Journal of Informatics and Web Engineering
Vol. 4 No. 1 (February 2025) eISSN: 2821-370X

Integrating Moral Values in AI: Addressing Ethical Challenges for Fair and Responsible Technology

Khushboo Shah[1*], Hiren Joshi[2], Hardik Joshi[3]
[1]Department of Computer Science, St. Xavier's College (Autonomous), Ahmedabad-9, Gujarat, India.
[2,3]Department of Computer Science, Gujarat University, Ahmedabad-9, Gujarat, India
**corresponding author: (khushboo.shah@sxca.edu.in; ORCiD: 0000-0001-7359-547X)*

Abstract – Today, Artificial Intelligence (AI) has become an integral part of our day-to-day life. From personal task to professional life everywhere we need AI. All the sectors like - transport, education, healthcare, agriculture, we find AI everywhere. Each coin has two side, similarly AI has its own pros and cons. Main challenge with AI is ethics. There is no doubt in the work efficiency of AI but ultimately, it's a machine without emotions hence whatever the decision it takes it purely without ethical values. Sometimes, this type of decision may lead to disasters, especially in the industry where human life is involved e.g. healthcare. The focus of this paper is to integrate moral values into AI systems. It also talks about different ethical frameworks like utilitarianism, deontology and virtue ethics along with the state-of-art work and knowledge gaps. This research also explored various case studies where AI implemented with ethics. Integration of moral values with AI has many issues like bias, transparency and accountability. Here author has proposed a new model named Ethical Alignment Algorithm (EAA). This model helps to integrate ethics with AI step-by-step. This approach will help AI to make fair, sensible and responsible decisions. This paper will also help researchers to work with a multidisciplinary approach. Different subject specialists can come together and make AI policies with ethics. EAA has the potential to make the AI systems not only advanced but with high moral values. In the end, the paper highlights current AI development and future scopes. The main aim of this research is to promote justice and fairness in AI decisions for the overall well-being of society.

Keywords— *Artificial Intelligence, Ethical Challenges, Ethical Frameworks, Future Research in AI, Moral Values in AI.*

Received: 27 August 2024; Accepted: 18 November 2024; Published: 16 February 2025

1. INTRODUCTION

Artificial Intelligence has become an essential part of our day-to-day life. Barely we see a field without AI integration. From agriculture to aviation, from academics to healthcare, everywhere AI has a powerful role [1], [2], [3]. Mainly AI's role is to create a machine which thinks, learns and takes decisions just like humans [4], [5], [6]. As the use of AI is increasing, society gets impacted, and many times people face ethical injustice. The main challenge with this is lack of moral values in existing AI systems. Moral values are the most important aspects in human society which help to take fair and unbiased decisions. Which tells what is right and what wrong [5]. Implementing moral values in Ai systems means making more reliable and responsible machines which take decisions ethically [4].

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

It is crucial to have AI with moral values because nowadays AI is extensively used in the areas where humans' life is directly involved; for example, healthcare, transportation and self-driving cars [4]. AI takes decisions and because humans' life is affected by the decisions, decisions should be fair without harm. AI without ethics may lead to disasters, privacy issues, gender or cast bias and many more [7]. If we ensure that AI system is making decisions with ethics, then we can reduce or nullify the risks of harmful results. For example, if AI system is there in employee hiring process, then it might be possible it takes bias decisions based on candidate's gender, caste or race. Another example is healthcare – it might be possible that AI without ethics make decisions by prioritizing profit over patient health [8]. In self-driving vehicles – if a car sees a human and animal on the road then who will be saved from an accident? So, there are many examples which really need AI with ethics.

This paper talks about how we can make AI systems with moral values. It also discusses various ethical frameworks like – utilitarianism, deontology and virtue ethics to understand human psychology and implementing them in AI [5]. The paper also reviews state-of-art AI systems, ethics in AI, their challenges and limitations in implementing it with AI [5]. It also discusses different case studies of different sectors with real-world problems along with the solution to include ethics in AI and what types of challenges can be faced during integrating it. The paper talks about the future of AI and why it is important to implement moral values with it for safe development. As we grow with AI, it is equally important to enhance ethical standards in systems. As we look at new trends and technologies which are helpful for society but without moral values it can harm humans hence there is a requirement of systematic, ethical, and unbiased AI systems in future [9]. This paper covers basics of AI, its ethics, methodology for implementations and its overall effect on human society [10]. The main aim of this research work is to implement moral values in AI for multidisciplinary systems. The subject experts from different fields can work to make new policies to implement these ethics in AI.

In response to the growing need for ethical considerations in AI systems, this paper introduces the Ethical Alignment Algorithm (EAA), a novel model designed to ensure that AI decisions are consistently aligned with human moral values. By looking at the need of ethics in AI systems, this paper proposes the model Ethical Alignment Algorithm (EAA). It's a novel approach to integrating moral values in AI systems. This model ensures that all the decisions taken by AI should be aligned with ethics. Step-by-step implementation process is also discussed in this paper for all the subject experts can work together to make robust and ethically sound multiciliary AI systems. Correct implementation of moral values to AI by following EAA model, will help the society to get fair, unbiased and accountable decisions from AI systems.

Finally, this paper discusses the responsibility of AI developers to make ethically sound AI systems. By focusing on the importance of ethical AI, paper encourages AI development which is advanced as well as ethically sound [5].

