BIAS IN ARTIFICIAL INTELLIGENCE

Researching how to combat bias in AI

Aashna Kumar

Links to my code

https://382cec73-b29d-4055-bdfa-b2b25ebc2636.usrfiles.com/ugd/382cec_8979fdd0d52046818b8de14d38983cd3.pdf

https://382cec73-b29d-4055-bdfa-b2b25ebc2636.usrfiles.com/ugd/382cec_55267c3da59a427fa11b64a661195271.pdf

https://382cec73-b29d-4055-bdfa-b2b25ebc2636.usrfiles.com/ugd/382cec_46176bbc1c3940eeb1d38ede7bdcfce2.pdf

https://382cec73-b29d-4055-bdfa-b2b25ebc2636.usrfiles.com/ugd/382cec_aac325dba839466ca2eb2a1efcbd0772.pdf

Researching Plan

Stage 1: Background
Familiarisation with the data-AI relationship and AI bias implications.

Stage 2:
Segmentation

Algorithm
Designing a machine learning algorithm that can combat bias in data.

Framework
Researching the causes of bias in data in various fields where AI is used.

Stage 3: Solution Development
Combining the algorithm & framework; testing with a prototype

Stage 4: Finalisation
Fine-tuning the solution with results from various trials

The Problem

As technology advances, we are increasingly integrating artificial intelligence (AI) into our decision-making processes and prioritization of individuals. This streamlined approach efficiently handles a multitude of factors and data. However, training an AI involves allowing it to form assumptions during decision-making, leading to biases against marginalized groups. This presents widespread consequences, affecting information retrieval through platforms like Google search and influencing fundamental aspects of our healthcare and justice systems.

The presence of AI is ubiquitous in contemporary daily life, extending from automated job screening algorithms in business to facial recognition technology in the criminal justice system. The impact of intelligent machines is profound, as evidenced by a 270% global growth in AI usage from 2015 to 2019, with the market projected to reach $267 billion by 2027 (Lin, 2020). Currently, 85% of organizations in various sectors, including technology, finance, healthcare, and government, are either assessing or actively employing AI in their operations (Magoulas and Swoyer, 2020).

As AI continues to permeate additional industries such as medicine and law, challenges like machine learning bias are anticipated to become more prevalent (Knight, 2017). Recent controversies, exemplified by disputes over racial bias in the COMPAS prediction algorithm, underscore the potential for artificial intelligence to unintentionally amplify bias in unforeseen ways.

The Problem

The field of Artificial Intelligence is experiencing rapid growth, with its applications becoming ubiquitous in various aspects of our daily lives. While it plays a crucial role in transforming human interactions and experiences, it is not without its imperfections. One notable issue is the potential bias in Computer Vision algorithms, often stemming from insufficient diversity in the training data. Our solutions address this challenge by employing different techniques to enhance data diversity. Another obstacle arises from the low quality of data, particularly evident in CCTV and bodycam footage. To tackle this, our solution introduces the process of upscaling, benefiting not only the model but also proving valuable for law enforcement.

Translation software also exhibits biases, with many systems incorporating gendered queries that result in translational biases. To mitigate such biases, our proposed solutions encompass a range of strategies. Ultimately, the central question we seek to address is: How can we ensure the inclusive accessibility of the rapid advancements in AI for everyone?

The Problem

In the era dominated by information technology, the interconnected realms of social media and the Internet of Things facilitate the widespread transmission of vast amounts of data globally. However, this surge in information dissemination has brought to the forefront the amplified negative consequences of misinformation, often driven by inherent biases. As artificial intelligence (AI) and machine learning (ML) increasingly shape our daily lives, it becomes imperative to scrutinize the biases inherent in computer algorithms. These biases wield significant influence, adversely impacting individuals and minority groups in critical sectors such as healthcare, justice, and networking.

Our examination reveals that bias manifests in two key aspects of the ML process: the assumptions embedded in algorithms and the biases introduced by human choices in selecting training data. Unfortunately, rectifying biases and assumptions in ML algorithms proves challenging, as these neural networks are trained rather than explicitly programmed. Moreover, as posited by Wolpert and Macready (1997), such assumptions enhance the performance of ML algorithms. Consequently, our focus shifts towards addressing bias in human-curated datasets.

The crux of the matter lies in the potential bias inherent in the datasets used to train ML models. If the dataset is skewed, the resulting model will inevitably reflect that bias. Training algorithms on datasets that inadequately represent the entire population can exacerbate existing inequalities and perpetuate systematic bias against underrepresented or misrepresented groups. Leveraging neural networks, we can identify, expose, and subsequently rectify biases ingrained in human-created datasets.

Hence, our defined problem centers on developing a machine learning algorithm capable of uncovering biases within datasets that unjustly establish correlations between unrelated or distantly related variables. This is particularly crucial in the context of variables related to identity politics, such as race and gender, or when dealing with underrepresented groups.

The Problem

What causes these issues to arise?

Machine learning models are not intentionally designed to exhibit bias; rather, they make assumptions based on the input data provided to them. If the training data is skewed against a specific demographic, the resulting models are likely to reflect these biases. Hence, it is essential for unbiased datasets to ensure proportional representation across different groups.

How should one manage sensitive data?

We broadly define sensitive data as information that has the potential to be discriminatory towards specific groups or individuals, encompassing factors like race, gender, personal details, and health-related information. Although one option to prevent bias is to disregard sensitive data, such an approach is overly simplistic. Sensitive data can still hold predictive value, and variables associated with it may also be linked to various other factors.

Nevertheless, there are ethical constraints to consider. Assuming individuals solely embody the characteristics of their respective groups is discriminatory. Consequently, manipulating sensitive variables should not unduly influence the output of the machine learning algorithm. Specifically, if two datasets are identical except for one altered sensitive variable, the disparity in the ML algorithm's outputs should not be statistically significant.

The Problem

AI's most intricate function lies in its ability to anticipate future events based on historical data. By taking the lead in the evolution of data analytics, AI facilitates "predictive decision-making" (Manheim & Kaplan 120). Despite its prowess, the inner workings of AI remain elusive and are susceptible to bias due to its restricted viewpoint. Relying solely on past datasets and input from developers, AI can give rise to discrimination as "big data collection and analyses codify historical and intentional discriminatory treatment" (Tschider 98). The presence of algorithmic bias in AI networks establishes a framework where personal information is leveraged against individuals, constituting a form of discrimination rooted in privacy violations. This becomes particularly perilous as inherent human biases, already problematic in society, are magnified when AI systems identify disparities and respond accordingly. Some noteworthy biases include:

RACIAL BIAS

Research indicates that numerous artificial intelligence programs exhibit racial bias, contributing to the amplification of existing systemic racism in different scenarios. One notable instance of this phenomenon is observed in risk assessments for criminal sentencing, commonly employed in courts using COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). Despite its purpose to evaluate the risk and criminal tendencies of defendants, COMPAS generates risk scores that manifest clear racial bias. In particular, Black defendants tend to receive higher risk scores compared to their white counterparts, who consistently receive significantly lower risk ratings (Angwin et al., 2016).

SOCIO ECONOMIC BIAS

Much like the issue of racial bias, another straightforward determinant is socioeconomic status. Unfortunately, in a society increasingly driven by capitalism, discernible patterns emerge within specific socioeconomic groups, yet attributing these trends directly to the status proves challenging. While humans may grapple with distinguishing cause from correlation, for an AI, the two concepts are interchangeable. The repercussions of this are particularly evident in the realm of healthcare. For instance, "health data could potentially be leveraged... to disqualify individuals for lower insurance premiums" (Manheim & Kaplan 121). Similarly, diagnostic analyses often rely on data from patients who can afford the associated services. Consequently, diseases and conditions more prevalent in lower socioeconomic communities may go misdiagnosed or undetected, leading to an increase in preventable tragedies and fostering distrust in the medical industry (Wang et al. 2020). Acting purely as a risk analyzer, AI exacerbates discrimination against already disadvantaged minority groups.

