Revista Review Index Journal of Multidisciplinary 2023; 3(2):24-35
e-ISSN: 2583-0031
Double-blind peer-reviewed | Refereed Quarterly Online Journal

RESEARCH PAPER

Artificial Intelligence Ethics and Fairness: A study to address bias and fairness issues in AI systems, and the ethical implications of AI applications

*Tejaskumar B. Modi

Assistant Professor, Kalol Institute of Management (MCA), Kalol, Gujarat, India

Abstract

Artificial intelligence (AI) has made incredible strides in several fields, revolutionising business and everyday life. Thoughts regarding the moral ramifications and fairness of AI systems have grown in prominence along with its fast development. This article explores the crucial issues of AI fairness and ethics, concentrating on ways to detect and reduce prejudice in AI systems while also discussing larger ethical implications. The paper emphasises the possible repercussions of biased decision-making while highlighting the many forms and sources of bias that might develop in AI models. There are several methods and strategies to deal with bias in AI systems, including data pretreatment, algorithmic changes, and transparency measures. In an effort to balance justice and efficacy, the trade-offs between fairness goals and overall model performance are examined. The article also emphasises how crucial it is for AI systems to be transparent and understandable in order to foster accountability. For the purpose of establishing ethical AI development and deployment practises, regulatory issues and ethical decision-making frameworks are also investigated. This study emphasises the need of ongoing research and development of moral AI systems to guarantee a just and equitable future for AI applications via in-depth analysis and case studies.

Keywords: *AI Ethics, Fairness, Bias, Ethical Implications, AI Applications, Ethical Decision-making*

Article Publication

Published Online: 30-Jun-2023

***Author's Correspondence**

Tejaskumar B. Modi

Assistant Professor, Kalol Institute of Management (MCA), Kalol, Gujarat, India

proftejasmodi@gmail.com

10.31305/rrijm2023.v03.n02.004

Introduction to AI Ethics and Fairness

Artificial intelligence (AI) has become a disruptive force in a variety of sectors, enhancing decision-making procedures and simplifying work to previously unheard-of levels of precision and efficiency. Thoughts on the ethical ramifications and fairness of AI technologies have risen to the fore of public debate as they continue to pervade our everyday lives. To fully use AI's potential while ensuring that its implementation is fair and impartial, these issues must be addressed.

The study of moral principles and values that guide the creation, implementation, and use of AI systems is referred to as AI Ethics. It entails evaluating the possible effects and social impact of AI applications while taking wider ethical implications into account in diverse circumstances (Floridi & Cowls, 2019). Protecting against the dangers of AI technology sustaining prejudice, discrimination, and damage to people or groups is the main goal of AI ethics (Jobin et al., 2019). As AI plays a more and bigger part in crucial decision-making processes including healthcare diagnostics, employment procedures, and criminal justice applications, ethical issues are crucial.

Fairness, one of the fundamental principles of AI Ethics, tries to guarantee that AI systems treat every person equally and do not discriminate against any specific group. In situations where AI systems affect access to opportunities or resources, fairness in AI is especially important since biassed decision-making might worsen already-existing social inequities (Dwork et al., 2012). Fairness in AI systems is a challenging goal that calls for careful consideration of the data used to train the models, algorithm design, and the decision-making criteria utilised (Ruggieri, 2014).

According to Bellamy et al. (2018), the idea of fairness in AI is multifaceted and entails tackling several biases, including social prejudice, algorithmic bias, and data bias. Due to historical imbalances in training data, some groups are overrepresented, which causes data bias. During the creation of AI algorithms, biases are typically added that reflect the preferences of the data sources or the goals of the model. These biases are referred to as algorithmic bias. Societal bias may emerge when artificial intelligence (AI) technologies unintentionally promote or magnify pre-existing societal prejudices in the larger community.

It is impossible to emphasise the importance of fairness in AI, since biassed AI systems may provide discriminating results and have a harmful effect on disadvantaged people (Baezconde-Garbanati et al., 2020). Fairness is necessary to maintain confidence in AI systems, which may prevent their broad use and acceptance. Therefore, addressing fairness issues is essential for building a suitable climate for responsible AI deployment in a variety of fields as well as for the welfare of humans.

This essay tries to investigate the complex issues surrounding AI Ethics and Fairness, including techniques for detecting and reducing bias in AI systems as well as the wider ethical implications of AI applications. We want to illuminate the prospects and difficulties in guaranteeing ethical and just AI development and deployment by critically analysing current research and case studies.

