IS 467 Midterm Report

Ethical Layer of System Card+ Framework for Responsible AI Development.

Group 16

Jackson Song, Jakub Szumny, Rishabh Shah, Srinath Nandigam

Abstract

Artificial Intelligence’s rapid growth and usage in many industries require an adaptable framework for responsible development. System Card+ Framework for Responsible AI Development as outlined by Dr. Haileleol Tibebu and Ioannis Kakadiaris provides just that, it provides a modular theoretical framework to evaluate Artificial Intelligence. This framework contains a five-layer benchmarking system, including performance accountability, fairness, inclusivity, ethical compliance, and legality. This paper will address the ethical compliance layer of this system, particularly the Data category. In this category it will discuss eight related research works regarding this topic. Then it will discuss the dataset that will be applied to this project. It will discuss the pre-processing steps taken and how it is applied to this project. The dataset we will be using analyzes financial risk assessment, compliance violations, and fraud detection trends among the Big 4 consulting firms from 2020 to 2025, with a focus on the impact of AI in auditing. It provides valuable insights into audit effectiveness, workload management, and ethical data practices. This paper will then present an exploratory data analysis of the aforementioned dataset. It will identify any trends, outliers, and distributions while discussing potential biases. Finally, it will discuss the plan for future work to be performed on this dataset, methodology, implementation strategies, and any key milestones in upcoming steps.

Introduction

Artificial Intelligence has become an integral part of modern industries, helping to shape decision-making processes in many fields such as finance, healthcare, education, and governance. As AI is used more often and as it expands, it is critical to ensure it is ethical and responsible. Ethical concerns surrounding AI primarily is from issues related to bias, fairness, transparency, and accountability. Without the proper ethical safeguards, these AI systems risk discrimination, which can reduce trust and violate privacy rights. To address these concerns, we need comprehensive frameworks that integrate ethical evaluation into the development of AI in its earliest stages.

The System Card + Framework for Responsible AI Development, proposed by Dr. Haileleol Tibebu and Ioannis Kakadiaris, offers a structured approach to assessing AI systems. Their framework consists of five layers: performance accountability, fairness, inclusivity, ethical compliance, and legality. Our research focuses on the ethical compliance layer, particularly with the Data category. Ethical data practices are fundamental to responsible AI, as they influence the quality, fairness, and even the interpretability of AI-driven outcomes. By analyzing the data used in AI systems, we can identify biases and assess transparency.

To explore these ethical considerations, we will be using the “Big 4 Financial Risk Insights (2020-2025) dataset from Kaggle. This dataset provides financial risk data from major global firms, offering valuable data for assessing ethical challenges in AI-driven decision-making in the field of finance. Our analysis will include dataset preprocessing, exploratory data analysis, and a discussion of potential biases in the data. We will also be examining how fairness and transparency can be integrated into the data-handling process to align AI systems with more ethical standards. By adding ethical checks at the data level, we aim to contribute to the development of AI systems that perform and are fair, transparent, and responsible. Furthermore, we will discuss key methodologies applied to the dataset, including preprocessing techniques such as data cleaning, handling missing values, and feature engineering. Our exploratory data analysis will help uncover trends, outliers, and distributions in the dataset while also evaluating any potential ethical concerns such as an imbalance in the data, and representational bias. We will use Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn to conduct statistical and visual analyses that visualize these trends effectively.

The insights we gain from this research will inform ethical best practices in AI data handling, specifically in financial risk assessment models. By adding these ethical considerations to our data analysis and modeling processes, we hope to provide a framework that enhances fairness and accountability in AI-driven financial decision-making. Our research will conclude with a discussion of future work, outlining plans for further analysis, model implementation, and more key milestones in the development of an ethical AI system.

Literature Review

As artificial intelligence becomes more common in vital areas like healthcare, finance, education, and governance, addressing ethical issues related to data handling is becoming increasingly important. With AI influencing so many aspects of people's daily lives, researchers have highlighted several key ethical concerns that need careful attention: bias and fairness, transparency and explainability, and consent combined with inclusive data representation.