The Problem

IMPACT OF BIAS

A significant portion of the global population encounters software biases, reaching up to 70% based on our surveys. These biases are frequently associated with factors such as ethnicity, language, age, gender, nationality, religion, or disability. The impact of this bias varies among individuals, manifesting as a diminished quality of service in applications like chatbots, image recognition, voice recognition (including speech-to-text), spam detection (in email and social media), and recommendation and advertising systems.

However, for certain individuals, bias assumes a more detrimental form, affecting resource allocation in areas like recruitment, criminal justice (including prediction of criminal activity), and determination of preferences in COVID vaccination. The persistence of these biases is attributed to their diverse sources, often rooted in human biases reflected in data rather than machine error.

GAPS IN DETECTION AND MITIGATION OF BIAS

The identification of bias has yet to become a standardized practice within the industry, resulting in biases often going unnoticed for extended periods before being acknowledged. While there have been technical interventions such as Dalex, AIF360, and BERT designed to assist AI creators in detecting and, at times, mitigating bias, these solutions either necessitate a complete overhaul of the current model or solely identify bias without offering mitigation measures. For many startups and developers, allocating significant resources to rectify existing models or procure new, unbiased data is impractical. The existing techniques are intricate and lack the comprehensiveness needed to accommodate all types and formats of AI models. Furthermore, certain approaches, like Group Benefit Equality, may inadvertently amplify bias through positive feedback. Simultaneously, these established techniques fall short in detecting implied bias present in AI models.

The background

The primary objective of our project is twofold: firstly, to counteract bias and foster equality in Artificial Intelligence (AI) and Machine Learning (ML), steering away from their misuse as tools that deepen societal divisions. Secondly, we emphasize the critical need to address the severe repercussions of biased models. Historical instances, such as the Correctional Offender Management Profiling for Alternative Sanctions (COMAS), highlight the dangers of discriminatory AI solutions. In the case of COMAS, employed in U.S. courts to predict repeat offender likelihood, the algorithm exhibited alarming bias, disproportionately producing false positives for black offenders (45%) compared to white offenders (25%).

Another notable example involves Amazon's 2015 hiring algorithm, which inadvertently discriminated against women due to skewed training data favoring male resumes. Similarly, gendered outputs from AI translators in the tech industry can perpetuate stereotypes by wrongly associating specific genders with certain activities.

The impact of biased algorithms extends to healthcare, where ML plays a growing role in critical tasks like skin cancer diagnosis, stroke detection in CT scans, and identifying potential cancers in colonoscopies. The consequences of biased training data are evident, with gender-imbalanced datasets hindering the accuracy of chest X-ray readings for underrepresented genders. Moreover, skin-cancer detection algorithms trained predominantly on light-skinned individuals raise concerns about their effectiveness for other complexions.

Our project addresses these issues by focusing on improving the representativeness of training data across all demographic groups. Through advancements in vision classifiers, generative models, and image regeneration & upscaling, we aim to enhance the fairness and inclusivity of Machine Learning, mitigating the risks associated with biased algorithms.

Different types of Biases

AI bias is an anomaly in the output of machine learning algorithms. These could be due to the prejudiced assumptions made during the algorithm development process or prejudices in the training data. AI systems contain biases due to two reasons:

COGNITIVE BIASES

These are effective feelings towards a person or a group based on their perceived group membership. These biases could seep into machine learning algorithms via either:

- designers unknowingly introducing them to the model a training data set which includes those biases
- a training data set which includes those biases

SAMPLE BIAS

If data is not complete, it may not be representative and therefore it may include bias.

Core Issues

Through extensive research and reviewing case studies I have compiled a list of issues that are pertinent

Defining Fairness

Within the field of computer science, three formal fairness criteria—namely independence, separation, and sufficiency—have been established. The Impossibility Theorem of Fairness, as outlined by Zhong (2020), asserts that satisfying all potential fairness criteria simultaneously is unattainable for an algorithm. Instead, the determination of fairness for a particular machine learning system falls upon the discretion of computer scientists. This determination is guided by considerations such as user experience, cultural, social, historical, political, legal, and ethical factors, which may involve tradeoffs (Google AI, n.d).

Despite these considerations, the task of defining fairness for artificial intelligence is compounded by the observation that AI developers lack training in ethical decision-making (Ebert, 2020). Matthew Stewart, a PhD researcher, notes that unlike doctors, computer scientists may not be inherently equipped to contemplate the ethical implications of their actions. The detachment of computer scientists from data subjects may lead to a perception that the impact on any individual is negligible and, consequently, overlooked (Stewart, 2020). This uncertainty and potential oversight underscore the need for a more deliberate integration of ethical considerations into the development process of artificial intelligence.

Human Biases

The incorporation of automatic cognitive biases is a fundamental aspect of expediting decision-making in humans. However, these decisions are not without their drawbacks, as they may inadvertently involve "racial or social class categories or other unfair stereotypes," as pointed out by Susan Fiske and Shelley Taylor (2020). Overcoming these inherent human biases proves to be a challenging task. Olga Russakovsky, an assistant professor in Princeton's Department of Computer Science, asserts that "debiasing humans is harder than debiasing AI systems" (Ghosh, 2021).

Cognitive biases are pervasive in society, influencing traditional human decision-making processes. These biases, deeply ingrained in societal norms, tend to seep into AI systems during training or user interactions. This integration of biases from human behavior can lead machines to inadvertently replicate and perpetuate human prejudices, as highlighted by Manyika, Silberg, and Presten (2019). Recognizing and addressing these biases are critical steps toward enhancing the fairness and equity of AI systems.

Biases in machine learning data and models

Concerning bias in data, issues of colorblindness and underrepresentation pose significant challenges in AI, resulting in discriminatory impacts on specific populations. For example, minorities, who often constitute a disproportionate portion of the lower class, are less likely to dedicate time to establishing a robust internet presence. This lack of representable online data can hinder a job candidate's access to opportunities, as the absence of an internet presence may signal a red flag to both AI systems and HR departments.

Furthermore, AI models tend to be blind to the complexities of class intersectionality, which involves intricate relationships between seemingly unrelated factors such as race, class, and gender. For instance, a woman may face a higher likelihood of holding lower-paying jobs compared to a man, and a person of color may be more likely to come from a lower tax bracket. Implementing AI that lacks an understanding of these nuanced data dynamics can lead to unexpected discriminatory outcomes. Addressing these intricacies is crucial for developing fair and unbiased AI systems.

Identifying the root problems

Through extensive research and reviewing case studies I have targeted the root problems

HISTORICAL DATA

Historic data is the use of outdated information fed into a system which serves the public. Creating a skewed version of society, an AI may create judgements that no longer reflect progressive society, and/or favour historically privileged groups. This can include employment rates by gender identity, income by race etc.