Understanding Bias in AI

Artificial intelligence (AI) technologies are permeating contemporary civilization more and more, impacting decision-making and reshaping many facets of human existence. These AI systems, however, are not immune to biases, which may unintentionally be included into their decision-making, training, or data sets. In artificial intelligence (AI), bias is the unjustified and systematic favouritism or discrimination of certain people or groups that results in unfair results. Exploring the many biases that might appear, such as data bias, algorithmic bias, and social prejudice, is essential to understanding the effects of bias in AI systems.

- **Data Bias:** The historical inequities and constraints in the training data used to create AI models are the source of data bias. The AI system will pick up on and reinforce prejudices during decision-making if the training data is biassed or not representative of the real-world population. The AI model may unintentionally prioritise male applicants over equally qualified female applicants in a recruitment system powered by AI that was trained on historical hiring data, for example, perpetuating gender discrimination (Crawford et al., 2019).
- **Algorithmic Bias:** Algorithmic bias is a result of how AI algorithms are built and designed. When creating an algorithm, biases may be introduced by the selection of characteristics, weighting variables, or optimisation goals. In predictive police systems, for instance, if an algorithm is tailored to concentrate only on historical crime data, it may disproportionately target certain neighbourhoods with higher crime rates, resulting in heightened monitoring and possible prejudice against such areas (O'Neil, 2016).
- **Societal Bias:** This refers to the prejudices that exist in society at large and may unintentionally be mirrored in AI systems. AI systems get their knowledge from past data, which may already be skewed by societal preconceptions. Because of this, the AI system may unintentionally reinforce and magnify social prejudices. In AI-driven loan approval systems, for instance, if prior lending choices were discriminatory towards certain racial or ethnic groups, the AI model may reproduce similar prejudices, resulting in increased financial inequalities (Larson et al., 2016).

Real-World Instances of Biased AI Leading to Negative Consequences:

- **Criminal Justice AI Bias:** AI algorithms have been used in the criminal justice system to predict a person's chance of reoffending and their suitable punishment. An artificial intelligence (AI) based recidivism prediction tool in the United States was shown to be biassed towards African Americans, incorrectly classifying them as high-risk almost twice as often as white offenders (Angwin et al., 2016). Because of this predisposition, sentencing choices may be unfair and arbitrary.
- **Facial Recognition Bias:** Studies have shown that persons of colour and women are more likely to be incorrectly identified by facial recognition systems. Researchers Buolamwini and Gebru (2018) found that commercial face analysis software made more mistakes when identifying the gender of women with darker skin tones than they did when identifying the gender of men with lighter skin tones. Misidentification in surveillance applications or prejudice in the recruiting process are two dire outcomes that might result from implicit biases.

To create AI systems that are fair and equitable for all people and do not perpetuate prejudice or injury, understanding bias in AI is essential. Developers and researchers of AI may make progress towards more equitable and accountable AI systems by first identifying the many forms of bias that exist and then taking steps to counteract them.

Implications of AI Bias

Numerous advantages have resulted from the fast development and broad acceptance of Artificial Intelligence (AI). However, there has been an equally rapid increase in worries about the ethical repercussions of biassed AI. Discrimination, stereotyping, and societal inequities are only some of the issues that may be exacerbated by biassed AI systems, which can have far-reaching implications. To mitigate the dangers and damages that AI bias may pose in important fields like healthcare, employment, criminal justice, and finance, it is crucial to have a firm grasp on its ramifications.

Ethical Implications of Biased AI:

a. **Perpetuating Discrimination:** Biased AI systems may encourage discrimination by favouring certain people over others on the basis of legally protected qualities, such as race, gender, or ethnicity. Unfair treatment and systemic injustices are possible outcomes of continuing prejudice (Sweeney, 2013).

b. **Reinforcing Stereotypes:** Stereotype reinforcement Artificially intelligent models may be biased if they learn on data that incorporates societal stereotypes. Therefore, people may make decisions based on these prejudices, which may further cement bias in society (Barocas & Selbst, 2016).

c. **Exacerbating Social Inequalities:** Communities already at a disadvantage may be hit worse by biased judgements made with the help of biased AI. The divide between the rich and the marginalised becomes even larger as a result (Diakopoulos, 2016).

Risks and Harms of Biased AI in Various Domains:

a. **Healthcare:** In the medical field, for example, bias in AI systems may cause incorrect diagnosis and discriminatory treatment of patients. For instance, if an AI system is trained with data that is skewed against underserved communities, it may have trouble correctly detecting specific illnesses in those groups (Obermeyer et al., 2019).

b. **Hiring and Employment:** Discrimination in the workplace may be perpetuated through biased AI used in the recruiting process. b. recruiting and Employment. To paraphrase (Datta et al., 2015), diversity and inclusion may be stunted if AI systems are educated on biased recruiting data from the past.

c. **Criminal Justice:** Unfair treatment of persons may result from the use of biased AI in a variety of criminal justice applications, including recidivism prediction and punishment. Artificial intelligence systems may favour more severe punishments for people of specific demographics, leading to unfair imprisonment rates (Chouldechova, 2017).

d. **Financial Services:** Discriminatory lending practises are possible in the financial sector due to the use of biased AI. Access to loans and other financial possibilities may be unequally distributed if AI models are biased against specific socioeconomic groups (Hajian et al., 2016).

