Developing a Curriculum for Ethical and Responsible AI: A University Course on Safety, Fairness, Privacy, and Ethics to Prepare Next Generation of AI Professionals

Ashraf Alam

Abstract In this scientific paper, the researcher develops a course on "Safety, Fairness, Privacy, and Ethics of Artificial Intelligence" (SFPE-AI) designed for university students. The course aims to provide students with a comprehensive understanding of the technical and ethical issues associated with the development and deployment of AI systems. The course is designed to be interdisciplinary, drawing on concepts and techniques from computer science, philosophy, and law. The curriculum is divided into four modules: safety, fairness, privacy, and ethics. To facilitate student learning, the course employs a variety of pedagogical tools, such as interactive lectures, case studies, group discussions, and hands-on projects. The case studies used in the course include real-world examples of AI applications and their associated ethical and societal implication, thus providing students with a diverse perspective on the challenges and opportunities associated with AI. After the completion of this course, students are expected to understand the technical and ethical issues associated with AI, design and develop AI systems that are safe, fair, private, and ethical, and critically evaluate the societal implications of AI. The SFPE-AI course is expected to prepare the next generation of AI professionals to build responsible and trustworthy AI systems. The course will also serve as a model for other universities and educational institutions looking to integrate the discussion of AI safety, fairness, privacy, and ethics into their curriculum.

Keywords Classroom · Curriculum · Teacher · University · Pedagogy · Artificial intelligence · Safety · Fairness · Privacy · Ethics · Educational technology · ICT in education · Learner

A. Alam (✉)
Rekhi Centre of Excellence for the Science of Happiness, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India
e-mail: ashraf_alam@kgpian.iitkgp.ac.in

G. Rajakumar et al. (eds.), *Intelligent Communication Technologies and Virtual Mobile Networks*, Lecture Notes on Data Engineering and Communications Technologies 171,
https://doi.org/10.1007/978-981-99-1767-9_64

1 Introduction

Artificial intelligence (AI) is a rapidly growing field that has the potential to revolutionize many aspects of our lives [1]. However, as AI becomes more advanced and more widely adopted, it is important to consider the safety, fairness, privacy, interpretability, human-AI interaction, and ethical implications of these technologies. Safety is a critical concern for AI systems, as they are increasingly being deployed in high-stakes situations such as transportation, health care, and finance. Ensuring the safety of AI systems requires a combination of technical measures, such as testing and validation, as well as governance and regulatory frameworks. For example, techniques such as "explainable AI" (XAI) can help make AI systems more transparent and interpretable, which can improve their safety by making it easier to understand and anticipate their behavior [2]. Fairness is another important consideration for AI systems, as they can perpetuate and even amplify existing biases in society [3]. For example, biased training data or algorithms can lead to unfair outcomes for certain groups of people. To address this, techniques such as "fairness-aware AI" have been developed, which aim to ensure that AI systems are fair and do not discriminate against certain groups of people [4]. Privacy is also a key concern for AI systems, as they often involve the collection and processing of sensitive personal data. Ensuring the privacy of individuals in the context of AI requires a combination of technical measures such as data anonymization and differential privacy, as well as robust governance and regulatory frameworks. Interpretability is a crucial aspect of AI, as it allows us to understand how an AI system is making its decisions. This is important for ensuring the safety and fairness of AI systems as well as for building trust in these technologies [5]. Techniques such as XAI can help make AI systems more interpretable and transparent, which can improve their accountability and reduce the risk of unintended consequences.

Human-AI interaction is another important area of research, as it deals with how people interact with and understand AI systems [6]. This includes issues such as user interface design, natural language processing, and human-AI collaboration. Ensuring that AI systems are easy to use and understand is crucial for building trust in these technologies and making them accessible to a wide range of people. Finally, ethics is a crucial area of consideration for AI, as these technologies have the potential to have a profound impact on society [7]. Ethical issues that need to be considered include issues such as autonomy, accountability, and social responsibility. Developing a robust ethical framework for AI is important for ensuring that these technologies are developed and used in a responsible and beneficial way [8].

In conclusion, AI is a rapidly growing field with enormous potential to improve our lives, but it is important to consider the safety, fairness, privacy, interpretability, human-AI interaction, and ethical implications of these technologies. Addressing these concerns will require a combination of technical measures, governance, and regulatory frameworks, as well as ongoing research and collaboration across multiple disciplines.

