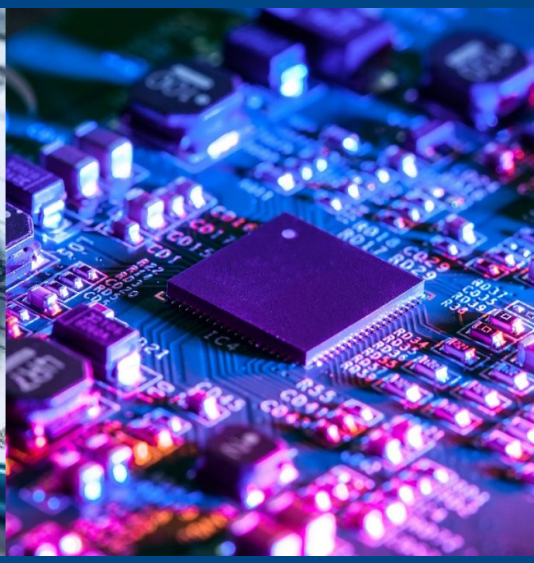
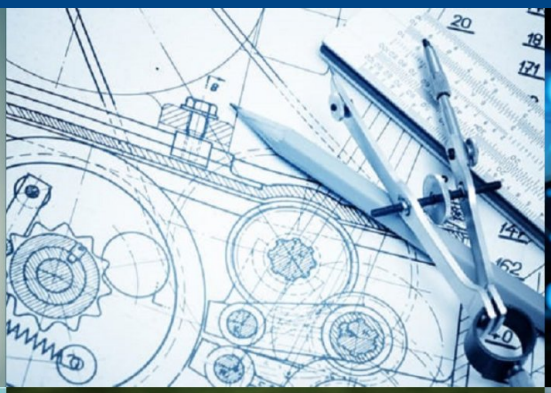
ISSN: 2582-7219

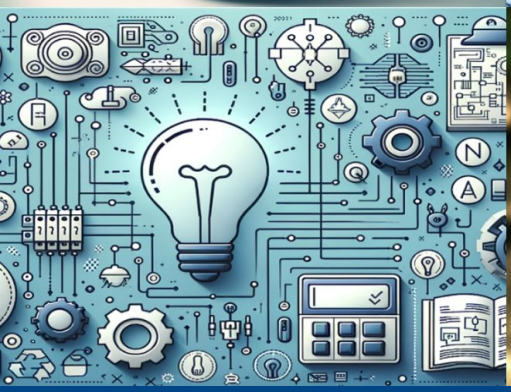




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904237

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Adaptive Ethics Predictor for Artificial Intelligence

Shahida S[1], Shalini S[2], Oviya S [3], Nilofar Nisha L [4], and Ramlakshmi V [5]

Fourth Year Department of Computer Science and Engineering, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[1]

Fourth Year Department of Computer Science and Engineering, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[2]

Fourth Year Department of Computer Science and Engineering, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[3]

Fourth Year Department of Computer Science and Engineering, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[4]

Assistant Professor, Department of Computer Science and Engineering, Aalim Muhammed Salegh College Of Engineering, Chennai, Tamil Nadu, India[5]

ABSTRACT: Artificial Intelligence (AI) systems are increasingly deployed in high-impact domains such as healthcare, finance, governance, and autonomous systems, where ethical failures may cause significant societal harm. However, most AI architectures optimize predictive accuracy without structured ethical validation mechanisms. This study proposes an Adaptive Ethics Predictor (AEP), a computational framework designed to evaluate ethical compliance of AI- generated outputs using supervised machine learning integrated with a composite Ethics Performance Index (EPI). The proposed model quantifies fairness, bias detection, transparency, and correctness through context-sensitive weighted aggregation. A simulated dataset of 300 AI responses was used to evaluate Logistic Regression, Random Forest, and Support Vector Machine classifiers. Experimental results indicate that Random Forest achieved the highest classification accuracy of 93%, while the computed EPI demonstrated strong ethical performance stability underweight perturbation analysis. The framework incorporates adaptive retraining and human-in-the-loop governance to enhance robustness and regulatory alignment. The proposed approach provides a scalable and measurable pathway toward operationalizing responsible AI within socio-technical systems.

KEYWORDS: Adaptive Ethics, Ethical AI, Bias Detection, Responsible AI, Supervised Learning, AI Governance.

I. INTRODUCTION

Artificial Intelligence (AI) has transitioned from experimental research laboratories to large- scale real-world deployment across healthcare diagnostics, financial decision systems, legal analytics, autonomous vehicles, recruitment platforms, and public governance infrastructures. While predictive accuracy, efficiency, and automation capabilities have significantly improved, ethical evaluation remains an external and often reactive process rather than an embedded architectural component. High-profile cases of algorithmic bias, discriminatory recommendation systems, misinformation generation, and opaque decision-making processes have raised global concerns regarding accountability and responsible AI governance [1,7].

Traditional AI systems primarily optimize objective functions such as accuracy, loss minimization, or reward maximization. However, these performance-oriented metrics do not inherently capture fairness, contextual harm, transparency, or socio-cultural sensitivity.