The effects of biased AI extend far and wide, touching people and communities in many ways. Promoting ethical AI research and deployment, promoting justice, and limiting possible risks and damages all require addressing bias in AI. Stakeholders may strive towards developing AI systems that are fair, transparent, and responsible by first recognising the ethical issues of biased AI and then analysing the possible ramifications in various circumstances.

Measuring Fairness in AI Systems

When developing ethical and accountable AI systems, fairness is a fundamental component. Avoiding the perpetuation of prejudice and discrimination requires checking that AI models and algorithms treat everyone equally and fairly. Quantifying and analysing the level of bias inherent in the decision-making processes is a challenging challenge when trying to measure fairness in AI. To evaluate the fairness of AI models, several metrics and measures have been suggested; however, defining fairness and selecting suitable metrics may be difficult, especially given the variety of application fields.

Different Fairness Metrics and Measures:

a. **Statistical Parity (Demographic Parity):** This metric assesses whether there is an equal distribution of favourable outcomes (such as loan approvals or job offers) across groups characterised by potentially discriminatory characteristics (such as race or gender). The ultimate goal is to level the playing field for all demographics (Zafar et al., 2017).

b. **Equalized Odds (Equal Opportunity):** True positive rates (i.e., recall or sensitivity) and false positive rates (FPR) are statistically equivalent across groups when equalised odds are used. It aims to reduce bias in both directions (Hardt et al., 2016) by reducing the number of false positives and false negatives.

c. **Individual Fairness:** In terms of individual fairness, we look at how people are alike and try to give them the same treatment. It necessitates that people who are otherwise comparable be treated the same, regardless of their social group (Dwork et al., 2012).

d. **Conditional Independence:** This measure determines whether the real value of one attribute (such as credit score) may be used to forecast the target variable (such as loan default) without regard to the sensitive attribute. It guarantees that the model's choice does not rely on the sensitive characteristic, but rather on the other aspects that are important in the context (Kusner et al., 2017).

Challenges in Defining Fairness and Domain Variability:

a. **Subjectivity of Fairness:** The Subjective Nature of fairness is Not Easily Defined. Fairness may be seen differently by various parties because of differences in viewpoint and cultural norms (Heidari et al., 2019).

b. **Trade-offs between Fairness Metrics:** Fairness metrics are frequently at odds with one another, and satisfying one fairness criteria may need sacrificing satisfaction of another. (Corbett-Davies et al., 2017) Finding a happy medium between competing fairness goals is difficult.

c. **Data Limitations:** The availability of relevant data for conducting fairness assessments may be constrained, particularly for underserved communities. Fairness evaluations may be influenced by biases introduced into the training data (Chouldechova, 2021).

d. **Application Domain Specificity:** To account for domain-specific factors and the variable effect of biases, fairness metrics may need to be modified for individual application domains (Ruggieri, 2014).

Evaluating the fairness of AI systems is crucial for preventing unfair or discriminating results. To evaluate the fairness of AI models, several metrics and measurements have been presented, each with its own set of advantages and disadvantages. However, the subjectivity of the notion and the particular obstacles given by diverse application domains need careful consideration when defining fairness and selecting acceptable indicators.

Sources of Bias in AI

Decisions and predictions made by Artificial Intelligence (AI) systems are trained on massive volumes of data. Inadvertently introducing biases into AI models via this data-driven learning process might have unintended negative consequences. Accurately addressing and mitigating these problems requires first pinpointing the origins of bias in AI.

Training data bias, algorithmic flaws, and human bias in the decision-making process are the three main causes of bias in AI systems. Human prejudices are also crucial in programming such biases into AI programmes.

