as well as ethical considerations.

considerations.

• Data Sheets for Datasets: Detailed documentation about where the training data was sourced, how it was collected,

cleaned, and if there is any bias or exclusion in it.

Benefits:

• Helps developers improve accountability.

• Allows auditors and regulators to judge system quality.

• Educates users on how the system could perform across various situations.

This trend fosters transparency and audibility in the life span of AI.

6.4.3 Open Source and Peer Review of AI Tools

Open sourcing AI systems makes them available for the wider community to inspect, criticize and improve.

Benefits of open-sourcing AI tools:

• Transparency: Anyone can look over the code and figure out how the system operates.

• Peer review: Skilled scholars can find bugs, security holes or bias.

• Reproducibility: Scientists can replicate findings and check claims.

• Collaboration: Contributors from around the world lead to better design and innovation.

But, for open-source AI to work, there needs to be responsible release techniques and processes:

• Usage guidelines,

• Disclosure of risks,

• Limits on abusive use (surveillance, disinformation).

Examples:

• OpenAI’s staged access early releases of GPT models

• Google’s TensorFlow and IBM’s AI Fairness 360 Toolkit • Big picture: Some in the tech industry are starting to think that A.I. researchers, like scientists who benefit from broad

access to research results, would benefit from looser rules for sharing their discoveries with one another.

6.4.4 Transparency in AI Procurement and Deployment

When AI systems are purchased or implemented by governments or institutions, transparency is key every step of the way.

Best practices include:

• Disclosing criteria for evaluation of an AI vendor.

• Reveal which AI systems are being used and what decisions they underpin.

• Requiring vendors to deliver risk assessments and fairness reports.

• Establishing public consultation when AI touches upon citizen rights (surveillance, eligibility for benefits).

Why this matters:

• Public institutions have to be accountable to the public.

• People should have the right to understand how AI is evaluating, monitoring or categorizing them.

Example:

• A number of cities (including Amsterdam and Helsinki) have established AI registries that enumerate all public AI systems introduced in the urban environmen

ui], its indication and risk level.

6.4.5 Educating Users and Stakeholders

Transparency isn't just making data available — it's ensuring that people can understand it and make use of it.

Educational strategies include:

• Plain-language explainers of AI systems and decisions.

• Public education on how AI is shaping daily life.

• Workshops or training sessions for staff, policymakers or citizens.

• User guides that describe how results from A.I. should be interpreted, or decisions appealed.

Benefits:

• Reduces fear and misinformation,

• Assists elderly in making informed decisions,

• Fosters trust between humanity and technology.

Transparency has value only when people are informed, enabled and involved.

6.5 Challenges and Limitations

While accountability and transparency in AI are essential, achieving them in practice is not easy. There are

several technical, legal, organizational, and social obstacles that can limit how much transparency is possible

or enforced. Understanding these limitations helps stakeholders plan for realistic, responsible AI governance.

6.5.1 Technical Complexity and Opacity of Advanced Models

Today's AI systems, especially deep-learning-based ones, can be pretty hard to interpret. These

models have thousands — or even millions — of parameters that multiply together in non-linear ways.

Challenges include:

• The black-box nature: It's difficult to explain why the model made a certain decision.

• Explanation tools are approximations: Although there are other research that LIME or SHAP can provide and explanation, these algorithms can only generate an approximation.

not always full clarity.

• Risk of oversimplification: Over-simplified explanations could mislead users into believing that the system is fewer than.

understandable than it is.

This makes the technological transparency a hard problem, in particular this happens in:

• Natural language processing systems (e.g., ChatGPT),

• Image recognition (e.g., face detection),

• Predictive modeling Example: Healthcare diagnostics.

6.5.2 Conflicts Between Transparency and IP/Security

Common justifications Organizations frequently cite intellectual property 14 reasons for not sharing how they decide.

their AI systems work.

Common tensions:

• Companies might be reluctant to share code, data or algorithms as trade secrets.

• An overreliance on the laser pointers could make it possible for hackers to tamper with or break into a system.

• Complete transparency may facilitate gaming (in exams, hiring tests or fraud.

detection).

Ensuring transparency yet maintaining business interest and ensuring the integrity of the system is an

ongoing challenge.

6.5.3 Cultural and Organizational Resistance

Its unfortunate, but many organizations are very reluctant to change particularly with respect to transparency and accountability.

Examples of resistance:

• No reward: Companies may value performance or speed over ethics.

• Fear of exposure: Transparency could expose errors, biases or unethical behavior.

• Hierarchical cultures: Some companies value top-down control and discourage scrutiny of

AI decisions.

• Lack of capabilities: Teams might not have expertise in ethical AI, fairness auditing or responsible governance.

This resistance paves the way to bad documentation, a human and inefficient oversight, and causes internal486 487 /503 starting fussiness.

governance.

6.5.4 Difficulty in Achieving Universal Standards

There is as yet no single worldwide standard for what constitutes an “accountable” or “transparent” A.I. system.

Barriers include:

• Distinct cultural norms: What is considered “fair” or “transparent” in one specific world does not necessarily in another.

another.

• Industry-specific needs: Health care, finance and education have different transparency needs.

• Evolving tech: Standards can’t keep up with the exponential pace of AI innovations.

• Misdefined terms: Words like explainability or fairness have different definitions across

disciplines.

In the absence of harmonized frameworks, AI developers and regulators sometimes encounter confusion or variation in

applying transparency principles.

6.5.5 Evolving Regulatory Landscape

AI is an emerging field in many countries, with few established regulations yet across much of the globe, and regulatory situations can be unclear or inhospitable

incomplete.

Key issues:

• New regulations on the horizon (eg, EU AI Act, NIST frameworks in US) but enforcement is

slow.

• No enforcement mechanism for current guidelines.

• Conflicting national laws make AI deployment across the world very complex.

• Businesses that span borders grapple with confusion over compliance.

Consequently, a lot of businesses apply a `wait-and-see´ strategy in which transparency regulations are implemented after legislation (or court rulings) have already been put into place.

become clearer.

Knowledge Check 1

Choose correct option:

1. Which of the following best defines accountability in AI?

A. Making AI models open-source for public use

B. Ensuring that AI systems work without human intervention

C. Assigning responsibility for the outcomes and impacts of AI systems

D. Publishing research papers on AI models

2. What is a major challenge in achieving transparency in deep learning models?

A. They are usually inaccurate

B. They require no training data

C. Their internal decision-making process is too complex to explain

D. They can only be used for image recognition tasks

3. What is the purpose of a “model card” in AI development?

A. To identify potential hackers targeting the model

B. To document the model’s structure and hyperparameters only

C. To summarize the model’s intended use, limitations, and performance metrics

D. To list the names of all developers involved

4. Which of the following is an example of a “human-in-the-loop” system?

A. An AI system that updates itself every hour

B. An AI that makes final decisions without supervision

C. A system where humans review AI recommendations before action

D. An AI tool used only for image filtering

5. What does "black-box model" refer to in AI?

A. A system built only for cybersecurity

B. An AI model that is simple and transparent

C. An algorithm with hidden code

D. A model whose internal logic is not easily understood

6.6 Summary

❖ As AI becomes more and more involved in making important judgments for society, it is crucial to guarantee accountability and

transparency becomes essential. This unit speculative artificial intelligence can incre ase ef ficiency, but also spent.

Introduce challenges regarding responsibility, fairness and trust—especially when decisions are automatized and conducted by.

difficult to explain.

❖ We investigated the meaning and scope of accountability, including to whom (e.g., developers,

institutions, and regulators that need to be responsible for AI systems and their results. Strategies for

accountability such as human-in-the-loop, traceability and ethical impact agendas3749.

assessments.

Transparency, meanwhile, means making AI systems understandable and open.

and what processes there are to encourage and enable thatAI (eg, eXplainable AI – XAI, documentation practices like model cards and data sheets, open-source peer

review are important tools in accomplishing this. And transparent systems enable users, regulators, and developers to

evaluate and trust AI decisions.

❖ But full accountability and transparency remain difficult to attain. The unit also examined technical,

aspects related to legal and organization limitations, e.g., complex developm ent model trade-offs betw een

intellectual property, and the absence of common standards.

❖ In summary responsible AI development is to weigh innovation against ethical guardrails, transparency and clearment of impact.

documentation, stakeholder engagement, and continuous review of new risks.

6.7 Key Terms

Responsibility: Responsibility of anyone involved with AI to take responsibility for the decisions made and their consequences.

Transparency – Extent to which AI processes and decisions can be placed u nder scrutiny and understood by

humans.

Explainable AI (XAI) – Methods to explain why AI made specific decisions in human terms.

Auditability – The capacity to track, inspect and review the way an AI system has arrived at a decision.

Model Card – A document that details an AI model's purpose, data usage, performance, and limitations.

Human-in-the-Loop – A system that relies on humans to vote to confirm or deny an AI decision.

Ethical Impact Assessment (EIA) – A process for considering the social and ethical impacts of an AI

system.

Black Box Model – An AI system that is not possible (or very hard) to understand its inner processes.

Governance – The policies, processes and structures that oversee the development and use of AI.

Regulatory environment — The various policies and assumptions governing AI systems by region and industry.

6.8 Descriptive Questions

What is Accountability in AI and Why Does it Matter?

Identify and discuss at least three major stakeholders who have responsibility for accountability in AI.

What is explainable AI (XAI)? What do we need transparency for?

Contrast the black box and white box models with an example.

Identify two of the challenges involved in trying to achieve complete transparency for AI systems.

What is the role of model documentation and datasheets in promoting AI transparency?

What is an EIA? How does it advance responsible AI deployment?

Share about the trade-offs in transparency and intellectual property.

6.9 References

1. Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. Fordham Law Review.

2. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv.

3. European Commission (2021). Proposal for a Regulation on a European Approach for Artificial

Intelligence (EU AI Act).

4. OECD (2019). Principles on Artificial Intelligence.

5. IBM Research. (2021). AI Fairness 360 Toolkit.

6. Mitchell, M. et al. (2019). Model Cards for Model Reporting. Proceedings of FAT*.

7. Gebru, T. et al. (2018). Datasheets for Datasets. arXiv.

8. NIST (2023). AI Risk Management Framework.

9. World Economic Forum (2020). Toolkit for Responsible AI.

10. Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: Evaluating Bias in Commercial AI Systems.

Answers to Knowledge Check

Knowledge Check 1

1. c) Assigning responsibility for the outcomes and impacts of AI systems

2. c) Their internal decision-making process is too complex to explain

3. c) To summarize the model's intended use, limitations, and performance metrics

4. c) A system where humans review AI recommendations before action

5. d) A model whose internal logic is not easily understood

6.10 Case Study

Who's to Blame? A Case of AI-Driven Loan Rejection

Background:

A bank unveiled an AI-based loan approval system that would see the joining of the parties involved in] "]}]}]}],"comments":[{"content":"Give a lending pitch deck for applysing.ai at Paris FinanXMLivitiation to contract quicker than ever.

decision-making process. Anjali, another applicant, had a similar experience when she wasn't extended a loan for housing without reason.

Asked for an explanation, customer service replied: "The algorithm did it. We cannot override it."

No additional information was given on the rejection criteria.

Issue:

Anjali was in good credit standing, had a stable income and no existing debts. She filed a legal complaint

from an unwillingness to take responsibility and be transparent." On further examination, the algorithm:

• 1st preference to urban candidates,

• Indirectly penalized certain incomes brackets,

• Failed to provide for human review or appeal.

Consequences:

• The bank was fined by the financial authority for operating in opaque fashions,

• The bank had needed to contain review by people in the loop,

• Trust in AI-driven decision-making by the public was flagged.

Key Lessons:

• Responsibility should be clearly codified — even with automated systems.

• Users should have access to explanation and appeal procedures.

• High-risk AI systems should be audited regularly and governed ethically.

Reflection Questions:

When in this AI system's lifecycle was accountability most absent?

What is a way that humans could have prevented this issue from ocurring?

Which transparency tools should already have been put in place?