2 Robust Ethical Framework

A robust ethical framework for AI should be based on a set of guiding principles that can be used to evaluate the design, development, and deployment of AI systems [9]. Such a framework should take into account the unique characteristics of AI, such as its ability to learn and adapt, and should be flexible enough to be applied to a wide range of applications [10]. One possible set of principles for an ethical framework for AI include.

Respect for autonomy: AI systems should respect the autonomy of individuals and not be used to control or manipulate them without their consent.

Fairness and non-discrimination: AI systems should be designed to be fair and not discriminate against any individuals or groups based on their race, gender, age, class, caste, or other characteristics.

Responsibility and accountability: AI systems should be designed to be transparent and interpretable, and the creators and users of these systems should be held responsible for their actions.

Privacy: AI systems should be designed to protect the privacy of individuals and not be used to collect or process sensitive personal data without their consent.

Safety and security: AI systems should be designed to be safe and secure and should not be used to cause harm or put individuals at risk.

Transparency: AI systems should be designed to be transparent and explainable, and the creators and users of these systems should be open and honest about how they work.

Human-AI interaction: AI systems should be designed to be easy to use and understand and should be designed to complement and enhance human capabilities, not replace them.

Social benefit: AI systems should be designed to be beneficial to society and should not be used to perpetuate or amplify existing societal problems.

Ethical decision-making: AI systems should be designed to make ethical decisions or to support human decision-making in an ethical way.

Continual evaluation: The ethical implications of AI should be continually evaluated as the technology and its applications evolve, and the framework should be updated as necessary.

It is important to note that these principles are not exhaustive and may not be applicable to all context or scenarios, but they provide a good starting point for creating an ethical framework for AI. Additionally, the implementation of these principles should involve active participation from multiple stakeholders such as technologists, policymakers, legal experts, ethicists, and representatives from affected communities to ensure that the framework is inclusive and comprehensive.

3 Explainable Artificial Intelligence (XAI)

Explainable artificial intelligence (XAI) is a rapidly growing field that aims to make AI systems more transparent and interpretable, so that their behavior can be better understood and anticipated. This is important for ensuring the safety, fairness, and accountability of AI systems, as well as for building trust in these technologies [11]. One of the main challenges in XAI is that many modern AI systems, such as deep neural networks, are highly complex and their behavior can be difficult to understand. One approach to addressing this is to develop techniques that can generate human-interpretable explanations of the decisions made by AI systems [12]. This can be done in a variety of ways, such as by identifying the most important features used by the AI system in making its decisions or by generating natural language explanations [13].

Another approach to XAI is to develop techniques that can directly modify the behavior of AI systems to make them more interpretable. This can be done by designing the AI system to be more transparent in how it makes its decisions or by developing techniques that can "debug" the AI system to identify and correct any errors or biases [14]. Additionally, XAI can be used in the context of interpretability of models, where it is essential to understand the model's decision-making process. It can be achieved by using techniques such as feature importance, model-agnostic interpretability, and also post-hoc interpretability [15].

There are also various methods for evaluating the interpretability of AI models, including user studies, human-in-the-loop evaluations, and metrics based on information theory [16]. These methods can be used to measure the effectiveness of XAI techniques and to identify areas where further research is needed. Another area of XAI is the generation of synthetic data or counterfactual examples, which can be used to understand the decision-making process of models and also to identify potential biases in the model [17].

In conclusion, XAI is an important field that aims to make AI systems more transparent and interpretable and has the potential to improve the safety, fairness, and accountability of AI systems. The field is rapidly evolving, and ongoing research is needed to develop new techniques and methods for generating human-interpretable explanations, modifying AI systems to make them more interpretable, evaluating interpretability, and identifying potential biases [18]. Additionally, collaboration between experts in AI, cognitive psychology, human–computer interaction, and other relevant fields is important to advance the field of XAI [19].

4 Fairness-Aware Artificial Intelligence (FAAI)

Fairness-aware artificial intelligence (FAAI) is a rapidly growing field that aims to ensure that AI systems are fair and do not discriminate against certain individuals or groups based on their race, gender, caste, region, religion, age, or other characteristics.