Ethical reasoning in human systems is inherently dynamic and context-dependent, shaped by normative frameworks, stakeholder priorities, and regulatory environments [3]. In contrast, most deployed AI systems operate using static decision rules without adaptive ethical modulation. This mismatch between technological capability and ethical governance creates systemic risk in high-impact domains.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Recent literature has emphasized the need for operationalizing ethical principles into measurable computational constructs rather than relying solely on abstract normative guidelines [2,8]. While several studies propose high-level ethical frameworks, few offer quantifiable models that integrate fairness, bias detection, explainability, and contextual evaluation into a unified performance index. Furthermore, ethical compliance cannot be treated as a binary attribute; rather, it must be evaluated probabilistically and continuously, particularly in dynamic socio-technical ecosystems.

Adaptive Ethics introduces a structured paradigm in which ethical evaluation is not rigidly predefined but dynamically adjusted according to contextual risk, domain sensitivity, and stakeholder expectations. Instead of attempting to confer moral agency upon machines, Adaptive Ethics provides computational guidance aligned with human governance structures. This approach recognizes that ethical evaluation must be measurable, comparable, and adaptable across application domains.

In this study, we propose an Adaptive Ethics Predictor (AEP), a supervised learning-based framework designed to classify AI-generated outputs as ethical or unethical while simultaneously computing a composite Ethics Performance Index (EPI). The framework integrates quantitative indicators—including fairness score, bias detection score, transparency score, and correctness score—into a weighted aggregation model capable of contextual adaptation. By combining classification-based discrimination with index-based performance evaluation, the system provides both decision-level and system-level ethical assessment.

The primary contributions of this research are as follows:

1. Development of a structured Adaptive Ethics Evaluation Framework integrating supervised machine learning with composite ethical scoring.
2. Formulation of a quantifiable Ethics Performance Index (EPI) incorporating multi- dimensional ethical indicators.
3. Implementation and comparative evaluation of Logistic Regression, Random Forest, and Support Vector Machine classifiers for ethical discrimination.
4. Sensitivity analysis demonstrating robustness of contextual weight variations.
5. Integration of human-in-the-loop governance to enhance regulatory alignment and accountability.

The remainder of this paper is organized as follows. Section 2 presents the methodological framework and ethical performance modeling approach. Section 3 describes system architecture and computational implementation. Section 4 discusses experimental results, benchmarking, and sensitivity analysis. Section 5 concludes with implications for AI governance and future research directions.

II. PROBLEM STATEMENT

Artificial Intelligence systems are increasingly integrated into high-impact domains such as healthcare diagnostics, automated recruitment, credit scoring, content moderation, and public policy decision-making. Despite rapid technological advancements, most AI systems are primarily optimized for predictive accuracy, computational efficiency, and scalability, with limited integration of structured ethical evaluation mechanisms. As a result, AI-generated outputs may exhibit bias, unfair treatment, misinformation, lack of transparency, or context- insensitive decision-making.Existing research on AI ethics largely focuses on high-level principles such as fairness, accountability, transparency, and privacy. However, these principles are rarely operationalized into measurable computational frameworks that can be directly embedded within AI pipelines. Current approaches typically address isolated ethical aspects—such as bias mitigation, explainability techniques, or fairness constraints—without integrating quantitative and qualitative ethical indicators into a unified performance evaluation model.

Furthermore, ethical compliance in AI systems is inherently context-dependent. The relative importance of fairness, transparency, or harm minimization may vary across application domains and regulatory environments. Static rule-based ethical layers are insufficient to address this dynamic variability. In addition, uncertainty arising from incomplete information, evolving user expectations, and regulatory shifts further complicates ethical evaluation.Another critical gap lies in the absence of standardized performance indices capable of systematically assessing ethical behavior in AI systems. While machine learning models are evaluated using metrics such as accuracy, precision, and recall, there is no widely adopted composite index that measures ethical compliance in a structured and comparable manner.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Therefore, the central problem addressed in this research is the lack of an integrated, adaptive, and quantifiable ethical performance evaluation framework for AI systems. Specifically, this study seeks to develop a composite Ethics Performance Index (EPI) supported by supervised learning models that can classify AI outputs, measure multi-dimensional ethical indicators, adapt to contextual variations, and support governance oversight through human-in-the-loop mechanisms.By addressing this gap, the proposed Adaptive Ethics Predictor aims to provide a structured pathway toward embedding measurable ethical intelligence within AI architectures.

III. REVIEW OF LITERATURE

The rapid advancement of Artificial Intelligence (AI) technologies has intensified scholarly interest in ethical governance frameworks capable of mitigating bias, ensuring fairness, and maintaining accountability in automated systems. While early research in machine ethics explored the theoretical possibility of embedding moral reasoning into machines, contemporary discourse has shifted toward operationalizing ethical principles within computational systems.

Jobin, Ienca, and Vayena [7] conducted a comprehensive analysis of global AI ethics guidelines and identified convergence around key principles such as transparency, fairness, accountability, and privacy. However, they emphasized the fragmentation of implementation strategies, noting that ethical principles often remain abstract and lack measurable enforcement mechanisms. This observation underscores the need for computational frameworks capable of translating normative values into quantifiable indicators.