2. LITERATURE REVIEW

Combining ethics with AI has gained major attention in the past few years. It led the researchers to contribute to literature which explores different aspects of ethics in AI. Implementing Different traditional ethical frameworks to AI is the focus of today's research. These frameworks include utilitarianism, deontology and virtue ethics. Binns [5] discusses how these frameworks can help to understand deployment of these designs into AI systems. The paper discusses that utilitarianism is all about maximizing overall happiness, deontology is about moral rules and virtue ethics talks about the character and intentions of the people involved. Each framework has unique insights and has distinct challenges when applied to AI systems.

There are facial recognition systems which are not accurate when it comes to race and gender biasness decisions. Buolamwini and Gebru [11] have highlighted how these systems work and what are the challenges with their decisions. Their research concludes the importance of inclusive training data and robust evaluation metrics. They concluded that ethics in AI is a major concern in all the industries.

Doshi-Velez and Kim [12] research work showing that not only fair and unbiased systems are required but also transparency and accountability also matter. In their research they have explored and gave their support to methods like interpretable models and pos-hoc explanations which enhance transparency and allow users to understand and trust AI decisions.

Domain centric research is also done by many researchers. For example, Morley et al. [13] showed the importance and challenges of ethics in AI in healthcare domain. Research focused on the need for patient-oriented AI systems which should take decisions mainly based on the safety of patient not only for profit of hospital. Similarly, Goodall [14] talks about ethical dilemmas in autonomous vehicles. When it comes to taking decision to save life during

the accident condition, how AI vehicle behaves. Also, he discussed and relate this problem with the classic example of 'trolley problem scenario'.

Though there are various studies conducted for ethical AI but still there are gaps in research. Traditional ethical frameworks help in understanding the meaning and value of ethics but integrating them with AI systems practically is still an open issue. Existing systems are still on paper as practical implementation is a real challenge.

Ethics is a subjective thing and implementing it in AI systems which operate in dynamic environments is a big problem. Moral values have a specific base, but it varies from person to person and situation to situation hence adapting one framework and implementing it with machine is not a simple job. State-of art works have not satisfactorily addressed how we can align AI systems ethics with human values in all the conditions.

Many studies have been done in AI for ethics, especially for bias and fair decisions but still not all biases are covered. There are biases which can be emerged only after deployment. Still there is a need for end-to-end bias mitigation techniques. This required proper data collection, model training and real-world applications. To implement ethics in AI systems, we need universal guidelines, but it is observed that still there is a gap in its acceptance globally. There are organizations and government bodies who have proposed policies but there is not any agreement for following them. This inconsistency may lead to varied decisions regionally and industry wise.

Hence, there is a major need of preparing one universal guideline and end-to-end bias mitigations AI systems.

3. AI AND ETHICAL FRAMEWORKS

Ethical frameworks are very important to guide AI systems which are going to interact with human beings and to make decisions. This framework ensures that AI systems work in alignment with human moral values. There are many frameworks that exist, but this research focuses on the main three: Utilitarianism, deontology and virtue ethics. These frameworks are popular and perfectly fit well with the existing challenges of implementing moral values with AI systems. All three frameworks have their unique features. First utilitarianism helps in decisions making that benefit the post people. Deontology is mainly about the rules principles which set the ethical boundaries. And last virtue ethics which is about the character and virtues of decisions makers. This research work has used all these three frameworks to implement ethics in AI systems for fair, responsible and ethical outcomes.

3.1 Utilitarianism

Utilitarianism is an ethical theory which is consequentialist. It decides whether the action is right or wrong based on its consequences. Its main aim is to maximize overall happiness and minimize overall loss. This principle is often co-related with the popular phrase "the greatest good for the greatest number"[4], [15], [16].

Working with AI models is easy but at the same time relying on its decisions completely needs special attention. AI decisions should be analyzed that how it impacts society and huma being. Utilitarianism technique helps to understand and evaluate the consequences of the decisions. For example, in health care, patients' health should be the priority. Here Overall happiness means patient's health along with the reasonable service and consultation charges. Recovery, improvement in health, a complete cure all these should be the priority when AI takes any decisions [9]. If we talk about autonomous vehicles, priority is safety of humans and minimum harm if there is any accident condition. Here Utilitarianism approach can help AI systems in decision making to reduce the impact of accident by choosing the right action [2].

Incorporating utilitarianism with AI systems has many challenges. One major challenge is predicting future consequences of AI actions. It can be uncertain and complicated. There is also a possibility that by applying strict utilitarianism approach, AI may overlook rights and needs of an individual [5]. It can violate the privacy and priority of an individual.

3.2 Deontology

Deontology is unlike utilitarianism. Its main aim is to focus on ethical rules rather than consequences of actions. [16]. Many deontologists believe that if actions are taken morally then need not worry about the consequences [9], [10]. This framework's prime focus is on roles, responsibilities and rights.

As far as implementing ethics to AI systems is concerned, it is important to make systems function based on moral principles. For example, if there is an AI system for recruitment process, then it needs to be made sure that AI's

evaluation of candidates should be based on skills, qualifications and potential, not based on any biasedness [10]. This type of fair decision helps to protect human rights and maintain fairness.

However, deontological ethics can lead to rigid decision-making. For instance, if an AI system strictly follows rules, it might struggle with complex situations where rules conflict or where rigid adherence to rules might cause harm. An AI system programmed to follow privacy rules strictly might not adapt well to scenarios where data sharing is necessary for achieving better outcomes [5].