Bias and fairness in AI systems remain a major concern. When AI is trained on biased data, it can lead to unfair decisions, reinforcing societal inequalities. Mehrabi et al. [5] explored different techniques to detect and reduce bias, including demographic parity, disparate impact analysis, and counterfactual fairness. Their study emphasized the importance of addressing bias early in the data collection and preprocessing phases to prevent harmful outcomes. Hanna et al. [3] highlighted how AI can misrepresent or exclude certain groups if datasets are not diverse. They found that many AI systems produce unfair outcomes because they fail to include data from different racial, gender, or socioeconomic backgrounds. Their “Critical Race Methodology” encourages developers to actively incorporate diverse perspectives in AI training datasets to reduce bias. Similarly, Binns et al. [1] emphasized that fairness in AI should be continuously monitored rather than assessed only once. They argued that ethical AI requires ongoing evaluation to ensure fairness, as biases can evolve. Together, these studies show that fair AI starts with diverse, well-balanced datasets and requires regular assessments to minimize bias.

Transparency and explainability are crucial for ethical AI. Users, developers, and policymakers need to understand how AI makes decisions to ensure accountability and trust. Mittelstadt et al. [6] stressed that AI must be transparent to build confidence among stakeholders. Their research found that when AI decisions are unclear, users often feel frustrated and distrust the system. Whittlestone et al. [7] expanded on this by discussing the tension between AI efficiency and ethical accountability. They found that while AI can process information quickly, it must also provide clear reasoning for its decisions to maintain public trust. Floridi et al. [2] further argued that AI cannot be truly ethical without proper governance and data-sharing policies. They warned that AI systems could easily become unreliable and untrustworthy without strict guidelines on how decisions are made. These studies suggest that AI should not operate as a “black box” but rather provide explanations that help users understand and trust its outputs.

Consent and inclusive data representation have also become central to ethical AI discussions. Ethical AI requires clear communication about how user data is collected, stored, and used. Binns et al. [1] studied public concerns regarding data privacy and found that many users feel powerless when their personal information is used without proper consent. Their research emphasized the importance of transparency in obtaining informed consent. Jobin et al. [4] reviewed global AI ethics guidelines and highlighted best practices such as data minimization, explicit user permission, and continuous communication with individuals about their data. They stressed that AI developers must clearly explain what data they collect and how it will be used. Additionally, Hanna et al. [3] argued that inclusive representation in AI is essential for fair decision-making. They found that AI often performs poorly for underrepresented groups because datasets lack diversity. To address this, they recommended that AI systems be trained on balanced and representative datasets to ensure equitable treatment of all users. These studies collectively show that informed consent and diversity in data representation are necessary to create AI systems that are both ethical and effective.

Despite significant progress in AI ethics, major gaps remain, particularly in how early ethical considerations are implemented. Fairness, transparency, and consent are often considered too late and can be found only after major data decisions have already been made. This delay makes it harder to correct ethical problems and increases the risk of unintended consequences. Early choices in AI development set the foundation for fairness, accountability, and user trust. If bias is present in training data, AI systems will learn and reinforce those biases, leading to discriminatory outcomes [5]. If transparency is not prioritized from the start, later efforts to explain AI decisions may be vague or ineffective, leaving users confused about how the system functions [6]. If user consent and data privacy are not properly addressed early on, AI applications may violate user trust and fail to comply with legal standards [4].

The consequences of these gaps are significant. AI systems that lack early ethical safeguards can contribute to discrimination, reduce public trust, and create unintended societal harm. For example, biased facial recognition systems have disproportionately misidentified individuals from certain racial backgrounds, leading to wrongful arrests and systemic discrimination [3]. Similarly, AI-powered hiring tools have been found to favor specific demographic groups over others, perpetuating workplace inequalities [7]. These ethical failures become deeply embedded in AI-driven systems, making them difficult to correct after deployment.

