International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

Corporate Ethics Policies: A Quantitative Framework for Evaluating AI Governance in Major Technology Firms

Mariyam Malik, Professor Dr. B Sasi Kumar
Dept of CSE
Dr VRK Women's College of Engineering and Technology, Hyderabad, India.

Abstract — Artificial Intelligence (AI) is now widely used in decision-making systems developed by technology giants to drive decisions, triggering concerns related to fairness, transparency, and accountability. In response, organizations such as IBM, Microsoft, and Google have published internal AI ethics policies aligned with international standards. Crucially, these policies prove descriptive in nature and lack measurable methods for evaluation. The work envisioned in this paper, purposefully, a quantitative framework to assess the alignment between corporate AI ethics policies and deployed machine learning systems. A supervised learning model is implemented as a case study using a public income prediction dataset containing sensitive demographic attributes. Fairness is evaluated using demographic parity, while model transparency is examined through SHAP-based feature attribution techniques. We apply additional plausible constraints to address privacy and accountability concerns by limiting sensitive identifiers and enforcing a modular, reproducible pipeline. An aggregated Ethics Compliance Score combines multiple ethical dimensions into a single evaluation measure, showing that ethical risks may persist even in accurate models. Unlike prior work that focuses 27 separately on principles or tools, the proposed framework links corporate ethics commitments directly to quantitative system-level indicators, providing a practical basis for internal AI governance.

Keywords— Corporate ethics, AI governance, Artificial intelligence ethics Ethical AI, AI policy, Corporate governance, Technology firms.

I. INTRODUCTION

Artificial Intelligence (AI) systems are increasingly embedded in applications that influence social and economic outcomes, including hiring, financial assessment, healthcare support, and online content moderation [1]. As these systems become more autonomous and widespread, concerns regarding ethical risks such as bias, lack of transparency, and misuse of personal data have gained significant attention [2]. These concerns are particularly relevant for large technology companies whose AI systems operate at global scale.

In recent years, companies such as IBM, Microsoft, and Google have released internal AI ethics policies and governance frameworks. These documents typically emphasize principles such as fairness, transparency, accountability, and privacy, and are often aligned with international guidelines proposed by organizations such as the OECD and UNESCO [3]. While these initiatives represent an important step toward responsible AI development, they are primarily policy-driven and lack concrete mechanisms for technical evaluation.

In practice, there is limited clarity on how ethical principles stated at the organizational level translate into deployed machine learning systems. As a result, it becomes difficult for organizations to verify compliance, for regulators to assess accountability, and for users to trust automated decisions. Most existing evaluations focus on model accuracy, while ethical performance is treated as a secondary or abstract concern.

The objective aligned in our work is to address this gap by presenting a measurable framework for evaluating corporate AI ethics policies at the system level. By integrating fairness metrics, explainability analysis, and governance-oriented design constraints, this work demonstrates how ethical commitments can be operationalized in real-world AI systems. A case study is presented to illustrate the application of the framework and to highlight practical challenges in ethical AI governance.

The highlighted contributions of this work are as follows:

- We map commonly stated corporate AI ethics principles (fairness, transparency, privacy, accountability) to concrete technical evaluation criteria at the model level.
- We implement a case study on income prediction using a public dataset with sensitive attributes, evaluating fairness through demographic parity and transparency through SHAP-based feature attribution.
- We propose an Ethics Compliance Score that aggregates performance, fairness, transparency, privacy, and

accountability into a single quantitative indicator to support corporate AI governance.

II. RELATED WORK

Ethical concerns that disrupt Artificial Intelligence have been discussed extensively in recent literature, particularly in the context of large-scale deployment by technology corporations. Early research in this area primarily focused on identifying ethical risks such as algorithmic bias, lack of transparency, and misuse of personal data [4]. Recent survey and review practices have shifted toward proposing governance frameworks and technical tools to address these challenges.