3.3 Virtue Ethics

Virtue ethics is a psychological approach where a person's character is prime objective. It doesn't care about the rules and consequences [16]. Theory of Aristotle says that ethical behaviour developed by honesty, compassion and courage. Actions are always right if the person has moral values [16].

Incorporating virtue ethics in AI means encouraging designs of AI systems which are more towards the good qualities like respect, honesty and empathy. For example, if there is an AI customer agent interacting with humans then it should offer help with kindness and understanding. This approach helps to build the user's trust and satisfaction.

Virtue ethics is a subjective behaviour which varies from situation to situation. One challenge in incorporating virtue ethics with AI is that it is difficult to define good behaviour in different situations. It is difficult to make AI systems that follow kindness and fairness with rules.

Table 1 illustrates the importance of applying ethical frameworks to AI. Without ethics, AI could cause harm, perpetuate injustices, or degrade societal values. However, when ethics are applied, AI can contribute positively to society by making decisions that are fair, just, and aligned with human values.

4. MORAL VALUES IN AI DEVELOPMENT

AI systems are mainly designed and developed to act and behave like humans. Nowadays, AI systems reason just like humans do. They are logic-based and take actions and decisions rationally. They also ensure that the reasons behind the actions should be correct [15]. But only being rational is not sufficient when it's a matter of human life. Some actions and decisions require ethics hence incorporating them into AI is crucial. AI systems should be aligned with human values, especially when making decisions for human life and society. Making ethical AI systems can give us fair and responsible machine-made outcomes. And it leads to less or no harm to humans.

4.1 Importance of Incorporating Moral Values in AI design

Incorporating moral values into AI systems is important because nowadays we have AI systems that take actions and make decisions which affect human lives. If actions and decisions are taken without moral values, then it can cause major harm to human lives and society. For example, AI in the healthcare domain where AI machines do surgery, make decisions on treatments, make bills, give suggestions, prescribe medicines, and whatnot [2]. If AI works only by keeping profit in mind and not thinking about the situation of the patient who is without moral values, then it can directly harm humans. Similarly, with autonomous vehicles. If there is a condition of accident and matter of saving lives then to whom AI will save, that's the biggest question [7]. Hence here also needs AI with knowledge and ethics that decisions should be taken which can harm minimum. There is also a case where AI without ethics can bias and invade privacy. Ethical AI should respect human dignity and rights [8]. AI algorithms should avoid discrimination and protect individual's privacy. If AI is Hiring candidate for job, then the decisions should be based on skill, efficiency and qualifications not on recommendations and race. So, AI with ethics is most important these days [17].

4.2 Methods for Embedding Ethical Considerations in AI Systems

There are various methods that can be used to embed moral values into AI systems. One is to integrate ethical guidelines directly into AI's decision-making algorithms. Through programming rules, these guidelines can be implemented to get fair, transparent outcomes. It also enhances the accountability of the algorithm [10]. ML algorithms with the knowledge of fairness can be designed to detect and reduce the biases in the training dataset. Also, ensures more impartial results [13], [18], [19].

The other method is incorporating human supervision into AI systems. It means rather than allowing machines to make decisions, there can be human operators who review and intervene in AI's decisions first and then allow AI to produce its final output. For example, in the case of an accident where AI is unable to make ethical decisions, the human operator can overrule the harmful result of the machine in emergency situations [12].

Taking different perspectives from different experts to design and develop AI systems for decision-making can also help to make better ethical machines. These systems can be transparent and give logical as well as ethical explanations of their decisions. This way users can understand how and why such decisions are made. It can help to build trust in AI systems [10].

Table 1. Comparing Utilitarianism, Deontology, and Virtue Ethics With Examples and Discusses The Consequences of Applying or Not Applying These Ethical Frameworks to AI Systems

Ethical Framework	**Definition**	**Example in AI**	**Consequence of Not Applying Ethics**	**Consequence of Applying Ethics**
Utilitarianism	Focuses on the greatest good for the greatest number. Decisions are based on the outcomes or consequences.	An AI system allocates resources in a hospital to maximize the number of lives saved, even if it means prioritizing some patients over others.	It is possible that decisions of AI harm minorities and benefit the majority which is unfair. Then there is a possibility of social conflicts.	It will maximize the overall happiness and well-being of society and citizens. And everyone will be benefited with minimum harm.
Deontology	Its focus is on rules, responsibilities, and duties irrespective of the results.	An AI system strictly follows privacy laws and does not share any personal data, even if sharing the data could save lives.	If AI follows rules strictly, then there can be a change in outcomes if there is any update or modification in rules and AI is not aware of it. In this type of case, the consequences can be negative and can harm society or individuals.	AI follows ethical guidelines and laws. It ensures that actions are morally right as per the defined rules. And this helps to gain the trust and reliability of people in AI systems
Virtue Ethics	Focuses on the character and virtues of the decision-maker rather than specific actions or consequences.	An AI mentor system encourages users to develop virtues such as honesty, empathy, and courage through interactions and feedback.	AI could promote or tolerate unethical behaviours or traits, leading to a degradation of moral standards in society and a lack of accountability.	AI fosters positive traits and behaviours, contributing to the moral development of users and promoting a more ethical society.

4.3 Case Studies/Examples of AI Systems with Moral Considerations

Many studies have been conducted which show how to integrate moral values into AI systems. For example, IBM developed an AI system called Watson. It was developed mainly to help doctors in diagnosing and treating cancer. Watson is incorporated with ethical guidelines which respect patients' privacy as well as recommending diagnosis and treatments based on latest medical research. Overall Watson gives fair results [20].

