Equalmodel

Team Email: eseseniordesign202021team06@gmail.com

Brian Handen, bhanden@seas.upenn.edu, NETS

Tashweena Heeramun, htash@seas.upenn.edu, SSE

Qingrong Ji, qji@seas.upenn.edu, NETS

Hyewon Lee, hyewon98@seas.upenn.edu, SSE

Trishla Pokharna, trishla@seas.upenn.edu, EE/Wharton

Advisors: Aaron Roth, aaroth@cis.upenn.edu; Chris Jung, chrisjung.sy@gmail.com

Equalmodel: A Solution to Implicit Bias in AI

Value Proposition:

Artificial intelligence (AI) is disrupting virtually every industrial sector as we know it. From automating fraud detection in the financial sector to revolutionizing healthcare with more accurate disease diagnosis, AI is here to stay. However, alongside this wave of innovation, algorithmic bias is coded into consequential artificial intelligence models, perpetuating societal, racial, and gender prejudices. These biases are not necessarily created by racist programmers that write explicit lines of code to generate unfair algorithms; rather, the bias is implicit through years of inequitable data collection that reflects long standing inequalities that exist in society. When the data does not reflect the population (for instance, when facial recognition software cannot identify people of color after being trained on a data set that has too many white males), the models spit out answers that favor the data, not necessarily the true population. This is where much of the controversy with AI/ML lies — these "black-box" algorithms that do not explain why or how they make their decisions can dictate much of a person's life, from what type of loan they receive to how much pre-trial bail they must pay. This phenomenon has led to the resurgence of a belief that decision making should be left solely to humans. Clearly, society is divided about how firms that do choose to utilize AI should be regulated and, more broadly, when AI should even be used.

Recently, healthcare has come into the forefront of the AI debate, because an algorithm used by hospitals and insurers involved in the care of 200 million Americans was recently discovered to refer care to White Americans over Black Americans. In fact, the healthcare given to Black Americans was reported to be on average $1,800 less than to a White American with the same number of chronic health problems under this widely adopted algorithm. Hence, there is a sore need of a technology and product that can ascertain that the algorithms that touch so many peoples' lives are not a source of racial, gender or other biases.

In the sector of healthcare insurance, the usage of AI can greatly impact the lives of everyday people. For example, for the case of accepting or rejecting a claim, using an AI-based system can help identify which claim amounts are "unusual" and help with the current cumbersome process of manual examinations by auditors. Since these auditing processes take a long time, the intervention by AI can greatly speed up these processes for both the insurer and the insured.

Additionally, customers may also end up paying more for insurance because policies are not tailored for their unique needs. AI can be used to price insurance policies more competitively and relevantly and recommend useful products to customers. Insurers can price products based on individual needs and lifestyle so that customers only pay for the coverage they need. This increases the appeal of insurance to a wider range of customers, some of whom may then purchase insurance for the first time.

Especially with President Biden's updated mandate of incurring a penalty for not being covered under health insurance and alongside the existence of COVID-19, increasing numbers of people are utilizing healthcare services and enrolling in insurance plans. Therefore, having access to more personalized, accurate insurance policies tailored by artificial intelligence will greatly benefit individuals who might be reluctant to become insured otherwise due to cost or other barriers.

Since health insurance and policies are personal and significant to many individuals, there is a need to certify that the algorithms underlying the AI are free of bias. We want to take steps to eliminate discrimination and disparities in AI that predicts "risk factors" when generating estimated costs for new patients' policies based on race, background, and more.

Our product, Equalmodel, will solve this problem as a post-processing software package for implicit bias reduction. Anyone developing AI based models for human applications can use our AIaaS (Artificial Intelligence as a Service) for continuous monitoring of their models to quantify how uncertain a prediction is for a certain demographic or subgroup. Equalmodel can even adjust predictions the consumers' models make to ensure that the prediction reflects the true answer of these subgroups, no matter how small the group is and thus eliminating discrimination against any demographic. We provide a platform and service for users to easily learn how to integrate Equalmodel onto their products and advise consumers that do not have the knowhow to diagnose bias and utilize the product. This service will audit the consumers' entire data collection and analysis process to ensure proper fairness in their models, following all necessary compliance guidelines under the Health Insurance Portability and Accountability Act (HIPAA). In short, our solution is two-fold: first, we identify bias so that humans know when it is necessary to be involved in a decision that AI cannot make; second, we adjust the model to remove bias in the predictions. This product can be built on top of any AI model, removing one of the biggest hurdles in AI adoption and thus making the use of our model much more ubiquitous.