To prevent such issues, researchers emphasize the importance of embedding ethical evaluations from the beginning of AI development. Mittelstadt et al. [6] stressed that proactive transparency measures help reduce ethical blind spots, making AI systems easier to regulate and audit. Jobin et al. [4] found that AI ethics guidelines should be integrated into the design phase to ensure fairness and privacy considerations are not overlooked. Whittlestone et al. [7] emphasized that AI decision-making must be an ongoing process, where fairness, transparency, and consent are continuously reassessed at multiple stages of system development to prevent ethical deterioration over time.

Governments and regulatory bodies are also recognizing the importance of integrating ethics early in AI design. Policies such as the European Union's AI Act and the United States' AI Bill of Rights advocate for strict regulations on AI fairness, transparency, and data privacy [4]. These initiatives emphasize the importance of legal accountability alongside ethical responsibilities, pushing companies and developers to prioritize responsible AI practices. By following these emerging legal frameworks, AI researchers and engineers can help create systems that align with both ethical and legal expectations, ensuring their technology benefits society as a whole.

Our paper aims to address these challenges by integrating ethical checks at the earliest stages of data processing. By embedding fairness evaluations, transparency requirements, and consent mechanisms early in development, we ensure AI systems adhere to ethical principles and regulatory standards. This approach minimizes the risks of bias, lack of transparency, and privacy violations. Ultimately, it contributes to the development of AI systems that are fair, trustworthy, and aligned with societal expectations.

Dataset Selection and Preprocessing

The dataset used in this project focuses on financial risk assessment, compliance violations, and fraud detection trends among the Big 4 consulting firms (Ernst & Young, PwC, Deloitte, and KPMG) from 2020 to 2025. The dataset captures key metrics including the number of audit engagements, high-risk cases, detected fraud cases, and compliance breaches. Additionally, it provides insights into the impact of AI on auditing, employee workload, and client satisfaction scores.

This dataset was chosen due to its relevance to the Data category within the Ethical Layer for AI-Based Decision Support Systems. The ethical challenges associated with data collection, processing, and utilization are particularly significant in financial auditing, where transparency, accountability, and fairness are crucial. By analyzing this dataset, the project aims to assess how AI-driven auditing practices align with Responsible AI standards, focusing on ethical data sourcing, consent protocols, and data diversity.

The dataset's comprehensive nature and focus on high-profile consulting firms make it valuable for financial analysts, auditors, data scientists, and risk managers. Additionally, it facilitates comparative analysis across various industries, including finance, technology, retail, and healthcare, allowing for broader insights into ethical auditing practices.

Preprocessing is an essential step to ensure that the dataset is accurate, consistent, and suitable for analysis. Given the complexity and sensitivity of financial data, thorough preprocessing is necessary to address potential issues related to data quality and integrity. The key preprocessing steps undertaken in this project are outlined below:

Missing values are common in financial datasets and can lead to biased results if not handled properly. In this project, numerical missing values were imputed using the median, as it is less sensitive to outliers compared to the mean. Categorical missing values were addressed using the mode to preserve data integrity. This approach ensures that the imputation does not distort the original data distribution. Duplicates and inconsistent records, particularly within audit engagement data, were identified and removed to maintain the accuracy and credibility of subsequent analyses. Ensuring data accuracy at this stage is vital as inaccuracies can significantly impact the reliability of subsequent models and analyses.

The selection of relevant features is vital to ensure that the analysis remains focused on ethical risk assessment and audit effectiveness. Features such as the number of audit engagements, high-risk cases, detected fraud instances, compliance breaches, and AI impact metrics were prioritized due to their relevance to the project objectives. These features provide a comprehensive view of financial risk management and the role of AI in auditing practices. Features that were deemed redundant or irrelevant, including non-informative categorical variables, were systematically excluded to streamline the analysis and reduce computational complexity. This enhances the model's efficiency and interpretability by focusing solely on variables that contribute to ethical and financial insights.

To ensure compatibility with analytical models, categorical variables (e.g., consulting firm names and client industry sectors) were encoded using one-hot encoding. This method effectively converts categorical data into binary vectors, maintaining interpretability while enabling model compatibility.