Ethics in Artificial Intelligence_V3_Unit 7 (3).docx

Ethics in Artificial Intelligence_MBA_2

Ethics in Artificial Intelligence_MBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127423170

Submission Date

Feb 3, 2026, 10:27 AM GMT+5:30

Download Date

Feb 3, 2026, 10:37 AM GMT+5:30

File Name

Ethics in Artificial Intelligence_V3_Unit 7 (3).docx

File Size

38.9 KB

24 Pages

4,686 Words

27,867 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 7: AI in the Workplace

Learning Objectives

Learn how AI is effecting different areas of the contemporary workplace at industry.

Determine which industries and positions are most vulnerable to automation and displacement by AI.

Examine the ethical dilemmas AI presents in hiring, surveillance and employee decision-making.

Analyze strategies that organizations can use to enforce such integration of AI in ethical and responsible ways.

Consider the part that co-operation between human and machine could play in improving productivity and driving innovation.

Think about how AI will transform the skills, roles and even culture of the workplace in decades to come.

Acknowledge the power of lifelong learning and reskilling in response to AI-induced work environments.

Content

7.0 Introductory Caselet

"The New Intern: Human or Machine?"

Case Study: EVA the AI Coworker

Background

It was designed to manage scheduling, documentation and generate performance reports. At the time, you just assumed it was a high-end productivity plugin for the company. However, within weeks, EVA was already sending a client emails, predicting potential delays on the current project and assigning tasks to your team based on their work analytics. During the next team meeting, one of the members joked, “When’s she getting promoted?” points to EVA The room fell silent – not because it was a good joke, but because no one could explain how EVA arrived at her conclusion and who, if anyone, was supervising her decisions. It made you think. How much decision-making power can an AI have daily? Will it replace you in the future? Who’s to blame if it makes a mistake? Efficiency was high, but the issues were rising. Thus, critical thinking prompt: L000054 630066rnHow does a company balance efficiency and human judgement while scaling the use of AI?

7.1 Introduction to AI in the Workplace

Artificial Intelligence (AI) is rapidly reshaping how people work, collaborate, and create value in

organizations. AI isn’t only a back-end tool — whether automating mundane tasks or helping make decisions, AI comes to life.

it’s solidifying and the human labour is starting to get involved.

Key Features of AI in the Workplace:

• Automation of Routine Tasks

AI can free you from data entry, customer questions, scheduling and reporting da**n it.

repetitive work.

• Decision Support Systems

AI is also deployed to crunch huge volumes of data in advising business decisions related with marketing, hiring etc.

logistics, and finance.

• Human-AI Collaboration

Today, however, AI tools in the workplace often act more like a co-pilot: They can assist people as they try to do parts of their jobs better but can't do much on their own.

• Integration with Digital Platforms

Tools such as CRMs, communication apps, productivity dashboards and much else besides have had AI baked in.

workflow management systems.

• Real-Time Feedback and Monitoring

Some AI models also provide workers with real-time feedback on performance, which may determine how they ultimately perform.

evaluated or managed.

Key Benefits:

• Increased efficiency and speed

• Reduction in human error

• More personalized for customers and users at an individual level

• Remote and hybrid work hours.

Emerging Concerns:

• Job insecurity among employees

• Transparency of AI-informed decisions

• The ethics of surveillance and data privacy

• Blurring man and machine as we design new forms of humans to go with A.I.

In short, A.I. is not just a tool but a worker, too. As companies get more responsibilities organically, the way these firms are run will have to be different."

employee training, and ethical responsibilities.

7.1.1 Overview of AI Applications in the Workplace

Workplace applications of AI can be split into the following categories:

Automation of tasks: AI helps automating repetitive tasks; such as data entry, meeting schedule, invoice

processing, and answering FAQs.

Predictive Analytics: AI create anticipation using historical data to predict trends, sales, customer behavior or

equipment failures.

Natural Language Processing AI chatbots and virtual assistants, who can comprehend and answer

human language (e.g., in customer support).

Computer Vision: Applications to quality control in production or surveillance in security.

Recommender Systems: Recommending products, content, decisions to the user based on user data.

Performance Metrics: Tools that monitor employee performance or customer happiness in real

time.

AI makes processes smoother, it helps us make better decisions and do customization of user experiences.

7.1.2 Types of Jobs Affected by AI and Automation

AI impacts employment in three key ways:

• Fully Automated Jobs:

It's #2: Work that is very repetitive and rule-driven – this type of work can be fully automated away with AI, or replaced by robots.

Examples:

o Data entry clerks

o Telemarketers

o Assembly line workers

• Partially Automated Jobs:

AI aids these jobs but they still require human oversight.

Examples:

o AI on radiology workflows

o Financial Analysts using prediction tools

o Customer service reps supported by chatbots

• AI-Augmented Roles:

Jobs augmented by AI, in which labor is freed up for more-complicated, creative or social tasks.

Examples:

Managers using AI to task projects

o HR pros who use AI to screen on resumes at first filter

As a rule of thumb, predictable tasks tend to be more automatable whereas creative and relational work tends toNot all tasks are equal.

augmented.

“Activity: Map Your Job's Automation Risk”

Instructions to Learners:

Select a real-world job role (it can be your own, a family member’s, or a job of interest). Break down

the role into its main daily tasks. Then, research whether these tasks can be fully automated,

partially automated, or are uniquely human using AI tools or academic resources.

• List 5–7 core tasks of the job.

• Categorize each task as:

o Fully Automatable

o Partially Automatable

o Non-Automatable

• Justify each category based on research or reasoning.

• Reflect on what new skills would help the worker remain employable as AI adoption increases.

Deliverable: A one-page report or table that summarizes the analysis and includes a paragraph on

reskilling recommendations.

7.1.3 Sector-Wise Impact: Manufacturing, Services, IT, etc.

AI's influence varies across sectors:

• Manufacturing:

o Robots & computer vision for automatic assembly, inspection and packing.

o AI-based machine predictive maintenance.

• Healthcare:

o AI facilitates diagnosis (for example, detecting cancer), patient monitoring and drug discovery.

• Banking & Finance:

o Identifying fraud, credit risk scoring, algorithmic trading and customer service chatbots.

• Retail & E-commerce:

o Machine learning: Personalized product recommendations, inventory prediction and automated checkout systems.

• Information Technology (IT):

cybersecurity, systems optimization and automatic code generation AI tools.

• Education:

o Intelligent tutoring systems, autograders, and connected learning platforms.

• Logistics & Transportation:

o More efficient delivery, driverless cars and predicting the supply chain.

The level and speed of adoption are a function of the sector’s need for efficiency, data utilization and human interaction.

7.1.4 Changing Nature of Work and Skills

Artificial intelligence is not concerned only with the displacement of certain jobs; it is also changing where we work. We were but one known way leads to do it, that debt automation Will drop the world you won't lose the people Who live loon sanctuary around чтиш но}]е Social (ч^Definition о.

demand shifts toward:

• Analytical thinking and problem-solving

•Be better educated in digital literacy and tech

• Emotional intelligence and communication

• Creativity and innovation

• Adaptability and lifelong learning

New roles are emerging:

• AI trainers and explainability coaches

• Data ethicists and A.I. auditors

• Human-machine collaboration managers

Today's workforce needs to re-skill and up-skill in an AI-led workplace.

7.1.5 Opportunities Created by AI in the Workplace

AI not only destroys jobs — it creates them as well, namely:

• New job titles: Title VI specialists, robotics engineers, data scientists, AI product managers.

• More jobs: Workers would have more time for actual work — not "busywork."

• Work-from-Anywhere empowerment : AI to support and drive remote working, as well as productivity monitoring.

• Diverse hiring: If democratised, AI can help companies look in more diverse places for candidates.

• Business growth: Owners of small businesses can scale with A. I.-powered tools.

Artificial intelligence can be used to design fair, efficient, safe workplaces.

7.2 Job Displacement and Automation

The rise of AI and automation is transforming the workforce by redefining job roles, reducing demand for certain tasks, and sometimes replacing entire occupations. While new opportunities are created, many workers face disruption, uncertainty, and the need to adapt.

7.2.1 Historical Context of Technological Unemployment

Labour-displacing technical progress is nothing new in this respect. “Each new wave of inventions every, say, major wave of innovation — the telephone, the internet — “begins as a way to get closer to one another.

Industrial Revolution, the mechanization of farming or the advent of personal computing — has ever destroyed jobs."""

in the short-run but frequently ushered in long-term economic change.

Historical examples:

19th century: The power loom was invented, which diminished the need for textile workers.

20th century: ATMs cut the number of human bank tellers but spawned more jobs in financial services.

21st century: Automation in factories drove job loss in them but grew demand for

logistics and robotics.

This is what sets AI apart: it can replace not only physical labor, but cognitive and decision_queue_processing labor as well.

making work, making its effects wider and faster.

7.2.2 Mechanisms of Job Displacement through AI

Job displacement caused by AI There are several ways in which AI displaces jobs:

• Task Automation:

AI technology does specific tasks faster, cheaper, and more accurately than we do (e.g., chatbots)!

replacing call center agents).

• Process Optimization:

AI automation improves workflow and reduces redundant roles (e.g., AI-powered inventory systems which reduce overposting by determining when necessary manpower is needed to restock).

logistics staff).

• Decision-Making Automation:

AI takes over many repetitive decisions (like approving loans or scheduling shifts) so that middle-

management.

• Self-service Technology:

Customers can complete tasks without the need for human interaction, thanks to kiosks, apps and virtual assistants.

Unless there is a new position or change in the role that they can take, then it typically results in job roles being duplicated.

7.2.3 Low-skill vs High-skill Job Impacts

The AI impact on low and high-skill jobs is divergent:

Full automation poses a higher risk for low-skill jobs.

Examples:

o Cashiers

o Data entry clerks

o Warehouse workers

o Drivers (with autonomous vehicles)

• High-skill jobs might be augmented instead of replaced.

Examples:

o Doctors putting AI to the test for diagnosis

o How lawyers are using AI for legal research

o AI-assisted simulators for designing by engineers

Middle-skill occupations in particular, those with some routine and semi-complex tasks (e.g., insurance processing), confront

a “shrunk” future – it can be either automated or re-imagined.

7.2.4 Gig Economy and Algorithmic Management

The gig economy — work done on a task-by-task, flexible and often tech-enabled basis — has been growing rapidly for despite.

platforms like Uber, Zomato and Upwork.

AI plays a central role in:

• Task assignment (matching workers to tasks)

• Surge pricing (dynamic wages)

• Performance tracking (ratings, reviews, and pace data)

• Worker ratings (auto-deactivations or penalties)

It is an example of algorithmic management – the use by AI of information to take managerial decisions, without human input.

involvement.

Concerns include:

• Some intransparency regarding decision making

• Loss of control for workers

• No right of appeal or ability to question automated decisions

Temporary work can be flexible, but gig work also means no job security, benefits or negotiating power.

Did You Know?

“Were you aware that some ride-sharing companies modify a driver’s visibility to available work based on the individual’s acceptance rate, even if he or she has not been told?” KOLLMORGEN via AP“This practice, called algorithmic nudging, is a subtle form of behavior modification that can really change the whole atmosphere for workers,” writes @nytimes’s Noam Scheiber—suggesting that AI systems are controlling worker behaviors through covert incentives and penalties with... important ethical questions raised about consent & fairness.

7.2.5 Psychological and Social Effects of Job Insecurity

6.1 Psychological and social costs of AI mediated displacement Indeed, whether real or perceived as Machine-Induced Displacement; this could bring about several major psychological and social costs such as:

• Angst and existential fear

• Who you are and self-worth – 2nd only to work

• Investment in skills or career experiences Analogous to reduced feelings of ownership over skills and perceived likelihood of having a professional career

• Growing inequality as the digital have-nots fall behind

• Mistrust of tech or corporate leaders

On the macro side there is a huge amount of long-term damage from mass involuntary unemployment, for example:

• Social unrest

• Emphasis on welfare or retraining programmes

• Political chatter about universal basic income or automation taxes

So wrangling with the human toll of AI is just as important an exercise as imaginatively deploying the tech itself."

technology.

