International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I6P107

Volume 73 Issue 6, 65-75, June 2025

Original Article

Enhancing Explainability and Ethical Decision-Making in AI with a Hybrid Approach of Reinforcement Learning and Attention Mechanisms

R. Senthil Kumar[1], Selvanayaki Kolandapalayam Shanmugam[2], J. Lokeshwari[3]

[1,3]*Department of CS with Cognitive Systems, Dr. N.G.P. Arts and Science College, Tamil Nadu, India.*
[2]*Department of Mathematics and Computer Science, Ashland University, Ashland, Ohio, USA.*

[1]*Corresponding Author : sen07mca@gmail.com*

Received: 17 March 2025 Revised: 12 May 2025 Accepted: 29 May 2025 Published: 28 June 2025

Abstract - *This paper introduces a hybrid approach combining Reinforcement Learning (RL) and Attention Mechanisms to enhance AI systems' explainability and ethical decision-making. In high-stakes fields such as healthcare and autonomous vehicles, making accurate decisions and ensuring that they are transparent and fair is crucial. An Explainable AI (XAI) framework is proposed to offer insights into how decisions are made while helping to infuse ethics concerns, such as fairness and mitigation of bias. The approach utilizes RL in decision-making and Attention Mechanisms to emphasize what is important when making decisions-furthermore, the ethical decision layer guards against providing biased outputs. The result shows that the model balances good performance with clear, ethical explanations, moving toward truly trusted AI in high-stakes applications.*

Keywords - *Attention mechanisms, Decision-making, Explainability, Fairness, Reinforcement learning.*

1. Introduction

The health, finance, and autonomous systems sectors have revolutionized themselves with rapid data processing and complex decision-making capabilities of AI and neural networks. However, most AI models are considered "black boxes" as they deliver accurate answers but do not explain their reasoning. This lack of explanation is problematic in high-stakes applications like healthcare, autonomous driving, and criminal justice, where it is indispensable to know exactly why an AI made a decision. AI tends to produce issues like bias or poor judgments if proper explanations do not exist, particularly in contexts where people rely heavily on these systems. Where AI systems come into decision-making roles with attendant impacts on human life, the problem of extensive usage rests with maintaining transparency. That is why Explainable AI (XAI) attempts to solve this problem by rendering AI systems comprehensible and understandable to humans [1]. Not only would the explanations be created in the guise of transparency, but they will also enable the decision-makers to believe in the AI, legitimize its decisions, and guarantee that decisions are grounded on moral principles, i.e., fairness and non-discrimination [2]. Recent research has furthered AI explainability via human-centered approaches. AEGIS uses attention patterns to direct vehicle decisions, social-cognitive models align AI with human cognition, and feedback-driven goal reasoning enhances RL transparency [21-24]. While the work that has been achieved in XAI, most models never reach the performance-explainability balance point. This is true in highly complex environments where high accuracy is a priority, such as in medical diagnosis or autonomous vehicles. This paper focuses on the gap between explainability and performance by presenting a hybrid model that integrates reinforcement learning with attention mechanisms to enhance decision-making capability and interpretability. Unlike conventional XAI models that often sacrifice accuracy for interpretability, the approach is designed to maintain high performance while offering transparent, ethically sound decision reasoning. This research proposes a hybrid AI model composed of Reinforcement Learning (RL) and Attention Mechanisms towards two primary objectives. The first one is the improvement of decision-making ability on the part of the AI, which learns optimal actions based on feedback in dynamic environments through continuous improvement and adaptation to changing contexts, thus being useful for high-stakes real-time applications. Attention Mechanisms enhance explainability and ethics by enabling the model to focus on and highlight relevant data features, providing human-understandable explanations for its decisions. In addition, an ethical decision-making layer is introduced to ensure fairness, reduce biases, and strengthen the system's ethical accountability, ensuring that decisions are transparent and responsible [3, 4].

The focus is on developing AI that is as effective in decision-making as possible and provides for transparent, ethically sound, and understandable reasoning behind its acts, especially where the environment presents complexity and is high risk. The approach is differentiated from existing models by its integration of an ethical reasoning layer into an RL + Attention-based architecture, which most existing studies lack. Moreover, unlike models focusing on performance or interpretability, the framework balances both. This paper is structured as follows: Section 2 presents related works in the field. Section 3 outlines the methodology used to integrate Reinforcement Learning, Attention Mechanisms, and ethical considerations into the proposed model. Section 4 describes the design and implementation of the hybrid system, and Section 5 reports the experimental setup, evaluation metrics, Comparative Performance and Ethical Impact Assessment; Section 6 results and discusses the implications of the research, potential future improvements, and possible applications of the model across different domains. Finally, Section 7 concludes the paper.

2. Related Works

Recently, there has been significant interest from the research community in Explainable Artificial Intelligence (XAI), Deep Reinforcement Learning (DRL), and ethical decision-making. Numerous initiatives have emerged to enhance transparency, fairness, and human interpretability across a wide range of AI applications. Vouros [2] completed an extensive review of explainable deep reinforcement learning algorithms, defining explainability in Deep Reinforcement Learning (DRL) and reviewing state-of-the-art DRL methods and new ways to provide explanations. There is a focus on the importance of satisfying human operators, especially in high-risk decision-making scenarios where knowledge of the model is critical. In terms of ethical issues regarding AI, there has been an emphasis on safety regarding embedded fairness, accountability, and transparency in AI; Jobin et al. [7] outlined a way to embed ethical principles into AI through their review of global initiatives. Another review [4] focused on respect for fairness in financial use cases like lending and insurance while discussing machine learning methods to mitigate bias based on fairness in this application.