The other example is autonomous vehicles. There are various autonomous vehicles with AI with ethics. Waymo and Tesla are the two most leading companies in this area. Both have implemented decision-making systems along with balanced safety in their vehicles. These frameworks were designed to handle the situation where decision is purely based on ethics like preventing accidents to save lives. Minimum harm in unavoidable accident situation is the main priority of these systems [12].

The next example is social media. Facebook, Twitter, and Instagram platforms use AI detectors which remove harmful contents. These systems are incorporated with ethical guidelines to provide censorship. It ensures that decisions should respect freedom of expression while protecting users from harmful content [14].

5. PROPOSED MODEL – ETHICAL ALIGNMENT MODEL

The Ethical Alignment Model (EAM) will help AI systems to follow ethical rules. It checks the AI's decisions using a simple process. The model uses an algorithm to compare each decision with set of ethical standards. This makes sure the AI respects human values.

5.1 The Ethical Alignment Algorithm (EAA)

The Ethical Alignment Algorithm consists following steps:

Step 1. Collect relevant input data that the AI system will use for decision-making.
Step 2. Identify the context and domain of the decision (e.g., healthcare, finance, etc.).
Step 3. Initialize ethical criteria based on the context. This includes setting up rules derived from ethical frameworks such as utilitarianism (maximizing overall good), deontology (adhering to duties), and virtue ethics (promoting moral virtues).
Step 4. Assign weights to each ethical criterion depending on its importance in the given context.
Step 5. Generate possible decisions or actions based on the input data.
Step 6. Evaluate each decision against the initialized ethical criteria.
Step 7. For each decision, calculate an Ethical Score with the help of following equation: $\sum_{i=1}^{n} \quad (Criterion_i \; X \; Weight_i)$
Step 8. Rank the decisions based on their Ethical Scores, prioritizing those with higher scores.
Step 9. Validate the top-ranked decisions against hard ethical constraints (non-negotiable rules, such as "Do not harm").
Step 10. If a top decision violates any hard constraints, discard it and consider the next one in the ranking.
Step 11. Implement a consensus mechanism where multiple ethical agents (representing different ethical perspectives) review the selected decision.
Step 12. If consensus is not reached, re-evaluate the decision by adjusting the weights of the ethical criteria and re-ranking the decisions.
Step 13. If the system operates in a dynamic environment, implement a feedback loop where the decision is monitored and adjusted in real-time based on its outcomes.
Step 14. Continuously update the ethical criteria weights based on new data or feedback, ensuring the system adapts to evolving ethical norms.
Step 15. Execute the validated and consensus-approved decision.
Step 16. Log the decision-making process and ethical evaluations for future audits and transparency.
Step 17. After execution, conduct an audit to ensure the decision adhered to the ethical criteria and analyze its impact.
Step 18. Document any ethical dilemmas or challenges faced during the process for continuous improvement.

The Ethical Alignment Algorithm (EAA) will help AI systems to make decisions based on ethics. It checks and ranks decisions using a set of ethical rules. It also uses strict checks and looks for agreement from different sides. This makes sure the AI follows human values. The EAA can be changed to work with different types of AI. It helps make sure technology is fair and responsible.

5.2 Application of the EAA in Diverse AI Domains

5.2.1 Scenario 1: Healthcare Diagnostics

Problem Statement:
In healthcare, AI systems can sometimes provide diagnoses that may not prioritize patient safety or ethical guide lines, leading to misdiagnoses or harmful outcomes.
Step 1 : Collect Relevant Input Data - Gather patient medical history, symptoms, and diagnostic tests.
Step 2 : Identify Context and Domain - Context is healthcare, focusing on patient diagnostics.
Step 3 : Initialize Ethical Criteria - Establish criteria such as patient confidentiality, informed consent, and prioritizing patient safety.
Step 4 : Assign Weights to Each Ethical Criterion - Weight patient safety as the highest priority, followed by confidentiality.
Step 5 : Generate Possible Decisions or Actions - AI suggests potential diagnoses based on input data.
Step 6 : Evaluate Each Decision - Assess each diagnosis against ethical criteria (e.g., risk to patient safety).
Step 7 : Calculate an Ethical Score - Use the formula to assign scores based on established criteria.
Step 8 : Rank Decisions - Rank diagnoses by their Ethical Scores, prioritizing safer options.
Step 9 : Validate Against Hard Constraints - Ensure no diagnosis violates non-negotiable ethical principles.
Step 10 : Discard Violating Decisions - Eliminate any diagnoses that compromise patient safety.
Step 11 : Implement Consensus Mechanism - Involve medical professionals to review the top diagnoses.
Step 12 : Re-evaluate if Consensus Not Reached - Adjust weights based on feedback from the medical team.
Step 13 : Incorporate Feedback Loop - Monitor outcomes of selected diagnoses to refine future decisions.
Step 14 : Continuously Update Weights - Adapt weights based on new data or emerging medical ethics.
Step 15 : Execute Approved Decision - Implement the chosen diagnosis and treatment plan.
Step 16 : Log the Process - Document the decision-making for future audits.
Step 17 : Conduct Post-Execution Audits - Review the effectiveness of the diagnosis and its impact on patient outcomes.
Step 18 : Document Ethical Dilemmas - Note any ethical challenges encountered during the diagnostic process.
Solution:
The EAA ensures that diagnostic decisions are aligned with ethical standards, significantly reducing the risk of harmful outcomes by prioritizing patient safety and informed consent.