7.3 Ethical Issues in the Workplace

While A.I. has the power to turbocharge productivity and reduce hiring biases, it also sounds ethical alarm bells.

questions. So many of these are things people worry about whether it's fairness, privacy, discrimination and dignity.

transparency. And it must be wielded judiciously — or risk causing businesses and governments to accidentally do more harm than they may even intend.

employees or violate rights.

7.3.1 Fairness in AI-Driven Hiring and Performance Evaluation

• Resume screening

• Personality assessments

• Video interview analysis

Concerns:

bias in training data: if there were historic hiring practices that were themselves biased, the AI might learn those biases and reproduce them.

• Unfair ranking:ka:ka fits relevant items but through the use of algorithmic criteria it ranks inappropriate ones higher at times.

• Opacity: Candidates are seldom told why they were passed over.

• Reliance on A.I. to a fault: Managers might discount the context or judgment of humans.

Principle: Fairness and equal opportunity

Best practices include ongoing auditing, human scrutiny of decisions and algorithm "explainability," they said.

7.3.2 Workplace Surveillance and Privacy

AI is used to track what workers do — particularly in remote work or gig work contexts. Examples

include:

• Monitoring the keys pressed and mouse movements

• Emails and video calls were monitored.

• Monitoring how much time you are spending on assignments or platforms.

Ethical concerns:

• Violation of Privacy and Integrity Personalities

• Stressful work environments

• Surveillance creep, or surveillance that shouldn't be done spreading

• Lack of informed consent

I get they pose it as a productivity question but mandating 24/7 access can erode trust and morale.

7.3.3 Discrimination and Bias in Workplace Algorithms

AI systems can also purpose or encourage discrimination by:

• Gender

• Race or ethnicity

• Age

• Disability

• Socioeconomic background

Real-world examples:

• Hiring systems driven by artificial intelligence rejected women applying for engineering jobs after determining that they were dominated on those job sites by men and — because it observed what the applicants did, not what they said about themselves — casually learned from them.

• Algorithms that shortchange gig workers in some locations

The effect is real, even in the absence of overt discrimination — and legally, as well as morally, thorny.

What is the open question here is the matter of morality: Not treating some others fairly and equally.

Businesses, meanwhile, have to manually audit their systems constantly by comparing new results with old ones — to ensure that the relevant outcomes are fair across racial and gender lines.

7.3.4 Human Dignity and the Role of Labor

AI often reduces humans to data points o r units of product ivity without considering and the value of io n work unk.com

creativity.

Ethical questions include:

• Is it ethical to replace people with machines so you can make a buck?

• Do workers count as partners in innovation — or just expendable human capital?

• But is there room for people to make important work and express themselves?

Human dignity is to consider workers as more than widgets in a factory.

well-being.

They need to watch over the rightful place of technology in society as an instrument for human well-being, and not merely increased efficiency.

7.3.5 Transparency and Consent in Automated Systems

Workers often are not aware of how AI systems are making decisions that affect them — in hiring,

promotions, or surveillance.

Concerns:

• Model isn't interpretable (can't peek inside the "black box")

• No means of appeal or to contest automated decisions

•Not consenting for GATHERING/DATA PROCESSOR

Ethical principles at risk:

• Autonomy

• Informed consent

• Right to explanation

Transparency means:

• Empowering a not-too-tech-savvy audience to understand AI tools.

• Telling workers when AI is on the scene

• Allow for appeals and human oversight.

7.4 Strategies for Ethical AI Adoption

Adopting ethical AI isn't just about compliance—it's about building trust, fairness and sustainability in.

how organizations use AI. This is to help guide companies as they incorporate AI in ways that protect employee

rights, promoting inclusion, and advancing human-AI collaboration.

7.4.1 Codes and Ethical Guidelines with Respect to AIin Social and Work Life

Ethical guidelines; These will include principles or rules to ensure that AI applications respect human rights and values.

Examples of widely recognized principles:

• Equitable: Impartial fair treatment regardless of gender, race or background

• Transparency: API/behavior and ease of decision-making

• Individual responsibility: The product of the system must not belong to more than one person

Privacy – Only the personnel's files must be kept private.

• Human Oversight: Humans need to be in the loop to monitor and act on important events, at least eventually

Sources of ethical codes:

• European Commission Ethics Guidelines for Trustworthy AI

• IEEE's Ethically Aligned Design

• OECD AI Principles

Institutions must develop codes of ethics at the local level, and imparl all those who are party to them.

follow them.

7.4.2 Employee Involvement and Communication

There is no top-down ethical A.I. — it has to be developed by implication and trust among employees.

Ways to involve employees:

• Clarity around what AI is being used and why

• Necessary: Consultation and feedback loops before implementing AI tools

• Work or training at how to teach workers best to work with AI

Involving workers leads to:

• Better adoption rates

• Reduced fear or resistance

• Ethical vetting at the outset

That is creating a culture of collaboration, rather than a culture of fear around machines taking our jobs.

7.4.3 Reskilling and Upskilling for AI Transitions

As AI continues to transform job descriptions, companies must support lifelong learning, or risk having world-weary ex-staffers on the payroll.

behind.

Reskilling = Preparing Workers for New Jobs

Upskilling –> training your current employees to remain competitive.

Examples:

• Training administrative staff in use of digital tools

• Learning warehouse workers to manage robots

Helping HR people make the most of AI recruiting platforms

ETHICAL AI ADOPTION WILL DEMAND A PEOPLE-FIRST APPROACH – Enable all employees to thrive in the new digital.

workplace.

7.4.4 Designing Human-Centered AI Systems

AI as design should be human centered, in that it builds systems whose purpose is to amplify rather than replace humans. It focuses

on:

• User friendly: If it's not easy to use systems

• Back up the bus: AI should be riding in back, not driving

• Trust in human judgment: Humans should make decisions

• Ethical design choices: No dark pattern/gray interfaces.

7.4.5 Organizational Accountability and Inclusive Policies

Let's call that AI's green new deal (what it would be in this case is a move to make AI people-centric): If we want the deployment of AI to be "ethical": Organizations should Develop and sustain clear frameworks and inclusive processes.

Key elements:

• AI ethics boards or review committees

• Continued audits of A.I. tools for bias, transparency and fairness

• Redressal (Appeal and rectification of AI Decisions) steps

• Companies willing to adopt and expand AI are wide-ranging

• Compliance with labor and data protection laws

We should have a practical ethic of adoption: as accessible, transparent and enforceable globally as we can make it.

7.5 Future of Work

AI, automations and the future of work AI and automations are affectingchanging shaping the workforce into as we see it today. While technology will

there will be job cuts, but its profit-driven impacts will hit the hardest in For redefining human roles, and for rebuilding officesupdatedAt:1433818729.

and the need for new modes of leadership and governance. Organizations, governments, and individuals must

respond soberly and responsibly to them.

7.5.1 Hybrid Work Models and AI Integration

The trend toward adaptable work, remote and in office, was accelerated by the COVID-19 pandemic

work. AI is necessary to enable these environments and to run them today.

Key integrations include:

• A.I powered schedule to manage office usage and team presence

• Collaboration tools based on natural language processing (e.g., AI note-takers, voice assistants)

• Distance work productivity tracking analytics

• VR and AR for team meetings with immersive capabilities

AI may make productivity, cooperation and flexibility easier, but it also raises existential ethical questions about the sort of world we want to live in and who gets to decide.

surveillance and work-life boundaries.

Did You Know?

“Did you know that AI can predict meeting fatigue in hybrid teams? Some workplace AI tools now analyze meeting frequency, speaking time, and response patterns to detect employee overload or burnout in remote settings—prompting team managers to adjust meeting loads and encourage breaks.”

7.5.2 Rise of Human-AI Collaboration

Instead of transforming how we work by replacing humans, the future of work will be all about how people can collaborate with

intelligent systems.

Examples of human-AI collaboration:

• Doctors are looking to A.I. for quicker diagnoses

• Writer and designer productivity enhanced with generative AI tools

• Engineers are turning to artificial intelligence to replicate models and sniff out glitches.

• Customer service agents collaborating with chatbots

Successful collaboration depends on:

• Designing AI to be Human Friendly and Supportive

• Educating workers how to work with A.I.

• Establishing trust between people and machines

This shift requires a new mindset, where humans see AI as its ally rather than a competitor.

7.5.3 Redefining Workplaces in the Age of AI

Changes include:

• A new kind of worker is being chained to digital tools.

the posit ions of theoccupants）。° Collaps ible， light-weight.

• Forms of project-based, not permanent, positions

• The need for lifelong learning in work and life contexts

Work has been redefined in terms of flexibility and in the service of data and skills that work for value, not an institution.

outcomes rather than hours worked.

This new conception of the nature of daily work raises questions about job design, worker rights and well-being in an AI world.

driven era.

7.5.4 Ethical Leadership in AI Adoption

“We need leaders for this AI future that will pave the way for ethical, inclusive and sustainable applications of AI.

Traits of ethical AI leaders:

The Future That Awaits: A world in which we’re using AI as a tool to elevate humanity, not as a weapon to bamboozle it.

Section 2: Cultural Competence - Responsiveness to various staffing expectations and situations

Clear communication: To say clearly and bluntly what it is you’re doing with AI.

Empathy It's a balance of having the ability to make it happen while also giving a damn what happens when you actually do.

AI responsibility (liabilities) and AI systems with responsibilities of their own

Trust, fairness and value are central to the ethical lead¬ership of transformation AI is bringing.

7.5.5 Policy and Global Responses to Workforce Automation

Governments and multilateral organizations are facing up to the threat posed by AI and job automation through a

range of policy tools:

Key approaches:

: Programs for Government-Sponsored Retraining and Reskilling

"Test drives" with universal basic income (UBI) as an answer to mass automation

Taxing automation so everyone can profit from progress.

AI regulations to ensure transparency, fairness and human rights of workers

International cooperation around ethical AI standards (UNESCO, OECD, EU AI Act)

These responses are building towards a future of work that is fair and those that embraces technology in the public interest.

Knowledge Check 1

Choose the correct options:

1. What is a major ethical concern with AI-powered hiring systems?

A. They are too slow

B. They require too much manual work

C. They may inherit bias from historical data

D. They increase interview rounds unnecessarily

2. Which of the following is most likely to be fully automated by AI?

A. Creative writing

B. Strategic planning

C. Data entry and form filling

D. Leadership coaching

3. What is the role of "human-centered AI" in the workplace?

A. To eliminate human involvement in decision-making

B. To increase profits by reducing staff

C. To design AI systems that enhance human abilities and values

D. To develop AI systems that work only in laboratories

4. What does the term "algorithmic management" refer to?

A. Managing software bugs in AI systems

B. Use of AI to control and evaluate employees' tasks and performance

C. Hiring only programmers for AI roles

D. Replacing employees with robotic arms

5. Which policy approach supports workers affected by automation?

A. Increasing working hours

B. Introducing outdated machines

C. Launching large-scale reskilling programs

D. Removing performance incentives

7.6 Summary

❖ The nature of work is being modified by AI in exciting and challenging ways. From automating

jobs to remote collaboration AI technologies are changing the way individuals work 1.

organisations operate and how employees interact with their work.

❖ Even though automation can eliminate some low-level or routinized jobs, it will also generate new opportunities in the field of high\HttpFoundation =14 and mid-level skill jobs.

skill domains, such as data science, human-AI interaction, and ethical surveillance. The future of work will be

blended by hybrid models, AI and an increasing focus on lifelong learning.

❖ But the ethical questions are inescapable. from algorithmic bias,44 workplace

, surveillance, and job insecurity demonstrate the importance of responsible AI use". Organizations must

training, and adopt ethics, transparency , employee participation, and support.

upskilling initiatives.

❖ The future of work will require ethical leadership, equitable practices and cooperative design go.