- **Biased Training Data:** One of the most prevalent causes of bias in AI systems is the use of biased training data. The AI model will learn and transmit any biases contained in the data used to train it (Caliskan et al., 2017) even if the data is not representative of the real-world population. For instance, if the AI model is trained on data that is biased towards a certain demographic, it may unwittingly favour applicants who belong to that group when making recruiting decisions.
- **Flawed Algorithms:** Bias-Inducing Flawed Algorithms due to their design and optimisation goals, flawed algorithms may induce bias into AI systems. The optimisation method, the model's design, and the features themselves may all introduce biases. Particularly worrisome is algorithmic bias, which may result in biased outcomes but isn't always obvious during development (Barocas & Selbst, 2016). In the absence of justice or equality in the algorithm's optimisation target, it is possible that precision will be prioritised above treating all persons equitably.
- **Biased Decision-Making Processes:** Biased results may occur even if the AI model is itself objective. Humans utilising AI-generated suggestions may still be impacted by their own preconceptions and biases. For instance, human decision-makers might make matters worse rather than better by acting on AI-generated outcomes if they do so with prejudice (Diakopoulos, 2016).
- **Role of Human Biases:** Human prejudices have a major influence in programming prejudice into AI programmes. Data used to train AI models may unwittingly reflect the biases of the people who collected and annotated the data as well as the people who developed the AI (Bolukbasi et al., 2016). The selection of training data, the labelling of data, and the fine-tuning procedure are all vulnerable to the introduction of human bias. For instance, a biased human decider may categorise data such as photographs or texts in a manner that reflects their own biases, which the AI model would then learn from.

Addressing Bias in AI

Addressing bias in AI is crucial for ensuring fair and equitable results for all users as AI systems spread across more and more fields. Discriminatory choices, increased societal inequality, and decreased faith in AI are all possible outcomes of unchecked bias in the field of artificial intelligence. Data pretreatment, algorithmic tweaks, and adversarial testing are only some of the tools and strategies suggested to reduce bias in AI systems. In addition, minimising prejudice and encouraging responsible AI development requires the participation of diverse and inclusive teams.

- **Data Preprocessing:** Data curation and preparation during data preprocessing is another method for minimising bias in training datasets. Methods for achieving this goal include the elimination of skewed data and the insertion of new information to compensate for gaps in coverage (Bolukbasi et al., 2016). In order to train AI models to recognise patterns without being impacted by previous biases, it is necessary to properly choose and analyse training data.
- **Algorithmic Adjustments:** Adjustments to algorithms attempt to make AI models more objective and unbiased. To prevent the model's predictions from being unfairly weighted in favour of or against specific groups, fairness restrictions may be included during the optimisation process (Zafar et al., 2017). Reweighting or resampling data may also be used to reduce bias in learning by giving equal weight to underrepresented groups.
- **Adversarial Testing:** In adversarial testing, artificial intelligence models are tested using intentionally biased and vulnerable adversarial inputs. Artificial intelligence models that display biased behaviour or produce biased results may be exposed via adversarial testing (Liu et al., 2019). Developers may identify and correct causes of bias in AI models by putting them through such scrutiny.
- **Importance of Diverse and Inclusive Teams:** When developing AI, it is essential to include diverse and inclusive teams in order to minimise prejudice. Teams of researchers, engineers, and designers are often involved in the creation of AI systems. Teams that are diverse in terms of demographics and life experiences are better able to spot and address any biases that may be present (Gebru et al., 2018). Also, more inclusive and equitable AI systems may be developed by teams that reflect the diversity of their end users.

Combating prejudice in AI is a complex problem that calls both innovative technological measures as well as policy reform. Data pretreatment, algorithmic tweaks, and adversarial testing are all strategies for making AI systems more equitable and less biassed. It's hard to stress the value of having a diverse and welcoming workforce while building AI. Assuring that AI is useful for all its users and does not perpetuate prejudices requires collaborative initiatives that incorporate people from a variety of backgrounds.

Trade-offs in Fairness

To avoid unfair treatment of any population by AI systems, fairness is an essential part of the AI design process. At the same time, pursuing perfect justice may come at the price of overall model performance and usefulness, therefore it's important to weigh the pros and cons before committing to a particular approach to attaining fairness in AI models. Developers face a formidable obstacle in finding a happy medium between these competing goals of AI systems' fairness and their efficiency.

Trade-offs Between Fairness Objectives and Model Performance:

a. **Accuracy vs. Fairness:** One of the main compromises in fairness is between precision and equity. Artificial intelligence models that are optimised for high accuracy may unintentionally perpetuate data biases, leading to unequal results for certain groups (Zemel et al., 2013). However, if fairness is highly valued, the model may become less effective since it will avoid making judgements that can be seen as biased.

b. **Equal Opportunity vs. Equalized Odds:** Different fairness aims, such as those emphasising equal opportunity or equalised odds, place emphasis on different elements of fairness. Equalised odds and equal opportunity both guarantee that the genuine positive and false positive rates are the same for all groups. Fairness gains in one area may have unintended consequences in another, making it difficult to achieve both goals at once (Pleiss et al., 2017).