5.2.2 Scenario 2: Autonomous Vehicles

Problem Statement:
Autonomous vehicles may encounter complex ethical dilemmas on the road, such as choosing between the safety of passengers and pedestrians in critical situations.
Step 1 : Collect Relevant Input Data - Gather data on road conditions, nearby vehicles, and potential obstacles.
Step 2 : Identify Context and Domain - Context is transportation, specifically autonomous vehicle navigation.
Step 3 : Initialize Ethical Criteria - Establish criteria focusing on minimizing harm and prioritizing passenger and pedestrian safety.
Step 4 : Assign Weights to Each Ethical Criterion - Weight pedestrian safety higher than other criteria.
Step 5 : Generate Possible Decisions or Actions - AI suggests maneuvers based on real-time driving data.
Step 6 : Evaluate Each Decision - Assess maneuvers against ethical criteria (e.g., likelihood of causing harm).
Step 7 : Calculate an Ethical Score - Assign scores to each maneuver using the established criteria.
Step 8 : Rank Decisions - Rank maneuvers by their Ethical Scores.
Step 9 : Validate Against Hard Constraints - Ensure no maneuver violates fundamental safety rules.
Step 10 : Discard Violating Decisions - Remove any maneuvers that could endanger lives.
Step 11: Implement Consensus Mechanism - Engage multiple ethical agents (e.g., traffic safety experts) for review.
Step 12 : Re-evaluate if Consensus Not Reached - Adjust maneuver weights based on feedback.
Step 13 : Incorporate Feedback Loop - Monitor real-time performance of selected maneuvers for adjustments.
Step 14 : Continuously Update Weights - Adapt weights as driving conditions change.
Step 15 : Execute Approved Decision - Implement the chosen driving maneuver.
Step 16 : Log the Process -Document decisions for audit trails.
Step 17 : Conduct Post-Execution Audits - Evaluate the maneuver’s impact on safety.
Step 18 : Document Ethical Dilemmas - Record challenges faced in decision-making.
Solution:
The EAA helps autonomous vehicles navigate complex ethical dilemmas by systematically prioritizing decisions that minimize harm to both passengers and pedestrians, ensuring ethical compliance in real-time.

5.2.3 Scenario 3: Recruitment Algorithms

Problem Statement:
Recruitment algorithms can inadvertently perpetuate biases, leading to unfair hiring practices that disadvantage certain groups of candidates.
Step 1 : Collect Relevant Input Data - Gather resumes, interview scores, and candidate demographics.
Step 2 : Identify Context and Domain - Context is human resources, specifically hiring practices.
Step 3 : Initialize Ethical Criteria - Establish criteria emphasizing fairness, diversity, and non-discrimination.

Step 4 : Assign Weights to Each Ethical Criterion - Prioritize diversity and fairness in the hiring process.
Step 5 : Generate Possible Decisions or Actions - AI evaluates candidates for job openings.
Step 6 : Evaluate Each Decision - Assess candidates against ethical criteria (e.g., potential bias).
Step 7 : Calculate an Ethical Score - Use the established formula to score candidates.
Step 8 : Rank Decisions - Rank candidates by their Ethical Scores.
Step 9 : Validate Against Hard Constraints - Ensure compliance with anti-discrimination laws.
Step 10 : Discard Violating Decisions - Remove candidates who do not meet fairness criteria.
Step 11 : Implement Consensus Mechanism - Involve HR professionals to review candidate rankings.
Step 12 : Re-evaluate if Consensus Not Reached - Adjust weights based on HR feedback.
Step 13 : Incorporate Feedback Loop - Monitor outcomes of hires to refine future algorithms.
Step 14 : Continuously Update Weights - Update weights based on demographic shifts or changes in ethical standards.
Step 15 : Execute Approved Decision - Select candidates for hiring based on ranked lists.
Step 16 : Log the Process - Document the recruitment process for transparency.
Step 17 : Conduct Post-Execution Audits - Review hiring outcomes for fairness and effectiveness.
Step 18 : Document Ethical Dilemmas - Note any challenges encountered in the recruitment process.
Solution: The EAA addresses biases in recruitment algorithms by implementing ethical criteria that prioritize fairness and diversity, ensuring equitable treatment of all candidates.

5.2.4 Scenario 4: Financial Decision-Making

Problem Statement: Financial decision-making processes may overlook ethical considerations, leading to irresponsible lending practices and negative social impacts.
Step 1 : Collect Relevant Input Data - Gather financial data from loan applicants.
Step 2 : Identify Context and Domain - Context is finance, particularly in lending decisions.
Step 3 : Initialize Ethical Criteria -Set criteria emphasizing fairness, transparency, and responsible lending.
Step 4 : Assign Weights to Each Ethical Criterion - Weight transparency and fairness higher than profitability.
Step 5 : Generate Possible Decisions or Actions - AI evaluates loan applications.
Step 6 : Evaluate Each Decision - Assess loan decisions against ethical criteria (e.g., risk of discrimination).
Step 7 : Calculate an Ethical Score -Score each application based on ethical criteria.
Step 8 : Rank Decisions - Rank loan approvals by their Ethical Scores.
Step 9 : Validate Against Hard Constraints - Ensure compliance with lending laws and ethical guidelines.
Step 10 : Discard Violating Decisions - Remove approvals that do not align with ethical lending practices.