7.7 Key Terms

Automation The use of AI or machines to perform a job that once required humans.

Humans and AI as Complements – An operation paradigm in which humans and AI work in complementarity.

strengths.

Algorithmic Management – Use of AI for managing workers for jobs or keeping an eye on their output.

evaluation, and scheduling.

Reskilling - Retraining workers who are at risk of losing their jobs to automation.

Hybrid Work Model – It is a flexible work structure with the mix of remote and in-office work.

Job Displacement – When you get fired because robotics or AI can do your job instead.

Workplace Surveillance – Auditing of staff through digital means or AI platforms.

Ethical Leadership – Fairness, transparency and human wellbeing oriented leadership style in the a) the 74 Clipper Race.

technology adoption.

Human-Centered AI – Developing AI technologies that are human-centric.

Policies that support everyone Organizational or governmental rules that make sure everybody is treated fairly and gets to participate

groups, particularly under conditions of technological change.

7.8 Descriptive Questions

Describe AI on the job in the new workplace.

What are the low-skill and high-skill job effects of AI?

What are the Ethical Implications of AI in Hiring?

How can companies incorporate employees into responsible AI implementations?

Explain what the psychological and social impacts of AI job insecurity are.

What is algorithmic management and what are its dangers?

Discuss on the importance of responsible leadership and ethical AI in working together.

Propose two approaches for building AI systems that are sensitive to humans?

What is the role of public policy in dealing with automation of workforce?

Can you articulate the future of work with regard to human AI collaboration and hybrid models?

7.9 References

1. Brynjolfsson, E., & McAfee, A. (2014). The Second Machine Age. Norton & Company.

2. World Economic Forum. (2020). The Future of Jobs Report.

3. European Commission. (2021). Ethics Guidelines for Trustworthy AI.

4. Binns, R. (2018). Algorithmic Accountability and Transparency in the Workplace.

5. OECD. (2021). AI and the Future of Skills.

6. IEEE. (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and

Intelligent Systems.

7. Accenture. (2020). Human + Machine: Reimagining Work in the Age of AI.

8. MIT Technology Review. (2022). Why Algorithms Can Be Biased and What We Can Do About It.

9. ILO (International Labour Organization). (2023). Reskilling for an Inclusive Future.

10. Harvard Business Review. (2021). How to Build Ethical AI in the Workplace.

Answers to Knowledge Check

Knowledge check 1

1. c) They increase interview rounds unnecessarily

2. c) Data entry and form filling

3. c) To design AI systems that enhance human abilities and values

4. b) Use of AI to control and evaluate employees' tasks and performance

5. c) Launching large-scale reskilling programs

7.10 Case Study:

AI in Recruitment – A Case of Unintended Bias

Background:

An AI hiring platform was used by a multinational corporation to select for candidates. The system

screened resumes, identified candidates for further consideration — and compiles a 'facial expression score' based on the person's video interview performance."

analysis and voice tone. At first HR was thrilled how much time they saved.

Issue:

It soon began to emerge that the algorithm was de-prioritising women because all data had been based on male hires.

for technical roles. Later it was discovered that AI learned from the past hiring data where most

applicants were men, reflecting a history of bias.

Furthermore, there was no discernable process by which and therecandidates are to be informed on how data would be utilised.

existed to contest denials imposed by the system.

Ethical Concerns Raised:

Bias in training data

Lack of transparency and consent

No review process by human or appeal.

Reduced diversity in hiring outcomes

Organizational Response:

The company halted the system and audited its algorithms.

It retranded the training set, and launched a human-in-the-loop review.

Candidates were now given rationales for decisions and tenancy of notice, defensive control from eviction.

human review.

Its future AI deployments would be managed by an in-house ethics team.

Discussion Questions:

When and how did ethical review break down in this case?

What would you have done differently to prepare launching of the system?

What does this case teach us about deploying HR tech in the future?

Ethics in Artificial Intelligence_V3_Unit 8.docx

Ethics in Artificial Intelligence_MBA_2

Ethics in Artificial Intelligence_MBA_2

ATLAS SkillTech University

Document Details

Submission ID

trn:oid:::3618:127350332

Submission Date

Feb 2, 2026, 11:32 AM GMT+5:30

Download Date

Feb 2, 2026, 1:02 PM GMT+5:30

File Name

Ethics in Artificial Intelligence_V3_Unit 8.docx

File Size

39.6 KB

25 Pages

4,623 Words

28,919 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 8: AI and Human Rights

Learning Objectives

1. Understand the relationship between artificial intelligence and international human rights principles.

2. Identify how AI technologies can either support or violate fundamental rights such as privacy, equality, and

freedom of expression.

3. Explore key risks of AI in areas like surveillance, discrimination, and access to justice.

4. Analyze real-world case studies where AI systems impacted human rights positively or negatively.

5. Examine legal, ethical, and institutional frameworks designed to protect human rights in the age of AI.

6. Recognize the responsibilities of governments, companies, and developers in safeguarding rights through

responsible AI deployment.

7. Evaluate strategies for creating human rights–centered AI systems, including transparency, accountability,

and participatory design.

Content

8.8 References

8.9 Case Study

8.0 Introductory Caselet

"The Border Bot: Fast Decisions, Slow Consequences"

Background:

A country is rolling out a new AI-based immigration screening system at its international airports, in an effort to help its officials process travelers more quickly

border control. The system evaluates passports, scans their facial characteristics, runs a check on a traveler's history and sends.

real-time "clear" or "flag"—all without human intervention.

Fatima, a conflict zone scholar, is flagged many times, denied entry and held for further separation.

questioning. She cannot learn why she was flagged — or even how the AI came to its decision. Later, she learns that

she was being watched due to an algorithm trained on skewed sources.

She is not alone. Cases like this have also been reported by advocacy groups that target people from certain regions or

ethnic backgrounds. The government insists the system is neutral—but human rights consequences are not.

growing.

Critical Thinking Question:

Is it possible to maintain efficiency and national security in the absence of fairness, dignity and human rights? What

safeguards should be in place?

8.1 Introduction to AI and Human Rights

Artificial Intelligence AI can enhance or threaten some basic human rights, depending on whether it is use.

is designed and deployed. AI begins to infiltrate government systems, as well as healthcare, education and finance — just })(Extended Page 3 on the other.to onlygmention two examples.

and law enforcement, it is taking on a growing role in how decisions are made that affect people's lives.

What Are Human Rights?

Human rights specify the minimum conditions necessary for people to live with basic dignity—whether they're in prison or not."

nationality, race, gender, or status. These include:

• Right to privacy

• Freedom of expression

• Right to equality and non-discrimination

• Right to work and equal treatment

• Judicial access and public services

Such rights are permitted by international laws, such as the Universal Declaration of Human Rights (UDHR).

and regional instruments such as the European Convention on Human Rights.

AI and Human Rights: Two Edged Sword

AI is on the one hand a defender and an enabler of human rights, but also a threat to certain human rights:

Potential to Promote Rights

Risks of Violating Rights

AI & Bot used in medical diagnostics can save lives

Minorities may be targeted unfairly through predictive policing

The nitty gritty on natural language tools that expand education community access Surveillance AI could invade the public's expectations of privacy

AI chatbots support and lift up the blind, deaf and others with special needs

Job discrimination Job discrimination is another unfortunate potential outcome of algorithmic bias.

Why Is This a Concern?

• Ai systems tend to work as “black boxes,” their logic is difficult to comprehend or challenge.

• The data that trained AI systems can carry biases, discriminatory practices or exclusions.

• People may not be able to defend themselves or when AI decisions are automatic and opaque

appeal outcomes.

Cases of Endangered Human Rights

• Privacy Rights: Facial recognition in public spaces

• Right to Travel: AI in border access protocols

• Right to a Fair Trial: AI risk-scoring in criminal justice

Equality and Non-Discrimination: Software as hiring and credit calculus

What Should Be Done?

Respecting human rights in AI demands:

• Transparent systems

• Human oversight

• Clear legal frameworks

• Ethical design

• Public awareness and participation

In sum, AI needs to be so developed and governed that it is human-centric, inclusive and accountable.

8.1.1 Understanding Human Rights in the Digital Age

The rights or set of entitlements of belonging to each person, that do not depend on beliefs or nationality.

human. These rights include:

• Right to privacy

• Freedom of speech and expression

• Right to equality and nondiscrimination

- The right to work, education and healthcare
- Right to fair treatment and due process

These rights will have new challenges and opportunities in the digital age. Deployments of things such as AI, big data, etc., are helping.

surveillance systems have introduced:

• AI used for detection of hate speech/accessibility improvements/and more.

• New opportunities for abuse (like facial recognition tracking peaceful protests)

Digital human rights now include:

• Data protection

• Freedom from digital surveillance

• Internet and information access

• Protection against AI-driven discrimination

The question is how we guarantee that advances in technology don't compromise human dignity or freedom.

8.1.2 The Intersection of AI and Human Rights

AI touches human rights in a billion directions – some good, some quite bad:

AI has the potential to promote human rights when it is:

• Utilized to detect disease early (right to health)

4

• For assistive devices for people with disabilities (right to participation)

To monitor and denounce human rights violations (freedom from torture, right to justice)

When AI is violating human rights : when:

• Employed for mass surveillance (invasion of privacy)

• Predictive policing, using biased data (non-discrimination violation)

• To classify or evaluate citizens (infringing on self-respect and autonomy)

The problem is that AI systems are used for a type of decision-making that traditionally was reserved for humans — hiring, to be exact."

policing, or loan approval. These systems can have 224 disastrous effects on the public when they are opaque, biased, or inadequately monitored.

widespread and harmful effects.

Did You Know?

“Did you know that AI systems used in border control can deny entry to travelers without any human

intervention or explanation?

In some countries, automated systems are making decisions about who can enter based on facial

recognition, biometric risk scores, and past travel data—often using opaque algorithms that can’t be

challenged.”

8.1.3 International Frameworks for Human Rights (e.g., UDHR)

Human rights are conducted by numerous international legal and ethical instruments. The most influential include:

Universal Declaration of Human Rights (UDHR) 1948 13.1 Everyone has the right to freedom of movement and residence within the borders of each state.

Adopted by the United Nations, the UDHR enumerates 30 basic rights, including a right to privacy.

equality, labor, schooling, and freedom of thought.

The ICCPR

Ensures rights like:

• Fair trial

• Freedom of expression

• Freedom from arbitrary detention

Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR)

In reference to data protection and digital rights, notably in the field of General Data Protection Regulation (EU) 2016/679.

Protection Regulation (GDPR).

the UN Guiding Principles on Business and Human Rights (UNGPs)

These would hold private companies responsible for ensuring human rights are protected in their operations — including AI.

development.

These frameworks are now being augmented or adapted to encompass the specific risks arising from AI in today's digital.

societies.

8.1.4 Risk Areas in AI Applications

If AI is not designed and deployed responsibly, there are several rights that it could threaten. Common risk areas include:

a. Privacy Violations

- Public facial recognition
- Mobile app tracking of your behavior
- Employers or governments watching you with the help of AI

b. Discrimination and Bias

- Biased hiring tools that favor one gender or race
- Credit scoring that harms poor communities
- Health algorithms learning on non-representative data

c. Lack of Transparency, Accountability

- A.I. systems without explanation making decisions
- Victims can't appeal computer-generated rejections

d. Freedom of Speech and Press

- Moderation algorithms that silence particular viewpoints
- AI bots pushing misinformation or propaganda

e. Access and Inclusion

- Exclusion of the least privileged without access to internet or digital literacy
- AI systems that are not available for people with disabilities

These are the risks that demonstrate the urgent requirement of governance, fairness, and ethical design in AI technology.

8.1.5 Role of Governments, Corporations, and Civil Society

It's a responsibility shared to make sure AI respects human rights:

Governments

Pass and enforce privacy laws (e.g., GDPR)

• Create AI national ethics frameworks

• Be transparent about using AI in the public sector

• Shield citizens from abuses by private actors

Corporations and AI Developers

• Do human rights impact assessments

• Do not train AI with biased or non-consensual data.