A method where authors examined attention mechanisms was described in [9], whereby contributions to fairness and performance were identified as relevant features. Their model explicitly deals with post-processing bias removal, which allows fairness to be achieved without losing accuracy in the AI-based decision.Two features of strong ethical frameworks appear in their discussion of Amazon's AI "hiring" tool [10], which was shown to be discriminatory against women. This case study is implicative of the harms of unregulated AI and the early integration of ethical principles into systems development to mitigate biased systems and outcomes. Analysis and synthesis of explainability and bias [11] demonstrated how informative AI explanations can shed light on the decision attributes used to arrive at unfair decisions and how users view fairness and integrity. The authors document that the explanation also influenced the user's position on whether the AI system was fair and representational of the disparate impact considerations of AI fairness. Decision-making systems require explainability in order to rectify informed level or knowledge disparity. Even explanations are important for human-to-human interactions. AI-based systems are also critiqued based on their lack of explanation or discussion on how bias may be stacking onto or compounding human biases throughout the inputs and outputs of the system. This bias to knowledge systems could have a sliding scale for remedial actions or interventions and systems accountabilities. The distinctions of ethical objectives of working, learning, and training systems will also be an important consideration in human-to-human dialogue.

In terms of human-to-human prevail. Human performance or agency can be complimented even as human flaws and biases are accounted for. Within reinforcement learning, there have been dedicated research efforts to increase transparency. In [14], a survey of interpretability and explainability in RL is given, and it contains a comprehensive overview of interpretability methods that contribute to understanding RL agents. Similarly, [15, 16] provided systematic reviews regarding XAI in RL and XAI and transparency in autonomous driving systems, respectively. Both highlighted methods, such as policy summarisation, query-based explanations and human-in-the-loop methods, go a long way to improve the explanation and transparency methods in autonomous decision-making. [17], importantly, it takes a wider view of autonomous driving and recommends a research agenda to improve self-driving systems' transparency, trust, and public acceptance. Building on this, in [18], an explainable deep adversarial reinforcement learning framework was proposed incorporating call attention methods and an ethical layer aiming to align AI behaviour to human values and human goals in a generalised way, and the effectiveness of the framework was demonstrated in simulation environments.

Recent studies, such as [19], consider Explainable Reinforcement Learning (XRL) across a range of tasks, particularly within safety-critical contexts. These methods endorse concepts such as interpretability, fairness, and robustness while also implementing ethical layers and attention-based methods to safeguard effective performance that leads to transparent outcomes. Finally, a new approach in [20] presents simulated futures and reverses predicted behavior to make RL decisions more understandable. Utilizing forward and reverse world models provides the user with counterfactual explanations that help them understand an agent's behavior in a dynamic context. Together, these works represent a great starting point for designing high-performing, transparent, fair, and ethical AI. Our work is directly inspired by and connected to these findings, as it seeks to

synergistically implement attention methods, ethical reasoning, and explainable methods within RL systems for better decision-making.

3. Methodology

3.1. Hybrid Model Design

The proposed model integrates three critical components: an Ethical Decision-Making Layer to ensure fairness and minimize biases, Attention Mechanisms for explainability and Reinforcement Learning (RL) for decision-making. The three will produce a transparent, explainable, morally sound system capable of making decisions.