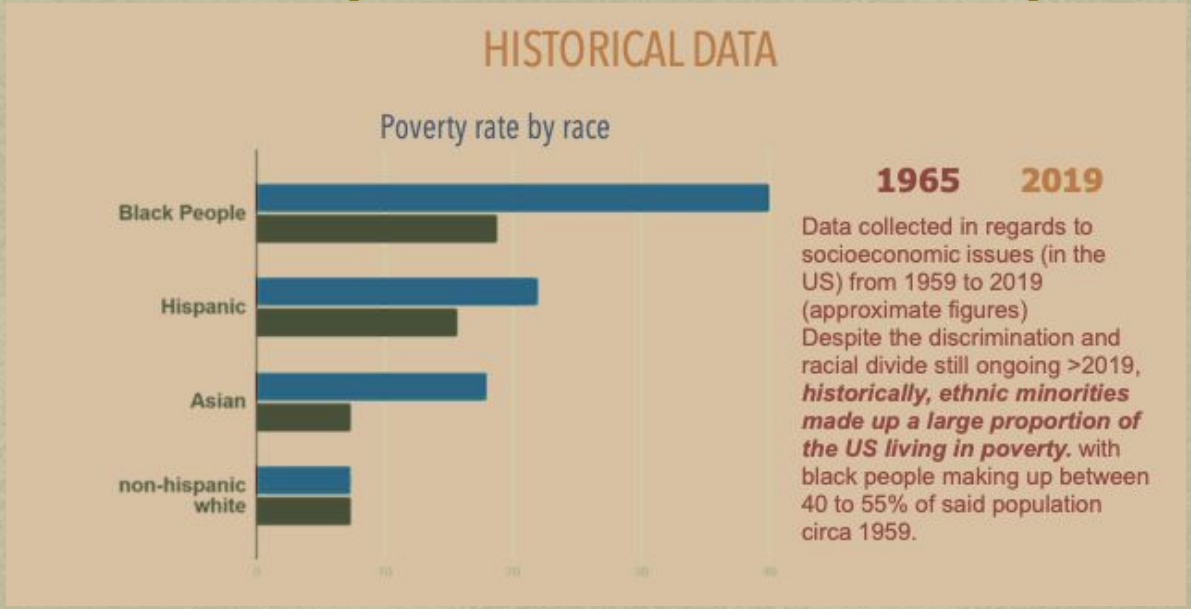
HUMAN BIASES

Humans create AI, and therefore a bias can transcend from discrimination within a team of people to discrimination within an AI. Lack of diversity within tech fields, and pre-existing discrimination can teach an AI what to favour and what to avoid.

SOCIOECONOMIC FACTORS

AI that serves the public: i.e healthcare and justice, needs up to date data that reflects society. Whilst this, in theory, tackles the issue of historic data; there are many aspects of society such as income, education, and place of birth that target marginalised groups and perpetuate generalisations based on data that reflects real societal issues.

Statistics (historical data)



Through the data seen, that displays socioeconomic disparity between race groups; a physical bias exists and can't be ignored. However, the statistics show that over the course of the last 60-70 years, the gap has considerably lessened. This means that historically, any generalisations that don't fit today, would very much reflect society back in the 1950's, therefore a bias will be almost impossible to avoid.

Statistics(historical data)

RACE AND ETHNICITY	1993	1995	1997	1999	2003	2006	2008	2010	2013	2015
American Indian or Alaska Native	0.2	0.3	0.3	0.3	0.3	0.4	0.3	0.2	0.2	0.2
Asian	9.1	9.6	10.4	11.0	14.2	16.1	16.9	18.5	17.4	20.6
Black	3.6	3.4	3.4	3.4	4.3	3.9	3.9	4.6	4.8	4.8
Hispanic	2.9	2.8	3.1	3.4	4.4	4.6	4.9	5.2	6.1	6.0
Native Hawaiian or other Pacific Islander	NA	NA	NA	NA	0.3	0.5	0.4	0.2	0.2	0,2
White	84.1	83.9	82.9	81.8	75.2	73.2	71.8	69.9	69.9	66.6
More than one race	NA	NA	NA	NA	1.4	1.4	1.7	1.4	1.5	1.6

Distribution of workers in SCIENCE and Engineering occupations, by race and ethnicity: Selected years,1993-2015
NA - not applicable / not found

Statistics(historical data)

What can we conclude from the table above?

The biases that can be formed through AI can be due to the lack of diversity within the teams that create them- seen through the staggering low percentage of ethnic minorities in STEM fields over the last 20 years. This means that biases are more likely to find their way into an AI if there is a shortage of representation in the creation process. These biases can be both malicious within the team, or that there was simply not enough research and input from other ethnicities and genders etc.

Through reports published by a New York University research Centre, they spoke about examples where this lack of diversity has had detrimental effects: such as facial recognition technology classifying racial minorities with features that are offensive, and online AI- controlled ‘chat bots’ initiating discriminating and hateful speech.

Socio Economic Bias (historical data)

Race/Ethnicity	1967	1972	1977	1982	1987	1992	1997	2002	2007	2012
All Races	17.0	14.6	14.1	13.9	12.6	11.0	11.0	10.5	8.7	6.6
White	15.4	12.3	11.9	11.4	10.4	7.7	7.6	6.5	5.3	4.3
Hispanic	No data	34.3	33.0	31.7	28.6	29.4	25.3	25.7	21.4	12.7
Black	28.6	21.3	19.8	18.4	14.1	13.7	13.4	11.3	8.4	7.5

Percentage of high school dropouts among persons 16 to 24 years old in the United States by race/ethnicity: 1967 through 2012

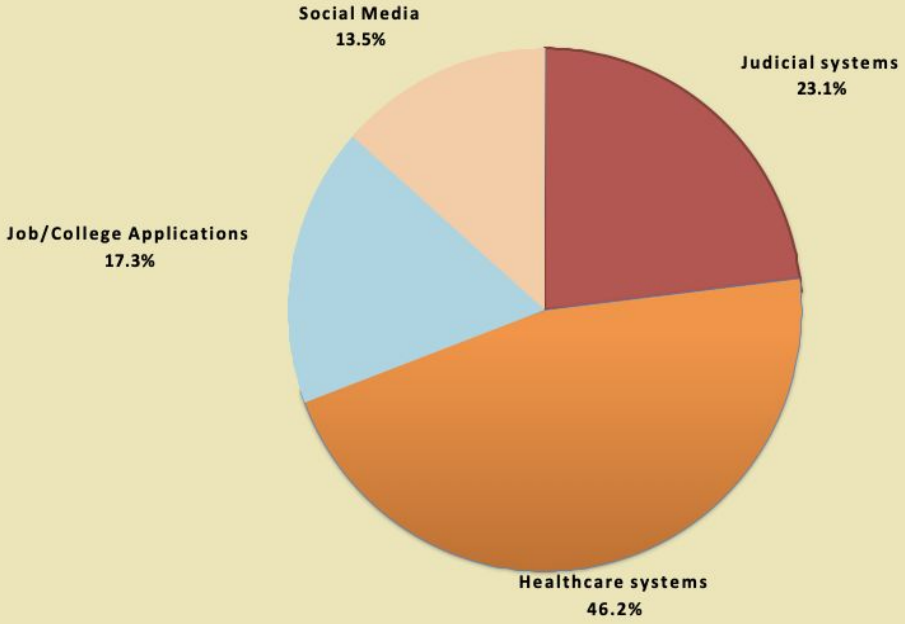
Socio Economic Bias (historical data)

What can we conclude from the table above?

Despite improvements in reducing socioeconomic disparities among different racial groups, AI systems remain susceptible to biases influenced by factors such as education, income, and household dynamics. The data highlights persistent disparities, particularly in dropout rates among Black and Hispanic students, which remain significantly higher than those for white students, especially during the 2010s. Moreover, the justice system has historically used lack of education and dropout rates as indicators for determining the causes of crime.

