Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building Ethics into Artificial Intelligence. *ArXiv*. /abs/1812.02953 <https://doi.org/10.48550/arXiv.1812.02953>

The Belmont Report is the starting point for ensuring ethics in artificial intelligence. In the Belmont Report there were principles stated that would ensure that ethics would sustain in human-AI interactions. The principles include three main requirements, those being: people’s personal autonomy should not be violated or in other words, they should be able to maintain their free will when interacting with the AI. Another requirement is that the benefits brought about by the technology should outweigh the risks. The final requirement is that the benefits and risks should be fairly distributed fairly among the users, which means that people should not be discriminated against based on their personal backgrounds such as race, gender, and religion. For ethics to be built into every artificial intelligence, then ethics should be part of the AI curricula.

Huang, C., Zhang, Z., Mao, B., & Yao. (2023). An Overview of Artificial Intelligence Ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799–819. <https://doi.org/10.1109/tai.2022.3194503>

This paper does an overview of the field of AI by summarizing and analyzing the ethical issue raised by AI, ethical guidelines and principles issued by different organizations, approaches for addressing ethical issues in AI, and methods for evaluating the ethics of AI.

Wang, Z., Huang, C., & Yao, X. (2023). *Feature Attribution Explanation to Detect Harmful Dataset Shift*. <https://doi.org/10.1109/ijcnn54540.2023.10191221>

This paper talks about detecting whether a distribution shift has occurred in a dataset of a machine learning model is important for the application and the performance of the model. If the model has too large of a shift in the data distribution it could make the deployed model fail. Data shifts can be defined in 3 separate categories. Those ways are Covariate Shifts: Changes and differences in the distribution of input variables between training and test data. Label Shift (Prior Probability Shift): Changes and differences in the distribution of target variable (i.e., class variable) between training and test data. And Concept Shift: Changes in the relationship between input variables and class variables. How they measured this change is by indicating how much each input contributes to the models output for a given data point and then checking if a input contributes too much to an models output. They found that their method of checking a data shift set is successful and is more effective than other state of the art methods.

Gagandeep, Kaur, J., Mathur, S., Kaur, S., Nayyar, A., Singh, S. P., & Mathur, S. (2023). Evaluating and mitigating gender bias in machine learning based resume filtering. *Multimedia Tools and Applications: An International Journal*, 1–21. <https://doi.org/10.1007/s11042-023-16552-x>

With the recent rise in everyday use of AI, companies have started to automate the process of sorting resumes by having an algorithm pick which resumes are classified as suitable enough to get to the next step. However, these training systems often incorporate high social biases in the models. This is because these filtering systems are trained on historical data which often happens to be men’s resumes that they use for this. This paper introduces a method to hide specific gender terms from data(Gender Masking) before finding the similarity with the job requirements. Some methods they theorized and tested is hiding gender specific information from the dataset, assign similarity scores to every application based on other applications to the same position, and assigning a similarity score to every application based on the job requirements in the job listing. They found that implementing any of the methods in a system that was trained on data that implemented Gender Masking, reduced the amount of gender bias in a model

Szczekocka, E., Tarnec, C., & Pieczerak, J. (2022). Standardization on Bias in Artificial Intelligence as Industry Support. *2022 IEEE International Conference on Big Data (Big Data), Big Data (Big Data), 2022 IEEE International Conference On*, 5090–5099. <https://doi.org/10.1109/BigData55660.2022.10020735>

This paper talks about standardizing the concepts of “Responsible AI” and tries to define the types of biases in Artificial Intelligence. Responsible AI is a concept that is difficult to define but related to the question of if AI is trustworthy and scalable? Some research defines the principles of responsible AI as: ethics, transparency, regulation and control, socioeconomic impact, design, and responsibility. While IBM considers explainability, fairness, robustness, transparency, and privacy as AI ethic principles and Microsoft has defined their six principles of responsible AI as: fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. These similar but differing principles by many different groups bring cause to standardizing these principles.

This paper also wants to define the types of bias that exist in artificial intelligence. For example, ISO proposed that several types of biases exists which consists of human cognitive bias: human bias that might impact the design and application of a system, data bias: data properties that if unaddressed lead to AI systems that perform better or worse for different groups, machine learning model architecture bias, and other biases that do not fall within the other categories which includes: bias in system design, requirement bias, and statistical bias.