This is a particularly important issue in sensitive applications such as criminal justice, health care, and finance, where AI systems have the potential to perpetuate existing societal biases [20]. One of the main challenges in FAAI is that many modern AI systems are highly complex and their behavior can be difficult to understand. This makes it difficult to identify and correct any biases in these systems. One approach to addressing this is to develop techniques that can identify and quantify biases in AI systems. This can be done in a variety of ways, such as by analyzing the training data used to train the AI system or by measuring the performance of the AI system on different groups of individuals [21]. Another approach to FAAI is to develop techniques that can directly modify the behavior of AI systems to make them more fair. This can be done by designing the AI system to be more transparent in how it makes its decisions or by developing techniques that can "debias" the AI system to correct any errors or biases [22].

One widely used technique for FAAI is the "pre-processing" technique, where the data is transformed or modified before training a model to remove or reduce any biases present in the data. Another technique is "in-processing" where the model is designed to be fair during the training process, for example, by adding a fairness constraint to the optimization problem [23].

Additionally, there are various methods for evaluating the fairness of AI models, including demographic parity, equal opportunity, and equalized odds [24]. These metrics can be used to measure the effectiveness of FAAI techniques and to identify areas where further research is needed. Another area of FAAI is the use of interpretable models, which can be used to understand the decision-making process of models and also to identify potential biases in the model [25]. Collaboration between experts in AI, statistics, sociology, and other relevant fields is important to advance the field of FAAI.

5 Developing Robust "Governance and Regulatory Frameworks" to Address Privacy Concerns for AI Systems

As artificial intelligence (AI) becomes more prevalent in society, it is increasingly important to develop robust governance and regulatory frameworks to address privacy concerns [26]. Privacy is a complex and multifaceted issue that is closely tied to the development and deployment of AI systems [27]. It is critical to ensure that these systems are designed, built, and deployed in a way that respects individuals' privacy and protects their personal information. One key aspect of privacy in AI is the collection, storage, and use of personal data. This includes not just data that individuals provide directly, but also data that is inferred or generated by AI systems [28]. It is crucial to ensure that personal data is collected and used in a way that is transparent, fair, and lawful. This includes providing individuals with clear and concise information about what data is being collected, how it will be used, and

who it will be shared with. It also includes ensuring that data is stored and processed securely and that individuals have the right to access, correct, and delete their personal data [29].

Another important aspect of privacy in AI is the use of personal data to make decisions that affect individuals. This includes decisions made by AI systems in areas such as credit, employment, housing, and health care. It is important to ensure that these decisions are fair, unbiased, and transparent, and that individuals have the right to contest and correct any decisions that are made about them. To address these concerns, it is necessary to develop governance and regulatory frameworks that are tailored to the unique challenges posed by AI [30]. This may include laws and regulations that govern the collection, use, and storage of personal data, as well as guidelines and best practices for the design and deployment of AI systems. Additionally, it may involve creating independent oversight bodies to ensure that these laws and regulations are being followed, and that any violations are quickly identified and addressed. In addition to the traditional legal frameworks, technical solutions like differential privacy, federated learning, and homomorphic encryption can also be used to protect personal data while still allowing AI models to be trained and used [31].

6 Interpretability Aspect of AI

Interpretability is a crucial aspect of artificial intelligence (AI) that refers to the ability to understand and explain the behavior and decisions of AI systems [32]. This is particularly important in sensitive applications such as health care, finance, and criminal justice, where the decisions made by AI systems can have a significant impact on individuals and society. One of the main challenges in interpretability is that many modern AI systems, such as deep neural networks, are highly complex and their behavior can be difficult to understand [33]. This makes it difficult to identify and correct any errors or biases in these systems. Additionally, the black box nature of these models makes it hard to understand how the model arrived at a particular decision and what features of the data were important in making that decision [34].

To address these challenges, researchers have developed a variety of techniques for making AI systems more interpretable. One approach is to use techniques that can identify and quantify the features of the data that are most important in making a decision [35]. For example, feature importance techniques like permutation feature importance, SHAP, and LIME can be used to understand which features of the data were most important in making a decision. Another approach is to use techniques that can directly modify the behavior of AI systems to make them more interpretable. This can be done by designing the AI system to be more transparent in how it makes its decisions, or by developing techniques that can "debias" the AI system to correct any errors or biases [36]. For example, decision trees and rule-based models are more interpretable than deep neural networks because they make decisions based on a set of simple if–then rules, which are easy to understand and explain [37]. Additionally,

there are various methods for evaluating the interpretability of AI models, such as the model's global and local interpretability. Global interpretability refers to the extent to which the overall behavior and decision-making process of the model can be understood, while local interpretability refers to the extent to which the model's behavior for a specific input can be understood [38].