Challenges in Achieving Perfect Fairness and Model Effectiveness:

a. **Data Quality and Representativeness:** Complete Justice demands complete and impartial information. Data may have inherent biases or reflect just a subset of circumstances in the actual world (Bolukbasi et al., 2016). It is difficult to guarantee full fairness without sacrificing model performance when the data is biased.

b. **Complex Decision Landscapes:** When making predictions, AI models often face complicated decision landscapes, where a number of variables interact. Fairness may have to be sacrificed for other traits in such settings, such as interpretability and resilience (Raghavan et al., 2018). Finding a happy medium between all these aims is not always easy.

c. **Contextual Fairness:** Fairness might differ depending on the area of application and cultural norms (Diakopoulos, 2016), thus it's important to keep that in mind. It may be impossible or even contradictory to strive for absolute justice in every situation.

Addressing Trade-offs:

Understanding the individual application environment is essential for striking a fair balance between fairness and model performance. When developing fairness goals, it is crucial to include and hear from a wide range of stakeholders. Understanding how biases and fairness issues impact predictions may be aided by the transparency and interpretability of AI models. In addition, unexpected repercussions may be identified and corrections made to reach a more balanced approach by constant monitoring and assessment of AI systems in real-world situations (Mehrabi et al., 2019).

A fundamental difficulty in developing AI is balancing fairness trade-offs. It takes complexity and awareness of context to strive for ideal justice without sacrificing model performance or usefulness. Developers need to think about how different fairness targets could affect the overall efficiency of the model since fairness is a multidimensional term. Responsible, transparent, and egalitarian AI systems may be achieved by understanding these trade-offs and incorporating a wide range of stakeholders in the decision-making process.

Transparency and Explainability

With the proliferation of AI systems across industries, there is a pressing need for openness and explanation. Users are better able to comprehend the reasoning behind AI-driven results when such systems are transparent and provide explanations for their judgements. Accountability, user trust, and regulatory compliance may all be improved with more openness like this. However, there are major obstacles to attaining transparency and explainability in sophisticated AI models, since doing so may reduce performance and forecast accuracy.

Need for Transparent AI Systems:

a. **Accountability:** Stakeholders may ensure that AI models are responsible for their actions if the system is transparent. Understanding the AI's thought process is crucial when it makes life or death judgements in fields like medicine, finance, or law enforcement (Lipton, 2016). Models that can be audited in this way may help uncover biases and mistakes more quickly.
b. **Trust and User Acceptance**: Users need to comprehend and trust the AI-generated results for AI systems to be broadly accepted and trusted. Users often struggle to make sense of complicated AI models because to their "black box" character, but this gap may be narrowed by disclosure and explanation (Guidotti et al., 2018).
c. **Regulatory Compliance:** Strict rules in several areas, like healthcare and banking, mandate that AI systems provide justifications for their judgements (Hutson, 2020). Transparent AI aids in fulfilling these regulatory needs and guarantees conformity.

Challenges in Interpretable AI Models:

a. **Model Complexity:** It's not easy to make sense of AI models like deep neural networks because of their complicated designs and the millions of parameters they use (Ribeiro et al., 2016). Simplifying such models for readability risks omitting important details.
b. **Interpretability-Accuracy Trade-off:** As a result of the trade-off between interpretability and prediction accuracy, several interpretable strategies aim to reduce model complexity (Lundberg & Lee, 2017). ones that are easy to understand may not be as effective as more intricate "black-box" ones.
c. **Comprehensibility vs. Completeness:** The trade-off between interpretability and completeness is a potential problem when designing AI models (Doshi-Velez & Kim, 2017). It's hard to find the sweet spot between readability and thoroughness.
d. **Feature Interaction and Non-Linearity:** When AI models depend on intricate relationships between features or display non-linear behaviour, it might be difficult to make sense of them (Ribeiro et al., 2016). Oversimplifying model behaviour may occur if these interactions are too simplified.

Addressing Challenges:

Researchers and developers are experimenting with a number of methods to overcome the obstacles standing in the way of full transparency and explainability in AI:

- **Model-specific Interpretability:** Create methods adapted to certain model architectures, using model-specific features to offer explanations (Ribeiro et al., 2016).
- **Post hoc Interpretability:** Adopt post hoc interpretability strategies that provide an explanation for the actions taken by AI models without impacting the models' underlying design or performance (Lundberg & Lee, 2017).
- **Local vs. Global Explanations:** Differentiate between explanations that are local to a certain prediction and those that are global to the model's overall behaviour (Guidotti et al., 2018).
- **Human-in-the-loop Approaches:** Having humans read and approve AI-generated explanations is important for ensuring that they are correct and make sense (Lipton, 2016).

Accountability and credibility in AI systems depend on their openness and capacity to explain their decisions. Accountability, consumer trust, and regulatory compliance may all benefit from having AI judgements explained. However, overcoming obstacles like model complexity, interpretability-accuracy trade-offs, and feature interactions is essential for establishing transparency in complex AI models. We may make progress towards more transparent and explainable AI systems without sacrificing performance by tackling these difficulties using model-specific and post hoc interpretability approaches and by integrating human experts in the process.