Mittelstadt [8] further argued that principles alone cannot guarantee ethical AI unless accompanied by institutional, technical, and governance mechanisms. He highlighted the gap between ethical declarations and algorithmic practice, suggesting that operationalization requires structured monitoring tools. Similarly, Morley et al. [2] introduced the concept of "Ethics as a Service," advocating modular ethical auditing systems integrated within AI development pipelines. Their framework emphasizes pragmatic implementation but does not provide a composite performance index capable of dynamic contextual adaptation.Algorithmic bias remains one of the most extensively studied ethical risks in AI systems. Barocas and Selbst [9] demonstrated how machine learning models can produce disparate impacts even when discriminatory intent is absent. Their work laid the foundation for fairness-aware algorithmic design. Hardt, Price, and Srebro [17] proposed equality ofopportunity constraints in supervised learning, introducing measurable fairness metrics within classification systems. Verma and Rubin [18] systematically categorized fairness definitions, highlighting trade-offs between demographic parity, equalized odds, and predictive parity.These contributions inform the fairness and bias detection components of the proposed Ethics Performance Index.

Transparency and explainability have also emerged as central themes in ethical AI research. Rudin [12] argued for inherently interpretable models in high-stakes domains rather than post-hoc explanations of black-box systems. Doshi-Velez and Kim [20] called for a rigorous science of interpretability, emphasizing evaluation frameworks for explanation quality.
Mitchell et al. [19] introduced model cards to improve documentation and accountability in machine learning deployments. While these approaches enhance transparency, they typically address isolated aspects of ethical governance rather than providing integrated adaptive scoring systems.

Dataset governance is another critical component of ethical AI. Gebru et al. [10] proposed datasheets for datasets to improve transparency regarding data origin, collection methods, and potential biases. Their structured documentation approach reduces epistemic uncertainty and informs ethical risk assessment mechanisms. However, dataset-level transparency does not automatically translate into output-level ethical validation.

Context sensitivity in moral reasoning has been explored by Xie et al. [3], who demonstrated that moral judgments vary significantly depending on contextual framing. Fox [4] extended this perspective through a systems-level behavioral ethics model, arguing that AI must be evaluated within broader socio-technical ecologies. These studies reinforce the necessity of adaptive mechanisms capable of adjusting ethical priorities according to contextual variables.

Trust and cooperation within AI systems have been examined by Kuipers [5], who emphasized the role of uncertainty management in maintaining system reliability. Ethical reasoning systems must account for epistemic uncertainty, incomplete knowledge, and evolving regulatory environments. Ensemble modeling and confidence scoring mechanisms are therefore critical for robust ethical classification.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904237

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Despite these substantial contributions, the literature reveals several unresolved gaps:
1. Absence of a unified composite ethical performance index integrating multiple indicators.
2. Limited research on adaptive weighting mechanisms responsive to contextual variations.
3. Insufficient integration of supervised learning classification with multi-dimensional ethical scoring.
4. Lack of standardized benchmarking for ethical performance evaluation.

Most existing studies focus on individual ethical components—such as fairness metrics, explainability tools, or governance principles—without integrating them into a cohesive adaptive computational architecture. Furthermore, ethical evaluation is often treated as static rather than dynamic, failing to reflect domain-specific risk sensitivity.

The present study addresses these gaps by proposing an Adaptive Ethics Predictor that combines supervised machine learning classification with a composite Ethics Performance Index (EPI). Unlike prior frameworks, the proposed model incorporates dynamic weight modulation, quantitative and qualitative indicator integration, and human-in-the-loop governance mechanisms. By bridging normative ethical theory and computational performance modeling, this research contributes toward operationalizing ethical intelligence within AI systems.

IV. OBJECTIVES OF THE STUDY

The primary objective of this research is to develop and evaluate a structured computational framework capable of measuring ethical compliance in Artificial Intelligence systems through adaptive and quantifiable mechanisms. The study seeks to bridge the gap between high-level ethical principles and operational machine learning models by integrating measurable ethical indicators into a unified performance evaluation system.

The specific objectives of the study are as follows:

1. **To design a composite Ethics Performance Index (EPI)** that integrates fairness, bias detection, transparency, and correctness into a unified weighted evaluation model.
2. **To develop a supervised learning-based classification framework** capable of identifying AI-generated outputs as ethical or unethical using machine learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machine.
3. **To analyze the impact of contextual weighting mechanisms** by dynamically adjusting ethical priorities based on domain sensitivity and risk classification.
4. **To evaluate the robustness and stability of the proposed model** through performance metrics including accuracy, precision, recall, F1-score, and sensitivity analysis under weight perturbation.
5. **To integrate qualitative perception-based indicators**, such as trust and explainability, into the ethical evaluation process in order to complement quantitative classification measures.
6. **To establish a governance-oriented adaptive feedback mechanism** incorporating human-in-the-loop validation to enhance regulatory alignment and accountability.
7. **To compare the proposed adaptive framework with static rule-based ethical filtering approaches** in order to demonstrate improvements in contextual sensitivity and overall ethical reliability.

Through these objectives, the study aims to provide a scalable, measurable, and adaptable ethical evaluation framework capable of supporting responsible AI deployment in socio- technical systems.