Several industry-driven initiatives have played a key role in shaping corporate AI ethics practices. IBM introduced the AI Fairness 360 toolkit to provide metrics and mitigation techniques for bias in machine learning models. Similarly, Microsoft's Responsible AI Standard outlines principles such as fairness, reliability, privacy, and accountability across the AI development lifecycle [5]. Google's AI Principles and Model Cards framework emphasize transparency and documentation to support ethical oversight. While these efforts demonstrate strong organizational commitment, they largely focus on guidelines and tools rather than system-level evaluation mechanisms.

From an academic perspective, fairness in applied machine learning studies using metrics such as demographic parity, equalized odds, and disparate impact. These metrics enable quantitative assessment of bias across sensitive attributes like gender and race [6]. However, most studies evaluate fairness in isolation, without connecting results to broader organizational ethics policies. As a result, the link between technical findings and corporate governance remains weak.

Explainable AI (XAI) techniques have also received significant attention as a means to improve transparency. SHAP, combined with LIME, is widely adopted to interpret predictions and feature influence on ML models [7]. Although explainability improves user trust and auditability, it does not directly ensure fairness or accountability unless combined with additional evaluation criteria.

More works integrated to attempted ethical considerations into AI governance frameworks, particularly in regulated domains such as finance and healthcare [8]. Moreover, newer approaches rely on qualitative assessments or compliance checklists, making them difficult to apply consistently across systems. There remains a lack of unified, quantitative frameworks that translate corporate ethics policies into measurable technical outcomes.

Our strategy rebuilds upon existing research by combining fairness metrics, explainability analysis, and governance-oriented evaluation into a single framework. Unlike prior work, the focus is not solely on ethical principles or technical tools, but on their practical alignment within corporate AI systems.

III. CORPORATE ETHICS FRAMEWORK MAPPING

Major technology companies have published internal AI ethics policies with the intention of guiding responsible system development and deployment [9]. Although the terminology and structure of these policies vary, there exists a significant overlap in the ethical principles they emphasize. This section outlines how commonly stated corporate ethics principles are fused into tangible technical criteria within the proposed framework.

A tied frequency raised principle across corporate policies is fairness. IBM, Microsoft, and Google explicitly acknowledge the risk of biased outcomes, particularly when AI systems are trained on historical or socially imbalanced data [10]. In this study, fairness is interpreted in terms of outcome distribution across sensitive demographic groups. Demographic parity is selected as the primary evaluation metric, as it provides a clear and interpretable measure of whether model predictions differ systematically based on attributes such as gender or race.

Transparency is another core component of corporate AI ethics frameworks. Organizations often state that AI systems should be understandable, auditable, and explainable to relevant stakeholders. However, policy documents rarely specify how this requirement should be implemented at the software level. To address this gap, our renewed framework incorporates feature attribution methods to examine how input variables influence model decisions. Explainability is treated not as a post hoc visualization, but as a required component of system evaluation.

Privacy is commonly addressed in corporate policies through principles of data minimization and responsible data usage. Rather than introducing additional privacy-preserving algorithms, this study focuses on practical design choices that limit ethical risk [11]. These include avoiding highly sensitive personal identifiers and ensuring that the dataset used for experimentation does not contain biometric or location-based information. While current focus does not eliminate privacy

concerns entirely, it reflects realistic constraints often applied in enterprise settings.

Accountability is typically framed at the organizational level, emphasizing documentation, auditability, and responsibility for system outcomes [12]. Translating this principle into technical practice remains challenging. In this framework, accountability is supported through reproducible model pipelines, clear separation of preprocessing and training stages, and using evaluation metrics that usually be reviewed after deployment. These elements enable traceability and facilitate post hoc analysis in the event of system failures or ethical violations.
By mapping corporate ethics principles to specific technical evaluation criteria, this framework signifies a salient design, assessing whether deployed AI systems reflect stated organizational values. This mapping serves as the foundation for the quantitative evaluation presented in subsequent sections.