Simply removing historical data from AI models does not guarantee the elimination of bias, as societal issues continue to disproportionately impact ethnic minorities. It is crucial to address and rectify these underlying issues to ensure a fair and unbiased AI system that accurately reflects the diverse realities of different communities.

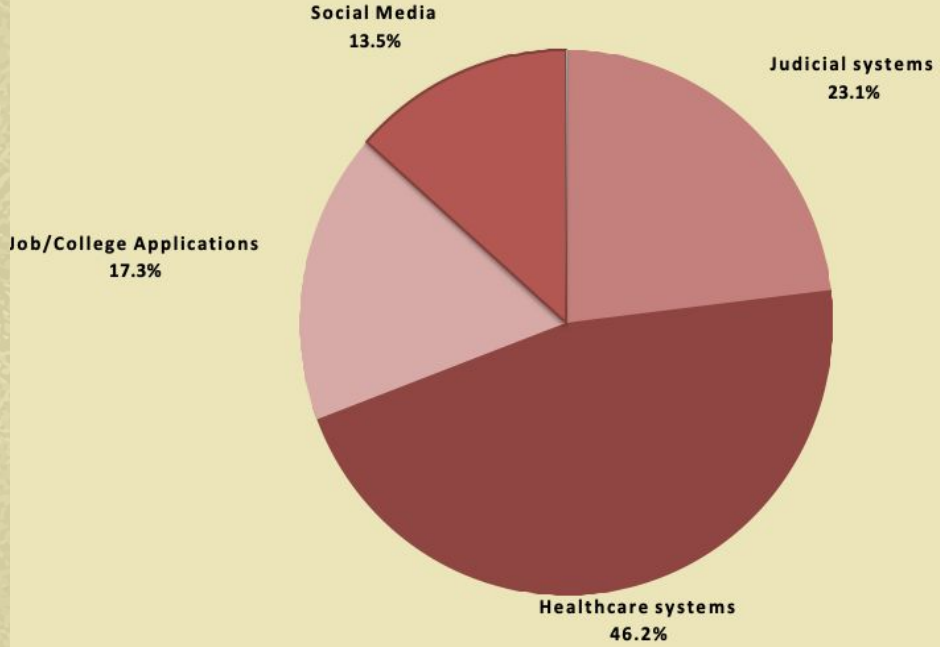
DATA ANALYSIS



We decided that was not the case since maximum people thought of that the healthcare system needed immediate attention for correction of bias, which we think can be influenced by the fact we are under a pandemic. It is most definitely true that the healthcare system is a significant sector and it is completely justified for people to feel that action needs to be taken against it.

However, very few thought that college and job application system needed attention. About 17% voted for this and 13% voted for social media and given that social media is not a requirement the difference between these two seemed alarming.

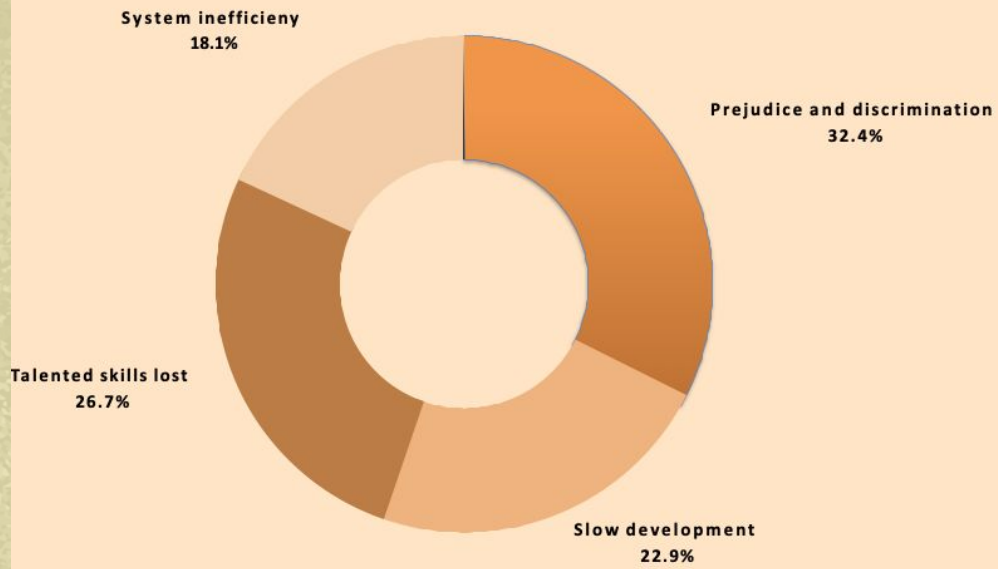
DATA ANALYSIS



Therefore, we think people are not aware of their applications possibly getting rejected because of bias in a system. It can be that he/she may not be preferred by the hiring managers but there are several scenarios where the person may have been highly qualified but were simply not chosen because of a bias and that must be demotivating.

So to mitigate the bias, a lot of these procedures take time as sometimes human intervention may be needed.

DATA ANALYSIS



This chart shows how people think bias in AI will affect the future

Finally, we asked how they think the bias will affect the future. Majority of the people thought that prejudice and discrimination would increase but another one that piqued our interest was that they also thought that it can possibly slow development.

This means that people are uncertain about the advancements in technology and losing confidence in them can certainly hinder development. That is a major problem as a lot of systems are reliant on technology.

So, taking account of all these factors, we thought that we can make a model that can mitigate bias significantly in some systems (particularly decision-making) so that technology can thrive and sustain among people.

Success Criteria

1. Flexible

Bias is present in a plethora of Machine Learning algorithms and datasets. Each model and situation is unique and requires an astute understanding of its intricacies and limitations. Our model seeks to provide a solution that can apply itself to as many situations as possible, maximizing its utility.

2. Accurate

A model can only be as useful as it is accurate. In order for our solution to be useful, it must be able to accurately detect when bias is present in a dataset.

3. Makes Few Assumptions

There should be as few assumptions made regarding data and indicators of bias (such as the inclusion of sensitive information) as possible. Real-life data is messy. There are times when assumptions can be made and there are times when they cannot, and it is extremely difficult for humans to discern between these two cases.

4. Indicative

If a dataset is biased, our model should be able to indicate how this may be fixed.

Hypothesis #1

By creating a versatile machine learning algorithm to refine raw datasets through combating bias, AI systems and applications will fulfill their primary purpose and promote social progress through diversity and inclusion of all demographics.

Core Focus: Input Data Refinement Instead of attempting to tackle the entire issue of AI in bias, we decided to focus on a single aspect: the data that is used to develop and train the AI. This division of the process, while not solving for bias, would help significantly reduce one area from which bias can arise from.

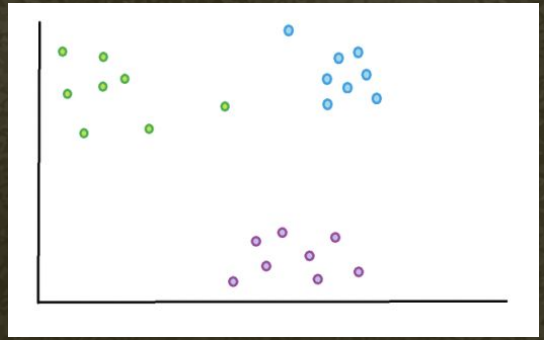
Hypothesis #1

1. Machine Learning Algorithm: Our machine learning algorithm is the heart of this project. It should eliminate the bias from the beginning stages of AI development by refining data.
2. Implementation: Our solution will not be all encompassing; ergo, steps for implementation should be established.
3. Solvency: Our solution should diminish AI bias in an effective, timely, and cost efficient manner.