In the medical field, interpretability is crucial as it allows medical experts to understand the decisions made by the AI system and ensure that they align with the medical knowledge. For example, in the case of radiology, interpretable models can help radiologists understand the model's decision-making process and identify the features that are important in making the diagnosis. This can help radiologists identify any errors or biases in the model and make adjustments accordingly.

7 Why Teach a Course on Safety, Fairness, Privacy, and Ethics" to Students?

Teaching a course on Safety, Fairness, Privacy, and Ethics of Artificial Intelligence (AI) to students is important for several reasons:

Preparing for the future: AI is rapidly advancing and is expected to play an increasingly important role in society. By teaching students about the safety, fairness, privacy, and ethics of AI, they will be better prepared for the future and the potential impact of AI on their lives.

Developing critical thinking skills: Understanding the safety, fairness, privacy, and ethics of AI requires students to think critically about the implications of AI on society. This can help students develop the critical thinking skills they need to evaluate information, make informed decisions, and solve problems.

Promoting responsible use of AI: By teaching students about the safety, fairness, privacy, and ethics of AI, they will be more likely to use AI responsibly and to recognize when AI is being used in a way that is harmful or unethical.

Promoting digital literacy: Understanding the safety, fairness, privacy, and ethics of AI is an important aspect of digital literacy. This will enable students to navigate the digital world and to understand the impact of technology on society.

Creating a more inclusive society: AI systems can perpetuate societal biases and discrimination if not designed and implemented carefully. Teaching students about the fairness and ethical considerations of AI can help create a more inclusive society by encouraging them to design and use AI in a way that is fair and unbiased.

Fostering a culture of responsibility: By teaching students about the safety, fairness, privacy, and ethics of AI, we can foster a culture of responsibility where individuals take responsibility for the impact of AI on society and work to ensure that AI is used in a way that is safe, fair, and respects privacy.

8 Curriculum Outline for a Course on "Safety, Fairness, Privacy, and Ethics of Artificial Intelligence"

I. Introduction

- Overview of the course
- Importance of understanding the safety, fairness, privacy, and ethics of AI
- Key concepts and definitions related to AI safety, fairness, privacy, and ethics.

II. Safety of Artificial Intelligence

- Overview of safety concerns in AI
- Techniques for evaluating and ensuring the safety of AI systems
- Case studies of AI safety incidents and lessons learned

III. Fairness in Artificial Intelligence

- Overview of fairness concerns in AI
- Techniques for evaluating and ensuring fairness in AI systems
- Case studies of AI fairness incidents and lessons learned.

IV. Privacy in Artificial Intelligence

- Overview of privacy concerns in AI
- Techniques for evaluating and ensuring privacy in AI systems
- Case studies of AI privacy incidents and lessons learned.

V. Ethics of Artificial Intelligence

- Overview of ethical concerns in AI
- Techniques for evaluating and ensuring the ethical use of AI systems
- Case studies of AI ethical incidents and lessons learned.

VI. Human-AI Interaction

- Overview of the interactions between humans and AI
- Techniques for designing human-AI interactions that are safe, fair, private, and ethical
- Case studies of human-AI interactions and lessons learned.

VII. Governance and Regulation of Artificial Intelligence

- Overview of governance and regulatory frameworks for AI
- Techniques for designing governance and regulatory frameworks that promote safety, fairness, privacy, and ethics in AI
- Case studies of governance and regulatory frameworks for AI and lessons learned.

VIII. Interpretability of Artificial Intelligence

- Overview of the interpretability of AI

- Techniques for making AI systems more interpretable
- Case studies of interpretable AI systems and lessons learned.

IX. Conclusion

- Summary of key concepts and takeaways
- Discussion of future directions in AI safety, fairness, privacy, and ethics.

Throughout the course, students will engage in hands-on exercises and projects, as well as discussions and debates to help them apply the concepts and techniques covered in the course to real-world scenarios.