• Think about explainability and fairness when you design AI systems

• More openness about how the data are gathered and decisions made

Civil Society and NGOs

• Highlight human rights risks in AI

• Advocate for vulnerable communities

• Watch for abuses, and hold institutions accountable

• Promote public involvement in developing AI policy

"Part of the elephant" All must work together to make sure AI is adopted ethically and inclusively, with vigorous oversight and

accountability mechanisms.

8.2 AI's Impact on Fundamental Rights

As AI systems become more integrated into our daily lives, they increasingly affect fundamental rights

guaranteed by international law. These impacts may be positive, such as improving accessibility or safety, or

negative, such as enabling mass surveillance or algorithmic discrimination. This section explores specific rights

at risk.

8.2.1 Right to Privacy

The privacy right keeps people safe from unnecessary intrusion in. personal life, data and.

communications. Laws securing this principle include:

- UDHR 12
- Article 17 of International Covenant on Civil and Political Rights (ICCPR)
- A list of data protection law such as GDPR (EU)

AI's risks to privacy include:

- Facial recognition in public places without consent

Smartphone app and smart device-based behavioral profiling

- Systems for identifying emotions in work or school settings
- Predictive analytics employed by the police or insurance companies

AI can quickly and invisibly churn through vast volumes of people's personal information, often without user knowledge, or even notifying someone that the AI is in use.

giving them a way to opt out.

"Activity 1: Evaluate a Public AI System for Privacy

Instructions to Learner:

Choose any AI system used in public life (e.g., facial recognition at airports, school surveillance, smart

traffic cameras). Conduct a short privacy impact assessment by answering the following:

1. What types of data does the system collect (e.g., images, behavior, location)?

2. Who operates the system (government, private company)?

3. Are individuals informed about the data collection?

4. Is there a way to opt out?

5. What human rights concerns might arise (e.g., chilling effect, profiling)?

Deliverable: Submit a 300-word analysis identifying the privacy risks and suggesting two concrete

measures to improve privacy protections.

8.2.2 Freedom of Expression and Access to Information

Freedom of expression includes:

- Freedom to speak and write, freedom of assembly and discussion
- Freedom to obtain and access information

AI plays a major role in:

- Content moderation (filtering for hate speech, misinformation, etc.)
- News recommendation algorithms
- Search engine ranking systems

Risks to this right include:

- Under-deletion of false positives and errors: over-deletion of legitimate content by AI given wrong classification
- Censorship via biased or opaque filtering algorithms
- Echo chambers formed by recommendation systems that exclude different views

Algorithms silencing marginalized voices or topics

While AI is useful for controlling poison content, it cannot be allowed to stifle dissent and undermine public access to

information.

8.2.3 Non-Discrimination and Equality

Article The non-discrimination principle guarantees that individuals and groups receive the same treatment irrespective of their race, sex, religion, handicap or other.

status. Discrimination by AI occurs when:

- Training data are subject to historical biases
- Non-fairness-aware algorithms are not verified between different groups
- Privacy and information sharing: o Developers do not consider a variety of user groups in systems design

Examples of AI-related discrimination:

- Job-search algorithms that favor male candidates
- Credit scoring systems that penalize applicants from certain ZIP codes
- Predictive policing tools that are biased against communities of color

• AI image generators that misrepresent some ethnic or gender groups

There is no fairness testing and limited transparency which makes it difficult for the victims to identify and question_recognise_and_question

algorithmic bias.

Did You Know?

“Did you know that a hiring algorithm developed by a major tech company downgraded resumes that

included the word “women’s” (as in “women’s chess club”)?

The AI model was trained on historically male-dominated hiring data, which led to gender bias in job

recommendations and resume scoring.”

8.2.4 Right to Work and Fair Labor Practices

The right to work is also affected by AI in some respects:

• It can result in job displacement through automation

• It could improve productivity by helping workers

• It can, by algorithmization of work as well, shape labor conditions.

AI-related labor risks include:

• Algorithmic management in gig economy platforms (e.g., Uber, Zomato)

• Unstable work schedules caused by artificial intelligence-guided shift scheduling

• Automated monitoring of performance that harasses workers

• Loss of job to mass automation in manufacturing, retail or logistics

AI should be developed to respect dignity of work, assure transparent assessment and encourage mutual trust.

retraining opportunities for displaced workers.

8.2.5 Rights of Vulnerable and Marginalized Groups

And AI can make existing disparities worse when it is not carefully implemented. Vulnerable groups include:

• Ethnic minorities

• Persons with disabilities

• LGBTQ+ communities

• Women

• Refugees and migrants

• Economically disadvantaged populations

Examples of risks:

• Medical AI systems taught on biased data may underdiagnose some racial and ethnic groups

• AI platforms include accessibility barriers that prohibit disabled users

• Language models can reflect gender biases or offensive terminology

• Refugees can be left without services because of algorithmic self-checking IDs

Protecting these groups requires:

• Inclusive data collection

• Bias testing and auditing

• Involvement of affected groups in the design of AI systems

• Legal barriers to harm

8.3 Ensuring AI Respects Human Rights

As AI systems gain influence over critical aspects of society—healthcare, policing, education, employment— it becomes essential to protect fundamental rights. Ensuring that AI respects human rights requires proactive design, regulation, and ongoing accountability at both national and international levels.

8.3.1 Human Rights by Design: Principles and Implementation

Human Rights by Design is a forward-looking methodology through which human rights-based values are built into AI

systems end to end—from data collection, model building and deployment, to updates.

Core principles include:

• Respect for choice and agency: AI ought to respect the risk-taking choices of individuals.

• Fairness and nondiscrimination: Systems must be tested “fairly across demographics,” the committee concluded.

• Privacy and data: AI must treat data with respect for privacy, consent and obligations of care.

• Explainability: People should be able to understand how decisions are reached.

• Inclusion: All relevant parties, including the most marginalized, need to be part of system design.

Implementation practices:

• Cross-functional group (tech, legal and ethics)

• Human rights impact assessments

• Inclusive design workshops

• Bias and risk evaluations at every step

This is akin to 'privacy by design' but applied to all rights rather than solely data protection.

8.3.2 Regulatory and Legal Safeguards

Legal systems are necessary to implement protections for human rights, and keep AI developers and downstream users.

accountable.

Key regulatory tools:

• General Data Protection Regulation (GDPR) – Strict ish controls on use of data and automated

decision-making

• EU AI Act (proposal) — Risk-based framework for categorising AI systems and regulating applications in high-risk domains

• Digital Services Act (EU) - Algorithmic transparency in platforms is part of the rules

• National AI strategies (e.g., Canada, Singapore) – Foster the Trust in and Collaboration with Artificial Intelligence

Legal safeguards must include:

• Right to an explanation of algorithmic decisions

• Right to contest/confront the outcomes of AI-driven systems

• Limitations on dangerous applications of AI (such as biometric surveillance)

• Tight rules on handling sensitive data

Nations must also ensure that the rules keep pace with the rapid development of AI, without. 12 Feb 2020 If we are unable to do so in a timely fashion, existential risks may be inherent AI progress by developing self-regulating open source software and conduct- ing safe tests across diverse venues.

freedoms.

8.3.3 Transparency and Accountability Mechanisms

Transparency is essential for understanding and questioning AI decisions that impact on human rights.

Effective mechanisms include:

- Documentation (e.g., datasheets or model cards) for modeling outputs.
- Decision logs for high-risk AI systems
- Public catalogues of deployed AI systems
- Transparency of AI use to users

Accountability mechanisms help ensure there's a responsible party when AI does something:

- Appointing AI ethics officers
- Establishing internal review boards
- Requiring impact audits and third-party auditors
- Creating grievance redressal processes

Transparency and accountability enable users to question unfair decisions, seek redress and trust AI

systems.

8.3.4 Role of Ethical Audits and Human Oversight

Ethical audits are formalised evaluations of whether an AI system complies with ethical and legal requirements.

Elements of an ethical AI audit:

- Testing of bias (by gender; race; age)
- Data source verification
- Audit of algorithmic reasoning

• Assessment of implications for human rights

Human oversight includes maintaining a “human-in-the-loop”, in particular for:

High-value choices (e.g., medical, criminal justice)

• Environments that are sensitive (e.g., hiring, immigration)

• Dispute Resolution: (e.g., automatic loan refusal)

“If you have oversight, it keeps AI from replacing the need for human judgment and gives you a way to intervene when

something goes wrong.

Did You Know?

“Did you know that some companies now use independent AI ethics audit firms to evaluate their

algorithms before deployment?

These third-party reviewers check for bias, fairness, transparency, and compliance with human rights

principles, especially for high-risk sectors like finance, law enforcement, and healthcare.”

8.3.5 Global Cooperation and Policy Harmonization

AI has no borders, so upholding human rights needs international cooperation and policy

alignment.

Global efforts and initiatives:

• UNESCO’s Recommendation on the Ethics of AI (2021) – Universal ethics_framework

• OECD AI Principles – Transparency, accountability and human-centred values should be the stressedIncrease transparency, accountability and promote human-centred values

• GPAI (Global Partnership on AI) – Multilateral collaboration for responsible AI

• Council of Europe on AI and HR – Legal compatibility checks

Goals of global cooperation:

• Avoid a “race to the bottom” in ethical standards

• The rights protected should be consistent across jurisdictions

• Disseminate best practices and technical solutions

• Help developing countries use AI responsibly

Harmonized policies can help ensure that AI respects rights everywhere — not only in places with strong enforcement.

8.4 Case Studies on AI and Human Rights

8.4.1 Case Study 1: AI and Freedom of Speech on Social Media

Context:

Facebook (later YouTube and Twitter) – the world's biggest social media platforms (now X) are built on content moderation moderated by AI

deraastrates systems able to detect and take down posts that voliol te plaiomega policies, like hate speech, misinformation seaya violence

content).

Issue:

Much of this AI moderation has mistakenly pulled down legitimate Material, such as political criticism, news pieces

reporting in conflict areas and posts from human rights activists.

Human Rights Impact:

• Freedom of speech: Excessive removals reduce individuals' opportunities to share opinions, particularly in

repressive regimes.

• Transparency: Users are frequently not told why their content was deleted.

• Lack of appeal: Content takedowns are frequently final, with no real way to challenge the AI's decision

decision.

Lesson:

needs to be accompanied by human reviewers, especially in politically sensitive contexts, and the platforms guiActive

are required to offer reasons and appeals.

8.4.2 Case Study 2: Surveillance AI and Right to Privacy

Context:

Not everyone has been able to say no. In cities including London, Beijing and New Delhi, the police have tested facial recognition technology in public

safe and crime free places.

Issue:

These are systems that often run without any kind of public consent, and there is evidence of misidentification indeed.

with communities of color and women.

Human Rights Impact:

• Right to privacy: We are constantly monitored without our knowledge.

• Freedom of movement and association: The awareness of being observed may lead to less involvement in

protests or public gatherings.

• Discrimination: If members of marginalized groups experience higher error rates, they may be unfairly targeted.

Lesson:

AI on surveillance should be strictly regulated and publicly consulted since it is only appropriate for transparent, clearly

defined legal purposes.

8.4.3 Case Study 3: Algorithmic Discrimination in Welfare Systems

Context:

With the implementation of automated welfare fraud detection systems in countries such as the Netherlands and UK, governments began to detect welfare fraud.

fraud by analyzing applicants' data.

Issue:

One of those systems, called SyRI (System Risk Indication) in the Netherlands, alerted authorities to people it identified as likely to commit fraud

risk based on things like postal code, ethnicity and previous unemployment status.

Human Rights Impact:

Equality and non- discrimination: The Low-income families, as well as immigrant ones disproportionately provoked by the algorithm.

neighborhoods.

• Right to due process: The affected were given no information about how decisions were made or how.

challenge them.

• Human dignity: Tens of thousands of people were deprived benefits due to flawed or biased profiling.