Regulatory and Policy Considerations for AI Ethics and Fairness

The necessity for strong legal and policy frameworks to address AI ethics and justice has grown more apparent as AI technology is deployed more widely. To avoid prejudice, bias, and other unintended outcomes, it is essential that AI systems be created and used in a fair and ethical way. In this setting, current norms and standards concerning the ethics and fairness of AI play a pivotal role in directing the development of these technologies. Effective adoption of ethical AI practises also requires the participation of governments, organisations, and industry players.

Review of Existing Regulations and Guidelines:

a. **National Regulations:** Several nations have started the process of creating and enforcing legislation to ensure the ethical and fair use of AI. Aiming to promote openness and accountability in AI systems, the European Union's General Data Protection Regulation (GDPR) contains requirements on automated decision-making and data protection (EU GDPR, 2016).
b. **Ethical Guidelines:** Many groups and institutions have formulated moral principles to guide AI research and development. Norms for transparent, accountable, and equitable AI development have been created by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE, 2016).
c. **AI Principles:** Many tech firms and trade groups have issued AI principles to help guide their AI-related work. To minimise prejudice and guarantee justice in AI systems, Google, for one, has pledged to adhere to its own AI Principles (Google, 2018).

Role of Governments, Organizations, and Industry:

a. **Government Oversight:** In order to ensure that AI is used in an ethical manner, governments must exercise oversight via the creation of thorough AI legislation and regulations. Data privacy, transparency, and accountability in AI systems may be regulated by them (Cheney-Lippold, 2018). Government organisations may also work with businesses to guarantee ethical and lawful use of AI.
b. **Corporate Responsibility:** Businesses creating AI technology should follow ethical AI procedures. They need to factor in fairness as they create AI models and infrastructure. To guarantee conformity with rules and regulations, businesses might form internal AI ethics committees (Liu et al., 2020).
c. **Industry Collaboration:** Sharing best practises, overcoming common obstacles, and encouraging an ethical AI development culture all need industry-wide collaboration (Floridi et al., 2018). Alliances in the business world may collaborate to create universally applicable codes of ethics.
d. **Public-Private Partnerships:** Together, governments, organisations, and industry stakeholders may tackle the issue of AI ethics and justice via the formation of public-private partnerships. d. Knowledge exchange, joint research, and innovative solutions are all aided by such alliances (Hagendorff, 2020).

The ethical development and use of AI technology are greatly aided by regulatory and policy issues. Transparency, justice, and accountability in AI systems may be fostered by building on existing standards and principles. Ethical AI practises need to be implemented by governments, organisations, and industry stakeholders. To properly address the difficulties and complexity of AI ethics and justice, collaboration amongst these organisations is crucial. We can pave the road for the ethical and trustworthy use of AI technology by developing rigorous regulatory frameworks and encouraging cooperation.

Ethical Decision-making in AI Development

Ethical standards must be taken into account throughout the development and deployment of AI technologies to guarantee that they will be utilised in a responsible manner. To help prevent prejudice, bias, and other negative social effects, creators of AI need ethical decision-making frameworks. Building trust with users and preventing negative results may be achieved by giving serious consideration to the social effect and possible implications of AI applications.

Understanding the larger implications of AI technology and developing frameworks for ethical decision-making are crucial to the development of ethical AI.

Frameworks for Ethical Decision-making in AI Development:

- **Principles-based Frameworks:** Guidelines-based frameworks establish broad ethical guidelines to direct AI research and development. Transparency, fairness, and accountability are only a few of the seven guiding principles outlined in the European Commission's "Ethics Guidelines for Trustworthy AI" (European Commission, 2019). Ethical principles may be included into the creation of AI with the use of such frameworks.
- **Impact Assessment Frameworks:** Potential ethical hazards and implications of AI applications are the primary focus of impact assessment frameworks. Developers are urged by these frameworks to consider the potential consequences of their AI solutions for all parties involved, especially those who are particularly vulnerable. Developers may avoid or lessen the severity of unforeseen outcomes by conducting impact assessments.
- **Human Rights-based Approaches:** Human rights-based frameworks place a premium on safeguarding civil liberties throughout the creation of intelligent machines. When creating AI systems, programmers think about protecting individuals' privacy, free speech, and equal treatment (Romeo Casabona, 2018). Artificial intelligence (AI) technology may be created in a manner that is consistent with society ideals if human rights are protected.