9 Learning Outcomes

The learning outcomes for a course on "Safety, Fairness, Privacy, and Ethics of Artificial Intelligence" for university students include the following:

1. *Understanding of the technical and societal implications of AI*: Students will be able to understand the technical capabilities and limitations of AI systems, as well as the potential impact of AI on society, including issues related to safety, fairness, privacy, and ethics.
2. *Knowledge of AI safety, fairness, privacy, and ethics*: Students will develop a comprehensive understanding of the key concepts and principles related to AI safety, fairness, privacy, and ethics.
3. *Ability to analyze and evaluate AI systems*: Students will be able to analyze and evaluate AI systems based on safety, fairness, privacy, and ethics criteria and to identify potential risks and ethical issues.
4. *Skills in designing and developing safe, fair, and ethical AI systems*: Students will be able to design and develop AI systems that are safe, fair, and ethical and to implement appropriate measures to mitigate risks and ensure compliance with relevant regulations and standards.
5. *Understanding of governance and regulatory frameworks*: Students will be able to understand the governance and regulatory frameworks related to AI and to analyze the effectiveness of different approaches.
6. *Ability to communicate effectively about AI safety, fairness, privacy, and ethics*: Students will be able to communicate effectively about AI safety, fairness, privacy, and ethics both with technical and non-technical audiences and to participate in interdisciplinary discussions and collaborations.
7. *Ability to use storytelling method to understand and communicate about AI safety, fairness, privacy, and ethics*: Students will be able to use storytelling method to understand and communicate about AI safety, fairness, privacy, and ethics and to evaluate the effectiveness of different storytelling methods.
8. *Critical thinking and problem-solving skills*: Students will develop critical thinking and problem-solving skills and will be able to apply them in real-world scenarios related to AI safety, fairness, privacy, and ethics.

In conclusion, the course will provide students with a solid foundation in the field of AI safety, fairness, privacy, and ethics and will equip them with the knowledge and skills needed to work with AI systems in a safe, fair, and ethical manner.

10 Pedagogical Tools and Teaching–Learning Materials

Pedagogical tools and teaching–learning materials (TLMs) that shall be used to teach a course on "Safety, Fairness, Privacy, and Ethics of Artificial Intelligence" for university students include.

Lectures and presentations: Use of lectures and presentations to introduce key concepts and provide an overview of the course material. This can include slides, videos, and interactive elements such as quizzes and polls.

Case studies: Use of real-world case studies to illustrate the importance of safety, fairness, privacy, and ethics in AI and to provide students with an opportunity to apply the concepts and techniques covered in the course to real-world scenarios.

Hands-on projects: Use of hands-on projects, such as programming assignments or design projects, to give students the opportunity to apply the concepts and techniques covered in the course in a practical context.

Guest lectures: Inviting experts in the field of AI safety, fairness, privacy, and ethics to give guest lectures on specific topics, to provide students with a broader perspective on the field.

Simulation and role-playing exercises: Use of simulation and role-playing exercises to help students understand the implications of AI in different scenarios.

Online resources: Use of online resources, such as videos, articles, and tutorials, to supplement the course material and provide students with additional information and resources on the topics covered in the course.

Discussion forums and debates: Encourage students to discuss the topics covered in the course and to express their opinions and ideas through discussion forums and debates.

Reading materials: Provide students with a recommended reading list of relevant books, articles, and research papers on the topic of AI safety, fairness, privacy, and ethics.

Ethical dilemma: Provide students with ethical dilemmas related to AI and have them work through resolving the situations.

Collaborative activities: Encourage students to work in teams or groups on projects and assignments, to foster collaboration and teamwork.

These pedagogical tools and teaching–learning materials can be used to create an interactive and engaging learning experience for students and help them understand

the importance of safety, fairness, privacy, and ethics in AI, and how to apply the concepts and techniques covered in the course to real-world scenarios.

11 Games and Activities

Games and activities that shall be used to teach a course on "Safety, Fairness, Privacy, and Ethics of Artificial Intelligence" for university students include