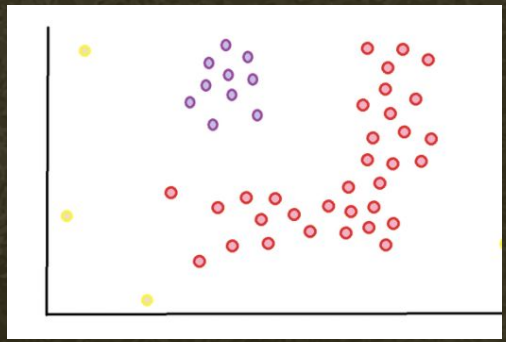
Solution #1

The Proposed Solution

Our approach addresses the core issue of bias within datasets by introducing an algorithm that mitigates the inclination of favoring specific individuals or groups over sensitive data. To identify the potential over-significance or overrepresentation of sensitive data within a dataset, a careful examination of "sensitive" variables is required. The variable undergoing scrutiny, known as the manipulated variable, often lacks explicit inclusion in the dataset. For example, in face datasets, ethnic labels may not be explicitly mentioned, necessitating the use of clustering to unveil this information.



Clustering is a process that involves automatically identifying natural groupings within data. Clustering algorithms interpret input data to uncover inherent clusters or groups in a feature space, classifying data points into distinct groups based on their similarities. These clusters are formed based on the manipulated variables. Various clustering algorithms, such as K-Means Clustering and DBSCAN, can be employed based on the problem type and scenario. However, determining the number of clusters is a critical hyperparameter, requiring manual specification. An effective method for determining this number is leveraging a validation set. The model undergoes multiple training iterations, each time with a different cluster count, and the model's performance is evaluated on the validation set. The optimal number of clusters leading to the best performance is then selected. Subsequently, each cluster is assigned a numerical index.



Solution #1

A Clustering Example

Consider a facial recognition application where our dataset comprises images of individuals, and our objective is to ensure a dataset that lacks disproportionate bias toward any specific race or gender. One approach involves examining each image in the dataset, specifically focusing on the faces, which are typically centrally located. By extracting pixel intensity values from a defined region around the center of each image, variations in complexion become apparent. This process facilitates the clustering of data into distinct groups based on differing pixel intensity values associated with individuals of varying complexions. While the challenge arises in determining the optimal number of clusters due to the continuous nature of pixel intensity values, this can be addressed using the 'elbow' method.

Solution #1

Equal Distribution of Clusters in Sample: Stratified Random Sampling

If our algorithm aims to train on a dataset with equal representation across all clusters, we employ a stratified random sampling technique. This involves randomly selecting a fixed and equal number of data points from each cluster. The result is the creation of a new dataset that ensures equal representation from all identified clusters in the initial step. This process eliminates exclusion bias, providing a dataset that allows our model to be trained without skewed representation.

Unequal Distribution of Clusters in Sample

However, not all situations call for equal representation amongst clusters. In certain situations, some clusters should be more represented in the data than others. In this case, a fixed but unequal percentage of data points is chosen from each cluster in a random manner.

Solution #1

Updating the Existing ML Algorithm: While clustering prevents the issue of disproportionation in datasets, there still presides the issue that certain clusters may be unjustly represented to correlate more/less with specific variables, which may greatly influence the output of an existing neural network. To circumvent this issue, we have proposed the following framework:

1. Enhance the dataset by adding a new column containing the index assigned to each data point's cluster. Modify the neural network to accommodate the updated dataset by introducing an additional neuron to the input layer. Regularly train the network. Subsequently, conduct testing in small batches, ensuring that data points within each batch belong to the same cluster.

2.

3. Record the mean output for each batch. Repeat the testing process, altering the assigned index value for each batch (e.g., changing "2" to "3"). Record the mean output for each batch with the adjusted index. Continue this iterative process until each batch has been assigned every possible index.

4.

5. Apply a One-Way Welch's Test for each batch, examining if the means of the samples are statistically different. The detection of significance in any of the One-Way Welch's Tests suggests the presence of bias in the dataset. Specifically, this framework indicates the existence of sensitive data within the dataset that significantly influences the output of the neural network under evaluation.

6.

Solution #1

In order to implement our algorithm, we will use an implementation framework similar to the EPIS framework designed by the University of San Diego; however, rather than having sustainability as our last step, we chose testing.

EXPLORATION

- Explore current algorithms that mitigate biases in AI and evaluate their effectiveness
- Find other clustering methods besides DB Scan and K-clustering
- Research different machine learning platforms in which we can program our algorithm.

PREPARATION

- Split team members into different roles:
  - Dataset finders
  - Supervisors
  - Researchers
  - Programmers
- Decide whether to use Jupyter notebook vs Google Colab for programming.
- Assign specific software components (for example: neural network) to programmers.

IMPLEMENTATION

- Execute the implementation and start development of the algorithm
- Shift around team roles if necessary
- Test each individual software component individually before running it altogether to save time
- Have meetings biweekly to review individual progress on the prototype.

TESTING

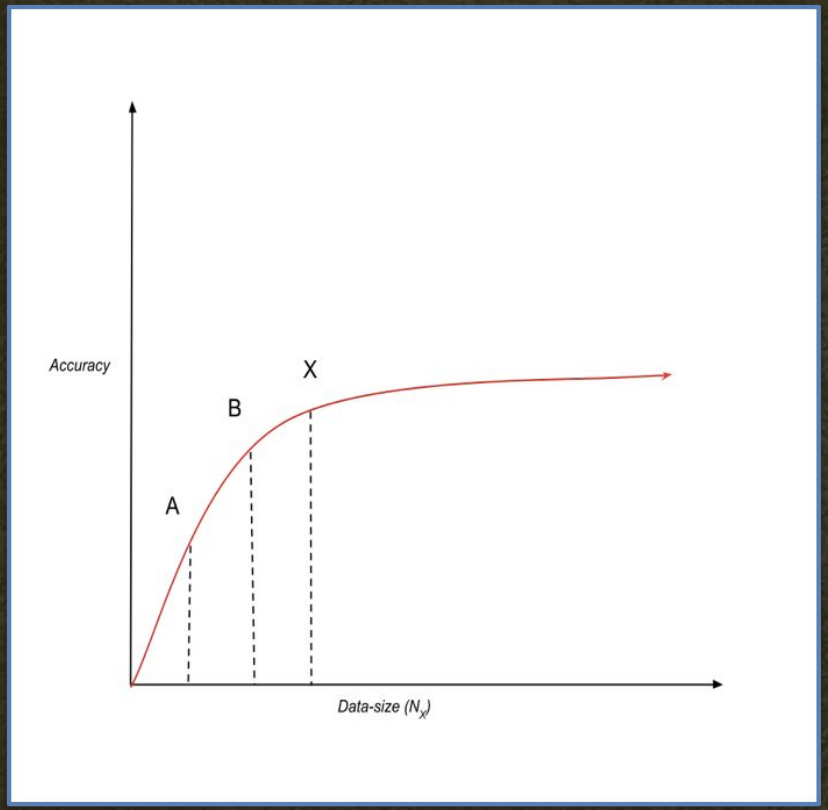
- Evaluate the effectiveness of the current model and continue to build the model
- Modify software components
- Try different testing datasets to improve the success rate
- Test the model on real machine learning models
- Refine the neural network

Hypothesis #2

The overarching objective is to enhance the transparency, security, and user-centric nature of technology. By augmenting and reconstructing classification models, we can reduce inaccuracies in identifying subjects based on gender and race, achieved through a diversified and equitable dataset. This strategy will fortify our Convolutional Neural Networks (CNNs) models, leading to more effective utilization of AI products. The introduction of augmented and generated data is anticipated to significantly elevate accuracy, opening up new possibilities for AI applications.

