This report examines the ethical considerations in using AI to analyze employee turnover at HumanForYou. It addresses data privacy, fairness, transparency, and accountability, ensuring responsible AI practices. Our approach prioritizes non-discriminatory analysis, employee well-being, and ethical decision-making to support fair and effective HR strategies.

AI for Employee Retention at HumanForYou

ETHICS REPORT

Table of contents

[1. Introduction 3](#_Toc193361800)

[1.1 Objective of the Report 3](#_Toc193361801)

[1.2 Relevance of AI Ethics in HR Analytics 3](#_Toc193361802)

[1.3 Ethical Framework Used: ALTAI 3](#_Toc193361803)

[2. Ethical Analysis Using ALTAI 4](#_Toc193361804)

[2.1 Respect for Human Autonomy 4](#_Toc193361805)

[ Avoiding intrusive decision-making 4](#_Toc193361806)

[ AI as a decision-support tool, not automation 4](#_Toc193361807)

[2.2 Technical Robustness & Safety 4](#_Toc193361808)

[ MODEL reliability and bias mitigation 4](#_Toc193361809)

[ ERROR minimization and ethical trade-offs 4](#_Toc193361810)

[2.3 Privacy & Data Governance 4](#_Toc193361811)

[ Data anonymization and security 4](#_Toc193361812)

[ Handling of sensitive attributes 4](#_Toc193361813)

[2.4 Transparency 4](#_Toc193361814)

[ Explainability of AI decisions 4](#_Toc193361815)

[ Interpretability methods used 4](#_Toc193361816)

[2.5 Diversity, Non-Discrimination & Fairness 4](#_Toc193361817)

[ Bias detection and mitigation 4](#_Toc193361818)

[ Fairness metrics and ethical considerations 4](#_Toc193361819)

[2.6 Societal & Environmental Well-being 4](#_Toc193361820)

[ Impact on workplace satisfaction 4](#_Toc193361821)

[ Computational efficiency and sustainability 4](#_Toc193361822)

[2.7 Accountability 4](#_Toc193361823)

[ Human oversight and responsibility 4](#_Toc193361824)

[ Ensuring HR professionals make the final decision 5](#_Toc193361825)

[3. Ethical Decision-Making Process 5](#_Toc193361826)

[3.1 Team Discussions on Ethical Concerns 5](#_Toc193361827)

[3.2 Justifications for Key Decisions 5](#_Toc193361828)

[ Feature inclusion/exclusion 5](#_Toc193361829)

[ Model selection and trade-offs 5](#_Toc193361830)

[ Limiting automation for fairness 5](#_Toc193361831)

[4. Conclusion & Recommendations 5](#_Toc193361832)

[4.1 Summary of Ethical Strengths 5](#_Toc193361833)

[4.2 Future Improvements and Monitoring 5](#_Toc193361834)

[5. Bibliography 5](#_Toc193361835)

ETHICS REPORT

AI for Employee Retention at HumanForYou

🔹Why is AI ethics essential in human resources?

AI-driven decisions in HR influence recruitment, promotions and dismissals. Ethical AI ensures fair treatment, greater transparency and respect for employees' rights, avoiding biases that could lead to discrimination.

# Introduction

## Objective of the Report

This report aims to analyse the ethical implications of using AI to predict employee turnover at HumanForYou. Using the Assessment List for Trustworthy AI (ALTAI) framework, we assess the compliance of our AI model with the ethical principles of autonomy, confidentiality, fairness, transparency, and accountability.

## Relevance of AI Ethics in HR Analytics

Integrating AI into human resources management presents opportunities but requires rigorous ethical analysis to avoid undesirable effects. By structuring our approach according to the ALTAI principles, we ensure that our AI system is both technologically robust and socially responsible.

## Ethical Framework Used: ALTAI

The Assessment List for Trustworthy AI (ALTAI) framework, developed by the European Commission, offers a structured methodology for assessing the trustworthiness of AI systems. The framework is grounded in seven fundamental principles: respect for human rights and governance, ensuring that AI respects fundamental rights and democratic values; technical robustness and security, ensuring that the system is reliable and resistant to attack; privacy and data governance, ensuring ethical handling of personal information; transparency, enabling the decisions made by AI to be understood; diversity, non-discrimination and equity, preventing bias and promoting inclusive use; societal and environmental well-being, taking into account social and ecological impacts; and finally, accountability, ensuring human supervision and recourse mechanisms in the event of problems. The application of this framework is pivotal in ensuring the continued effective functioning of our AI model in conjunction with the preservation of recognised ethical standards.

# Ethical Analysis Using ALTAI

## Respect for Human Autonomy

AI should not replace human decision-making but rather assist it. In HR, it is crucial to avoid intrusive decision-making that could lead to excessive employee monitoring. The **HumanForYou** AI model is designed as a **decision-support tool**, enabling HR managers to make informed choices while maintaining human oversight. AI should provide **predictive insights**, not enforce **automated employment decisions**.

## Technical Robustness & Safety

A reliable AI model must be resistant to errors and minimize biases. This requires:

* Rigorous data validation from HR records, surveys, and time-tracking logs.
* Bias detection techniques, ensuring attributes such as gender, marital status, or job role do not disproportionately influence predictions.
* Regular AI model audits, ensuring fairness over time.