1. *Ethical dilemma game*: Create a game that presents students with ethical dilemmas related to AI and have them work through resolving the situations.
2. *Scenario-based role-playing*: Create scenarios that depict the potential consequences of AI and have students role-play different characters in the scenario, to help them understand the implications of AI in different contexts.
3. *Debate*: Organize debates on current controversies related to AI safety, fairness, privacy, and ethics and have students participate in the debates and express their opinions on the topic.
4. *Escape room*: Create an escape room experience that simulates a complex AI scenario where students need to use their understanding of safety, fairness, privacy, and ethics to solve the puzzles and escape.
5. *Board game*: Develop a board game that simulates the decision-making process in AI development and have students play the game to understand the importance of safety, fairness, privacy, and ethics in AI.
6. *AI simulation*: Create a simulation of an AI system and have students experiment with different settings and configurations to understand the impact of different choices on safety, fairness, privacy, and ethics.
7. *Ethical hackathon*: Organize an ethical hackathon where students work in teams to identify and fix ethical issues in existing AI systems.
8. *Group discussions*: Divide students into small groups and give them a case-study or scenario related to AI safety, fairness, privacy, and ethics and have them discuss the situation and come up with solutions or recommendations.
9. *Case-study analysis*: Have students analyze real-world case studies related to AI safety, fairness, privacy, and ethics and have them present their findings and recommendations to the class.
10. *Jeopardy game*: Create a Jeopardy-style game where students answer questions on AI safety, fairness, privacy, and ethics and get points based on their answers.

These games and activities can be used to create an interactive and engaging learning experience for students and help them understand the importance of safety, fairness, privacy, and ethics in AI, and how to apply the concepts and techniques covered in the course to real-world scenarios.

12 Storytelling Method

Storytelling can serve as a powerful tool for teaching a course on "Safety, Fairness, Privacy, and Ethics of Artificial Intelligence" to university students. Here are a few ways storytelling can be used in such a course:

1. *Case studies*: Share real-world case studies of AI systems that have faced ethical challenges. Use storytelling to present the background, the situation, the challenges faced and the resolution.
2. *Scenarios*: Create fictional scenarios that depict the potential consequences of AI, both positive and negative, and use storytelling to present the scenarios in an engaging and relatable way.
3. *Stories from experts*: Invite experts in the field of AI safety, fairness, privacy, and ethics to share their experiences and stories of challenges they have faced and how they overcame them.
4. *Role-playing*: Have students act out scenarios in small groups, to help them understand the implications of AI in different contexts.
5. *Interactive storytelling*: Use interactive storytelling tools such as virtual reality, augmented reality, and gamification to create an immersive learning experience for students.
6. *Storytelling through data*: Use data storytelling techniques like data visualization, infographics, and videos to help students understand the impact of AI on society and how to measure safety, fairness, privacy, and ethics.
7. *Storytelling through analogies*: Use analogies and metaphor to help students understand the complex concepts of AI safety, fairness, privacy, and ethics.
8. *Ethical storytelling*: Use storytelling to explore ethical dilemmas related to AI and guide students through the decision-making process.

By using storytelling, students can relate to the material in a more personal way and better understand the implications of AI on society. Storytelling can also be used to create an engaging and interactive learning experience, making the course more memorable and effective.

13 Case-Study Method

The case-study method can be used to teach a course on "Safety, Fairness, Privacy, and Ethics of Artificial Intelligence" for university students in a number of ways, including the following:

1. *Real-world examples*: Case studies can provide students with real-world examples of AI systems and applications, including both successes and failures, and can help to illustrate the key concepts and principles related to AI safety, fairness, privacy, and ethics.

2. *Analysis and evaluation*: Students can analyze and evaluate the case studies using the concepts and principles learned in class and can identify the potential risks and ethical issues associated with each case.
3. *Problem-solving*: Case studies can serve as a starting point for problem-solving exercises, where students can work in groups to develop solutions to the issues identified in the case studies and to propose recommendations for addressing the risks and ethical issues.
4. *Debate and discussion*: Case studies can be used to spark debate and discussion among students, as they can have different perspectives and opinions on the issues and risks identified in the case studies.
5. *Guest speakers*: Inviting experts or practitioners who have been involved with the specific case studies being discussed can provide students with valuable insights and perspectives on the real-world challenges of AI safety, fairness, privacy, and ethics.
6. *Current events*: Case studies can be selected to reflect current events and the most recent developments in the field of AI, which can make the course more engaging and relevant for the students.

In conclusion, the case-study method can be an effective way to foster problem-solving skills, promote critical thinking and debate, and stay current with the field's developments.

14 Example of a Case-Study: The Case of Amazon's AI-Powered Recruitment Tool

In this case-study, the students can learn about:

The application of AI in recruitment: Amazon developed an AI-powered recruitment tool that uses machine learning algorithms to analyze resumes and job applications, in order to identify the most qualified candidates for open positions.