Text-based solutions, compatible with voice assistants and translators, have the potential to greatly assist minorities and individuals in non-English-speaking countries, offering seamless access to related facilities. Additionally, the implementation of upscaling holds promise in resolving issues within specific industries, with a particular focus on its implications in the justice system and in regions where crime is a persistent challenge. Through these approaches, our aim is to minimize challenges and ultimately alleviate bias in AI, ensuring a safer, faster, and more equitable progression of the industry.

Diversifying Datasets



Although literature has shown that an Artificial Intelligence 'model' cannot be biased, the data used to train it certainly can be, thereby **negatively changing the output of the model.** As the inequality of data-size per class increases, the overall accuracy of the model decreases. Eg: If data-size of class A (N_A) > data-size of class B (N_B), then the model will most certainly perform poorly with objects from Class B. This is why AI products fare abysmally when utilized by users from underrepresented communities.

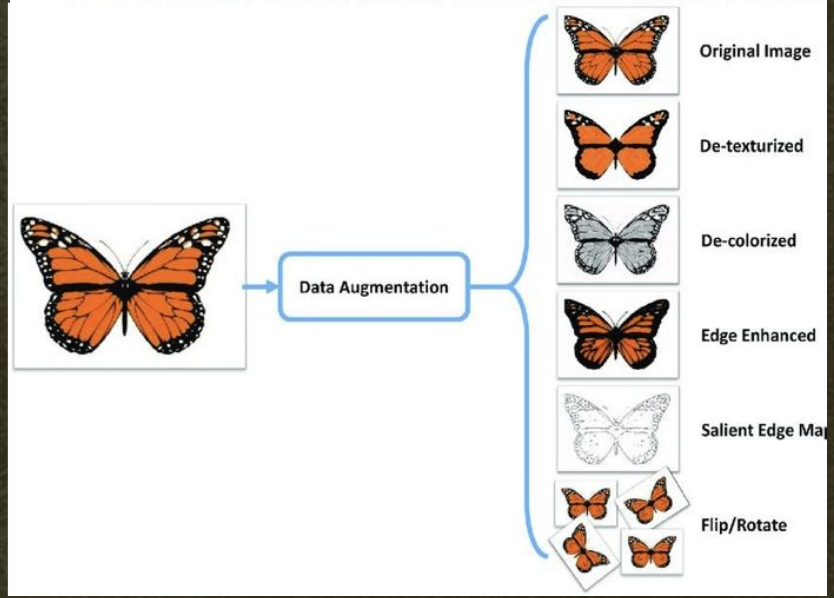
There is also a logarithmic relation between size of training data and accuracy of the model, which means that with increase in data-size (N_X), the accuracy will also increase to an extent, after which it stabilizes. Hence, we targets and proposes to **equalize data-size across all classes ($N_A = N_B$)** and then maximize the number of images per class, which depends model to model. This can be done using the following slides:

Augmentation

Data augmentation is the process of creating data using physical methods like changing RGB values, rotating, enlarging, minimizing, cropping, scaling, and many more.

The uses of augmentation are **wide-ranging.** Although it is used for increasing data-size in the industrial background for large neural networks (VGGNet, DeepVideo, AlexNet, etc.), we used it for our small datasets and noticed a huge jump in accuracy.

Augmentation enables the computer to interpret **some** form of the same image as a new one, hence broadening its capabilities



```
from keras.preprocessing.image import ImageDataGenerator, array_to_img,
img_to_array from keras.preprocessing import image
import matplotlib.pyplot as plt

datagen =
   ImageDataGenerator( ro
   tation_range = 45,
   width_shift_range = 0.2,
   height_shift_range = 0.2,
   shear_range = 0.2,
   zoom_range = 0.2,
   horizontal_flip =
   True,
   fill_mode = 'constant', cval = 125
)

img_path = ('imgpath.ext')
img = image.load_img(img_path, target_size =
(100,100)) img = img_to_array(img)

x = img /
255.0
print(x.shape
)
plt.imshow(x)
plt.show()

input_batch = x.reshape((1,
*x.shape))
print(input_batch.shape)

i = 0
for output_batch in datagen.flow(input_batch,
   batch_size = 1): plt.figure()
   imgplot =
   plt.imshow(image.img_to_array(output_batch[0])) i
   += 1


   if i ==
      10:
      break


   plt.axis =
   'off'
   plt.show()
```

```
Import tensorflow as tf

#mirroring(1)
flipped =
tf.image.flip_left_right(image)
visualize(image, flipped)

#greyscale(2)
grayscaled =
tf.image.rgb_to_grayscale(image)
visualize(image, tf.squeeze(grayscaled))

#saturation(3)
saturated = tf.image.adjust_saturation(image,
3) visualize(image, saturated)

#brightness(4)
bright = tf.image.adjust_brightness(image,
0.4) visualize(image, bright)

#cropped to the center (5)
cropped = tf.image.central_crop(image,
central_fraction=0.5) visualize(image,cropped)
plt.axis('off')
#the following functions can also be used

tf.image.stateless_random_brightness

tf.image.stateless_random_contrast

tf.image.stateless_random_crop

tf.image.stateless_random_flip_left_r

ight

tf.image.stateless_random_flip_up_dow

n  tf.image.stateless_random_hue

tf.image.stateless_random_jpeg_qualit

y

tf.image.stateless_random_saturation
```

Generative Models

Generative models, like augmentation help diversify datasets and thereby increase the accuracy of the model. Unlike augmentation however, they **use neural networks** and generate images which **do not directly** stem from the previous images.

Flow-Based Networks

Flow-based generative models: A flow-based generative model is constructed by a sequence of invertible transformations. There is a huge variety to these networks (Real NVPs, NICE, MADE, WaveNET, etc.), but we would recommend using Glow *(Kingma and Dhariwal, 2018)* due to its simplicity to use and efficiency in diversifying datasets.

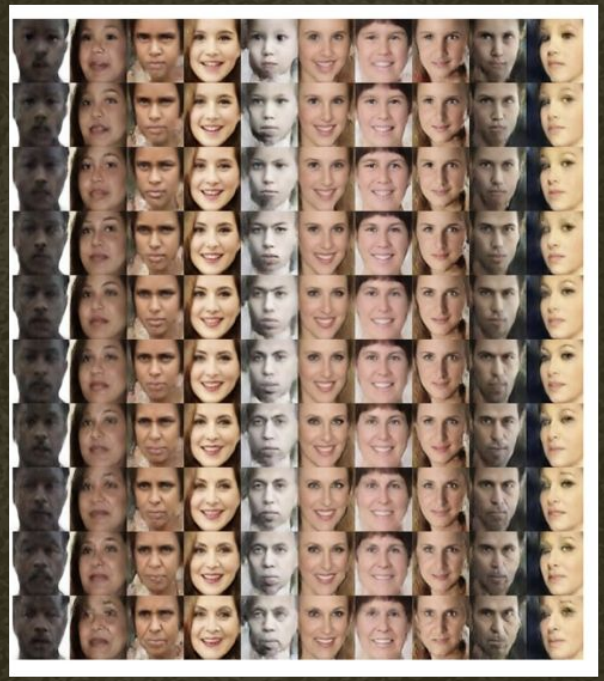
Steps to implement:

1) **Activation normalization (actnorm):** It performs an affine transformation using a scale and bias parameter per channel, similar to batch normalization, but works for mini-batch size 1. The parameters are trainable but initialized so that the first minibatch of data have mean 0 and standard deviation 1 after actnorm.
2) **Invertible 1x1 conv:** Between layers of the RealNVP flow, the ordering of channels is reversed so that all the data dimensions have a chance to be altered. A 1×1 convolution with equal number of input and output channels is *a generalization of any permutation* of the channel ordering.
3) **Affine coupling layer**

Generative Adversarial Networks

GANs have been in the academic spotlight in recent times, due to their versatility and wide-ranging uses. A generative adversarial network (GAN) has two parts:

- The generator learns to generate data. The generated examples become negative training examples for the discriminator.
- The discriminator learns to distinguish the generator's fake data from real data. The discriminator penalizes the generator for producing incorrect results, which the generator uses as feedback while producing the next image.