Stakeholders:

Companies that use AI technologies with societal impact are stakeholders in this space. It is important to note that rather than stating that these AI firms have a *responsibility* toward their respective stakeholders about the ethics of their algorithms, we believe that a better and more compelling term for firms is *risk mitigation*. Corporations want to minimize risk, mainly the risks incurred by the revelation of any hidden or implicit biases, which not only triggers legal action, restricts the use of AI, and damages the firm's reputation, but also greatly harms the people affected by these biases.

Within the myriad of fields where AI is being used, Equalmodel aims to initially target the healthcare sector through focusing on providing services to companies that provide AI diagnostic services to health insurance providers that use AI, forming a specialized and trustworthy product for the niche group of stakeholders. Healthcare, of the many sectors that we considered, has the largest need and opportunity for our product because uncertainty estimation is necessary to gauge risk in clinical decisions and insurance policies. Therefore, stakeholders who are worried about exposure to expensive legal risks and do not have the means to procure more fair datasets (which is an expensive task and sometimes impossible to do) will benefit the greatest with this product. Equalmodel will provide third party verification, which establishes a sense of trust among customers who use our stakeholders' products and potentially attracts even more customers for our stakeholders who were initially hesitant about utilizing AI.

Furthermore, Equalmodel has the potential to venture into sectors apart from healthcare in order to branch out our scope in the future. Areas such as real estate, loans, and mortgages are increasing their usage of AI; for example, AI is being used in lead generation for finding potential buyers that fit with an agent's listing, loan predictions to determine who would be able to pay back their debt to credit unions and banks on time, and the process of providing mortgage rates based on one's financial history.

Additionally, we aim to consider the different categories that our stakeholders fall into (for-profit, nonprofit, government agencies, etc.) and tailor our services to suit their needs. For example, agencies in the U.S. government such as transportation, agriculture, and labor are all beginning to use AI to identify and eliminate outdated regulations and find potential new contracts, and we want to help reduce any existing bias in these processes. Overall, all of these sectors are using AI to make decisions on people, and we believe that they will benefit tremendously from having a bias detection and reduction post-processing software that Equalmodel provides for ensuring that their algorithms are free of potential controversies.

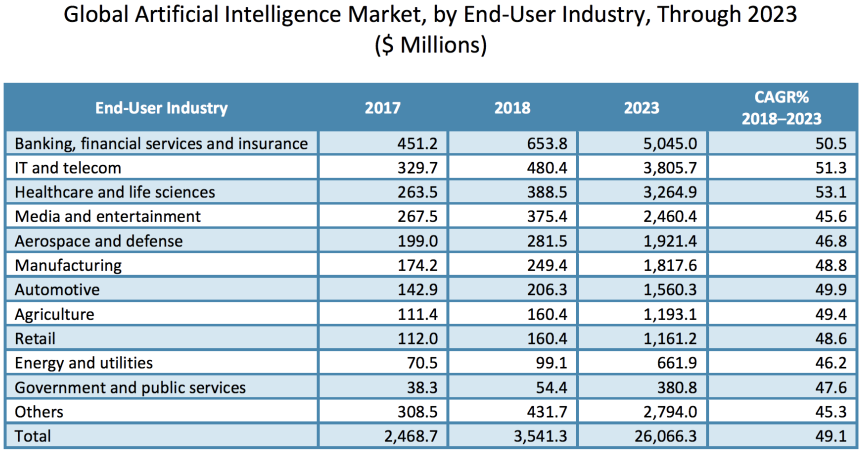
Customer Segments:

Although our model is technically generalizable towards any Machine Learning model, we believe that it is necessary to penetrate one industrial sector before we can sell our service to other sectors. As mentioned in our stakeholder discussion, our initial priority customer segment would be within healthcare, specifically health insurance providers that utilize AI to determine a patient's risk. Other segments in healthcare insurance include claims management, ER visit predictions, cost estimation for customers, and AI chatbots. Examples of companies in these spaces are IQVIA (provider of data science services to healthcare companies), Komodo (using AI for patient history tracking to reduce disease burden), and Cardinal Analytx (provider of healthcare recommendations in advance of medical crises using predictive modeling).

Market Size and Growth:

The total addressable market of the AI industry is $22.29 billion, and is expected to reach $126 billion by the end of 2025, signaling massive growth in this area and, in extension, even a bigger need for Equalmodel. The serviceable available market for artificial intelligence utilization in the healthcare sector may seem niche at first, but growth in the AI health market is expected to reach $6.6 billion by the end of 2021 (compound annual growth rate of 40 percent) according to research conducted by Accenture. As the years progress, the healthcare market will expand its usage of artificial intelligence, whether it be within the areas of dosage error reduction or preliminary/image diagnosis.