Bias in AI systems: Amazon's recruitment tool was found to have a gender bias, as it was trained on resumes submitted to the company over a 10-year period, which were mostly from men. As a result, the tool was less likely to recommend female candidates for open positions.

Fairness in AI: The case raises important questions about fairness in AI systems, and the potential risks and challenges associated with bias in AI-powered decision-making.

Ethical considerations: The case highlights the ethical considerations that must be taken into account when developing and deploying AI systems, particularly in relation to issues of fairness, privacy, and transparency.

Mitigating bias: The case-study can be used to discuss ways to mitigate bias in AI systems, such as using diverse training data, and developing techniques for identifying and addressing bias in machine learning models.

Current events: This case-study is based on a real event that occurred in 2018, it is still relevant today as it highlights the importance of preventing bias in AI systems and the need for AI governance and regulations.

Guest speaker: A guest speaker from Amazon, who was involved in the development and deployment of the recruitment tool, could be invited to speak about their experience and the lessons learned, providing valuable insights into the challenges of implementing AI in practice. Overall, this case-study can be used to teach students about the practical challenges and ethical considerations associated with the application of AI in recruitment and to encourage them to think critically about the implications of bias in AI systems and the need for fairness-aware AI.

References

1. Alam A (2022) A digital game based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning. In: 2022 international conference on sustainable computing and data communication systems (ICSCDS). IEEE, pp 69–74
2. Currie G, Hawk KE (2021, March) Ethical and legal challenges of artificial intelligence in nuclear medicine. Sem Nucl Med 51(2):120–125
3. Alam A (2022) Investigating sustainable education and positive psychology interventions in schools towards achievement of sustainable happiness and wellbeing for 21st century pedagogy and curriculum. ECS Trans 107(1):19481
4. Saghiri AM, Vahidipour SM, Jabbarpour MR, Sookhak M, Forestiero A (2022) A survey of artificial intelligence challenges: analyzing the definitions, relationships, and evolutions. Appl Sci 12(8):4054
5. Alam A (2022) Social robots in education for long-term human-robot interaction: socially supportive behaviour of robotic tutor for creating robo-tangible learning environment in a guided discovery learning interaction. ECS Trans 107(1):12389
6. Vellido A (2019) Societal issues concerning the application of artificial intelligence in medicine. Kidney Dis 5(1):11–17
7. Alam A, Mohanty A (2022) Metaverse and Posthuman animated avatars for teaching-learning process: interperception in virtual universe for educational transformation. In: international conference on innovations in intelligent computing and communications. Springer, Cham, pp 47–61
8. Al-Hwsali A, Al-Saadi B, Abdi N, Khatab S, Solaiman B, Alzubaidi M, Abd-alrazaq A, Househ M (2022) Legal and ethical principles of artificial intelligence in public health: scoping review
9. Alam A (2020) Pedagogy of calculus in India: an empirical investigation. Periódico Tchê Química 17(34):164–180
10. Siala H, Wang Y (2022) SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc Sci Med:114782
11. Alam A (2022) Positive psychology goes to school: conceptualizing students' happiness in 21st century schools while 'minding the mind!' Are we there yet? Evidence-backed. School-Based Positive Psychol Intervent ECS Trans 107(1):11199
12. Nasim SF, Ali MR, Kulsoom U (2022) Artificial intelligence incidents & ethics a narrative review. Int J Technol Innov Manage (IJTIM) 2(2)