Images Generated using GANs

NOTE: The results produced by GANs and Flow Based networks vary negligibly. It is the functioning of the model which varies

Upscaling

Upscaling is the process of increasing the resolution of a graphic or audio. Although upscaling has very little effect on models with less data (like ours), it substantially increases accuracy in situations involving a lot of data (in the industry). Upscaling can be performed on text, audio and video. Using upscaling for CCTV footage can be especially beneficial to the **law and justice system** and **developing countries** where crime rate is high.

```
import tensorflow_hub as
hub import tensorflow as
tf
generator = tf.keras.models.Sequential([
    hub.KerasLayer("https://tfhub.dev/captain-pool/esrgan-tf2/1",
    trainable=True), tf.keras.layers.Conv2D(filters=3, kernel_size=[1, 1],
    strides=[1, 1])
])
```

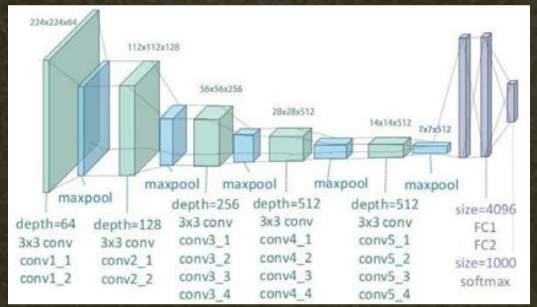
NOTE: This uses a pre-trained model which simplified the process for us.

Different Models

A few models, **provided that computational resources are not a problem**, fare extremely well in providing high accuracies, precision, and F1 scores with diversified datasets in the industrial landscapes.

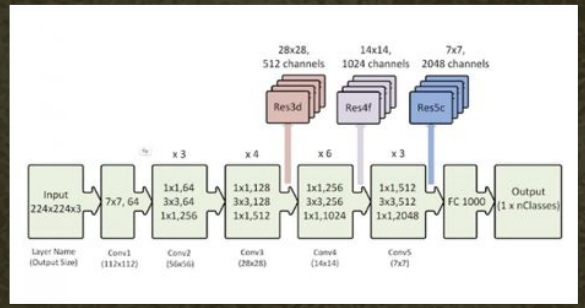
VGG-19

VGG19 is a variant of VGG model which in short consists of 19 layers (16 convolution layers, 3 Fully connected layer, 5 MaxPool layers and 1 SoftMax layer).
Accuracy on ImageNet ~92%



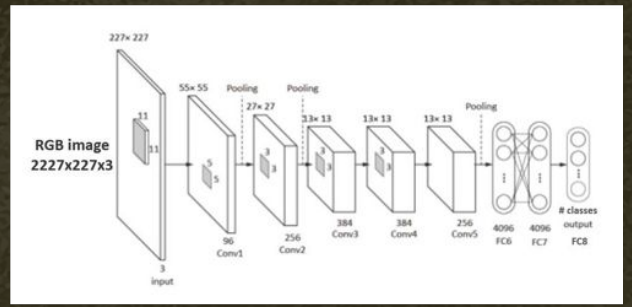
ResNet-50

ResNet-50 is a convolutional neural network that is 50 layers deep.
Accuracy on ImageNet ~93%



AlexNet

AlexNet contains eight layers; the first five were layers, some of them followed by max-pooling layers, and the last three fully connected layers. It uses the non-saturating ReLU activation function, which shows improved training performance over tanh and sigmoid.
Accuracy on ImageNet ~80%



Note: Top-5 has been taken as a metric to measure accuracy here

Reconstruction and Upscaling in the Criminal Justice System

Along with pre-existing cases in developing countries, recent protests against cases of police brutality have brought light to the criminal justice system, which disproportionately affects minorities. One reason for this could be because the **lack of evidence,** or **poor quality of recording apparatus** such as CCTVs, BodyCams, and microphones. Much like upscaling, reconstruction can be used to feed the computer with proper, non-erroneous data and also generate images using CNNs to help authorities. Upscaling and reconstruction can be used in the following ways:

1. A study conducted by **Carnegie Mellon** researchers has shown that reconstruction can be used for generating facial images taking audio clips as input using GANs.
2. Upscaling can be used to **increase resolution of video and audio files.**
3. Recent advancements have also enabled us to recreate complete high resolution videos using a few images.

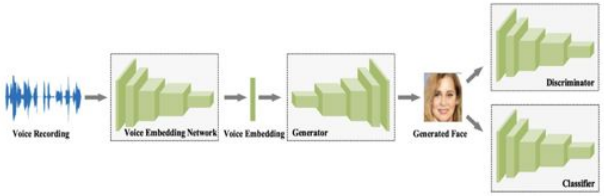


Figure 1: The proposed GANs-based framework for generating faces from voices. It includes 4 major components: voice embedding network, generator, discriminator, and classifier.

Figure 3: (a)-(e) The generated face images from regular speech recordings with different durations. (f) the corresponding reference face images. These 4 speakers (from top to bottom) are Danica McKellar, Cindy Williams, Damian Lewis, and Eva Green.

Voice Embedding Network			Generator		
Layer	Act.	Output shape	Layer	Act.	Output shape
Input	-	$64 \times t_0$	Input	-	$64 \times 1 \times 1$
Conv $3_{/2,1}$	BN + ReLU	$256 \times t_1$	Deconv $4 \times 4_{/1,0}$	ReLU	$1024 \times 4 \times 4$
Conv $3_{/2,1}$	BN + ReLU	$384 \times t_2$	Deconv $3 \times 3_{/2,1}$	ReLU	$512 \times 8 \times 8$
Conv $3_{/2,1}$	BN + ReLU	$576 \times t_3$	Deconv $3 \times 3_{/2,1}$	ReLU	$256 \times 16 \times 16$
Conv $3_{/2,1}$	BN + ReLU	$864 \times t_4$	Deconv $3 \times 3_{/2,1}$	ReLU	$128 \times 32 \times 32$
Conv $3_{/2,1}$	BN + ReLU	$64 \times t_5$	Deconv $3 \times 3_{/2,1}$	ReLU	$64 \times 64 \times 64$
AvePool $1 \times t_5$	-	64×1	Deconv $1 \times 1_{/1,0}$	-	$3 \times 64 \times 64$
Discriminator			**Classifier**		
Layer	Act.	Output shape	Layer	Act.	Output shape
Input	-	$3 \times 64 \times 64$	Input	-	$3 \times 64 \times 64$
Conv $1 \times 1_{/1,0}$	LReLU	$32 \times 64 \times 64$	Conv $1 \times 1_{/1,0}$	LReLU	$32 \times 64 \times 64$
Conv $3 \times 3_{/2,1}$	LReLU	$64 \times 32 \times 32$	Conv $3 \times 3_{/2,1}$	LReLU	$64 \times 32 \times 32$
Conv $3 \times 3_{/2,1}$	LReLU	$128 \times 16 \times 16$	Conv $3 \times 3_{/2,1}$	LReLU	$128 \times 16 \times 16$
Conv $3 \times 3_{/2,1}$	LReLU	$256 \times 8 \times 8$	Conv $3 \times 3_{/2,1}$	LReLU	$256 \times 8 \times 8$
Conv $3 \times 3_{/2,1}$	LReLU	$512 \times 4 \times 4$	Conv $3 \times 3_{/2,1}$	LReLU	$512 \times 4 \times 4$
Conv $4 \times 4_{/1,0}$	LReLU	$64 \times 1 \times 1$	Conv $4 \times 4_{/1,0}$	LReLU	$64 \times 1 \times 1$
FC 64×1	Sigmoid	1	FC $64 \times k$	Softmax	k