Importance of Societal Impact Considerations:

- **Avoiding Harm:** To protect people and communities from possible damage, it is important to think about how AI applications may affect society as a whole. Artificially intelligent systems or algorithms that are biased may have discriminatory impacts, further isolating and marginalising certain groups of people (Selbst et al., 2019). Making moral choices helps in spotting and fixing such problems.
- **Building Trust:** Ethical AI development helps build trust between the IT community and its end consumers. Users are more likely to accept the system's conclusions when AI technologies are applied clearly and ethically (Rudin et al., 2019). For artificial intelligence (AI) applications to be widely used, people must have faith in them.
- **Responsible Innovation:** Responsible innovation is encouraged by societal impact concerns. Designers of AI systems should consider how such systems could alter social dynamics and power dynamics (Larson et al., 2018). Developing AI responsibly requires foreseeing possible outcomes and making adjustments appropriately.
- **Public Perception and Acceptance:** The public's view and adoption of AI technologies are influenced by their potential effects on society and by ethical concerns. AI systems that are designed ethically have a better chance of being accepted by both users and politicians (Jobin et al., 2020). The public's backing is essential to the effective rollout of AI programmes.

Responsible AI development and deployment requires thoughtful consideration of ethical considerations. Developers may be guided in creating AI systems that adhere to ethical principles and values with the aid of proposed frameworks for ethical decision-making. Developers may discover and correct any biases, maintain justice, and prevent damage to people and communities by thinking about the social effect and potential repercussions of AI applications. The widespread use and acceptance of AI technology across many fields is facilitated by responsible AI development, which in turn fosters trust among users.

Conclusion

In conclusion, AI ethics and justice are crucial factors in the creation and implementation of AI systems. As AI spreads into more fields, it will become more important to eliminate discrimination and promote equality within these systems. Discrimination, the reinforcement of prejudices, and a widening of social gaps may all result from biased AI. To address these problems, it is essential to pinpoint their origins, such as biased data and algorithms. Responsible decision-making is required to address ethical concerns raised by AI's potential uses. Ethical systems prioritise openness, equity, responsibility, and respect for human rights. Adherence to these principles encourages responsible innovation, protects users, and builds trust with them. Fairness trade-offs are difficult to solve. There has to be a fine balance between fairness goals since striving for complete fairness might hurt the performance of the model as a whole. Having varied teams work on AI and taking into account different points of view may help in decision making. Integral to developing reliable AI systems are their explicability and openness. Explaining AI's reasoning improves transparency and confidence in the system. For AI to gain wider acceptance and adoption, it is essential that the general public get a deeper knowledge of AI ethics and justice. In conclusion, AI ethics and justice are crucial for the safe and ethical development of AI. We may develop AI systems that are more open, egalitarian, and in line with social ideals if we investigate strategies to reduce prejudice, embrace ethical decision-making, and comprehend societal consequences. To guarantee that AI technologies serve mankind and respect the ideals of justice, accountability, and transparency, achieving ethical AI is not simply a technological problem but a moral obligation.

References

[1] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Baezconde-Garbanati, L., Grana, R. A., & Cruz, T. B. (2020). Artificial Intelligence Ethics in Radiology: An Opportunity for Inclusive Imaging. Journal of the American College of Radiology, 17(7), 873–877. https://doi.org/10.1016/j.jacr.2020.02.015

[3] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104(3), 671-732. https://doi.org/10.15779/Z38BG31

[4] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Veale, M. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. In Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (pp. 19-21). https://doi.org/10.1145/3231335.3231341

[5] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS) (pp. 4349-4357). https://proceedings.neurips.cc/paper/2016/hash/086b536f2d26726c759c7a7563db32c9-Abstract.html

[6] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, 77-91. https://doi.org/10.1145/3177317.3177353

[7] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora necessarily contain human biases. Science, 356(6334), 183-186. https://doi.org/10.1126/science.aal4230

[8] Cheney-Lippold, J. (2018). We Are Data: Algorithms and the Making of Our Digital Selves. NYU Press.

[9] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163. https://doi.org/10.1089/big.2016.0047

[10] Chouldechova, A. (2021). The Mythos of Model Interpretability. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 203-213). https://doi.org/10.1145/3287560.3287598

[11] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 797-806). https://doi.org/10.1145/3097983.3098095

[12] Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., ... & West, S. M. (2019). The AI Now report 2019. AI Now Institute, New York, NY. https://ainowinstitute.org/AI_Now_2019_Report.pdf

[13] Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. In Proceedings on Privacy Enhancing Technologies, 2015(1), 92-112. https://doi.org/10.1515/popets-2015-0006

[14] Diakopoulos, N. (2016). Accountability in Algorithmic Decision Making. Digital Journalism, 4(6), 703-718. https://doi.org/10.1080/21670811.2016.1178596

[15] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608

[16] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214-226). https://doi.org/10.1145/2090236.2090255

[17] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214-226). https://doi.org/10.1145/2090236.2090255