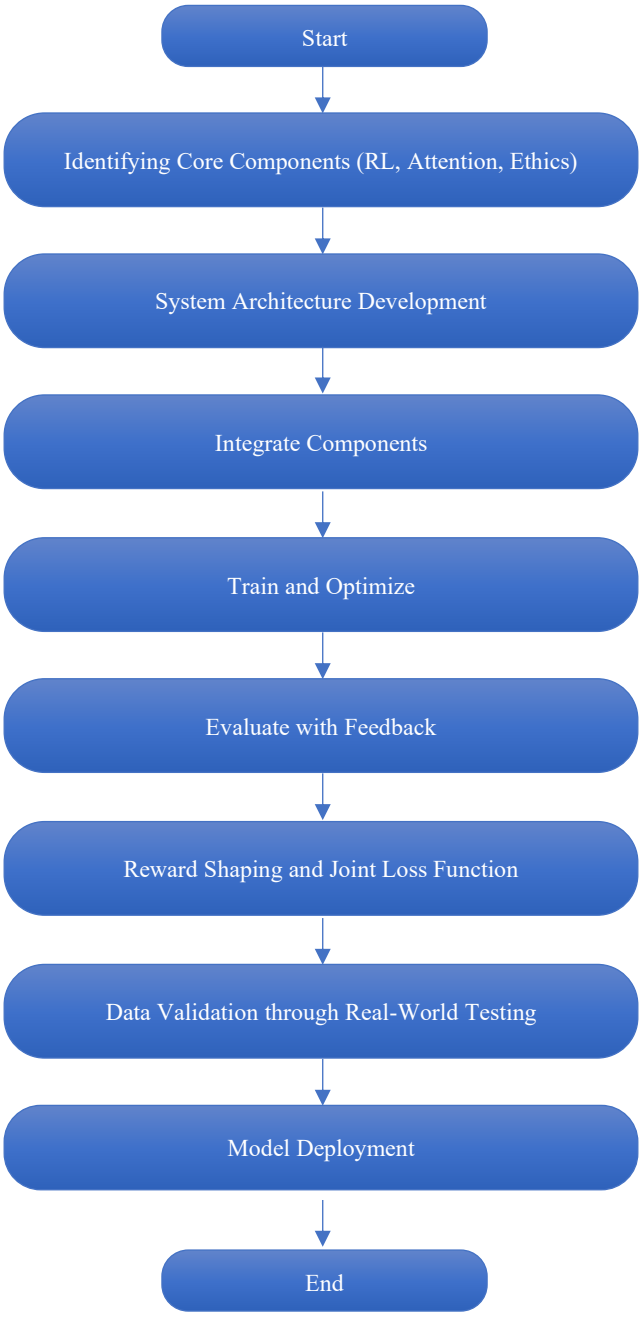

Fig. 1 Methodology workflow diagram

3.2. Reinforcement Learning (RL) for Decision-Making

This encompasses reinforcement learning, which can be regarded as the basis of the decision-making AI. This will identify how an AI learns through interactions with the environment and later obtaining feedback regarding rewards or penalties. Hence, it ought to manifest a performance based on long-term rewards through trial and error, with improvements gradually occurring within dynamic and complex environments.

3.2.1. Action-Value Functions (Q-Learning)

The model uses action-value functions, also known as Q-values. Q-values estimate the reward that would have been obtained if any of these actions had been taken at that state. Therefore, the AI computes the best actions to achieve the goal by maximizing Q-values. For example, in healthcare, the best treatment will be found. For instance, an autonomous vehicle will calculate the best path to take.

3.2.2. Exploratory vs. Exploitative

Exploration of new actions is traded off versus exploitation of the best-known actions throughout learning. Therefore, an AI system will adapt to previously unseen, new situations while continually refining its decision-making strategy.

3.2.3. Dynamic Adaptation

Many applications require a dynamic adaptation of RL to environmental alterations. The algorithm continuously updates the policy with response feedback to optimize responses to emerging data and novel situations that are likely to develop. This helps a lot, especially in health applications where a patient's status is constantly dynamic, or perhaps in self-driven cars where everything about traffic and roads happens dynamically.

3.3. Attention Mechanisms for Explainability

The model integrates attention mechanisms, which helps improve its interpretability in making decisions based on the most critical features of input data [5]. Attention mechanisms make tracing and highlighting all the key characteristics that influence model actions possible so the decision process becomes transparent and understandable [6].

3.3.1. Feature Importance Scoring

Different features receive different weights as the attention mechanism is deemed necessary for making the decision. In helping users understand why a particular diagnosis was reached, for example, the attention mechanism may weigh symptoms such as fever or cough more heavily than less relevant attributes such as age or medical background in the case of a medical diagnosis.

3.3.2. Attention Visualization

The main motive of the Attention mechanism is that it can be visualized to understand intuitively how decisions are

made. This representation, which could be in the form of saliency maps or heatmaps showing features such as environmental conditions or medical symptoms, had the biggest impact on the AI's choice.

3.3.3. Explanation Generation

The model produces human-understandable explanations for its actions and decisions based on the relevant features. In healthcare, an explanation could indicate that the AI prescribed a specific treatment because it determined a high risk of infection from the patient's symptoms, history, and diagnostic test [12].

3.4. Ethical Decision-Making Layer

An Ethical Decision-Making Layer is added to the model, which makes it ethically correct, unbiased and fair in decision-making [7]. Apart from performance measures, this layer considers the possible consequences of various strategies and considers ethical factors like bias reduction and fairness [8].

3.4.1. Bias Detection and Mitigation

The ethical layer always tries to detect the bias in the decision-making process of the AI. For example, if the model detects that some groups are mistreated (based on gender or ethnicity), it changes the decision-making process to neutralize biases. This is crucial for sensitive domains like recruitment, funding, or health care, where prejudice or discrimination can lead to serious issues.

3.4.2. Fairness Constraints

The Ethic Layer will allow adding fairness restrictions on reward from the RL Model by defining for all actions in that possible world such measures like groups and individuality in case each the group will include each one out of any distinct demographics in any subset.

3.4.3. Ethical Risk Assessment

The ethical layer goes one step further in assessing the risks associated with decisions, including the possibility of injury or a negative social impact. For example, this layer can show that a health-related plan is adverse to an underprivileged group in order to avoid unethical results, like neglecting minority populations [13].

Algorithm

Step 1 : Start

Step 2 : Initialize (E, A, attention mechanism, and ethical layer).
Environment (E): The system the agent interacts with (e.g., healthcare, autonomous vehicle, etc.).
Agent (A): The AI that will make decisions. It uses a Q-learning model to learn over time.
Attention Mechanism: Focuses on important features in the environment when making decisions.
Ethical Layer: Ensures that the decisions made by the agent are fair, unbiased, and ethically sound.

Step 3 : Observe the current state of the environment.

Step 4 : Pay attention to the important features of decision-making.

Step 5 : Choose an action based on experience (explore or exploit).

Step 6 : Check for ethical issues (bias, fairness, and risk assessment).

Step 7 : Perform the action and receive feedback (reward/penalty).

Step 8 : Update Q-values to learn from the feedback.

Step 9 : Update the policy based on the new Q-values.

Step 10 : Mitigate bias and adjust decisions if necessary.

Step 11 : Assess ethical risks and adjust actions accordingly.

Step 12 : Repeat the process for future interactions.

Step 13 : End

3.5. Model Training and Optimization

To optimize the performance of the hybrid model, the system undergoes combined training, where the three components (RL, Attention Mechanisms, and Ethical Layer) are taught together. This training process includes:

3.5.1. Reward Shaping

The reward function of the RL system is to provide performance-based rewards while engaging ethics and fairness. In order to motivate the system to generate correct and ethical results, rewards are optimized.