IV. METHODOLOGY

The resonating study presented here is to examine how corporate AI ethics principles can be cross-checked at the level of an implemented machine learning system. To achieve this, a structured methodology is followed that integrates model development with ethical assessment. Rather than treating ethics as a separate or final-stage consideration, ethical evaluation is embedded throughout the system design and analysis process.

1. Dataset and Problem Definition
A publicly available income prediction dataset is used as a case study. The task involves classifying individuals into income categories based on demographic and socio-economic attributes. The iconized dataset deals in sensitive parameters including gender and race, which makes it suitable for evaluating fairness-related concerns. This widely favoured dataset also projects the results to be compared with existing literature and ensures transparency in experimentation.

2. Data Preprocessing
The dataset contains a mixture of numerical and categorical features. variables in strings, encoded use one-hot encoding, while numerical features are retained in their original form. Preprocessing is implemented as a separate pipeline stage to ensure consistency between training, testing and validation phases as a known ML practice. Ethically focused attributes are not purged, as they are required for post hoc fairness quality, but they are implicitly used as target variables.

3. Model Selection and Training
A Random Forest classifier, identified for our AI package due to its ability to handle heterogeneous data types and its relatively interpretable structure compared to deep learning models. The trained model uses a standard train–test split to evaluate generalization performance. Hyperparameters are chosen conservatively balancing predictive performance against interpretability, reflecting practical considerations in enterprise AI deployment.

4. Fairness Evaluation
Fairness is evaluated using demographic parity, which measures differences in expected outcome ratios across demographic groups. In this study, demographic parity is applied separately to gender and race attributes in order to quantify disparities in predicted high-income classifications. For each sensitive attribute, the proportion of positive predictions is computed within each group, and the resulting rates are compared to identify potential imbalances in model behaviour. Demographic parity is selected because it provides an intuitive and easily interpretable measure of group-level disparities that can be communicated to corporate governance stakeholders. This metric, however, does not capture all forms of bias, and the present analysis focuses on identifying differences in outcomes rather than enforcing strict parity constraints.

5. Transparency and Explainability
To address transparency requirements, model predictions are analysed using feature attribution techniques. SHAP is employed to estimate the involvement of each feature to model outputs. These explanations enable inspection of dominant decision factors and support auditing of the model's behaviour. Explainability is treated as an evaluation tool rather than a mechanism for improving performance.

6. Ethics Compliance Scoring
To enable unified assessment, individual ethical dimensions are normalized and aggregated into an overall Ethics Compliance Score. In this framework, fairness, transparency, privacy, and accountability are each represented by a scalar value between zero and one, where affected values indicate better ethical performance. The fairness score is derived from demographic parity gaps across gender and race, with larger disparities leading to lower scores. Transparency is treated as fully satisfied when SHAP-based explanations are computed for the trained model and made available for inspection. Privacy and accountability are scored based on practical design choices, omitting highly sensitive personal identifiers and a workable modular, reproducible pipeline with clearly separated preprocessing and explainability analysis also allows

stakeholders to verify that training stages. The overall Ethics Compliance Score is calculated as the arithmetic mean (AM) for the four component scores, reflecting an unweighted trade-off between the considered ethical dimensions. This scoring mechanism is intended to support comparative evaluation across models or system versions rather than to provide an absolute certification of ethical compliance.

V. IMPLEMENTATION AND RESULT ANALYSIS

The proposed framework is implemented using Python in a Google Colab environment to ensure reproducibility and accessibility. Our Experiments were run using standard machine learning libraries, and the implementation follows a modular pipeline structure to support transparency and auditability. This section presents the experimental results and discusses their ethical implications rather than focusing solely on predictive performance.

1. Model Performance

The proposed income prediction model scored high accuracy of 85.42%, indicating strong predictive capability and effective feature learning from the dataset. While accuracy is not the primary focus of this study, reporting performance is formalised to ensure that ethical evaluation is conducted on a functional system. The results indicate that acceptable predictive performance can be achieved without resorting to complex or opaque modelling techniques.