Numerical variables, including metrics related to workload and client satisfaction, were normalized to standardize their scales. This normalization mitigates the risk of bias during model training, especially when combining variables of varying magnitudes. Additionally, scaling ensures consistency when comparing metrics across different firms and industries.

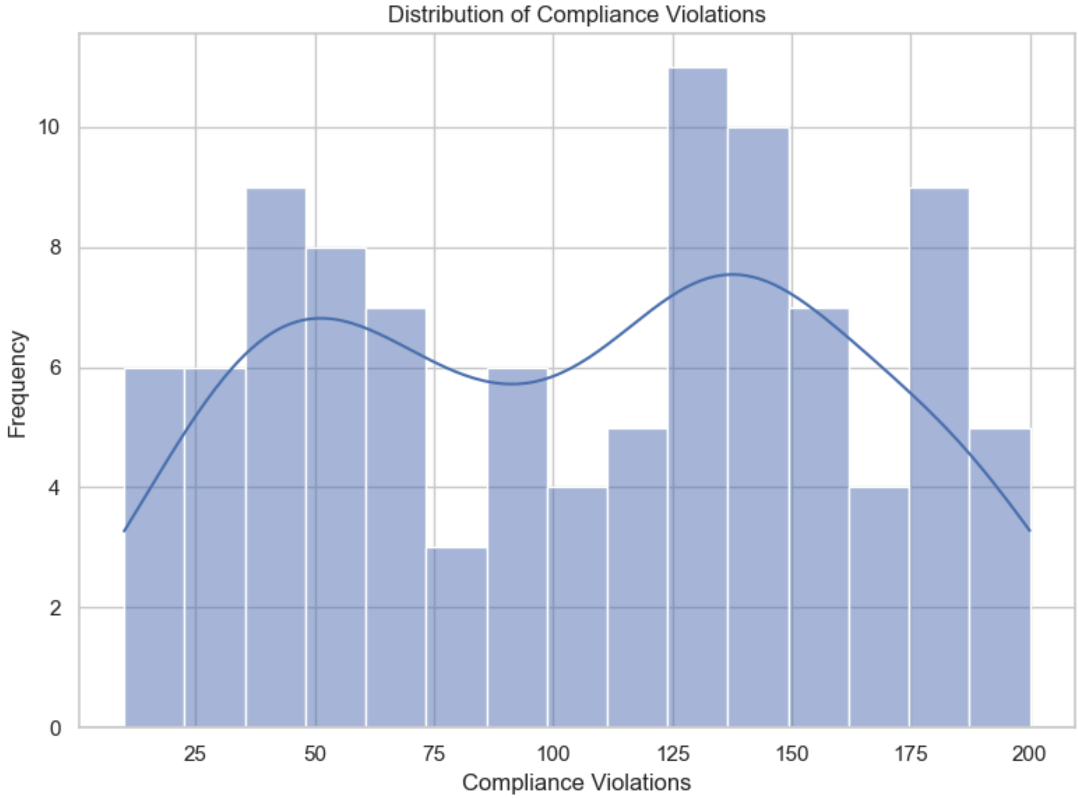
Ethical preprocessing practices were followed to ensure compliance with data privacy and fairness standards. Consent verification was conducted to ensure that data usage aligns with the original terms under which it was collected. Additionally, diversity checks were performed to ensure adequate representation across different industry sectors and firm sizes, thereby minimizing potential biases in the analysis. Ethical data management practices are particularly crucial when analyzing high-stakes financial data to maintain transparency and accountability.

By implementing these preprocessing steps, the project guarantees that the dataset is accurate, consistent, and responsibly prepared for further analysis. This careful approach aligns with the principles of Responsible AI by upholding transparency, fairness, and accountability throughout the data handling process. Through meticulous preprocessing, the project ensures that ethical considerations are embedded at every stage of data analysis, fostering trust and reliability in AI-driven auditing practices.