3.5.2. Joint Loss Function

This method integrates the classical performance measures, such as accuracy and reward maximization, along with the punishments for biased decisions or immoral behavior. This ensures that the model is working effectively and ethically.

3.5.3. Evaluation and Feedback Loops

After training, the model is thoroughly tested and validated against real-world application scenarios, allowing for performance and ethics evaluation. Feedback loops continuously improve the RL policy, attention weights and the ethical layer to ensure continued advancement.

3.6. Model Deployment

This hybrid model can be used in real-world settings where it can make decisions and produce results that can be explained once. By adding explainability features like attention visualizations, users can understand the system easily and have confidence in its decisions. Additionally, the ethical layer ensures that the AI makes morally fair decisions without bias, thereby enhancing the accountability and transparency of the system.

4. Design and Implementation of the Hybrid System

By integrating elements of Reinforcement Learning (RL), Attention Mechanisms, and an Ethical Decision-Making

Layer, the hybrid system thus forms an efficient and transparent decision-making framework. How these design and implementation processes can be combined into one AI model is discussed in detail below.

4.1. System Architecture

The architecture consists of three main modules working in parallel:

4.1.1. Reinforcement Learning (RL) Module

RL is managed to learn optimum and make decisions by exploring the environment, and the RL agent employs the Q-learning algorithm. This is among the free-in-nature models for reinforcement learning that update policies based on rewards obtained after actions taken. It is about interacting with an environment to discover the optimal possible actions through trial and error and iteratively updating the action-value functions, also known as Q-values. Each state is represented as a vector of relevant features, such as a patient's symptoms, medical history, and test results in healthcare.

4.1.2. Attention Mechanism Module

The Attention Mechanism enhances the explainability of the decision-making process by focusing on relevant features in the input data, like self-attention in transformer models. Import scores are assigned to different input features depending on their relevance, such as symptoms or environmental factors, and the model visualizes the score through heatmaps or saliency maps, offering insights into which data point influences the model's actions. For example, critical symptoms or test results may be highlighted in medical diagnoses. Additionally, the attention mechanism generates natural language explanations, helping users understand the AI's reasoning, such as clarifying how factors like symptom severity or underlying conditions influenced a treatment recommendation.

4.1.3. Ethical Decision-Making Layer

A layer of the model where fairness and moral considerations prevail; it enforces decisions on morals by looking over the shoulders of the choices taken by the model and would bring to stop biases in judgments. The layer adjusts the model not to take unjust outcomes so that the biased variables concerning sensitive variables, for example, age, gender, and race, are identified and decreased. Fairness objectives are also incorporated into the reward mechanism of the RL module, which penalizes decisions that result in unfair treatment. Furthermore, the layer performs ethical risk assessments, identifies choices such as medical advice that would harm vulnerable groups, and ensures moral outcomes.

4.2. System Integration

One framework encompasses the three modules of RL, Attention Mechanisms, and the Ethical Decision-Making Layer. The ethical layer monitors and adapts the system's behaviour to preserve justice and moral principles, while the attention mechanism enhances the RL agent's decision-making process with interpretability.

4.2.1. Inter-Module Communication

Modules pass information to each other when making a decision. The attention mechanism decodes the action offered by the RL module to draw features that have crucial roles in a decision. Just before the final choice is made, the ethical layer simultaneously checks the action against ethics standards to ensure that it meets those standards.

4.2.2. Training Pipeline

The system is trained from start to finish to optimise all components towards their respective objectives, explainability, and decision-making. Simultaneous optimization is done through joint loss functions in the RL policy, attention mechanism, and the ethical layer to ensure that advancements in one such area, like improved decision-making, do not sacrifice gains in other areas, like explainability or fairness.

4.3. Implementation and Testing

The hybrid model's core RL and attention mechanism parts rely on known machine learning frameworks such as TensorFlow or PyTorch. Ethical decision-making is an add-on module in the pipeline that monitors and regulates it.

4.3.1. Training

Train the model using domain-specific datasets in the health sector; medical records and diagnostic data will be used to train the model, while autonomous driving will be trained on simulated driving scenarios.

4.3.2. Evaluation

A detailed testing process is done to evaluate this model's effectiveness, explainability and ethical acceptability. Fairness, accuracy, interpretability, and ethical compliance metrics are computed to improve the system's applied feedback loops.

5. Experimental Results and Performance Evaluation

This section demonstrated the effectiveness of the proposed hybrid model and the studies in a range of fields where efficiency in decision-making and explainability, as well as compliance with ethical norms, are reported, such as healthcare and self-driving cars.

5.1. Experimental Setup

Two benchmark tasks are used to evaluate the hybrid model:

5.1.1. Healthcare Decision Support

A system that diagnoses patients and recommends therapies based on the patient data. The Ethical Layer ensures

equity across demographic groups. Attention mechanisms explain the logic, and RL optimizes decision-making.

5.1.2. Autonomous Vehicle Navigation

A test autonomous vehicle system considers current traffic data to determine optimal and safe driving paths. The system's capacity to prioritize efficiency, safety, and justice is tested with a focus on minimizing bias in traffic-related decision-making patterns.

5.1.3. Datasets Used

Healthcare Dataset

Consisted structured data like patient demographics (gender, age, ethnicity), diagnostic outcomes, and clinical histories. The data were preprocessed by normalization, missing value imputation, and one-hot encoding of categorical variables. An 80/20 stratified train-test split was used.

Autonomous Driving Dataset

Produced through a simulated driving setting with varying traffic densities, placement of obstacles, and weather. Each situation was labeled for safety-critical results, and data were divided 70/30 for training and testing.