L. Valtonen and S. J. Mäkinen, "Exploring the Relationships between Artificial Intelligence Transparency, Sources of Bias, and Types of Rationality," 2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Kuala Lumpur, Malaysia, 2022, pp. 1296-1300, doi: 10.1109/IEEM55944.2022.9989994.

This report researches the way that AI impacts a person’s choices. The experiment that this report does is to find the difference in the decision-making reasoning between an AI-guided environment and a human guided environment. They found that in the AI-guided environment, there tended to be over-reliance or authoritative stigmatization where AI is concerned. They concluded that AI is seen as authoritative in high and low stake decision making settings and that it does not suppress choice but if there is lack of transparency AI can suppress rationality in the decision maker.

El Annas, M., Benyacoub, B., & Ouzineb, M. (2023). Semi-supervised adapted HMMs for P2P credit scoring systems with reject inference. *Computational Statistics*, *38*(1), 149–169. <https://doi.org/10.1007/s00180-022-01220-9>

This paper focuses on the concept of reject inference focusing specifically on the application into P2P credit scoring. Reject inference is the concept of including the rejection data into the data set to avoid selection bias. This is because many credit card models that is used to approve who gets a loan or not does not include rejection data in their data sets. This can cause selection bias by the algorithm to only accept a certain selection of people. This can unintentionally cause discrimination in who gets a loan accepted. This paper proposes a new solution to this problem by proposing a semi-supervised hidden Markov model to evaluate the usage of semi-supervised machine learning for reference inference in credit scoring. They generated several samples with varied rejection rates and then tested them and found that the test passed in applicability, stability, and performance.

Anderson, B., Newman, M. A., Grim II, P. A., & Hardin, J. M. (2023). A Monte Carlo simulation framework for reject inference. *Journal of the Operational Research Society*, *74*(4), 1133–1149. <https://doi.org/10.1080/01605682.2022.2057819>

This paper uses a Monte Carlo simulation to simulate rejected applicants in credit scoring models. A Monte Carlo simulation is a mathematical technique that predicts possible outcomes of an uncertain event. This can be used to analyze past data and predict a range of future outcomes based on course of action. They created a methodology to illustrate how to simulate rejected applicants so that reject inference techniques can be studied and selected to reduce the amount of selection bias in algorithms. They created two Monte Carlo simulations; one uses a accepts:rejects ratio for each reject inference technique and the other model uses the same methodology as the first simulation and studies the performance of the three reject inference methods. The simplest reject inference method is called simple augmentation which builds a model on accepted applicants and applies this model to rejected applicants. The next reject inference method is parceling which is a hybrid method that combines proportional assignment and augmentation. Proportional assignment is the random portioning of rejects into good and bad classifications based on the number of good and bad accepted applicants within the scoring parcels. The final reject inference method is called fuzzy augmentation. This method uses rejected applicants twice in the final model. This rejected applicant is separated into two parts: a partial good and a partial bad which are generated in the form of a weight variable that has two observations. One observation has a good outcome loan status and the other has a bad outcome loan status. They found that the first Monte Carlo simulation with simple augmentation, AUC, accuracy, and sensitivity produced the same accepts:rejects ratio with the number of rejects being close to the ratio in the observed data. For Parceling and Fuzzy augmentation, there was an opposite trend in the accepts:rejects ration for AUC as compared to simple augmentation. They found that the more rejected applicants are used in simple augmentation , the better the AUC. They concluded that the Monte Carlo simulation can be utilized to access reject inference.

**AI Models**

<https://medium.com/@burak.bakkaloglu/ai-algorithms-in-hiring-processes-f54c7f899d93#:~:text=Most%20of%20the%20hiring%20AI,Supervised%2C%20unsupervised%20or%20reinforced%20learning>

Talks about how Hiring AIs are trained

Gender Masking AI Algorithm

* KNN(K- Nearest Neighbor) Algorithm
  + Supervised Learning
  + Makes no data Assumptions
  + Uses proximity to make classifications or predictions about the grouping of an individual data point