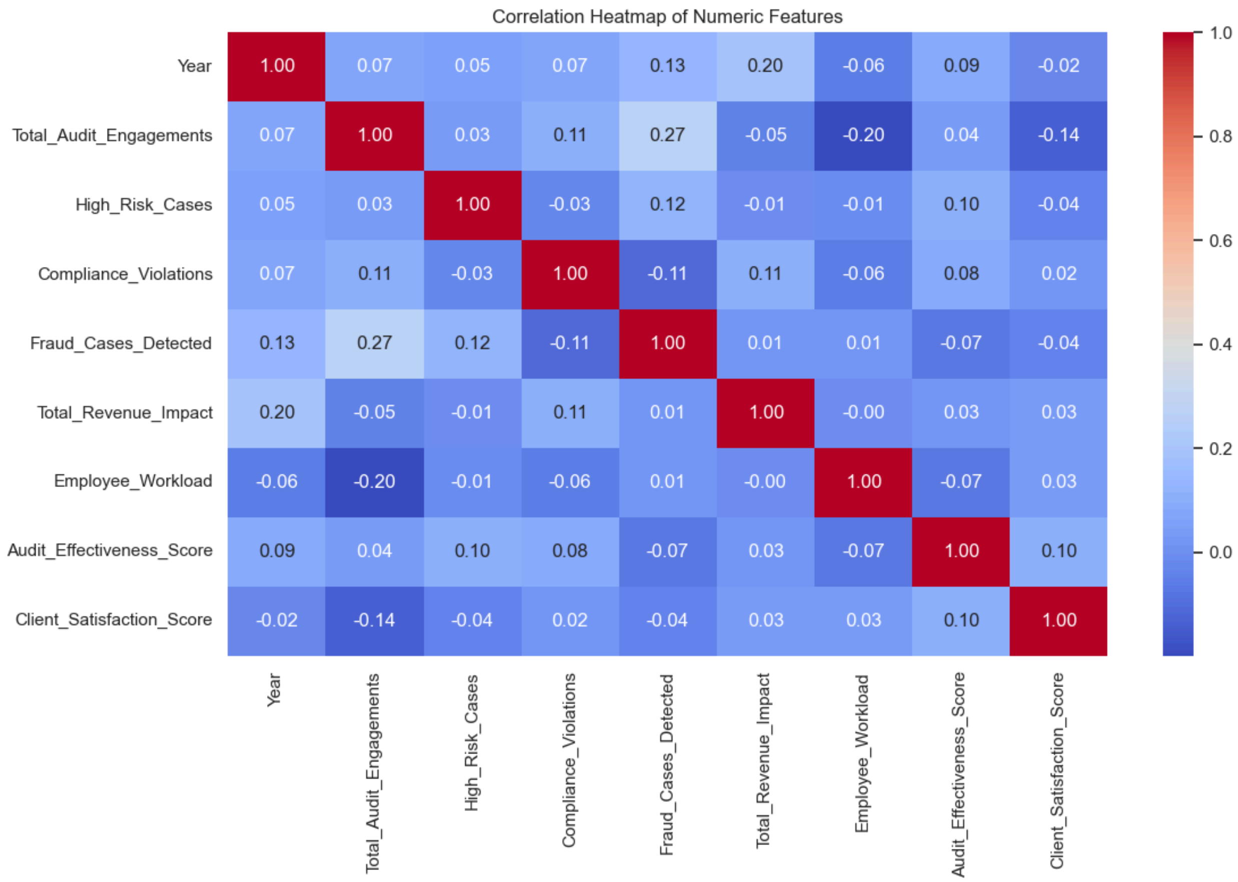
Exploratory Data Analysis (EDA)

To better understand the ethical issues in our dataset, we looked closely at the "Big 4 Financial Risk Insights (2020–2025)" dataset from Kaggle. It includes 100 rows of data from the world's biggest auditing firms like PwC, Deloitte, EY, and KPMG. Each row represents a year's worth of data and includes important features like how many audits were done, how many were high-risk, how many compliance violations happened, and whether or not AI was used. It also includes things like fraud cases, revenue impact, and how satisfied clients were. This kind of data helps us spot problems early, especially when it comes to fairness, transparency, and accountability in auditing with AI.