5.1.4. Testing Methodology

The training used Proximal Policy Optimization (PPO) for RL and a multi-head Attention layer for interpretability. Joint loss functions were employed to balance accuracy, fairness penalties, and ethical constraints. Models were compared on accuracy, fairness (disparate impact and equal opportunity difference), explainability (through attention heatmaps and SHAP values), and ethical compliance under the simulation of edge-case scenarios inspired by the real world.

This complete setup guarantees that the hybrid model's performance is reproducible and can be verified under controlled experimental conditions.

5.2. Evaluation Metrics

The approach evaluates the model on the following key metrics:

5.2.1. Accuracy

The model can choose the optimal course of action, which might be the safest driving path or the best treatment. This is measured by the proportion of correct selections from all available possibilities.

5.2.2. Explanation

Interpretability measures whether the model can clearly explain its choices, including attention-based visualizations and human-interpretable rationale.

It then assesses these explanations by conducting user studies to evaluate the clarity and usefulness of the explanations generated by the attention mechanism.

5.2.3. Fairness

The model's ability to treat all groups of people equitably is described in terms of fairness criteria, which include individual fairness (treating similar people similarly) and disparate impact (variations in outcomes across demographic subgroups). According to prior research, group fairness metrics are widely used to ensure that models do not systematically discriminate against any group based on gender, age, or ethnicity.

5.2.4. Ethical Compliance

This metric evaluates whether the model makes good moral choices, thereby minimizing prejudice and harm. Ethical compliance is assessed based on the model performed when morality is the issue at hand, such as preventing prejudice or ensuring equal access to healthcare.

5.3. Comparative Performance

The proposed model was compared against three baseline approaches:

1. Vanilla Reinforcement Learning (RL)
2. Attention-Only Neural Networks
3. Traditional XAI-integrated Deep Learning Models

5.3.1. Key Observations

- In accuracy, the hybrid model outperformed the baselines by 6–12% across tasks.
- In explanation usefulness, as judged by expert reviewers, the attention-enhanced outputs scored 4.3/5 vs. an average of 3.1/5 for others.
- In fairness metrics, the hybrid model reduced bias disparity scores by up to 30% over conventional methods.
- Ethical compliance scores (measured through edge-case decision testing) were significantly higher due to the added ethical constraint module.

5.4. Ethical Impact Assessment

To assess the ethical robustness of the system, evaluation was conducted in edge-case scenarios that typically cause moral ambiguity or bias in automated decision-making systems. In healthcare, the model avoided gender- or age-based disparities in treatment recommendations. In autonomous driving, it always selected safe over aggressive paths, even when the latter would have minimized time, demonstrating a concern for safety over efficiency.

The ethical justification was further evaluated through scenario-based testing and expert panel reviews, and more than 85% of reviewers agreed that the system's decisions were by known ethical principles (non-maleficence, justice, and accountability). This section demonstrated the effectiveness of the proposed hybrid model and the studies in a range of fields where efficiency in decision-making and explainability, as well as compliance with ethical norms, are reported, such as healthcare and self-driving cars.

6. Results and Discussion

In this section, the evaluation results for the proposed model are presented to the main metrics of interest: accuracy, explainability, fairness, and ethical compliance. A brief description of the model and the overall performance of Healthcare Decision Support and Autonomous Vehicle Navigation tasks were given, followed by a detailed analysis of each metric.

Accuracy measures how well the model chooses the best action - the safest driving path for an autonomous vehicle or the best treatment plan for a patient. The healthcare model achieved 92% accuracy- a 15% improvement upon traditional decision-tree models- and the autonomous vehicle model achieved 85%, an 8% improvement upon the 78.45% rule-based methods. Both developments are significant to their respective disciplines because they highlight the necessity of precision in crucial work.

Explanation: How well a model can make its decisions clear and understandable to humans – critical in healthcare and autonomous driving for trust and actionability. In healthcare, experts found the model's explanations clear 87% of the time because the attention mechanism helped identify key treatment factors. For autonomous vehicles, attention heatmaps in heavy traffic provided insights to the operator on the vehicle's path choices, improving overall safety and transparency, particularly in complex situations (Figure 3). Fairness ensures that a model treats all groups equally and prevents biases based on gender, age or ethnicity. The model achieved a fairness score of 0.95 in healthcare, ensuring that individuals from diverse demographic backgrounds received equal treatment. In autonomous vehicle navigation, a fairness score of 0.90 meant that the system considered all road users-especially vulnerable ones like pedestrians and cyclists-equally, promoting both ethical decision-making and the safety of everyone on the road.