The TAM in other sectors that Equalmodel will focus on in the future is also sizable. The AI market in finance is sized at $7.91 billion, in which AI is used to look at credit accounts, investment accounts, loans, fraud detection, and more. The real estate AI market is also expected to grow at a CAGR of 6% from 2020-2025.



Global Artificial Intelligence Market, by End-User Industry, Through 2023
($ Millions)

End-User Industry	2017	2018	2023	CAGR% 2018–2023
Banking, financial services and insurance	451.2	653.8	5,045.0	50.5
IT and telecom	329.7	480.4	3,805.7	51.3
Healthcare and life sciences	263.5	388.5	3,264.9	53.1
Media and entertainment	267.5	375.4	2,460.4	45.6
Aerospace and defense	199.0	281.5	1,921.4	46.8
Manufacturing	174.2	249.4	1,817.6	48.8
Automotive	142.9	206.3	1,560.3	49.9
Agriculture	111.4	160.4	1,193.1	49.4
Retail	112.0	160.4	1,161.2	48.6
Energy and utilities	70.5	99.1	661.9	46.2
Government and public services	38.3	54.4	380.8	47.6
Others	308.5	431.7	2,794.0	45.3
Total	2,468.7	3,541.3	26,066.3	49.1

Source: BCC Research

Competition:

There are a few startups in this space that are actively solving algorithmic bias problems. The two most prominent companies in this space are Arthur AI and Fiddler AI. Arthur AI provides real-time visibility into model performance and outcomes, alerting users when data drift, algorithmic bias, or underperformance is detected. Currently, Arthur AI focuses on healthcare and financial services. Fiddler AI allows users to monitor the key operational challenges in AI - data drift, outliers and model decay. Fiddler AI's use cases include fraud detection and churn analysis. Both of these companies focus on multiple sectors at a time, but Equalmodel aims to focus on a narrow but significant customer base within the healthcare sector in order to start out as a more targeted product. Additionally, Equalmodel aims to use a tiered model (refer to our Revenue Model Section) such that non-profits can easily integrate fairness algorithms into their respective products, making fairness much more accessible.

Apart from startups, Big Tech, like Facebook, IBM, and Microsoft, is also pursuing solutions to algorithmic bias across their company divisions. Microsoft has recently published an open source platform called FairLearn that provides fairness algorithms for developers to use. However, Equalmodel's algorithms are novel in that they can quantify uncertainty in machine learning predictions, which is an essential feature that FairLearn lacks. Furthermore, although there is API documentation to allow developers to understand how to mount their packages, FairLearn lacks the personalized technical expertise that we hope to provide to individual customers that do not have the background to understand how to use our product.

Our Competitive Advantage, Summarized

	Specialized Towards a Vertical	Technical Consultancy Service	Uncertainty Estimation Techniques	Open-Source & Free of cost
Microsoft - *FairLearn*				✓
Arthur AI	✓			
Fiddler AI	✓			
IBM - *AI Fairness 360*				✓
Equalmodel	✓	✓	✓	

Cost:

Because our product's nature as a post-processing algorithm software package, there are no outstanding material expenses on our behalf. Our main expenses will come from hiring specialized technical consultants with fairness algorithms knowledge and algorithm maintenance costs.

Revenue Model:

Our revenue model is based on a licensing operational strategy, alongside a tiered revenue model. Any customer that uses our product is charged a monthly licensing operational fee on the most basic tier of our offerings. To access our second tier that offers a technical consultancy service, there are additional fees placed upon the most basic tier. To keep our product accessible to non-profit organizations as well as for profit companies, we charge a discounted price to non-profit organizations.