2. Fairness Analysis

(a) Gender-Based Fairness

Demographic parity was evaluated across gender groups to assess fairness in model predictions. The observed demographic parity values were 0.0456 for females and 0.2062 for males. The difference indicates a disparity in positive prediction rates between genders, highlighting the importance of fairness-aware evaluation in socio-economic decision systems.

(b) Race-Based Fairness

Fairness across racial groups was analysed using demographic parity metrics. The results show varying parity values across different races, with Asian-Pac-Islander (0.2033) and White (0.1643) groups receiving higher positive prediction rates compared to Black (0.0571) and Amer-Indian-Eskimo (0.0455) groups. These variations suggest the presence of demographic bias, reinforcing the need for ethical oversight in AI-based income prediction models.

The fairness results demonstrate valuables incorporating explicit bias evaluation into AI software assessment. Without such analysis, these disparities would remain undetected under standard performance metrics.

3. Explainability Results

To examine transparency, SHAP-based feature attribution is applied to the trained model. The resulting feature importance visualization shows that attributes related to education, occupation, and working hours contribute significantly to income predictions. Crucial factors are raised reasonably; the limitations reflect practical constraints often encountered in sensitive attributes that do not dominate the decision-making process.

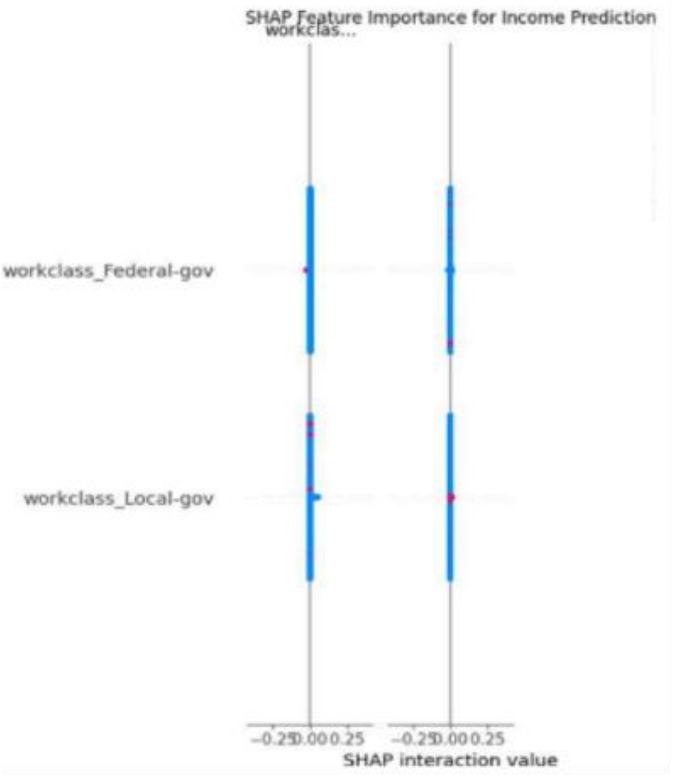

Fig. 1 SHAP interaction value

The SHAP analysis, shown in figure 1, supports the auditability of the system, gives valuable insights into designed model behaviour. It should be, however, taken care that explainability does not guarantee ethical correctness, and interpretations remain dependent on the quality of underlying data.

4. Ethics Compliance Score

An AI Ethics Compliance Score was computed by aggregating the component scores for performance, fairness, transparency, privacy, and accountability. The implemented model scored a high accuracy of 85.42%, which corresponds to a performance score of 0.854. The fairness score, derived from demographic parity gaps across gender and race, was 0.819, while transparency received the maximum score of 1.0 due to the input of SHAP-based feature attribution in the evaluation pipeline. Privacy and accountability were scored at 0.9 and 0.8, respectively, reflecting the omission of direct identifiers and the workable modular, reproducible workflow. Averaging these five dimensions yielded an overall Ethics Compliance Score of 0.875, indicating that the system aligns well with the considered ethical criteria while maintaining high predictive performance.