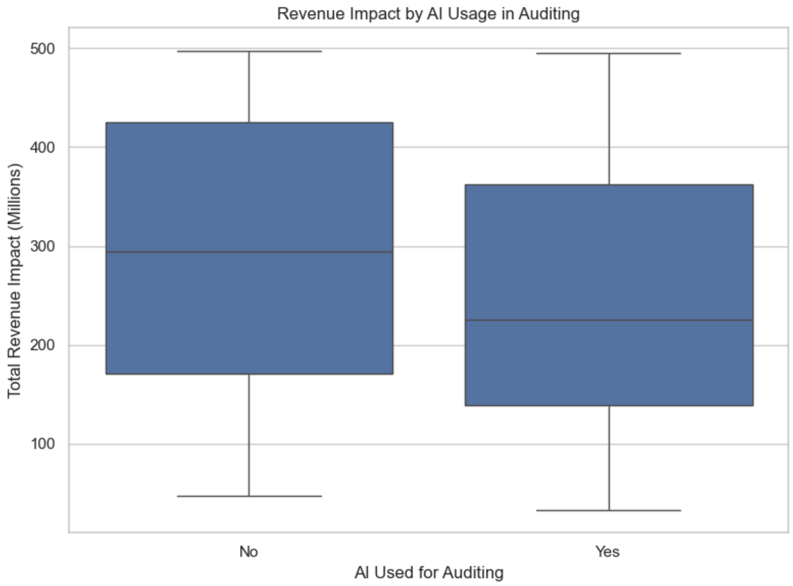
One of the first things we looked at was the number of compliance violations. This is important because it tells us how often these firms are breaking the rules. We made a histogram (shown below) to see how the violations are spread out across the dataset. Most companies had between 30 and 120 violations per year, but there were a few cases with way more, over 150! These outliers are a big deal. They might mean that certain firms have deeper problems, and it's something we want to look at more when we talk about bias and accountability.



Next, we made a heatmap (see below) to find out how different features in the dataset are related to each other. For example, we saw that compliance violations were strongly connected to both high-risk cases and fraud cases. That makes sense, if a firm is handling a lot of risky audits, there's a better chance they'll make mistakes or break rules. Another important finding was that audit effectiveness had a negative relationship with compliance issues. In simple terms, better audits usually meant fewer violations. This tells us that doing a good job in the audit process matters when it comes to being ethical.



We also wanted to see if AI had any impact on financial outcomes, so we made a boxplot (shown below) comparing firms that used AI with those that didn't. It turns out that firms using AI in their audits had a higher median revenue impact, meaning they were catching more important things or being more efficient. But we also noticed that the results varied a lot more when AI was used. This shows us that while AI can be helpful, it doesn't always guarantee success. From an ethical point of view, that could be risky, especially if a company relies too much on AI without understanding its limits.



Lastly, we checked out which industries and firms showed up the most in the dataset. We noticed that the healthcare and finance industries were more common than others, and PwC had more entries than some of the other firms. This could cause problems if an AI model trained on this data thinks that the way one industry or one company behaves is “normal.” If we’re not careful, that could lead to unfair results for industries or firms that weren’t well-represented in the data.

In conclusion, this EDA helped us find several important ethical risks in the data. We found outliers in compliance violations that could mean some firms are being less responsible. We saw strong relationships between risky audits and violations, showing how important good auditing is. And we learned that AI can help, but it also brings more unpredictability. These insights will help guide the rest of our project as we build out our evaluation criteria. By making sure we catch these problems early, we can work toward creating AI systems that are more fair, transparent, and ethical, especially in important fields like finance and auditing.

Proposed Plan for the Remaining Work

Moving forward, our research will focus on refining our ethical analysis of AI in financial auditing by implementing new methodologies for bias detection, fairness evaluation, and also for transparency assessment. The next steps will have three main phases which are dataset refinement, ethical benchmarking, and model evaluation.