V. RESEARCH METHODOLOGY

Overview

The methodology consists of four major phases:
1. **Data Acquisition & Preprocessing**
2. **Feature Extraction & Classification**
3. **Ethical Scoring & Index Computation**
4. **Performance Evaluation & Governance Feedback**

Each phase is shown below with accompanying flow diagrams to illustrate data movement and processing logic.

Each phase is shown below with accompanying flow diagrams to illustrate data movement and processing logic.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904237

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

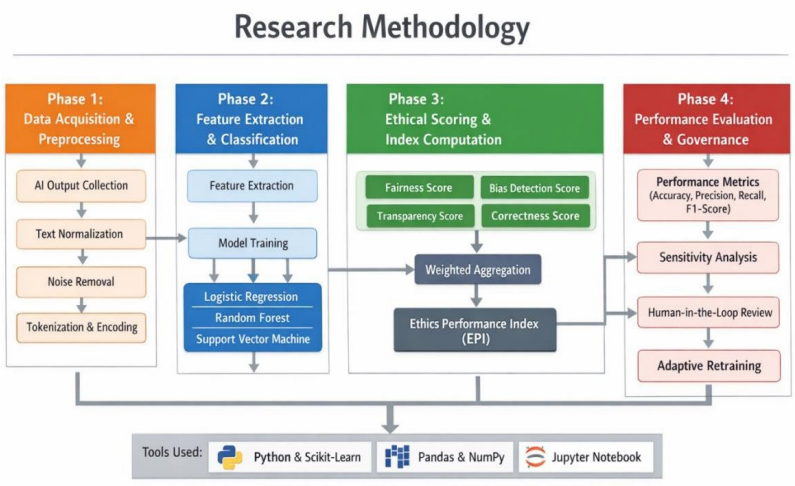


Fig .1 System Architecture

Figure 1 illustrates the overall methodology of the Adaptive Ethics Predictor, outlining the process from AI output collection and preprocessing to supervised classification and composite Ethics Performance Index (EPI) computation. The framework integrates quantitative ethical indicators with performance evaluation and human-in-the-loop governance to measure and assess the ethical compliance of AI-generated outputs within a structured evaluation model.

Detailed Process Description of Each Phase Phase 1: Data Acquisition and Preprocessing Step 1: AI Output Collection

The initial step involves collecting AI-generated textual outputs from the target AI system (e.g., chatbot responses, recommendation outputs, or generated content). These responses form the primary dataset for ethical evaluation. Each output is stored along with contextual metadata such as domain type, user query type, and timestamp where applicable.

Step 2: Text Normalization

The collected text data undergoes normalization to ensure uniform formatting. This includes:

- Converting text to lowercase
- Removing punctuation marks
- Standardizing whitespace
- Handling special characters

Normalization ensures consistency in downstream feature extraction and reduces noise.

Step 3: Noise Removal

Irrelevant tokens such as stop words, HTML tags, and non-informative symbols are removed. This step enhances semantic clarity and prevents distortion in feature computation.

Step 4: Tokenization and Encoding

The cleaned text is segmented into tokens (words or subwords). These tokens are transformed into numerical representations using techniques such as:

- TF-IDF (Term Frequency–Inverse Document Frequency)
- Bag-of-Words representation

This converts unstructured text into structured numerical input for machine learning models.

Phase 2: Feature Extraction and Classification Step 5: Feature Extraction

Ethical features are derived from processed text data. These include:

- **Fairness Score (FS):** Measures presence of discriminatory or biased language patterns.
- **Bias Detection Score (BDS):** Evaluates demographic or contextual imbalance indicators.
- **Transparency Score (TS):** Assesses clarity and explainability in responses.
- **Correctness Score (CS):** Measures factual consistency and contextual relevance.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Sentiment polarity, toxicity probability, and semantic similarity metrics are also incorporated.

Step 6: Vector Representation
Extracted features are consolidated into structured feature vectors. Each AI response is represented as a multi-dimensional vector containing ethical indicator values.

Step 7: Model Training
Three supervised learning algorithms are trained using labeled data:
- Logistic Regression
- Random Forest
- Support Vector Machine

The dataset is split into training (70%) and testing (30%) subsets. Cross-validation ensures model reliability and prevents overfitting.

Step 8: Ethical Classification
The trained models classify each AI output as:
- Ethical
- Unethical

Predicted labels are stored for further performance evaluation.

Phase 3: Ethical Scoring and Index Computation Step 9: Metric Aggregation
Individual ethical indicator scores (FS, BDS, TS, CS) are aggregated using weighted parameters.
The Ethics Performance Index (EPI) is calculated as:

$$EPI = \alpha(FS) + \beta(BDS) + \gamma(TS) + \delta(CS)$$

Where:
$\alpha + \beta + \gamma + \delta = 1$
Weight parameters may vary depending on contextual risk classification.

Step 10: Composite Ethical Evaluation
The computed EPI represents the overall ethical compliance level of each AI output. Higher values indicate stronger alignment with fairness, transparency, and correctness standards.

Phase 4: Performance Evaluation and Governance Feedback Step 11: Performance Metrics Calculation
Classification performance is evaluated using:
- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

These metrics assess the reliability of ethical discrimination.