Outcome:

That system was suspended when a Dutch court ruled that it violated privacy and non-discrimination rights.

Lesson:

Public sector AI must be transparent, auditable and not perpetuate systemic biases.

8.4.4 Case Study 4: Censorship and Content Moderation Algorithms

Context:

In authoritarian countries, it's not uncommon for governments to pressure platforms into training AI systems to squelch politically timing social media content.

scrutinized phrases including mentions of demonstrations, minority groups and opposition politicians.

Example:

Content in the posts shared there was blocked by AI filters on platforms in some https://apnews.com/ccd31a6ad051f4387691819017cc898cregions that included references to:

15

• Tiananmen Square (China)

• Kurdish rights (Turkey)

• LGBTQ+ advocacy (several countries)

Human Rights Impact:

• Freedom of information: Users are not allowed to access true content.

• Freedom of expression: Political opposition is muzzled.

• Marginalization: Minority groups are not visible online.

Lesson:

Platforms must resist government censorship control, uphold international human rights standards, and offer

user rights protections worldwide—not only in countries with democracies.

8.4.5 Case Study 5: Predictive Policing and Racial Profiling

Context:

A number of police forces throughout the US and UK have started using so-called predictive policing tools, which draw on historic crime data.

to forecast crime hotspots and suspects.

Issue:

However, since historical records are to some degree biased about where police patrol — and its demand on their behavior — the AI identified black guns at twice the rate they are really carried by civilians.

and Latino areas for more policing.

Human Rights Impact:

Discrimination: Reinforcement of racist profiling and over-policing of certain populations.

Right to equality before law: Minority groups are being discriminated against without reason.

Erosion of trust: Impacted communities lose faith in the police and justice.

Outcome:

These programs were suspended by several departments because of public backlash and no evidence that they work.

Lesson:

AI in criminal justice should be rigorously audited, incorporate community input and be evidence-based.

not bias.

Knowledge Check 1

Choose the correct options:

1. Which human right is most directly impacted by facial recognition technology used in public

spaces?

A. Right to education

B. Freedom of expression

C. Right to privacy

D. Right to property

2. What is the primary risk of using biased historical data in training AI algorithms?

A. System crashes

B. Data storage issues

C. Algorithmic discrimination

D. Faster decision-making

3. What does the principle of "Human Rights by Design" focus on?

A. Teaching human rights to software developers

B. Embedding human rights considerations in AI development

C. Designing human rights courses using AI

D. Protecting robots from human interference

4. In the case of predictive policing, what human right is at the highest risk?

A. Freedom of religion

B. Right to equality and non-discrimination

C. Right to education

D. Right to access entertainment

5. Why is transparency important in AI systems that affect human rights?

A. To allow governments to hide how systems work

B. To speed up AI decision-making

C. To help users understand, challenge, or appeal decisions

D. To reduce electricity usage in computing systems

8.5 Summary

❖ Artificial Intelligence powers how societies operate, governments govern and people live and

interact. Although it has the potential to improve human rights— with better health, participation in society and improved living》 standards for many people —this will■ not happen uncomplicated by ideological battles.

tools, ineffective service delivery—and serious risks to rights.

Key rights affected include:

• Privacy rights and the rise of AI surveillance

• The right to non-discrimination, when biased algorithms mirror social inequalities

• The freedom of speech: when content moderation suppresses legitimate speech

• The due process right when there is a lack of transparency, lack of means to appeal decisions

❖ Ensuring AI is consistent with human rights necessitates:

❖ Integrating “human rights by design” into supply chains of technology development

❖ Establishing strong legal and regulatory protections

❖ Transparency, accountability and oversight provisions include:

❖ Supporting international cooperation to promote convergence of standards among countries

❖ Engaging civil society and affected communities in the development of ethical AI

❖ AI can be developed in manners that are respectful, protective, and accountable through frameworks that carefully consider identity considerations.

and promote human dignity worldwide.

8.6 Key Terms

Human Rights – Fundamental freedoms and protections that everyone is entitled to, including freedom of expression, privacy,

and equality.

What to do About Algorithmic Discrimination In the cases where AI systems create unfair conditions for some constituents (usually because of the fact that they are + minorities or marginalized).

biased data or design.

Human Rights by Design – A design principle that embeds human rights standards into AI systems

from the beginning.

Surveillance AI — AI as a tool to watch what people are doing, often using facial recognition, sensors or digital tracking data.

tracking.

Freedom of Speech –The right to express ones ideas without fear of censorship or limit on personal activity.

Predictive Policing – Police using AI to predict future crime from historical data.

Bias in AI –AI systems have exhibited unfair treatment to certain people or groups, often on account of biased training data.

Ethical Audits – Formal reviews of AI systems to determine whether they comply with ethical +#+ legal obligations

standards.

Transparency – The capacity to comprehend how an AI system was trained and why it is making decisions.

Content Moderation –Methods (typically by AI) for filtering, blocking or excluding content on platforms eg

as social media.

8.7 Descriptive Questions

For which AI there are supplementary : To what extent can and should the right of privacy be design; infrastructure that supports algorithmic decisionmaking?

Elaborate on “Human rights by design” in AI building.

Examine the role of governments and civil society for safeguarding digital human rights.

What dangers do such algorithms pose to discrimination in public services?

Explain how surveillance technologies could impact freedom of movement and association.

How can we protect against abuse of AI in criminal justice systems?

What can international collaboration do to promote the ethical development of AI?

Why is transparency necessary for AI systems with human rights implications?

What are some of the lessons from case studies on predictive policing?

Give recommendations for how AI systems can be more inclusive of marginalized communities.

8.8 References

1. United Nations (1948). Universal Declaration of Human Rights (UDHR).

2. European Commission (2021). Proposal for a Regulation on Artificial Intelligence (EU AI Act).

3. Council of Europe (2022). Recommendation on the Impact of AI on Human Rights.

4. Access Now (2020). Human Rights in the Age of Artificial Intelligence.

5. UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence.

6. World Economic Forum (2022). AI Governance: A Holistic Approach.

7. Amnesty International (2021). Surveillance Giants: How the Business Model of Big Tech Threatens

Human Rights.

8. Algorithm Watch (2022). Automating Society Report.

9. OECD (2021). Principles on Artificial Intelligence.

10. Berkman Klein Center (Harvard University). Artificial Intelligence and Human Rights Toolkit.

Answers to Knowledge Check

Knowledge Check 1

1. c) Right to privacy

2. c) Algorithmic discrimination

3. b) Embedding human rights considerations in AI development

4. b) Right to equality and non-discrimination

5. c) To help users understand, challenge, or appeal decisions

8.9 Case Study: Algorithmic Discrimination in Welfare Distribution

Algorithmic Discrimination in Welfare Distribution

Background:

To combat the fraud and streamline the process, one country's national government began using an AI system to evaluate a petition from out of town.

welfare eligibility. The system scoured personal information such as income, address, family history and even a love of birds to work out whether applicants could be trusted.

education level and previous claimant usage to flag "high-risk" applicants for manual adjudication.

Problem:

The system was subsequently found to be biased, with a bias against players that included the subjcategory of individuals:

From low-income neighborhoods

With immigrant backgrounds

Who were unemployed or working part time

Many had support shut off with little or no explanation, and appealing was hard or impossible. The

patterns. system was discovered to have learned from biased historical data that patterned systemic

discrimination.

Human Rights Violations:

• Right to not be discriminated against (Article 7, UDHR)

• Right to social security (Article 22 of the UDHR)

• Right to due process and transparency

Outcome:

Amid public outcry and scrutiny, the system was deactivated. Courts ruled the algorithm

violated constitutional rights. Those affected were given some form of financial restitution while the goverment

tightened transparency and audit rules around future digital systems.

Key Takeaways:

Let’s make the algorithms in AI systems used by government transparent, fair, and answerable.

And any system for automatic decision-making must include human review.

Protecting our most marginalized, at-risk communities with inclusive policies and barriers to entry.

Ethics in Artificial Intelligence_V3_Unit 9.docx

Ethics in Artificial Intelligence_MBA_2

Ethics in Artificial Intelligence_MBA_2

ATLAS SkillTech University

Document Details

Submission ID
trn:oid:::3618:127350331

Submission Date
Feb 2, 2026, 11:32 AM GMT+5:30

Download Date
Feb 2, 2026, 1:03 PM GMT+5:30

File Name
Ethics in Artificial Intelligence_V3_Unit 9.docx

File Size
39.7 KB

25 Pages

4,844 Words

29,244 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Unit 9: Ethical AI Development and Governance, Worldwide AI Policies

Learning Outcomes

1. Understand the foundational principles for developing ethically sound AI systems.

2. Recognize the crucial role of policymakers and regulators in AI oversight.

3. Explore global frameworks and models that support AI governance.

4. Compare key AI policy actions from the United States and the European Union.

5. Identify emerging legislative trends in AI across different countries.

6. Summarize key ethical, regulatory, and governance insights related to AI development.

7. Apply conceptual knowledge through analysis of real-world AI policy case studies.

Content

9.0 Introductory Caselet

“The Machine and the Monk: A Dialogue on Intelligence”

Background:

Meera, a Bengaluru-based final-year computer science student secures an internship as part of a fellowship at an AI research

lab in San Francisco. The lab is working on neural networks so advanced they can simulate emotional

intelligence.

Eager and also apprehensive, Meera starts to wonder—can a machine truly fathom the complexity of human values?

Over a weekend, she goes to the Zen monastery just outside of town. There she encounters a Japanese monk

called Tenzin, who had also studied quantum computing before becoming a monk.

While walking through a peaceful bamboo grove, Meera explains her reservations.

Tenzin smiles and says,

“One knife and one scalpel, they,re both sharp. One can cure, the other can kill. No, the tool itself is not their point of difference — but I love how you think!

in the intention that its use suggests.”

He continues,

“Wisdom is the intelligent use of intelligence, by proper rule to not do what you can do.”

They sit in silence while the wind moves through the bamboo, and Meera for once thinks that perhaps the real frontier of

AI may not be intelligence — but it could be ethics.

Critical Thinking Question:

In a world where machines can learn better than people, who should teach them.MESSAGEEND\lang1033\langfe1033\lQuote Who should "teachers" be when it comes to teaching machines (I mean as in AI learning) into thinking.

ethics—and how?

9.1 Principles for Ethical AI Development

ARTIFICIAL INTELLIGENCE IN SOCIETY As artificial intelligence moves into the mainstream, making its way into a growing number of areas • through health care and finance to education AI in journalism.

governance — there's an increasing demand that it is used responsibly. Ethical AI development means

the designing, building and using of AI systems in line with moral values, human rights and societal well- being.

being. It is not solely about what AI can do, but what it should. The challenge is to build AI systems that we can trust,

just, and good for all people in a non-harmful and nondenigrating way.

Here are the key principles by which we should be guided for ethical AI development:

9.1.1 Core Principles: Fairness, Transparency, Accountability

Fairness

Fairness in AI is about the system treating everyone fairly, whether they be black or white, male or female, and so on."

socio-economic status. For instance, an AI used in recruiting should not be biased against female candidates.

simply because the training data exhibited a pattern of past male hires. Equitable practice Enable more proportional share of benefits to reduce bias and create better.

justice.

Transparency

Transparency entails making AI systems comprehensible. One should be able to know how and why a decision

was made by the AI. For instance, if a lending application was denied by an AI system then the individual ought to be able to

understand the reasoning behind it. Users can get the purpose of.

system.

Accountability

The concept of accountability means that someone should be held accountable for an AI system's decisions. If the AI makes a

error — like erroneously denying healthcare services — should humans be held responsible. Developers,

companies, and regulators ought to take to make sure that the AI is used responsibly and ethically." It also includes having processes

in place for redressing harm caused by AI.

9.1.2 Human-Centered and Trustworthy AI

Human-centered AI is the principle that the system should be designed with human values, needs and dignity at its core. It

are supposed to augment human decision-making, not replace it altogether. Reliable AI is when people can trust

the system to work as intended and not cause them harm or violate their data.