Step 11 : Implement Consensus Mechanism - Consult financial ethics experts to review decisions.
Step 12 : Re-evaluate if Consensus Not Reached - Adjust criteria weights based on expert input.
Step 13 : Incorporate Feedback Loop - Monitor loan performance and impact on communities.
Step 14 : Continuously Update Weights - Adapt weights as financial regulations change.
Step 15 : Execute Approved Decision - Approve loans based on ethical assessments.
Step 16 : Log the Process - Document lending decisions for audits.
Step 17 : Conduct Post-Execution Audits - Evaluate the ethical implications of loan approvals.
Step 18 : Document Ethical Dilemmas - Record challenges encountered in financial decision-making.
Solution:
The EAA enhances ethical decision-making in finance by prioritizing transparency and responsible lending, thus reducing the risk of unethical practices and promoting social responsibility.

The Ethical Alignment Algorithm provides a structured and versatile framework for embedding ethical considerations into AI decision-making processes across various domains. By illustrating its application in healthcare, autonomous vehicles, recruitment, and finance.

6. ETHICAL ALIGNMENT IN AI

6.1 Case Studies

6.1.1 Bias in Healthcare Algorithms

In healthcare, AI algorithms are increasingly used to assist in medical diagnoses and treatment recommendations. However, a study [21] revealed that a widely used healthcare algorithm in the U.S. disproportionately favored white patients over black patients, assigning lower risk scores to black patients with the same health conditions. This resulted in fewer healthcare resources being allocated to minority patients.

Challenge: Ethical dilemmas in healthcare often revolve around bias and fairness. The EAA, in this case, would ensure that all decisions (such as patient prioritization) are ranked and evaluated based on a set of moral values (e.g., fairness, non-maleficence). Using ethical scoring and continuous feedback, the EAA could flag bias and dynamically adjust its criteria to align with human values, ensuring equitable healthcare access.

6.1.2 The Trolley Problem in Self-Driving Cars

Autonomous vehicles face ethical dilemmas such as the "trolley problem" - where an AI must choose between two unfavorable outcomes, like deciding whether to swerve to avoid hitting pedestrians, which might endanger passengers [21].

Challenge: The ethical decisions in this scenario can involve weighing harm to passengers versus harm to pedestrians. The EAA model works on two main principles. First, it follows utilitarian principle whose prime focus is to minimize overall harm. Second, it applies deontology rules which means do not cause harm intentionally. This model has also a real-time feedback loop mechanism. It means, it can adjust its decisions based on new inputs and allow it to handle complex situations ethically.

6.1.3 Amazon's Biased Hiring Algorithm

Amazon was using an AI recruitment system to hire new employees. After sometimes it was found that, that AI recruitment system was biased against women. Hence it was scrapped in 2018 [22]. Algorithm was trained with resumes submitted in last 10 years, but majority resumes were of men. And as a result, system learnt to recruit

men. This way system downgraded resumes that included anything related with women, like college name or activity related to women.

Challenge: Bias in recruitment algorithms can lead to unjust ice to qualified female candidates. The EAA model provides a solution to this problem by assigning weights to ethical criteria such as fairness and equal opportunity to all genders. Continuing monitoring of EAA ensures to identify and mitigate all ethical constraints like gender and race.

6.1.4 Bias in Loan Approval Algorithms

AI is often used in financial institutions for credit scoring and loan approval. Studies have shown that some algorithms unintentionally discriminate against minority applicants, lowering their credit scores despite similar financial backgrounds to non-minority applicants [23].

Challenge: Here, the key issue is fairness in financial access. The EAA would help evaluate loan decisions by ranking them according to fairness, transparency, and accountability. A feedback mechanism would monitor long-term outcomes and adjust criteria if certain demographic groups are being unfairly treated, ensuring ethical decision-making.

6.2 Comparison with Existing Ethical AI Frameworks

Table 2 shows the comparison of the EAA model with other existing AI ethics frameworks and approaches.

Table 2. EAA Model Comparison With Other Existing AI Ethics Frameworks and Approaches

	Existing Method	**EAA Advantage**
IBM AI Fairness 360 Toolkit	IBM's AI Fairness 360 toolkit gives tools and measures to find and reduce bias in AI systems [24]. It helps developers check for fairness, but it does not provide a clear way to include broader ethical principles in decision-making.	The Fairness 360 toolkit only focuses on reducing bias. In contrast, the EAA model includes broader ethical principles, like fairness, "do no harm," and virtue ethics. It is useful for more complex situations. The EAA also has a real-time feedback mechanism which allows it to adjust to new situations.
Google AI Principles	Google's AI Principles give basic rules to make fair and responsible AI systems. Their focus is on fairness and accountability along with privacy [25]. But these rules are very general and sometimes do not explain exactly how to incorporate ethics in AI's decision.	Unlike Google's Principles, EAA provides detailed step-by-step algorithms. This framework evaluates and ranks decisions using ethical scores. Also, EAA includes a mechanism which ensures that decision is aligned with different ethical perspectives. And this feature help AI systems to be transparent and more responsible.
Differential Privacy	Differential Privacy is a method that helps AI systems to learn from the dataset and maintain the privacy of it [26]. It protects user's data but does not follow any ethical rules. And hence it results in unfair and sometimes harmful decisions.	The EAA model provides privacy-protection methods which incorporate moral values. So, when it is used with Differential Privacy, it checks if the decision is fair and responsible or not. This way, it is possible to get fair results without any harm.