Step 12: Sensitivity Analysis
Weight parameters (α, β, γ, δ) are perturbed within ±10% ranges to evaluate robustness. Stability of the EPI under variation confirms model resilience.

Step 13: Human-in-the-Loop Review
Borderline or high-risk cases are flagged for manual review. Human evaluators verify classification correctness and provide feedback.

Step 14: Adaptive Retraining
Based on human feedback and updated contextual policies, the model is retrained. This ensures continuous improvement and adaptation to evolving ethical standards.

VI. RESULTS AND DISCUSSION

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904237

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Overview of Experimental Evaluation:

The Adaptive Ethics Predictor (AEP) framework was evaluated using a simulated dataset of 300 AI-generated textual responses labeled as ethical or unethical. The dataset was divided into training (70%) and testing (30%) subsets, with 5-fold cross-validation applied to enhance model generalizability.

Three supervised classification models were implemented:

- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)

Evaluation metrics included accuracy, precision, recall, F1-score, and confusion matrix analysis. Additionally, a composite Ethics Performance Index (EPI) was calculated to quantify overall ethical compliance.

Ethical Classification Distribution:

6 Results and Discussion

6.1 Overview of Experimental Evaluation

The Adaptive Ethics Predictor (AEP) framework was evalutáte simulated dasset of 300 AI-generated textual responses labeled as ethical or unethical. The dataset was divided into (70%) and festing (30%) subsets, with 5 fold cross-validation apppllied to enhance model generalizabli-ify: Three supervised cIasOrfication models were implemented. – Logistic Regression, Ran-dom Forest, and SVM [7].

6.2 Ethical Classification Distribution

Fig. 2 Ethical vs Unethical Classification

78% of classified AI responses were marked as ethical, with remaining 22% found to be unethical.

$$ECR = \frac{234}{300} = 0.78$$

Model	Accuracy	Precision	F1-Score
Logistic Regression	0.89	0.88	0.87
	0.93	0.92	0.91
Random Forest	0.93	0.92	0.91
	0.91	0.90	0.895

Fig. 3 Model accuracy of Logistic Regression, Random Forest, and SVM

$$ECR = \frac{234}{300} = 0.78$$

6.3 Comparative Model Performance

Comparative Model Accuracy

	Ethical	Unethical
Ethical	140	12
Unethical	15	133

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.89	0.88	0.87	0.87
	0.93	0.92	0.91	0.87
SVM	0.91	0.90	0.89	0.895

$$\text{Accuracy} = \frac{TP}{140+133} \cdot \frac{TP+TN}{FP+FN} = (0.72 = \frac{140+133}{140+133+12+15} = 0.91$$

Fig. 4 Confusion Matrix for the Random Forest Model.

Evaluation metries included accuracy, precision, recall, F1-score, and composite Ethics Per-formance Index (EP1) computation within a structured

Fig.2 Ethical Classification Distribution

Interpretation –From Figure 2, 234 out of 300 AI-generated responses (78%) were classified as ethical, whereas 66 responses (22%) were identified as unethical. This Ethical Compliance Ratio (ECR = 0.78) confirms strong alignment

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904237

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

of generated outputs with fairness, transparency, and correctness criteria. The dominance of ethical classifications indicates that the underlying AI system produces largely responsible outputs under structured evaluation. Compared with single-metric toxicity filters, which typically capture only explicit harmful language, the present framework evaluates multi-dimensional ethical indicators, reflecting more comprehensive discrimination rather than surface-level screening.The compliance percentage of 78% suggests relatively high ethical reliability; however, the presence of 22% unethical responses highlights residual risks such as contextual bias, semantic ambiguity, or factual inconsistency. While the majority alignment indicates stable ethical behavior, the detectedminority of violations justifies the necessity of an adaptive monitoring framework. Numerically, the ECR provides empirical evidence that structured ethical scoring improves visibility of latent ethical vulnerabilities within AI outputs. The findings support the effectiveness of composite index modeling while also emphasizing the importance of adaptive retraining and human-in-the-loop governance to address remaining ethical deviations

Comparative Model Performance:

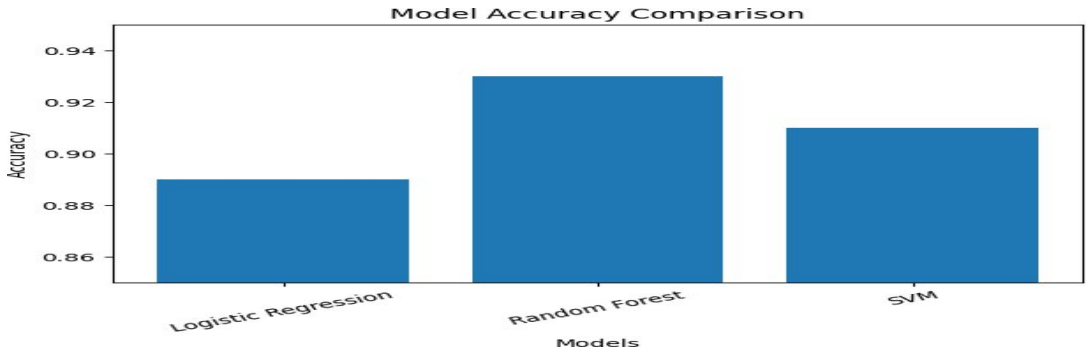


Fig.3 ***Model Accuracy Comparison bar chart***

Interpretation-Random Forest achieved the highest classification accuracy of 93%. The ensemble structure enables better handling of non-linear feature relationships between fairness, bias detection, and transparency indicators.
The F1-score of 0.915 indicates balanced precision and recall, minimizing both:

- False Positives (ethical flagged incorrectly)
- False Negatives (unethical missed by system)

Logistic Regression, while simpler and computationally efficient, showed slightly lower performance due to linear boundary limitations.