For instance, In health industry, An AI recommending healthcare policy for patient diagnoses must support doctor's perception.

judgment, not override it. The system must be safe, humane and with a view to improve human well-being.

A trusted AI also features attributes such as explainability, ensuring users can see how it functions.

9.1.3 Privacy and Data Protection by Design

This requirement suggests AI systems must be designed in a manner that would prevent peoples'personal data from the outset as it's only possible.

beginning—not as an afterthought. The system should only gather the data it needs, keep it secure, and provide for family-controlled repositories."

so users determine how their data is used.

For instance, a fitness app that recommends exercises based on AI should request permission before seeing location_permission_maintain logs.

data or personal health information. It should also give people the option to delete their data. Privacy by

design ensures AI respects the rights of the individual and engenders user trust.

Did You Know?

“The concept of "Privacy by Design" was first introduced in the 1990s—long before AI became popular.

It means privacy is built into a system from the very beginning, not added later. It became legally binding

in the European Union under the General Data Protection Regulation (GDPR) in 2018.”

9.1.4 Inclusivity and Non-Discrimination

Dec. 11, 2019 Inclusivity means AI systems must work well for people of all backgrounds — which includes different languages,.

cultures, genders, and abilities. Non-discrimination means the system should not differentiate by being unfair to any one or

deny people on grounds such as disability, race or age.

Just for example, a voice recognition system better be able to handle accents from around the world. If it only

works well with some voices (e.g., male American English), but is biased against others.

It makes AI better suited to everyone, by serving us all fairly and equitably.

9.1.5 Safety, Security, and Sustainability

Safety

Safety is the guarantee that AI systems will not harm people or the environment. For example, a self-driving

car needs to be tested comprehensively to make sure it can avoid accidents in different scenarios."

Security

Safety includes preventing AI from being exploited, hacked or manipulated. If an AI system is hacked, it

could potentially be used to spread misinformation, steal data or disrupt services. SysNBD and Secure AI protect also the model, because;linebreak they are executed on encrypted quantities.

its users.

Sustainability

Sustainability tries to minimize the environmental cost of AI. For example, training

high energy AI models. Developers need to figure out how AI can be more energy efficient and

environmentally friendly.

9.2 Role of Policymakers and Regulatory Bodies

Policymakers and regulators are central to the choice about whether and how AI is deployed in society. Their responsibility

is to formulate laws and guidelines, action plans and strategies that make AI safe, fair and good for all humans. While

tool builders to ensure that the tools they build for AI do not harm preventive institutions.

people, abuse data or widen inequality. Their job entails defining morality, guaranteeing that the law upholds standards coming from a natural right standpoint,:^(11) defending public morality and regulating private morality.

transparency, and encouraging responsible innovation.

Let's see who is playing what part on what stage:

9.2.1 National Governments and Legislative Oversight

AI is primarily regulated by national governments in their respective territories. They create

laws and rules that govern the development and use of AI. This can include rules on:

• Privacy of data (for example: Digital Personal Data Protection Act of India)

• AI adoption in sensitive areas, such as health care, policing and finance

• Responsibility in the event of damage resulting from AI systems

Governments also create dedicated regulatory agency or task force for the regulation of AI application. Legislative

oversight ensures that AI systems do not fall into a legal vacuum.” Parliaments and the role of legislators in

approving legislation to codify acceptable uses of AI and protect citizen rights.

9.2.2 Role of International Organizations (e.g., OECD, UNESCO)

International organizations contribute to setting global norms on AI ethics and governance. These bodies bring

from numerous countries into common standards. Their work is helping to ensure that AI does not develop into a

instrument for exploitation, war or spying on a world scale.

• The OECD (Organisation for Economic Co-operation and Development) hawks fanciful principles such as

in the name of human-centered values, as well transparency and accountabil ity when it comes to AI.

UNESCO (United Nations Educational, Scientific and Cultural Organization) today published

advice on the ethical application of AI, atoreoentoof human rights and environmental

sustainability.

These types of organizations are critical as AI technology flows across countries, and the global work makes certain that

cooperation and shared responsibility.

9.2.3 Public-Private Partnerships in AI Regulation

(William) “AI development is predominantly driven by the private sector, who rely on these.

governments. The counterbalance of regulation and innovation is achieved through public-private partnerships.

In these partnerships:

• Tech companies and governments partner to develop ethical standards.

• Corporations provide technical expertise to educate policymakers about AI.

• Joint programs could be funded to develop trustful AI applications, e.g., privacy preserving facial recognition

safeguards.

This partnership prevents profit from taking precedence over people, and government from making sole decisions!

rules without understanding the technology.

9.2.4 Ethical Advisory Committees and Think Tanks

Governments and intergovernmental organizations commonly resort to independent advisory bodies to develop ethical.

decisions about AI. These include:

• Ethical advisory boards composed of scholars specializing in law, technology, sociology and philosophy.

• Research organizations, or think tanks, that make policy recommendations.

These bodies evaluate risks and benefits of new AI technologies, as well as offer advice on things like bias in

algorithms, AI as a tool of war, or dangers of surveillance. They are instrumental in policy being tied What they aid, at their best, is keeping policy rooted in the concerns of the real world.

and moral reflection.

9.2.5 Funding, Innovation, and Ethics in Public Policy

Governments are the largest funders of AI research and innovation especially in enterprises that are less likely to do so themselves.

business-friendly, but it is crucial to society — like AI in public health care, agriculture or.getOwnProperty; What are the…

climate change.

At the same time, public policy must ensure that this money is directed at ethical development:

• Issing grants with the goal of supporting AI projects that advance fairness, transparency or environmental sustainability

• Establishing innovation hubs with ethical standards

• Ensuring startups and researchers adhere to fundamental principles of using AI responsibly

This is how the AI sector can continue to grow and at the same time remain consistent with social and ethical responsibility.

9.3 Frameworks for AI Governance

AI governance is the rules, processes and practices that have to be implemented in order that develop- ment, deployment andaleu2 use of AI technology methods are ensure wereigher :

used responsibly. It’s more like making a structure where innovation can take place safely, without hurting people.”

or society. Governance is necessary because AI can be complicated, unpredictable and make erroneous decisions.

important decisions on their own.

AI governance models are designed to provide roadmaps for developers, users, businesses and governments in.

embedding AI systems in ethical precepts, human rights and the rule of law.

9.3.1 Introduction to AI Governance Models

2 AI governance models: what are they and how can you implement them? These models can be

constructed by states, industries or supra-national organizations. Their primary aim is to optimize the gains of AI

(efficiency, automation, discovery) along with the risks (bias, job loss, surveillance, etc.).

Types of models:

• Regulatory systems: Concentrate on establishing laws and rules.

• Ethical models: Emphasize values such as fairness and justice.

• Industry-driven models: Stress voluntary requirements or best practices.

These models may differ from country to country and sector to sector, but they all seek to ensure trust, safety, and fairness in AI.

9.3.2 Risk-Based and Rights-Based Governance Approaches

Risk-Based Governance

This is a strategy to reduce the level of risk regarding various AI applications. For example:

• A recommendation system for music is low stakes.

&bull A system used in policing that scans faces is high risk.

Depending on the risk, the AI system could be subject to more-stringent rules or require special approval. This method is used in the

EU's AI Act, where the AI systems can be categorized as minimal risk, limited risk, high risk and

unacceptable risk.

Rights-Based Governance

It is a modest route of safeguarding human rights – privacy, dignity, freedom of speech and

equality. It asks:

• Does this AI system honor the user's right to understand how decisions are made?

Does it shield people from surveillance or discrimination?

Rights-based approaches gives more emphasis to justice, fairness and individual liberty regardless of the political systems.

level of technical risk.

9.3.3 Technical Standards and Certification Systems

Technical standards are explicit rules and thresholds that AI systems must reach objectively to be considered safe, and valid.

ethical. These standards may cover:

• Accuracy

• Security

• Fairness

• Data protection

• Explainability

For instance, an AI medical device might need to satisfy health tech standards on par with what Emmers, Cicilline and lawmakers proposed in the Food and Drug

Administration (FDA) requires.

The certification system is one where the AI system would be tested and certified before being deployed.

market. Just as electrical devices bear safety markings, AI systems might display markers to show that they have been.

reliable and ethically acceptable.

These are tools that can give developers and users confidence in the trustworthiness of an AI product in compliance with legal and ethical norms.

9.3.4 AI Lifecycle Governance: From Design to Deployment

AI governance is not something done once, but rather needs to span the life cycle of the system:

Design stage: Ethical goals are established; risks are mitigated.

Development Data collection and model training responsibly.

Testing stage: The system is tested for fairness, safety, and efficiency.

Deployment phase: AI is deployed into the world with monitoring and human supervision.

Post-deployment: Systems are subject to periodic review and may be updated or decommissioned if they become harmful.

Lifecycle governance safeguards that ethical perspectives are already integrated in to the system from the start, rather than sidelined 25 Lifetime of health research depending on data is certainly a facet of change.

added later as a correction.

“Activity: Mapping the Ethical Lifecycle of an AI Chatbot System”

Choose any one AI application (such as a chatbot, facial recognition system, or recommendation engine).

Create a lifecycle map showing the five key stages:

1. Design

2. Development

3. Testing

4. Deployment

5. Monitoring

For each stage, list at least one ethical risk and one mitigation strategy (e.g., ensuring fair data, human

oversight, etc.). Present your lifecycle as a diagram or table, and add a short paragraph explaining how

governance improves system trustworthiness.

9.3.5 Challenges in Global AI Governance Alignment

Artificial intelligence is deployed across countries and industries, but each country may have different rules and values for the technology. This creates

major challenges:

• No consistency: One country might approve facial recognition, another might ban it.

• Empty laws: Some countries have yet to pass any AI laws.

• Cross-border concerns: AI systems developed in one country can affect people in another.

•Corporate influence: Big tech firms may try to press their own standards and weak that devising those guidelines in the meantime.

governments.

Global AI governance is challenging, because it requires collaboration across countries and cultures.

legal systems. It must also balance innovation and control, local rights and global corruptions.

9.4 Landmark AI Policy Actions: US & EU

The U.S. and E.U. are world leaders in creating regulation and policy to mandate the safe and secure

and responsible use of AI. Their approaches help to guide how AI is governed in the rest of

world. While the US leans towards guidance, innovation and partnership, the EU is more oriented towards strict.

regulation and citizen rights.

The following are key policies and frameworks from both regions.

9.4.1 The US Executive Order on Safe, Secure, and Trustworthy AI

making." In October 2023, the President of the United States signed an Executive Order (EO) on AI that specifically mentioned the need to produce wicked problems likewise natural for AI reasoning systems.

safe, secure, and trustworthy.

This EO lays out wide-ranging government-wide actions to:

Promote responsible AI development

• Uphold the rights and safety of citizens

• Ensure US leadership in global AI innovation

Key areas covered include:

• Requiring that AI systems be tested for safety before being widely circulated

• Privacy protections, particularly relating to biometric data

• Combating prejudice and algorithmic bias

• Promoting innovation with research grants and public-private partnership

• Establishing standards for the federal government's use of AI tools

This Executive Order does not establish new rules but encourages federal agencies to make specific efforts to address

Threats posed by AI to the public good.

Did You Know?

"The 2023 US Executive Order on AI marked the first time the US government required AI developers

working on foundation models (like large language models) to report safety test results to federal

authorities—even before releasing them to the public. This was a major shift from previous self

regulated practices."

9.4.2 NIST Framework for AI Risk Management

AI Risk Management has been developed by the National Institute of Standards and Technology (NIST)

Manager's Framework (AI RMF) to assist organizations in addressing the risks related to AI systems.

The system is voluntary as well as flexible and intended to suit public and private sectors.

Its goals are to:

• Boost the trustworthiness of AI systems

• Encourage responsible development and use

• Help spot and address risks: of bias, lack of transparency or security threats

The NIST structure consists of four core functions:

Map AI risks

Measure and analyze those risks

Control and safeguard them

Oversee the process with oversight appropriate for it

It

offers pragmatic approaches for developers, authorities and companies to reconcile innovation with

responsibility.