## ****Privacy & Data Governance****

Protecting personal data is a critical concern. AI systems must ensure:

* Data anonymization, ensuring that all personal identifiers are removed or obfuscated to protect employee privacy, and implementing strict access controls to limit exposure to sensitive data.
* Secure storage of HR data, preventing unauthorized access.
* Strict compliance with regulations (e.g., GDPR for European employees).

Additionally, sensitive attributes such as gender or marital status must be analysed to prevent potential biases in predictive models.

## Transparency

The explainability of AI decisions is essential for building trust. **Interpretability methods**, such as **SHAP values or feature importance analysis**, help users understand how algorithms reach their conclusions. **Bias audits** should be conducted periodically to ensure AI-driven recommendations remain **clear and justifiable** to HR managers.

## Diversity, Non-Discrimination & Fairness

AI in HR must be designed to detect and mitigate biases. This includes:

* Fairness metrics (e.g., disparate impact analysis).
* Bias mitigation strategies, such as re-weighting biased training data.
* Regular audits to ensure fair treatment across different demographic groups.

For HumanForYou, ensuring that AI does not discriminate based on gender, education level, or travel frequency is crucial, as these attributes could disproportionately impact attrition predictions.

## Societal & Environmental Well-being

The impact of AI on employee well-being must not be overlooked. Ethical AI should:

* Enhance workplace satisfaction by identifying and addressing employee concerns.
* Promote work-life balance, ensuring AI does not encourage excessive work hours.
* Optimize energy efficiency to reduce environmental impact.

## Accountability

The final responsibility for decisions must always rest with human professionals. Human oversight is essential in AI-driven HR analytics, ensuring:

* HR managers can contest AI recommendations and override automated suggestions.
* Employees have access to appeal mechanisms, preventing unfair AI-driven outcomes.
* Regular reviews of AI decision-making, maintaining alignment with ethical principles.

# **Ethical Decision-Making Process**

## Team Discussions on Ethical Concerns

During the project, our team engaged in discussions to address the ethical concerns arising from the implementation of our tool in the HR analytics of HumanForYou. The main points discussed were data privacy, fairness, and transparency. In data privacy we need to ensure that employees data remains anonymised and protected from unauthorised access, for fairness, we need to analyse carefully if there are any features that can lead to discriminatory outcomes if we keep them (gender or martial status for example). Finally, on the transparency, we need to have outcomes that can be understandable and interpretable for HR professionals. Those discussions lead our approach to designing our AI model, ensuring compliance with fairness, accountability, and privacy principles.

## Justifications for Key Decisions

Features selection was a critical point on the ethical consideration. Some features were kept or excluded according to their impact on fairness and privacy. We kept the features more neutral and that stick to the ethical principles, for example “Years at Company” which is a key factor in attrition prediction that does not introduce bias, “Job Satisfaction” which provides valuable insight into employee retention likelihood “Performance Rating” which is justified as it reflects work-related performance metrics.

In another hand, we drop some attributes that can introduce a bias or doesn’t match the ethical conditions. In those attributes we have “Gender” and “Marital Status” which is excluding to prevent discrimination against certain demographic groups, “Exact Location Data” which is the precise addresses and useless since we can only use “distance from home” to protect privacy. The last dropped attribute is “Overtime” and “Late Check-ins” which is dropped to prevent reinforcing negative employment practices (e.g., penalizing employees for personal constraints).

Concerning the different models at our disposition to train our AI, we have evaluated them based on accuracy, interpretability, and fairness. We have tried five different models and only three were compliant to our ethical standards, the logistic regression model, the decision tree model, and the random forest model. We retained logistic regression model for its simplicity, transparency, and ease of interpretation, like our second model, the decision tree. The third model is the random forest that we choose for its balance between accuracy and interpretability, it provides robust feature importance analysis while mitigating overfitting risks.

The two models that we have not select is support vector classifier a K-nearest neighbours. The first one was not chosen for its complexity in tuning hyperparameters and the difficulty in interpreting its decision boundaries in a real-world HR context. The second one heavily relies on distance metrics, which could introduce location-based biases and lacks clear interpretability for HR decision-making.

Finally, to prevent AI from making autonomous hiring or termination decisions, we established clear human oversight mechanisms:

* AI serves as a recommendation system: HR managers receive AI insights but make the final decisions.
* Bias detection tools integrated: AI outputs are regularly audited for discriminatory patterns.
* Employee appeal process: Employees can contest AI-influenced decisions, ensuring fairness.

By limiting automation and maintaining human intervention, we ensure that AI supports rather than dictates employment decisions, aligning with ethical HR practices.

# Conclusion & Recommendations

Ethical AI in HR analytics must balance **accuracy, fairness, and transparency**. By adopting **trustworthy AI principles**, **HumanForYou** can improve employee retention strategies while ensuring **non-discriminatory practices**. Our approach prioritizes **bias mitigation, privacy protection, and human oversight**, reinforcing AI as a **support tool rather than a decision-maker**.

## Summary of Ethical Strengths

* Bias mitigation through careful features selection.
* Transparency ensured via interpretability methods.
* Human oversight in decision making.
* Privacy protection through anonymisation and data security.

## Future Improvements and Monitoring

* **Regular audits** to detect and mitigate bias.
* **Enhanced explainability** of complex models.
* **Continuous employee feedback loops** to improve AI recommendations.

# Bibliography

Interpretability: <https://www.ibm.com/think/topics/interpretability>

Ethical principles: <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>

Project: <https://github.com/equisdel/INDIAI/tree/main>