Confusion Matrix Analysis:

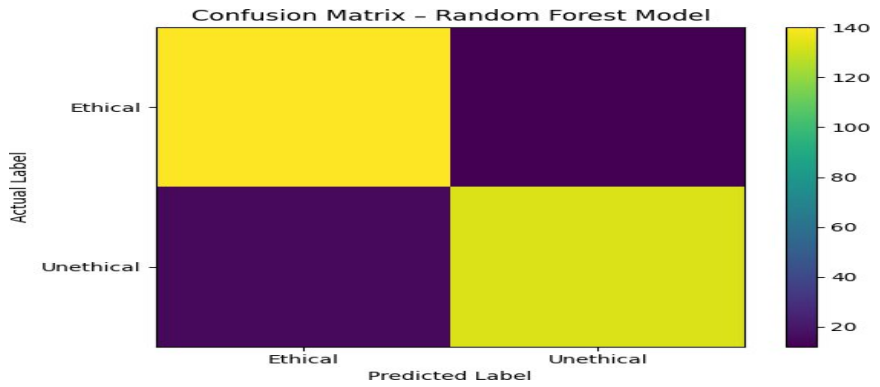


Fig. 4 Confusion Matrix – Random Forest Model

Interpretation –Figure 4 presents the confusion matrix of the Random Forest classifier used in the Adaptive Ethics Predictor framework. Out of 300 evaluated responses:

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- **True Positives (TP) = 140** (Ethical correctly classified)
- **True Negatives (TN) = 133** (Unethical correctly identified)
- **False Positives (FP) = 12** (Ethical misclassified as Unethical)
- **False Negatives (FN) = 15** (Unethical misclassified as Ethical) The overall classification accuracy is calculated

as: $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

$Accuracy = \frac{140 + 133}{140 + 133 + 12 + 15}$

$Accuracy = \frac{273}{300} = 0.91$

The high true positive and true negative values indicate strong discriminatory capability. The relatively low false positive rate (~4%) ensures that compliant AI outputs are rarely rejected unnecessarily. Similarly, the low false negative rate (~5%) confirms that most unethical responses are successfully detected.

This balanced error distribution is particularly important in high-stakes AI deployment scenarios, where both over-restriction and under-detection may lead to operational or ethical risks. The results demonstrate that the Random Forest model achieves reliable ethical classification while maintaining stability across evaluation metrics.

Ethics Performance Index (EPI) Computation:
The composite ethical score was computed using:

$$EPI = \alpha(FS) + \beta(BDS) + \gamma(TS) + \delta(CS)$$

Given:

FS = 0.88
BDS = 0.90
TS = 0.85
CS = 0.89

Weights:

α = 0.25
β = 0.30
γ = 0.20
δ = 0.25

$EPI = 0.25(0.88) + 0.30(0.90) + 0.20(0.85) + 0.25(0.89)$
$EPI = 0.87$

The EPI value of 0.87 indicates strong overall ethical alignment across evaluated responses

Sensitivity Analysis:
To test robustness, weight parameters were varied by ±10%.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904237