[18] European Commission. (2019). Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[19] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.6ce0c5cd

[20] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Tamburrini, G. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

[21] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2020). Datasheets for Datasets. arXiv preprint arXiv:1803.09010. https://arxiv.org/abs/1803.09010

[22] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Diverse and Inclusive Voices in AI Ethics. AI Magazine, 39(4), 61-65. https://doi.org/10.1609/aimag.v39i4.6950

[23] Google. (2018). AI at Google: Our principles. https://www.blog.google/technology/ai/ai-principles/

[24] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), 51(5), 1-42. https://doi.org/10.1145/3236009

[25] Hagendorff, T. (2020). The Ethics of AI Ethics — An Evaluation of Guidelines. Minds and Machines, 30(1), 99-120. https://doi.org/10.1007/s11023-019-09517-8

[26] Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2125-2126. https://doi.org/10.1145/2939672.2945388

[27] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323). https://proceedings.neurips.cc/paper/2016/hash/0b5fe4c13aa79a57ef971b7a007bdbc9-Abstract.html

[28] Heidari, H., Feller, A., & Hajian, S. (2019). A moral framework for understanding of fairness in machine learning. In Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency (pp. 409-418). https://doi.org/10.1145/3287560.3287596

[29] Hutson, M. (2020). Why is AI Explainability Important? Healthcare Analytics News. https://www.hcanews.com/news/why-is-ai-explainability-important

[30] IEEE. (2016). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. https://standards.ieee.org/standard/7012-2020.html

[31] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

[32] Jobin, A., Ienca, M., & Vayena, E. (2020). Artificial Intelligence: The Global Landscape of Ethics Guidelines. In The Oxford Handbook of Ethics of AI (pp. 59-80). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.4

[33] Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1-33. https://doi.org/10.1007/s10115-011-0463-8

[34] Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4066-4076). https://proceedings.neurips.cc/paper/2017/hash/16e82bb7633e9f8914f7d40a82d26145-Abstract.html

[35] Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2018). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[36] Lipton, Z. C. (2016). The Mythos of Model Interpretability. In Proceedings of the 2016 Conference on Artificial Intelligence and Statistics (AISTATS) (Vol. 51, pp. 1010-1018). https://proceedings.neurips.cc/paper/2016/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[37] Liu, B., Jiang, X., Liao, L., Zhao, T., Dey, A. K., Le, T. T., & Chen, C. (2020). Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-13). https://doi.org/10.1145/3313831.3376188

[38] Liu, C., Gao, J., Chen, P., & Han, Y. (2019). Bias also matters: Bias correction for deep neural network based age estimation with anchor age groups. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 108-114). https://doi.org/10.1145/3287560.3287592

[39] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4765-4774). https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[40] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys (CSUR), 52(4), 1-35. https://doi.org/10.1145/3287560

[41] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

[42] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On Fairness and Calibration. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1337-1345). https://doi.org/10.5555/3305890.3306030

[43] Raghavan, M., Kleinberg, J., Pardos, Z. A., & Wallach, H. (2018). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 113-119). https://doi.org/10.1145/3278721.3278779

[44] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

[45] Romeo Casabona, C. M. (2018). Human Rights and Ethical Considerations in the Application of Artificial Intelligence to Health Care. In Artificial Intelligence in Health Care (pp. 73-84). Springer, Cham. https://doi.org/10.1007/978-3-319-94857-8_7

[46] Rudin, C., Wang, C., Coker, B., & Doshi-Velez, F. (2019). The age of secrecy and unfairness in recidivism prediction. Harvard Data Science Review, 1(2). https://doi.org/10.1162/99608f92.4a4cf2b8

[47] Ruggieri, S. (2014). Discrimination-aware data mining. ACM SIGKDD Explorations Newsletter, 15(1), 1-10. https://doi.org/10.1145/2722877.2722883

[48] Sweeney, L. (2013). Discrimination in online ad delivery. In Privacy, Security, Risk and Trust (PASSAT), 2013 International Conference on and 2013 International Conference on Social Computing (SocialCom) (pp. 248-255). IEEE. https://doi.org/10.1109/PASSAT/SocialCom.2013.113

[49] Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web (pp. 1171-1180). https://doi.org/10.1145/3038912.3052660

[50] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. In Proceedings of the 30th International Conference on Machine Learning (Vol. 28, pp. 325-333). https://doi.org/10.5555/3042817.3043053

How Cite this article?

Modi, T. B. (2023). Artificial Intelligence Ethics and Fairness: A study to address bias and fairness issues in AI systems, and the ethical implications of AI applications. *Revista Review Index Journal of Multidisciplinary, 3*(2), 24–35. https://doi.org/10.31305/rrijm2023.v03.n02.004