7. CHALLENGES AND LIMITATIONS

Incorporating moral values with AI systems is not easy. It has many technical challenges. A major challenge is translating ethical guidelines to clear rules for computer understating. AI systems work with data and algorithms. Humal morals are subjective and very complex hence it is difficult to embed them with machines. There is not any universal standard for morals because it varies from situation to situation and person to person. Different

cultures and societies have different views on morals. In this condition it is extremely challenging to make one efficient and responsible ethical AI system.

Another challenge is the possibilities of ethical dilemmas and conflicts. Many times it happens that AI systems face situations where moral values clashes. For example, during the condition of accident, it is difficult for AI system to choose to save passengers or pedestrian. Because both lives are equally important. Because in this type of situation, decisions depend on various factors. How to feed the data to AI and how to give instructions through algorithms for this type of situation is quite challenging. This dilemma highlights the limitations of existing AI systems which are not capable of handling complex moral decision-making [12].

Existing technologies also have limitations in embedding ethical guidelines. Most of today's AI machines rely on ML models that are trained with diverse and huge datasets. Generally, these training datasets contain biases like stereotypes, unfair representation etc. When AI systems get trained with this type of dataset, it may reflect and use those biases in predictions. Despite making so many efforts to mitigate these biases, it remains as it is and it ruins the moral integrity of AI outcomes [17],[14],[10].

In conclusion, integrating moral values with AI systems is a main goal but it has lots of technical challenges, ethical dilemmas, and limitations in existing technologies. Addressing these challenges requires collaborations of different fields, continues monitoring and refinement of ethical frameworks [10].

8. FUTURE DIRECTIONS

As AI continuously evolving, there is a huge scope of advancement in it. Incorporating moral values into these systems is one of the main areas of research. Creating more sophisticated and responsible algorithms with ethical guidelines is the most promising sector. By using advanced ML techniques, these AI systems can predict more fair and unbiased results. Most essential part of future scope is to make AI systems which can make decisions that aligned with human moral values [2].

Emerging trends and technologies also show the possibilities of new advancement in this field. Integration of AI with different areas like psychology and neuroscience, can lead the better AI systems which can understand human emotions and moral values. This type of system can also help to predict and avoid ethical dilemmas even before they happen [5], [9], [11].

One more emerging trend is the development of Explainable AI (XAI). This technology is more transparent and helps users to understand the reasons behind the specific decision. This way it will be possible to build the user's trust in AI systems and their predictions. Decisions taken ethically by AI system along with transparency can take it to new heights of relations between user and machines [10].

Ethics is a subject specific. It varies from person to person and situation to situation. Hence, there is a scope of development of such AI systems which can adjust their results dynamically depending of the context. Incorporating moral values in AI system is not just enough but with that it is important research scope to make system that much capable to understand and apply these moral rules in various situations efficiently [17], [14]. Moreover, various experiments need to be conducted to test these ethical AI systems with real-world scenarios. This will help scientists and researchers to understand the behavior of these systems in different conditions [8].

Prime research scope is to work with different fields like computer science, neuroscience, ethics, law, sociology, psychology. Because only this collaboration can make it possible to develop efficient, transparent, and fair AI systems. Diverse perspectives help to train the machines with wide range of data which result in more accurate results which can be closely aligned with human moral values.

9. CONCLUSION

This paper has explored the importance of incorporating moral values in AI systems. It discussed how different ethical frameworks can be applied to enhance AI decision-making power. In this paper three main ethical frameworks have been discussed: utilitarianism, deontology, and virtue ethics. The paper showed the significance of embedding these ethics to AI systems to get the results close to human moral values. Because nowadays, we find AI is everywhere and, in each sector, we use AI for decisions. Hence it is important to have ethically aligned AI systems.

This paper has proposed a novel model called EAA. This model is a step-by-step description and explanation of how to incorporate moral values with AI machines. EAA provides a structured approach to incorporate ethics to

get fair, transparent results which can be aligned with human moral values. By embedding this EAA model with existing AI machines, we can prevent harm, reduce biases and promote fairness in outcomes.

The paper also discusses the consequences of unethical and biased decisions made by AI systems. How it affects human lives and society. Why it is essential to make ethical AI machines and how we can build the trust of users into AI systems.

In the end, the paper showed that developing ethical AI is not that easy. It has many technical challenges. With the collaboration of multidisciplinary, policy makers and stakeholders, it is possible to make efficient and transparent ethical AI systems. The EAA model represents the significant step towards the development of AI systems enriched with moral values for betterment of human lives and society.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for the careful review of this article.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Khushboo Shah: Led all aspects of the research, including Conceptualization, Methodology, Data Curation, Validation, Project Administration, Writing – Original Draft Preparation, and Supervision.
Hiren Joshi: Provided feedback during the review process.
Hardik Joshi: Contributed to the review and editing of the manuscript.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

The paper follows The Committee of Publication Ethics (COPE) guideline.

REFERENCES

[1] E. Brynjolfsson and K. McElheran, “The rapid adoption of data-driven decision-making,” in *American Economic Review*, May 2016, vol. 106, no. 5, pp. 133–139, doi: 10.1257/aer.p20161016.

[2] E. Brynjolfsson and A. McAfee, “The Business of Artificial Intelligence,” *https://hbr.org/*, Jul. 18, 2014.