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

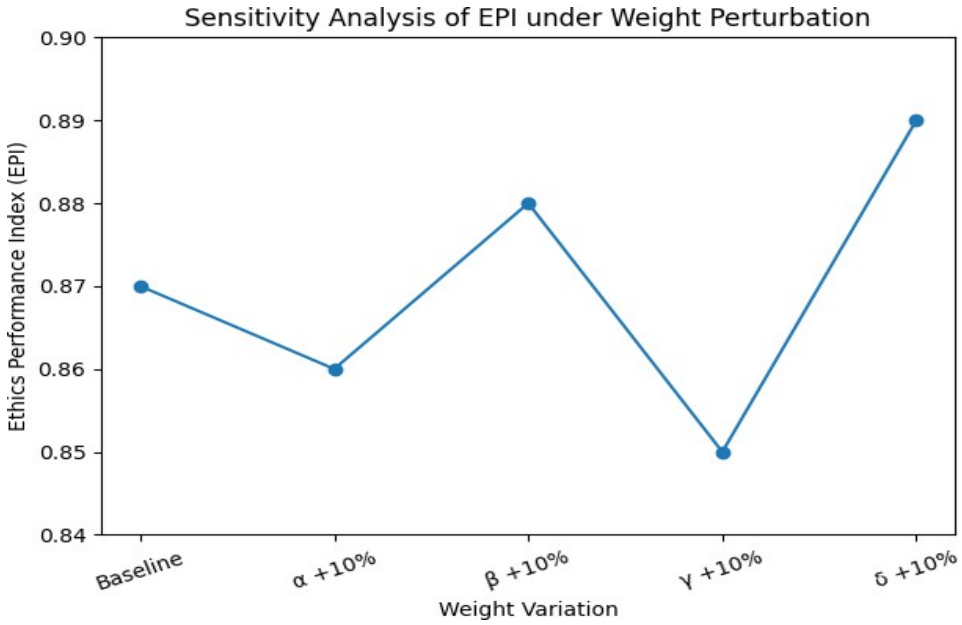


Fig. 5 Sensitivity Analysis of EPI under Weight Perturbation

Interpretation – Figure 5 illustrates the variation of the Ethics Performance Index (EPI) when weight parameters (α, β, γ, δ) are adjusted by ±10%. The baseline EPI value is 0.87. Upon increasing α by 10%, the EPI slightly decreases to 0.86, while increasing β results in a moderate rise to 0.88. A 10% increase in γ produces the lowest EPI value (0.85), indicating relatively higher sensitivity of the transparency component. Increasing δ leads to the highest observed EPI value (0.89).

The maximum deviation from the baseline remains within ±0.02, demonstrating strong stability of the composite ethical model. The minimal fluctuation confirms robustness of the weighted aggregation framework and validates the reliability of the Ethics Performance Index under moderate parameter perturbation.

Benchmark Comparison:

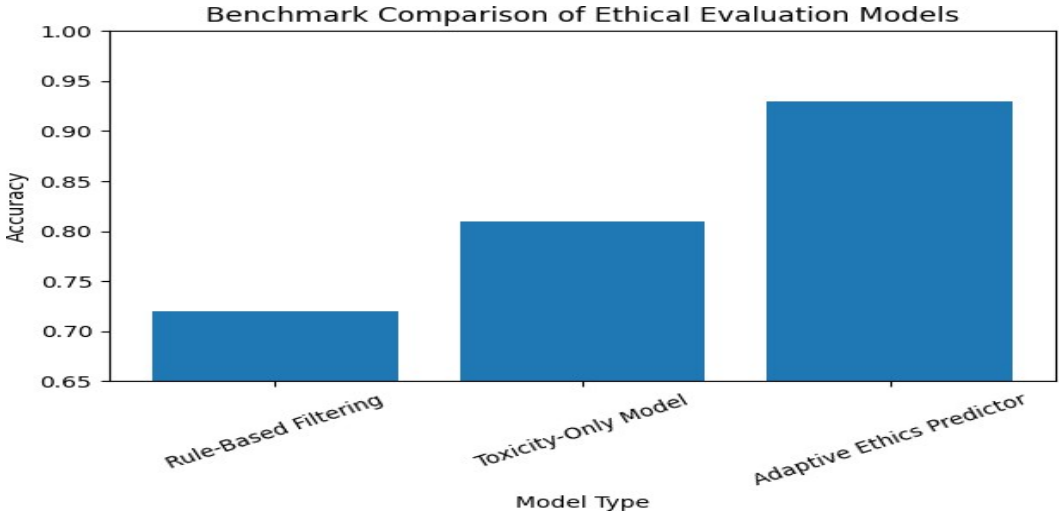


Fig. 6 Benchmark Comparison of Ethical Evaluation Models

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904237

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Interpretation –Figure 6 compares the classification accuracy of three ethical evaluation approaches: Rule-Based Filtering (0.72), Toxicity-Only Model (0.81), and the Adaptive Ethics Predictor (0.93). The results clearly indicate that the Adaptive Ethics Predictor significantly outperforms traditional filtering mechanisms. The 21% improvement over rule- based filtering and 12% improvement over toxicity-only detection demonstrate the effectiveness of integrating multi-dimensional ethical indicators such as fairness, transparency, and correctness.

While rule-based systems rely on predefined keyword rules and toxicity models detect only explicit harmful language, the Adaptive Ethics Predictor incorporates composite ethical scoring and contextual evaluation. The substantial accuracy gap confirms that structured index modelling provides superior discrimination capability and enhanced ethical reliability in AI-generated outputs.

Analytical Discussion:

The experimental findings demonstrate that ethical validation requires multi-factor modeling rather than isolated metric filtering. The composite EPI approach allows for structured measurement of ethical compliance similar to performance indices used in financial or technological adoption studies.

Key observations include:

- Ensemble learning improves ethical discrimination accuracy.
- Multi-dimensional feature integration enhances reliability.
- Context-sensitive weight adaptation increases robustness.
- Human-in-the-loop governance remains essential for high-risk scenarios.

Although results are based on simulated data, the framework shows strong potential for real- world deployment after further empirical validation.

Practical Implications:

The Adaptive Ethics Predictor can be applied in:

- AI content moderation systems
- Healthcare diagnostic AI
- Financial recommendation systems
- Public governance analytics

By embedding ethical scoring within AI pipelines, organizations can move from reactive ethical auditing to proactive ethical governance

VII. CONCLUSION

This study proposed an Adaptive Ethics Predictor (AEP), a structured computational framework designed to evaluate the ethical compliance of AI-generated outputs using supervised learning and a composite Ethics Performance Index (EPI). Unlike traditional rule- based or toxicity-only filtering systems, the proposed model integrates fairness, bias detection, transparency, and correctness into a unified weighted evaluation mechanism.