9.4.3 The EU Artificial Intelligence Act (EU AI Act)

The world's first comprehensive legal framework for AI came into being in 2024 with the passing of the EU AI Act.

Its primary objective is to ensure core rights, safety and democracy, with a space for innovation.

The Act applies to:

- AI systems developers and deployers in the EU
- Systems employed outside the EU but affecting EU citizens

The Act has specific rules, responsibilities and penalties included with it to make enforcement much stricter than US

policies.

It categorizes AI systems by type of risk and dictates how they should be regulated.

9.4.4 Key Provisions: Risk Categorization, Bans, and Obligations

Risk Categorization

AI Act by the European Union categorizes AI systems into four classifications:

Banned (e.g., social scoring by governments) – Unacceptable Risk

High Risk — Stringent requirements (e.g., AI in hiring, healthcare, policing)

Small Risk – Requirements : Transparency (bots should say they are AI).

Low-risk – Regulation is not required (for example, spam filters or video games)

Banned Practices

And some applications of AI are banned outright:

- Biometric public space surveillance in real time
- Manipulative AI that preys on vulnerable individuals
- Predictive policing with data on personal behavior

Obligations for High-Risk AI

Organizations using high-risk AI must:

• Perform risk assessments

• Ensure data quality

• Maintain documentation and logs

• Provide human oversight

• Permit audits and regulators' enforcement

Penalties for infringements can be as much as €30 million or 6% of worldwide annual turnover.

9.4.5 US-EU Comparison

Aspect

United States

European Union

Nature of

Framework

Voluntary guidance (Executive Order,

NIST RMF)

Binding legislation (EU AI Act)

Focus

Innovation, safety, and national security

Human rights, risk mitigation and consumer_segments.

safety

Risk-Based

Regulation

Less formalized

Clear and structured risk categories

Use of Bans

Rare

Accountability

Powerful prohibitions on harmful uses of AI

Encouraged through agency actions

Enforced through penalties and audits

Implementation

Through federal agencies and partnerships

Through EU-wide legal enforcement

mechanisms

The US and EU both want ethics-based AI, but the US proves more flexible and innovation-oriented.

while the EU perspective is more restrictive and right-based.

9.5 Global Legislative Trends

With emerging technology that includes artificial intelligence becoming increasingly powerful and ubiquitous, a number of countries have start.

the need for legislation, policy and guidelines to manage their use. These initiatives are intended to even out the playing field of AI

services, economic growth and innovation) against the risks (in bias, surveillance or harm to privacy).

jobs and privacy).

Some countries are moving more quickly than others, of course, but if we remove paralysed from the equation almost every part of world today is in an active discussion or drafting mode.

AI-policy, and international collaboration is increasingly needed to cope with the (cross-border) effects of AI.

9.5.1 AI Mentions in National Legislatures Across Continents

over national laws, bills or official strategies over the past few years 2.ctrlSome in recent year 3.

related to AI. These efforts often include:

• National AI strategies: When AI research, investment and education will bear fruit (e.g., India, Canada wxT).

UAE)

Privacy laws: For other privacy requirements in AI systems (e.g., Travel (GDPR in Europe, LGPD in Brazil))

• Ethical AI guidelines: Principles to develop responsibly

Examples:

• In the United States, Congress talks a lot about AI, with attention on risk management and

innovation.

• AI is getting officially regulated in the European Union with the EU AI Act.

• India published a national AI strategy seeking inclusionary growth with AI in sectors such as agriculture and

education.

It's a sign of an increasingly urgent acknowledgment that the law, and courtroom practices, need to find their way back from tech landia.

9.5.2 Asia-Pacific: China, Japan, South Korea

China

China has a concentrated and rapid advocacy of AI policy, spurred by its aspiration for global leadership in the field of AI by

Key features include:

• Government investment in AI startups driven by government

• Controls on deepfakes and recommendation algorithms

• Tighter regulation regarding what AI does in media and public discourse

China's AI emphasis is on national security, economic power and social governance.

Japan

Human-centered and ethical AI is the way in which Japan perceives AI. Its policies emphasize:

• Transparency

• Accountability

• Public trust

Japan endorses co-regulation between government and industry, and is engaged in several international AI

ethics forums.

South Korea

South Korea is spending heavily on AI in fields like manufacturing, education and public services. It has:

• A national AI strategy that prioritizes innovation

• Privacy protection laws

• Government-backed codes of conduct for developers of AI

These countries exhibit a combination of state-led, ethics-based, and innovation-driven approaches in the Asia Peter Tague 2-Pacific region.

Pacific region.

9.5.3 Africa and Latin America: Emerging Approaches

Africa

The digital infrastructure in most African countries remains underdeveloped, but interest in AI is increasing. Key

features include:

• AI in agriculture, healthcare and climate risk management

• Regional collaborations such as the African Union's AI strategy

• Worries that foreign tech companies could be importing AI systems without local supervision

Hurdles include lack of appropriate legal frameworks, funding and human resources.

Latin America

It is governments in countries such as Brazil, Mexico and Chile that are making progress on:

• Developing policies and ethical guidelines around AI

• National policies on AI strategy (e.g., Brazil and its policy of an AI Strategy in 2021)

• Public input into policy design

Latin America's focus on inclusion, fairness and social dev't but with problems similar to Africa

in enforcement and infrastructure.

Rooms are reportedly early but promising experiments in AI governance.

9.5.4 Regional Cooperation Initiatives (e.g., G7, G20, OECD)

As AI affects multiple countries at once, regional and global collaboration is required.

G7

The G7 (Group of Seven major economies) encourages:

• Principles for trustworthy AI

• Popular methods of regulating AI

• Sharing of information and ethical values

In 2023, the G7 launched the Hiroshima Process to develop a shared set of rules for AI that can create.

G20

The G20 countries focus on:

• AI as the engine of global economic growth

• Developing standards for inclusive, human-centered AI

• Striking a balance between innovation and privacy and data protection sovereignty

OECD

The Organisation for Economic Co-operation and Development has:

• A popular bouquet of Ai Principles

• A global AI Policy Observatory that monitors and advises national policies

• Standardization and measurement for AI impact

Such partnerships offer opportunities for countries to learn from one another and can help stave off decentralized or contradictory regulations.

9.5.5 Future Outlook: Toward Global AI Governance

As AI systems become increasingly powerful and global in reach, we are under intensifying pressure to build a global.

framework for AI governance. This may include:

• A U.N.-sanctioned treaty or declaration concerning the ethical use of AI

• International norms on the safe use and protection of AI systems

• Joint research and testing institutes for AI risk assessment

• Cross-border enforcement and accountability regulatory agreements

However, challenges remain:

• Different cultural and political priorities

• Lopsided power between tech giants and small countries

• Battle between nations for AI leadership

But that is improving, and the trend now is for more global discussion even, collaboration in it?."

AI policy.

Knowledge Check 1

Choose the correct options:

1. Which of the following is not one of the core principles of ethical AI?

a) Fairness

b) Speed

c) Accountability

d) Transparency

2. The EU AI Act classifies AI systems into how many risk categories?

a) Two

b) Three

c) Four

d) Five

Which organization released the AI Risk Management Framework in the US?

a) UNESCO

b) OECD

c) NIST

d) G7

4. In AI governance, a ‘rights-based’ approach primarily focuses on:

a) Algorithm efficiency

b) Reducing costs

c) Protecting human freedoms and dignity

d) Increasing speed of development

9.6 Summary

❖ AI is already a fabric of our society, its strength must however be balanced

with responsibility. Fair,Fair, transparent, and ethical is based on principles such as:transparent,and ethical AI development requires guidelines Principles.

responsibility, and protection of privacy and human rights. key role and who policy makers and regulatory bodies have an essential

purpose of AI — by way of regulation, international collaboration, public-private co-operation and standards development.

partnerships, and ethical advisory boards.

❖ AI Governance Frameworks that mitigate risks and promote human rights across the full life-cycle of AI, ranging from creation allto adoption and to retirement.

design to deployment. US and EU approaches demonstrate differing governance styles—one)}}

flexible and innovation-centric, the other restrictive and rights-oriented. Worldwide, other nations' laws and National Security Laws Act 50 of 1950 (July) applied AI regulationElsewhere in the world, more countries are putting into place AI

related laws, and international bodies like the G7, G20 or OECD strongly recommended regional cooperation.”

While there are differences, the future will probably converge on common norms for trusted and user centred

AI.

9.7 Key Terms

Ethical AI– How we align our values with designing and using AI in ethical, fair-anded guidelines.

societal well-being.

Accountability – Making people or institutions accountable for what AI does and for unintended consequences

systems.

Transparency—Ensuring that AI systems can be understood and explained by those who use them, as well as by regulators.

Risk-Based Governance – A model to govern AI that depends on how risky it is to people and

society.

Rights anuary 4, 2020 Rights-Based Approach – A method of governance that ensures human rights protection for example privacy and which is based on the fact that everyone is entitled to inherent rights.

equality, and freedom.

NIST AI RMF -Approach developed by the US National Institute of Standards and Technology to handle

AI risk.

Sources EU AI Act - An umbrella legislation by the European Union to regulate AI on a risk-based approach.

Certification Systems – Mechanisms used to verify that AI systems are safe, ethical and usable before they are deployed.

deployment.

PPP – A relationship between government organizations and private sector companies to foster the execution of a public project, programme or general service.

responsible AI use.

Global AI Governance — The push by nations and groups for synchronized rules and headfirst into the imbroglio of American startups that are promoting products around thorny questions of data privacy and civil rights.

AI development and dissemination worldwide.

9.8 Descriptive Questions

Why are the fundamental principles of ethical AI development important?

Explain what the role of national governments is in regulating and monitoring AI technologies.

What are the differences between a risk-based and rights-based governance approach to dealing with AI?

What's in the EU Artificial Intelligence Act?

Contrast the US and the EU in AI policy.

Comment on the importance of Public-Private Partnerships in advancing responsible AI development.

How are developing areas, such as Africa and Latin America thinking about governance of artificial intelligence?

Explain the difficulties in achieving alignment across the global in AI governance.

Why are technical standards and certification mechanisms introduced in the AI-Content?

How are global institutions such as the G7 and OECD shaping AI policy?

9.9 References

1. European Commission. (2024). The Artificial Intelligence Act (EU AI Act) – Official Text and

Explanatory Memorandum.

2. White House. (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of

Artificial Intelligence.

3. NIST (2023). AI Risk Management Framework. National Institute of Standards and Technology, U.S.

Department of Commerce.

4. OECD (2021). OECD Principles on Artificial Intelligence.

5. UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence.

Answers to Knowledge Check

Knowledge Check 1

1. b) Speed

2. c) Four

3. c) NIST

4. c) Protecting human freedoms and dignity

9.10 Case Study

Regulating AI in Healthcare – Lessons from the EU and the US

Background:

An AI-based diagnostic tool for early detection of lung

cancer using CT scans. The product was also set to roll out in Europe as well as the US.

EU Context:

This product was labelled as a “high-risk AI system” under the EU AI Act. Before deployment, the

company had to:

• Perform an extensive risk analysis

• Confirm the model was trained on balanced, reliable data

• Create human oversight systems

• Be certified by the appropriate EU authorities

US Context:

There was no such law in the US, but the company tracked with those NIST AI Risk Management safety and security guidelines.

Framework and obtained FDA approval. That required voluntary testing, transparencyporate2020 11.

auditing to ensure patient safety.

Challenge:

Complex documentation and approval procedures caused delays for the company in Europe.

But once it was approved, the system enjoyed a great deal of public trust. In the US, although its debut was

speed, data biases and the absence of clear accountability when errors occurred.

Key Learning:

This particular case exemplifies the delicate equilibrium between innovation and regulation, as well as the necessity of carefully-thinking through how to handle it on a global level.

aligned governance models that safeguard users while promoting development.