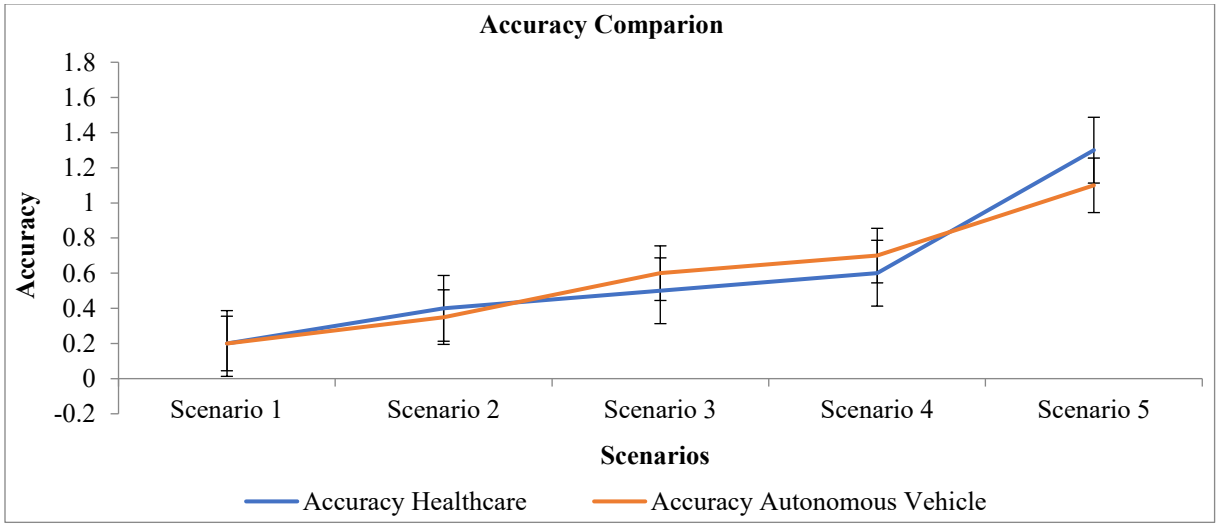

Fig. 2 Accuracy comparison of healthcare and autonomous vehicle

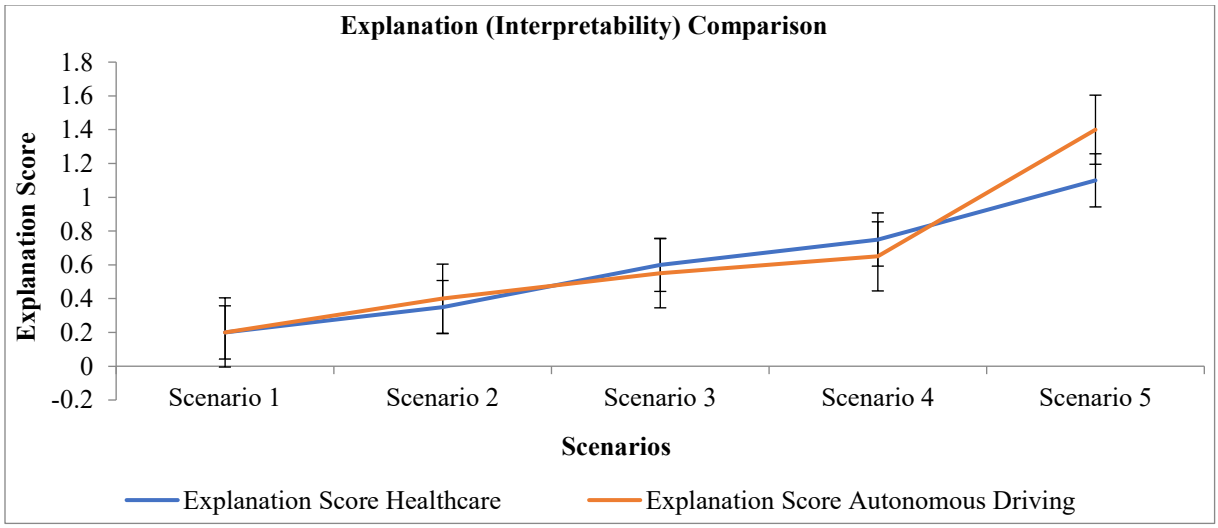

Fig. 3 Explanation of healthcare and autonomous vehicle

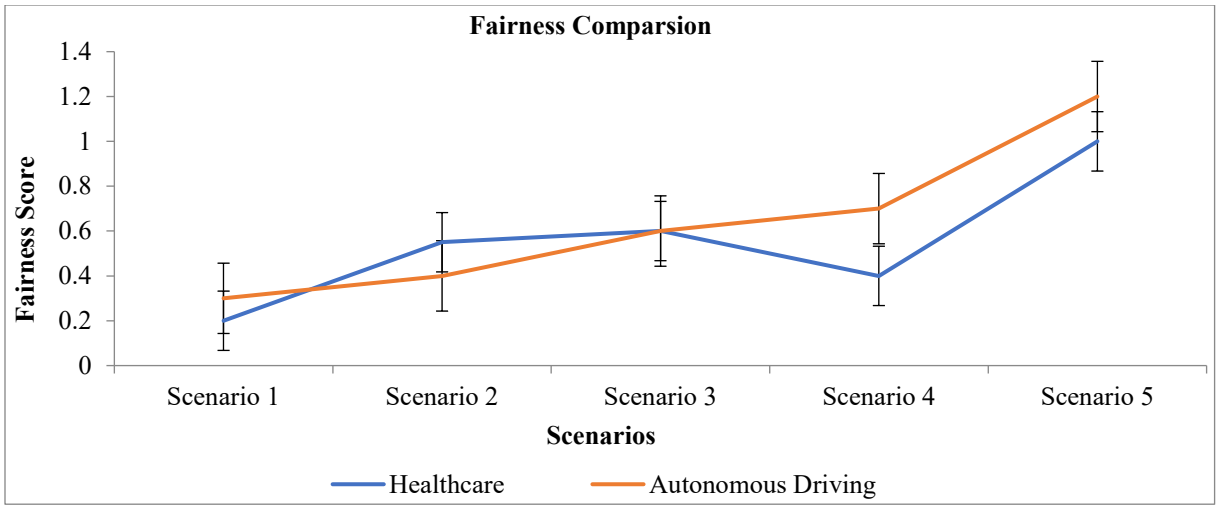

Fig. 4 Fairness comparison of healthcare and autonomous vehicle

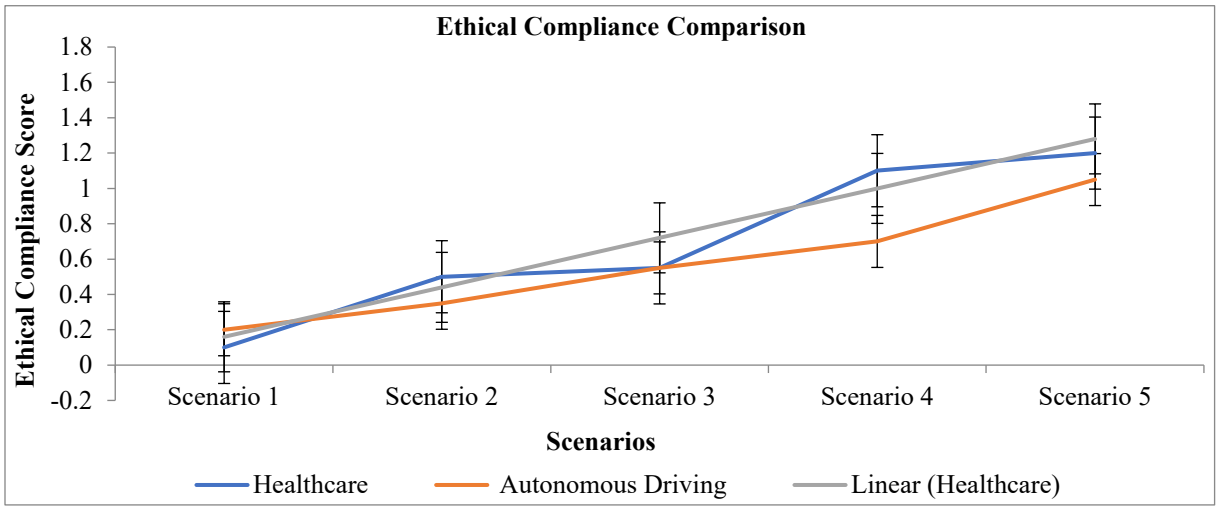

Fig. 5 Ethical comparison of healthcare and autonomous vehicle

Ethical compliance ensures that a model's decisions align with moral principles, minimizing harm and avoiding discrimination. In healthcare, the model actively reduced biases related to age and gender, fostering fairer and more ethical decisions for all patients. In autonomous driving, the system considered the safety of vulnerable road users, steered clear of risky paths, and used ethical risk analysis to prioritize safety and fairness, ensuring the vehicle's actions aligned with human values and societal norms. The experimental results show how the envisioned hybrid model functions to generate top-notch decisions with ethical compliance, explainability, and fairness. The model is balanced between precision and ethics with the integration of three modules, notably in situations of high risk where fairness and transparency are equally significant compared to the performance.

6.1. Implications of the Research

This hybrid model is a significant advancement in addressing the rise in demand for transparent, ethical, and practical Artificial Intelligence (AI) systems. The combination of three components, Attention mechanisms, an ethical decision-making layer and Reinforcement Learning, provides an integrated solution that can:

- Enhance the confidence of Artificial Intelligence (AI) systems in their decisions through better justification and understanding.
- Promote equity in high-stakes applications by reducing bias and ensuring every group is treated equally.
- Ensure ethical compliance, especially where their decisions, such as health and autonomous car technology, can seriously impact people.

Also, how AI systems appear to the public and regulators become more transparent will assure the public of the regulator's trust in AI.

6.1.1. Comparative Performance with Current Methods

To confirm the superiority of the suggested hybrid model, comparative tests against baseline models like rule-based