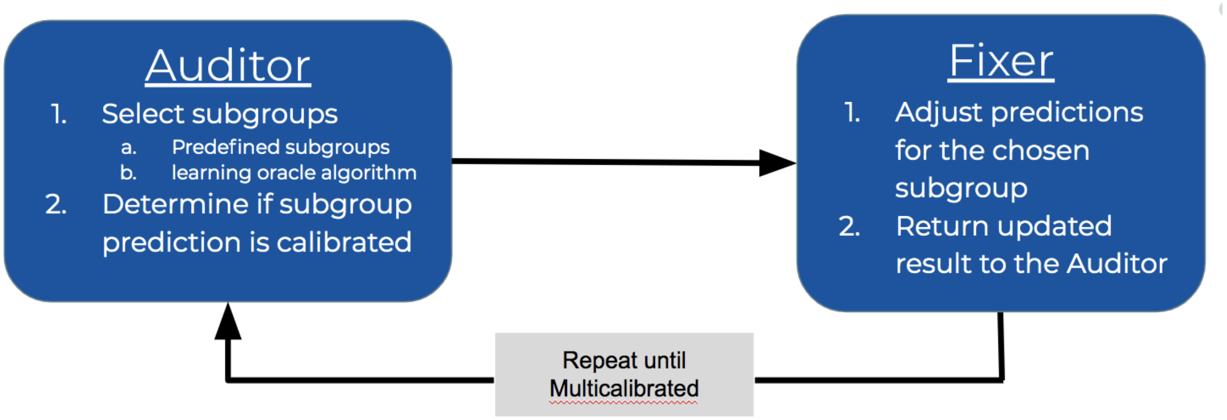
Progress and Future Goals:

The goal of our algorithm is to reduce the implicit bias present in current black-boxed artificial intelligence models by ensuring all predictions are calibrated for each possible subgroup or demographic of individuals. This calibration aims to achieve a consistent degree of confidence in a prediction, so that the confidence of a prediction on someone for whom there may be a lot of data should be identical to that of the under-represented populations.

Our algorithm is composed of two key components, the Auditor and the Fixer. The Auditor is in charge of selecting subgroups based on either a predefined list or an oracle that can find these groups itself. It also analyzes each subgroup to determine whether its predictions are calibrated. The Fixer then adjusts the predictions for each subgroup via gradient descent in order to better calibrate them. This is done by ensuring both means and variances become consistent.

So far, we have succeeded in making the Auditor work with both the predefined subgroups and the subgroup auditor. We can identify uncalibrated subgroups and adjust predictions according to the mean. In the near future we intend to finish implementing the variance adjustments in the Fixer and possibly implement higher level moments for a creator degree of certainty. We also want to expand our testing dataset onto real medical data so as to ensure all hyper parameters are best tuned for this field. Ultimately, we hope to combine our algorithms into a simple to use package by March and create a website that describes our product as well as how to use it.

Overview of Algorithm



Auditor

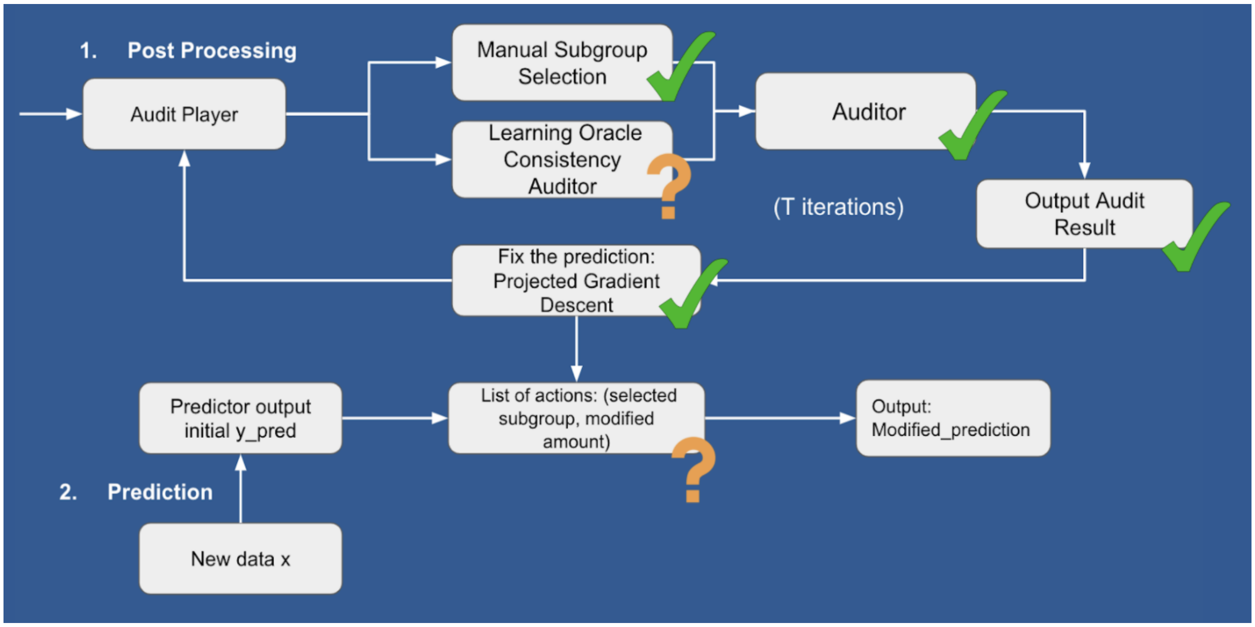
1. Select subgroups
   a. Predefined subgroups
   b. learning oracle algorithm
2. Determine if subgroup prediction is calibrated