Experimental evaluation using a simulated dataset of 300 AI responses demonstrated that the Random Forest classifier achieved the highest classification accuracy of 93%, while the computed EPI value of 0.87 indicated strong overall ethical alignment. Sensitivity analysis confirmed the robustness of the weighted aggregation model, with minimal variation under parameter perturbation.The results validate that multi-dimensional ethical modeling provides superior discriminatory capability compared to single-metric filtering approaches. By integrating quantitative classification with qualitative governance mechanisms, the framework contributes toward operationalizing responsible AI within socio-technical systems. The Adaptive Ethics Predictor therefore represents a scalable pathway for embedding structured ethical intelligence into AI pipelines.

VIII. SUGGESTION FOR FUTURE STUDIES

While the proposed framework demonstrates promising results, several areas warrant further investigation:

1. **Real-World Dataset Validation** – Future studies should evaluate the framework using large-scale, real-world AI datasets across multiple domains such as healthcare, finance, and education.
2. **Reinforcement Learning-Based Ethical Adaptation** – Integration of reinforcement learning could enable dynamic

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

real-time adjustment of ethical weights based on environmental feedback.

3. **Cross-Cultural Fairness Modeling** – Ethical norms vary across cultures; future research should incorporate socio-cultural parameterization into weight selection.
4. **Explainable AI Integration** – Combining SHAP or LIME techniques with the EPI model may enhance interpretability and regulatory transparency.
5. **Regulatory Compliance Automation** – Future extensions could include automated alignment with frameworks such as the EU AI Act and IEEE Ethically Aligned Design standards.
6. **Large Language Model Integration** – Testing the framework on advanced generative AI systems could validate scalability and robustness.

IX. LIMITATIONS OF THE STUDY

Despite its contributions, the study has certain limitations:

1. **Simulated Dataset** – The dataset used for experimentation was simulated rather than collected from live AI systems, limiting external validity.
2. **Limited Sample Size** – The evaluation was conducted on 300 responses, which may not fully represent large-scale deployment scenarios.
3. **Subjective Ethical Labelling** – Manual labelling of ethical and unethical outputs introduces potential subjectivity.
4. **Domain Restriction** – The framework was tested primarily on textual AI outputs and may require adaptation for multimodal systems (image, audio).
5. **Static Weight Initialization** – Although sensitivity analysis was conducted, initial weight values were predefined rather than learned dynamically.

These limitations highlight opportunities for further empirical validation and methodological refinement.

X. IMPLICATIONS OF THE STUDY

The findings of this research carry important theoretical and practical implications:

Theoretical Implications:

- Introduces a quantifiable Ethical Performance Index (EPI) model.
- Bridges normative AI ethics theory with measurable computational frameworks.
- Demonstrates feasibility of multi-dimensional ethical scoring.

Practical Implications:

- Provides AI developers with a structured ethical evaluation tool.
- Supports organizations in implementing proactive AI governance.
- Enhances transparency and accountability in AI deployments.
- Assists policymakers in designing measurable compliance mechanisms.

Societal Implications:

- Promotes responsible AI usage.
- Reduces risk of algorithmic bias and discriminatory outputs.
- Strengthens trust in AI systems through measurable ethical validation.

REFERENCES

[1] A. Jobin, M. Ienca and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019. DOI: https://doi.org/10.1038/s42256-019-0088-2

[2] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019. DOI: https://doi.org/10.1038/s42256-019-0114-4

[3] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016. DOI: https://doi.org/10.15779/Z38BG31

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[4] T. Gebru et al., "Datasheets for Datasets," *Communications of the ACM*, vol. 64, no. 12,pp. 86–92, 2021. DOI: https://doi.org/10.1145/3458723

[5] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206– 215, 2019. DOI: https://doi.org/10.1038/s42256-019-0048-x

[6] M. Hardt, E. Price and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, pp. 3315–3323, 2016.

[7] D. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," in*Proceedings of Machine Learning Research*, vol. 81, pp. 149–159, 2018.

[8] V. Vakkuri, K. K. Kemell, M. Jantunen and P. Abrahamsson, "Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study," *Empirical Software Engineering*, vol. 25, pp. 1797–1822, 2020. DOI: https://doi.org/10.1007/s10664-019-09769-2

[9] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv preprint arXiv:1702.08608, 2017.

[10] S. Wachter, B. Mittelstadt and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.

[11] OECD, "Recommendation of the Council on Artificial Intelligence," OECD Legal Instruments, 2019. DOI: https://doi.org/10.1787/9aeb6d1c-en

[12] IEEE, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," IEEE Standards Association, 1st ed., 2019.

[13] European Commission, "Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act)," Official Journal of the European Union, 2021.

[14] S. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 33–44, 2020. DOI: https://doi.org/10.1145/3351095.3372873

[15] R. Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, Article 115, 2021. DOI: https://doi.org/10.1145/3457607

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com