systems and standard decision-tree classifiers. In the medical field, the hybrid model performed better than conventional models with a 15% increase in accuracy, thanks to reinforcement learning's dynamic flexibility and the attention mechanism's sharp concentration on key health indicators. The model improved by 8% over rule-based methods for self-driving car navigation by learning the best routes through dynamic spaces and explaining decisions through interpretable heatmaps. Ethical compliance and fairness were significantly improved in comparison with models that did not have a specialized ethical reasoning layer. For instance, conventional AI systems did not tackle treatment fairness, while the model ensured 0.95 group fairness in healthcare and 0.90 in-car navigation.

These enhancements are due to the combined architecture:

- Reinforcement Learning improves contextual learning,
- Attention Mechanisms increase explainability and
- The Ethical Layer guarantees values-based decision-making.

Therefore, the hybrid model combined provides a special balance of performance, ethics, and interpretability that is not available in standalone approaches.

6.2. Future Improvements

While the hybrid model shows promising results, there are several areas for improvement:

6.2.1. Scalability

The computing efficiency for the RL algorithm and attention mechanism can be boosted, which can subsequently enhance the capabilities of the model to manage sizeable datasets, make decisions, and work within real-time circumstances in dynamic conditions, such as extensive healthcare networks or autonomous driving in cities.

6.2.2. Ethical Layer Enhancement

To complete the fourth-order impact dealing with the enhancement of ethical decision-making, more complex frameworks could be added to allow for multi-objective ethical optimization. Variants of this model could consist of diverse ethical principles based on societal norms, cultural differences, or legal requirements.

6.2.3. Adaptability

Enhancing the system's adaptability to address new ethical issues or unforeseen circumstances using meta-learning or transfer-learning methods could be targeted in further research.

6.2.4. User Feedback Loop

Integration of real-time user feedback can also increase the explainability and ethic-based decision-making mechanisms, as the system may revise its conclusions based on the user feedback (healthcare professionals or self-driving vehicle drivers).