13. Alam A (2022) Mapping a sustainable future through conceptualization of transformative learning framework, education for sustainable development, critical reflection, and responsible citizenship: an exploration of pedagogies for twenty-first century learning. ECS Trans 107(1):9827
14. Huang C, Zhang Z, Mao B, Yao X (2022) An overview of artificial intelligence ethics. IEEE Trans Artif Intell
15. Alam A (2022) Employing adaptive learning and intelligent tutoring robots for virtual classrooms and smart campuses: reforming education in the age of artificial intelligence. In: Shaw RN, Das S, Piuri V, Bianchini M (eds) Advanced computing and intelligent technologies. Lecture notes in electrical engineering, vol 914. Springer, Singapore
16. O'Sullivan S, Nevejans N, Allen C, Blyth A, Leonard S, Pagallo U, Holzinger K, Holzinger A, Sajid MI, Ashrafian H (2019) Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. Int J Med Robot Comput Assist Surg 15(1):e1968
17. Alam A (2022) Platform utilising blockchain technology for eLearning and online education for open sharing of academic proficiency and progress records. In: Asokan R, Ruiz DP, Baig ZA, Piramuthu S (eds) Smart data intelligence. Algorithms for intelligent systems. Springer, Singapore
18. del Pero AS, Wyckoff P, Vourc'h A (2022) Using artificial intelligence in the workplace: what are the main ethical risks?
19. Alam A (2023) Cloud-based e-learning: scaffolding the environment for adaptive e-learning ecosystem based on cloud computing infrastructure. In: Satapathy SC, Lin JCW, Wee LK, Bhateja V, Rajesh TM (eds) Computer communication, networking and IoT. Lecture notes in networks and systems, vol 459. Springer, Singapore
20. Köbis L, Mehner C (2021) Ethical questions raised by AI-supported mentoring in higher education. Front Artif Intell 4:624050
21. Alam A (2022) Cloud-based E-learning: development of conceptual model for adaptive e-learning ecosystem based on cloud computing infrastructure. In: Kumar A, Fister Jr I, Gupta PK, Debayle J, Zhang ZJ, Usman M (eds) Artificial intelligence and data science. ICAIDS 2021. Communications in computer and information science, vol 1673. Springer, Cham
22. Juric M, Sandic A, Brcic M (2020, Sept) AI safety: state of the field through quantitative lens. In: 2020 43rd international convention on information, communication and electronic technology (MIPRO). IEEE, pp 1254–1259
23. Alam A (2021) Possibilities and apprehensions in the landscape of artificial intelligence in education. In: 2021 international conference on computational intelligence and computing applications (ICCICA). IEEE, pp 1–8
24. Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399
25. Alam A (2022) Impact of university's human resources practices on professors' occupational performance: empirical evidence from India's higher education sector. In: Inclusive businesses in developing economies. Palgrave Macmillan, Cham, pp 107–131
26. Ashok M, Madan R, Joha A, Sivarajah U (2022) Ethical framework for artificial intelligence and digital technologies. Int J Inf Manage 62:102433
27. Alam A (2022) Educational robotics and computer programming in early childhood education: a conceptual framework for assessing elementary school students' computational thinking for designing powerful educational scenarios. In: 2022 international conference on smart technologies and systems for next generation computing (ICSTSN). IEEE, pp 1–7
28. Dhirani LL, Mukhtiar N, Chowdhry BS, Newe T (2023) Ethical dilemmas and privacy issues in emerging technologies: a review. Sensors 23(3):1151
29. Alam A (2020) Challenges and possibilities in teaching and learning of calculus: a case study of India. J Educ Gifted Young Sci 8(1):407–433
30. Arslan J, Ounissi M, Jiménez G, Kar A, Racoceanu D (2022) Responsible artificial intelligence: a review of current trends. Winter School AI4Health

31. Alam A (2020) Possibilities and challenges of compounding artificial intelligence in India's educational landscape. Int J Adv Sci Technol 29(5):5077–5094
32. Alam A, Mohanty A (2022) Business models, business strategies, and innovations in EdTech companies: integration of learning analytics and artificial intelligence in higher education. In: 2022 IEEE 6th conference on information and communication technology (CICT). IEEE, pp 1–6
33. Lukkien DR, Nap HH, Buimer HP, Peine A, Boon WP, Ket JC, Minkman MM, Moors EH (2023) Toward responsible artificial intelligence in long-term care: a scoping review on practical approaches. Gerontologist 63(1):155–168
34. Alam A, Mohanty A (2022) Foundation for the future of higher education or 'misplaced optimism'? Being human in the age of artificial intelligence. In: International conference on innovations in intelligent computing and communications. Springer, Cham, pp 17–29
35. Kaur D, Uslu S, Rittichier KJ, Durresi A (2022) Trustworthy artificial intelligence: a review. ACM Comput Surv (CSUR) 55(2):1–38
36. Alam A (2020) Test of knowledge of elementary vectors concepts (TKEVC) among First-semester bachelor of engineering and technology students. Periódico Tchê Química 17(35):477–494
37. Bakiner O (2022) What do academics say about artificial intelligence ethics? An overview of the scholarship. AI Ethics:1–13
38. Alam A (2021) Should robots replace teachers? Mobilisation of AI and learning analytics in education. In: 2021 international conference on advances in computing, communication, and control (ICAC3). IEEE, pp 1–12