[3] R. Gorwa, R. Binns, and C. Katzenbach, “Algorithmic content moderation: Technical and political challenges in the automation of platform governance,” *Big Data Soc.*, vol. 7, no. 1, Jan. 2020, doi: 10.1177/2053951719897945.

[4] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach Third Edition*. 2010.

[5] L. Floridi and J. Cowls, “A Unified Framework of Five Principles for AI in Society,” *Harvard Data Sci. Rev.*, Jun. 2019, doi: 10.1162/99608f92.8cd550d1.

[6] J. Chin, “Editorial : Artificial Intelligence and Cybersecurity in Pervasive Computing,” *J. Informatics Web Eng.*, vol. 3, no. 3, 2024.

[7] V. Eubanks, *Automating inequality : how high-tech tools profile, police, and punish the poor*. Picador, St. Martin’s Press, 2019.

[8] V. N. February, S. Tan, S. Chong, K. Wee, and L. Chong, "Personalized Healthcare : A Comprehensive Approach for Symptom Diagnosis and Hospital Recommendations Using AI and Location Services," *J. Informatics Web Eng.*, vol. 3, no. 1, 2024.

[9] L. Floridi and J. W. Sanders, "On the Morality of Artificial Agents," 2004.

[10] V. Dignum, *Responsible Artificial Intelligence: How to Develop and Use Ai in a Responsible Way*. Springer Verlag, 2019.

[11] J. Buolamwini, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *," 2018.

[12] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Feb. 2017, [Online]. Available: http://arxiv.org/abs/1702.08608.

[13] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Sci. Eng. Ethics*, vol. 26, no. 4, pp. 2141–2168, Aug. 2020, doi: 10.1007/s11948-019-00165-5.

[14] N. J. Goodall, "Machine_Ethics_and_Automated_Vehicles," pp. 93–102, 2014, doi: 10.1007/978.

[15] W. Burgard, "Artificial Intelligence," in *The Cambridge Handbook of Responsible Artificial Intelligence*, Cambridge University Press, 2022, pp. 11–18.

[16] S. Stos, "Utilitarianism, Deontology and Virtue Ethics: Teaching Ethical Philosophy by Means of a Case Study," *NeilsonJournals Publ.*, 2018, doi: https://doi.org/10.4135/9781529734836.

[17] R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," 2018.

[18] P. A. Rizk et al., "Machine Learning–Assisted decision making in Orthopaedic Oncology," *JBJS Reviews*, vol. 12, no. 7, Jul. 2024, doi: 10.2106/jbjs.rvw.24.00057.

[19] S. Amershi *et al.*, "Software Engineering for Machine Learning: A Case Study," in *Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2019*, May 2019, pp. 291–300, doi: 10.1109/ICSE-SEIP.2019.00042.

[20] IBM, "5725-W51 IBM Watson for Oncology Product Life Cycle Dates," 2023. [Online]. Available: https://www.ibm.com/docs/en/announcements/watson-oncology?region=CAN.

[21] S. M. Ziad Obermeyer, Brian Powers, Christine Vogeli, "Dissecting racial bias in an algorithm used to manage the health of populations," 2019. https://www.science.org/doi/epdf/10.1126/science.aax2342 (accessed Sep. 22, 2024).

[22] J. Dastin, "Insight - Amazon scraps secret AI recruiting tool that showed bias against women," *www.reuters.com*, 2018. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/ (accessed Sep. 22, 2024).

[23] A. Cristina, B. Garcia, M. Gomes, P. Garcia, and R. Rigobon, *Algorithmic discrimination in the credit domain : what do we know about it ?*, vol. 39, no. 4. Springer London, 2024.

[24] R. K. E. Bellamy, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *ieeexplore*, 2019, doi: 10.1147/JRD.2019.2942287.

[25] Google, "Our Principles." https://ai.google/responsibility/principles/ (accessed Sep. 22, 2024).

[26] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential Privacy and Machine Learning : a Survey and Review," *arxiv*, pp. 1–30, 2014.

BIOGRAPHIES OF AUTHORS

	Dr. Khushboo Shah is an Assistant Professor at St. Xavier's College (Autonomous), Ahmedabad, Gujarat, India, serving since 2017, with over 15 years of teaching experience at various institutes. She has authored 4 research papers and 2 books. Her academic contributions and professional profile are accessible via her ORCID profile and the college website, though the latter may not reflect her latest accomplishments. Research Areas: NLP, ML, AI E-mail : khushboo.shah@sxca.edu.in
	Dr. Hiren Joshi is the Head of the Department of Computer Science at Gujarat University, Ahmedabad, Gujarat, India, serving as a professor since 2005. He has published 38 research papers, 2 books, and 1 book chapter. A life member of CSI and ISTE and a professional ACM member, Dr. Joshi is a prominent academician. His research contributions are available via Scopus (ID: 57189001503) and ORCID (0000-0003-0878-0753). Learn more on his profile. Research Areas: NLP, ML, AI Email : hdjoshi@gujaratuniversity.ac.in
	Dr. Hardik Joshi is an Associate Professor in the Department of Computer Science at Gujarat University, Ahmedabad, Gujarat, India with a teaching career spanning since 2003. He has authored 8 research papers, 2 books, and 2 book chapters. Dr. Joshi has also served as a jury member and chairperson for various events and competitions. His research contributions are accessible via Scopus (ID: 57210543228) and ORCID (0000-0002-5431-0173). Learn more on his profile. Research Areas: NLP, ML, AI E-mail : hardikjoshi@gujaratuniversity.ac.in