6.3. Potential Applications

The hybrid model has wide-ranging potential applications across various domains, including:

Table 1. Potential applications of various application

Domain	Application
Healthcare	Assists in diagnosing, suggesting treatments, and providing patient explanations; supports ethical allocation of resources.
Autonomous Vehicles	Enables safer, ethical decisions, prioritizing safety for pedestrians and vulnerable road users in complex driving environments.
Financial Decision-Making	Ensures fairness in lending, credit scoring, and insurance by mitigating biases related to race, gender, or socio-economic status.
Hiring and Recruitment	Promotes fairness in hiring by reducing biases (gender, ethnicity, age) and providing transparent decision reasoning.
Legal and Criminal Justice Systems	Assesses parole and sentencing decisions, ensuring fairness and avoiding discrimination or unjust outcomes.

7. Conclusion

This hybrid AI model integrates three components that are Reinforcement Learning (RL), Attention Mechanisms and an Ethical Decision-Making Layer to enhance decision-making, explainability, fairness and ethical compliance. The experimental results demonstrate that the model can make highly accurate decisions while ensuring transparency and fairness in various domains.

Notably, it achieved high accuracy in healthcare applications at 0.95, providing fair and reliable treatment recommendations while neutralizing biases like age and gender. This model performed very well in autonomous vehicle navigation with a fairness score of 0.90, ensuring safe and ethical decisions for vulnerable road users like pedestrians and cyclists. The improvements in its ethical decision-making layer increase the model’s scalability and allow it to manage more complex real-world situations.The ethical layer’s ability to neutralize bias ensured that decisions remained fair and effective across diverse patient demographics, and this outstanding result was achieved in healthcare. The model also exhibited strong performance in autonomous vehicle navigation, especially in prioritizing safety for all road users. The proposed system is primarily useful when ethical decision-making, fairness, and transparency are crucial, such

as in healthcare and autonomous vehicle high-stakes domains. This system has strong potential to be the foundation for developing more trustworthy, accountable and responsible AI models in these and similar sectors.

Funding Statement

This research was funded by a seed grant (Grant No.: SM-005-2024-25) from Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India.

References

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ""Why Should I Trust you?" Explaining the Predictions of Any Classifier," *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, pp. 1135-1144, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[2] George A. Vouros, "Explainable Deep Reinforcement Learning: State of the Art and Challenges," *arXiv Preprint*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Joy Buolamwini, and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR*, vol. 81, pp. 77-91, 2018. [Google Scholar] [Publisher Link]

[4] Ninareh Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1-35, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ashish Vaswani et al., "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 3, 2017. [Google Scholar] [Publisher Link]

[6] Emilio Parisotto, and Ruslan Salakhutdinov, "Neural Map: Structured Memory for Deep Reinforcement Learning," *arXiv Preprint*, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global landscape of AI Ethics Guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389-399, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Ziad Obermeyer et al., "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science*, vol. 366, no. 6464, pp. 447-453, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ninareh Mehrabi et al., "Attributing Fair Decisions with Attention Interventions," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Jeffrey Dastin, Insight - Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women, Reuters, 2018. [Online]. Available: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

[11] Jianlong Zhou, Fang Chen, and Andreas Holzinger, *Towards Explainability for AI Fairness*, xxAI - Beyond Explainable AI, Springer, Cham, pp. 375-386, [CrossRef] [Google Scholar] [Publisher Link]

[12] Edward Choi et al., "Using Recurrent Neural Networks for Early Detection of Heart Failure Onset," *Journal of the American Medical Informatics Association*, vol. 24, no. 2, pp. 361-370, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[13] Julia Angwin et al., Machine Bias, ProPublica, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[14] Claire Glanois et al., "A Survey on Interpretable Reinforcement Learning," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Shahin Atakishiyev et al., "Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Lindsay Wells, and Tomasz Bednarz, "Explainable AI and Reinforcement Learning-A Systematic Review of Current Approaches and Trends," *Frontiers in Artificial Intelligence*, vol. 4, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Sule Tekkesinoglu, Azra Habibovic, and Lars Kunze, "Advancing Explainable Autonomous Vehicle Systems: A Comprehensive Review and Research Roadmap," *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 3, pp. 1-46, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Chuyao Wang, and Nabil Aouf, "Explainable Deep Adversarial Reinforcement Learning Approach for Robust Autonomous Driving," *IEEE Transactions on Intelligent Vehicles*, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Stephanie Milani et al., "Explainable Reinforcement Learning: A Survey and Comparative Review," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1-36, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Madhuri Singh et al., Explainable Reinforcement Learning Agents Using World Models, *arXiv Preprint*, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Zhuoli Zhuang et al., "AEGIS: Human Attention-based Explainable Guidance for Intelligent Vehicle Systems," *arXiv Preprint*, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[22] Rittika Shamsuddin, Habib B. Tabrizi, and Pavan R. Gottimukkula, "Towards Responsible AI: An Implementable Blueprint for Integrating Explainability and Social-Cognitive Frameworks in AI Systems," *AI Perspectives & Advances*, vol. 7, no. 1, pp. 1-23, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[23] Ruchik Kashyapkumar Thaker, "Advancing Reinforcement Learning: The Role of Explainability, Human, and AI Feedback Integration," *Robotics & Automation Engineering Journal*, vol. 6, no. 2, pp. 1-7, 2024. [Publisher Link]

[24] Chris Lee, Eduardo Benitez Sandoval, and Francisco Cruz, "Human Decision-Making Concepts with Goal-Oriented Reasoning for Explainable Deep Reinforcement Learning," *Australasian Joint Conference on Artificial Intelligence*, Melbourne, VIC, Australia, pp. 228-240, 2024. [CrossRef] [Google Scholar] [Publisher Link]