5. Limitations

This implementation has several limitations. First, the fairness assessment relies primarily on demographic parity, which does not capture all relevant aspects of bias and may conflict with other fairness notions in certain contexts. Second, privacy and accountability are addressed through design choices—such as data selection and pipeline structure—rather than through formal technical guarantees like differential privacy or cryptographic audit mechanisms. Third, the case study focuses on a single publicly available dataset, which may limit the generalizability of the findings to other domains and data distributions. These real-world corporate systems and highlight opportunities for future work.

VI. CONCLUSION AND FUTURE WORK

This paper presented a quantitative framework for evaluating corporate AI ethics policies through measurable technical indicators. By translating high-level principles of fairness, transparency, privacy, and accountability into specific evaluation criteria, the framework provides a structured approach to assessing whether deployed machine learning systems reflect stated organizational values. The case study on income prediction demonstrated that ethical risks, such as group-level disparities in positive outcomes, may remain present even in technically well-performing models. The proposed Ethics Compliance Score aggregates multiple ethical dimensions into a single metric that can support internal governance, model comparison, and continuous improvement, rather than serving as an absolute certification of ethical approval. Future work will extend this approach by incorporating additional fairness metrics, integrating formal privacy-preserving techniques, and introducing longitudinal monitoring of deployed systems. Further evaluation acrossdiverse application domains and regulatory contexts would strengthen the generalizability of the framework and help organizations respond to evolving expectations around AI governance.

REFERENCES

1. Parmar, Hemlata, and Utsav Krishan Murari. "Human-AI Synergy in Ethical Content Moderation: Navigating Fairness, Accountability, and Transparency Challenges." Ethical AI Solutions for Addressing Social Media Influence and Hate Speech. IGI Global Scientific Publishing, 2025. 191-212.
2. Kamila, Manoj Kumar, and Sahil Singh Jasrotia. "Ethical issues in the development of artificial intelligence: recognizing the risks." International Journal of Ethics and Systems 41.1 (2025): 45-63.
3. Lund, Brady, et al. "Standards, frameworks, and legislation for artificial intelligence (AI) transparency." AI and Ethics 5.4 (2025): 3639-3655.
4. Bahangulu, Julien Kiesse, and Louis Owusu-Berko. "Algorithmic bias, data ethics, and governance: Ensuring fairness, transparency and compliance in AI-powered business analytics applications." World Journal of Advanced Research and Reviews 25.2 (2025): 1746-1763.
5. Sharma, Ravishankar, Samia Loucif, and Aijaz A. Shaikh. "Implementing responsible AI: A life-cycle reference model based on FATES principles." Journal of Information & Knowledge Management 25.01 (2026): 2550078.
6. Ford, Jacob. "Fairness in focus: quantitative insights into bias within machine learning risk evaluations and established credit models." Management System Engineering 4.1 (2025): 8.
7. Kalasampath, Khushi, et al. "A literature review on applications of explainable artificial intelligence (XAI)." IEEE access 13 (2025): 41111-41140.
8. El Arab, Rabie Adel, Omayma Abdulaziz Al Moosa, and Mette Sagbakken. "Economic, ethical, and regulatory dimensions of artificial intelligence in healthcare: an integrative review." Frontiers in Public Health 13 (2025): 1617138.
9. Jedličková, Anetta. "Ethical approaches in designing autonomous and intelligent systems: a comprehensive survey towards responsible development." AI & society 40.4 (2025): 2703-2716.
10. Oberhauser, Hubert Josef. Bias in Artificial Intelligence: Exploring its role in institutional discrimination and strategies for mitigation. MS thesis. Universidade NOVA de Lisboa (Portugal), 2025.

11. Renuka, Oladri, et al. "Data privacy and protection: Legal and ethical challenges." Emerging threats and countermeasures in cybersecurity (2025): 433-465.
12. Gamasan, Reina, and Ramon George Atento. "Localized quality management system implementation and operational performance in a Philippine maritime manning office: The roles of human factors and organizational practices." Journal of Enterprise Strategy and Management Innovation (JESMI) 1.1 (2026).