Fixer

1. Adjust predictions for the chosen subgroup
2. Return updated result to the Auditor

Repeat until Multicalibrated

Detailed Algorithmic Design and Progress



1. Post Processing

Audit Player

Manual Subgroup Selection ✓

Learning Oracle Consistency Auditor ?

Auditor ✓

(T iterations)

Output Audit Result ✓

Fix the prediction: Projected Gradient Descent ✓

2. Prediction

Predictor output initial y_pred

New data x

List of actions: (selected subgroup, modified amount) ?

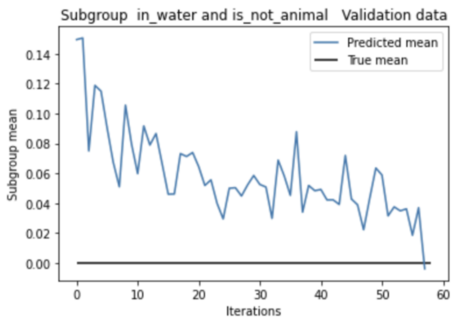
Output: Modified_prediction

Experiment Results

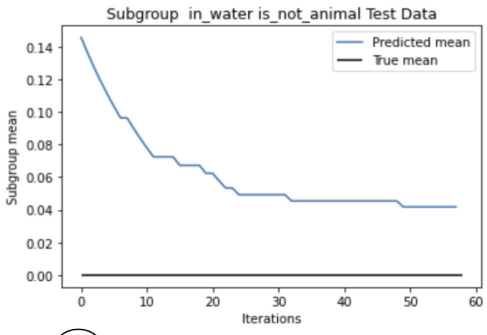
The algorithm was tested on the CIFAR-10 dataset, which contains a set of photos that has objects. The initial AI model we feed into our post processing algorithm is a convolutional neural network that predicts whether a given photo contains a subject that can fly.

Fixer:

The following plots show the convergence of predictions to the true subgroup mean during post-processing for a predefined subgroup of subjects that live in water and are not animals. The right plot shows that post-processing reaches convergence with fresh test data.

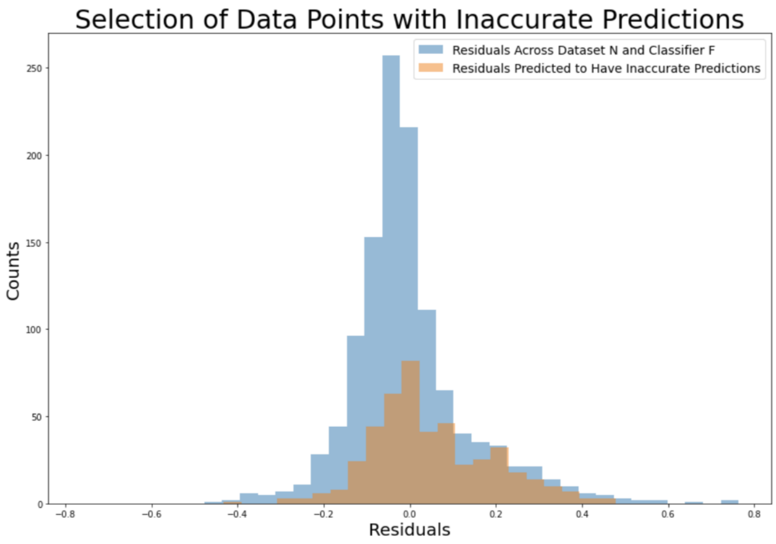


Mean predictions adjusted during post-processing for validation data



Mean predictions adjusted during post-processing for test data

Auditor:
This algorithm, tested on the same dataset, is designed to predict data points that will likely have inaccurate predictions. The plot below shows that, as data points have higher residuals (i.e. worse predictions) the auditor will more likely select these points over those that have lower residuals.


Selection of Data Points with Inaccurate Predictions
Residuals Across Dataset N and Classifier F
Residuals Predicted to Have Inaccurate Predictions
Counts
Residuals