While we have done an initial exploratory data analysis on our chosen dataset, we still need to do further work to enhance the data quality and eliminate any possible ethical risks. To do this we will implement more advanced data cleaning techniques for any issues with the data such as missing values, outliers, and inconsistencies, all of which affect fairness assessments. We will also look into any class imbalances, specifically in compliance violations, fraud cases, and audit risk categorizations to determine if any of these categories are underrepresented. Also, we plan to apply feature engineering, to extract any additional insights, such as aggregating financial risk trends over time and normalizing variables to try to reduce bias in a model's predictions.

To ensure that our AI financial analysis aligns with ethical standards, we will make benchmarking criteria based on fairness, transparency, and accountability. We will need to use statistical fairness metrics, such as equalized odds to identify biases in the audit outcomes. Also we will analyze the subgroups to try and determine whether AI models are disproportionately affecting certain industries and firms. We will also need to evaluate how the AI’s involvement in audits can influence any patterns of decision-making and whether there are any strong biases.

To further explore ethical concerns, we plan to implement an AI model to analyze financial risk and assess its ethical implications. This will involve model development, so we will need to train a machine learning model, possibly using logistic regression, on the chosen dataset to try and predict audit risks and compliance violations. Also we will try to compare model performance between AI and non-AI audits to see if the AI introduces any disparities in the financial risk assessments. We will have to evaluate the model's interpretability to improve the transparency in AI decision-making.

The final proposed stage of our plan is to use our findings to create ethical recommendations for AI implementation in financial risk assessment. To do this we will have to summarize any key insights from our analysis and model evaluations, propose more guidelines for the improvement of fairness and accountability in the AI audits, and lastly identify any regulatory and governance measure that can enhance ethical compliance in these financial AI applications.

By following this plan, we aim to develop an ethical framework that ensures AI in financial auditing is more transparent, fair, and responsible. Our research will hopefully contribute valuable insights into how AI can be bettered without not following any ethical standards in financial decision-making.

Group Work Distribution

Worked together on researching and did 2 pages each for the paper

Srinath – Abstract + Data Selection and Pre-Processing (2 pages total)

Jakub - Introduction + Proposed plan for remaining work (2 pages total)

Jackson - Literature review (2 pages total)

Rishabh - Exploratory Data Analysis (2 pages total)

Reference list

[1] Binns, Reuben, et al. “It’s Reducing a Human Being to a Percentage.” *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18*, 2018, pp. 1–14, doi.acm.org/10.1145/3173574.3173951, https://doi.org/10.1145/3173574.3173951.

[2] Floridi, Luciano, and Mariarosaria Taddeo. “What Is Data Ethics?” *Academia.edu*, 3 Dec. 2016, www.academia.edu/30234860/What_is_Data_Ethics.

[3] Hanna, Alex, et al. “Towards a Critical Race Methodology in Algorithmic Fairness.” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 22 Jan. 2020, arxiv.org/pdf/1912.03593.pdf, https://doi.org/10.1145/3351095.3372826.

[4] Jobin, Anna, et al. “The Global Landscape of AI Ethics Guidelines.” *Nature Machine Intelligence*, vol. 1, no. 9, 2 Sept. 2019, pp. 389–399, https://doi.org/10.1038/s42256-019-0088-2.

[5] Mehrabi, Ninareh, et al. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys*, vol. 54, no. 6, July 2021, pp. 1–35, dl.acm.org/doi/10.1145/3457607, https://doi.org/10.1145/3457607.

[6] Mittelstadt, Brent, et al. "Explaining Explanations in AI." *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 2019, arxiv.org/pdf/1811.01439.pdf, https://doi.org/10.1145/3287560.3287574.

[7] Whittlestone, Jess, et al. "The Role and Limits of Principles in AI Ethics." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 27 Jan. 2019, https://doi.org/10.1145/3306618.3314289.

[8] A. Soundankar, "Big 4 financial risk insights (2020-2025)," Kaggle, https://www.kaggle.com/datasets/atharvasoundankar/big-4-financial-risk-insights-2020-2025 (accessed Mar. 23, 2025).