3D Reconstruction using audio input

Translational

In Neural Machine Translation (NMT), gender bias has been shown to reduce translation quality particularly when the target language has grammatical gender. Ideally we would reduce system bias by simply debiasing all data prior to training, but achieving this effectively is itself a challenge. Rather than attempt to create a 'balanced' dataset, we use transfer learning on a small set of trusted, gender-balanced examples. This approach gives strong and consistent improvements in gender debiasing with much less computational cost than training from scratch. Regularized training is a well-established approach for minimizing catastrophic forgetting during domain adaptation of machine translation.

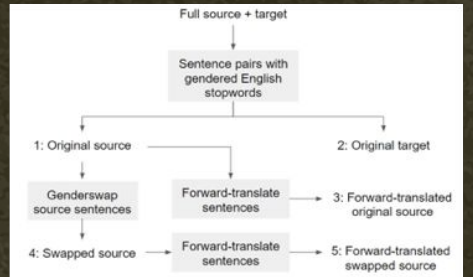


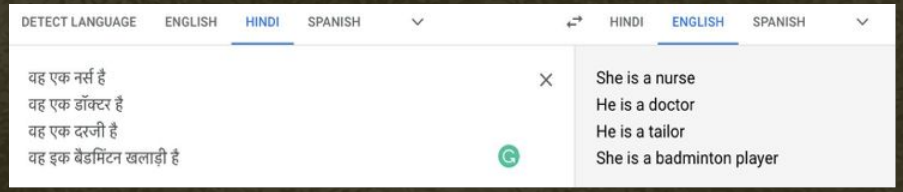
Figure 1: Generating counterfactual datasets for adaptation. The **Original** set is 1||2, a simple subset of the full dataset. **FTrans original** is 1||3, **FTrans swapped** is 4||5, and **Balanced** is 1,4||2,5

Fig 1

This lets us compare four related sets for gender debiasing adaptation, as illustrated in Figure 1:

- **Original**: a subset of parallel sentences from the original training data where the source sentence contains gendered stopwords.
- **Forward-translated (FTrans) original**: the source side of the *original* set with forward-translated target sentences.
- **Forward-translated (FTrans) swapped**: the *original* source sentences are gender-swapped, then forward-translated to produce gender-swapped target sentences.
- **Balanced**: the concatenation of the *original* and *FTrans swapped* parallel datasets. This is twice the size of the other counterfactual sets.

Translational



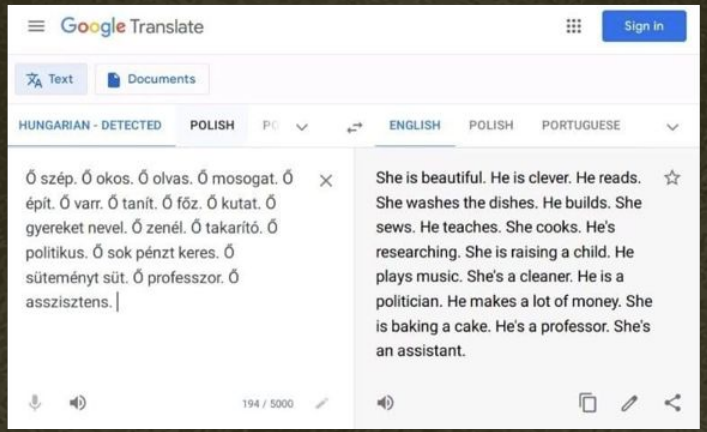


Fig *

The images (languages hindi and hungarian) display a real-life translational bias.

Instead of programming the model to output results arising from gendered queries (those which have a gender associated with them IN DATA, eg: male doctors, female nurses), it is essential to break the model's association to a gender (male doctors, nurses; female-doctors, nurses). This process is also called rewriting and uses language models, and generally requires rewriting datasets along with re-training the pre-existing model (like shown in fig *).

"The bias reduction of the existing Turkish-to-English system improved from 60% to 95% with the new approach. Our system triggers gender-specific translations with an average precision of 97% (i.e., when we decide to show gender-specific translations we're right 97% of the time)."

Hypothesis #3

Based on our investigation, we propose a hypothesis that posits any bias in an AI model manifests through outcomes strongly influenced by factors unrelated to the model's essential purpose. Consequently, an unbiased model would yield results minimally affected by changes in irrelevant factors. Though detecting implicit bias is often challenging, our hypothesis suggests that the existence of such bias may be discerned through its correlation with various other factors. Analyzing these factors in combinations or groups could unveil the presence of bias.

Furthermore, we postulate that rectifying a biased model is a feasible task, achievable through retraining on new, diverse, and unbiased data. Considering the model's data requirements, the creation of additional data based on previously inputted sample/actual data emerges as a practical approach. Consequently, if diverse and substantial volumes of data can be generated, the accuracy of AI models can be improved, mitigating concerns of overfitting or bias.

Hypothesis #3

EXPERIMENT METHODOLOGY

To test our hypothesis, we designed a series of experiments that employed our solution completely, or in part. In general, the workflow of our experiments was along the following lines. These experiments have been discussed later in this presentation.

Generate Random Data with GAN or Fetch random data from existing Large Datasets

Pre-Process the data to enhance diversity and to meet compatibility standards

Send the Data to Real Models or Simulations

Analyze the Results and find bias Qualitatively and Quantitatively

Hypothesis #3

We predict that a comprehensive solution for bias in AI involves 1: a human-centered ethics awareness campaign, 2: a framework for applying AI to identify bias in machine and human decision-making, and 3: a technical model for minimising implicit bias in data. These components will effectively mitigate issues with defining fairness, inherent human biases and bias in technical models respectively.

We propose a holistic solution to address bias in AI, consisting of three key components: 1) a human-centered ethics awareness campaign, 2) a framework employing AI to identify bias in both machine and human decision-making, and 3) a technical model aimed at minimizing implicit bias in data. These elements collectively aim to effectively address challenges related to defining fairness, inherent human biases, and bias within technical models.

The decision to incorporate both humanist and technical approaches stems from the recognition that enhancing the processes underlying AI development and testing goes beyond merely increasing human supervision as a solution to bias. Humans possess inherent cognitive biases and limitations in data processing, making it crucial to alleviate cognitive strain and expedite decision-making through automated cognitive biases, a fundamental aspect of human function today. However, as Eric Colson notes, the speed and almost unconscious nature of decisions do not always equate to optimal or accurate outcomes (Colson, 2019). These rapid decisions often come with the risk of incorporating "racial or social class categories or other unfair stereotypes" (Susan Fiske and Shelley Taylor, 2020), particularly concerning given AI's significant role in crucial social and life events such as hiring (Zhang et al., 2019).

The autonomy of cognitive biases not only makes pre-training data-processing and developmental interactions with AI more susceptible to translating people's biases into algorithms (Manyika et al., 2019) but also renders the removal of biases through ethics training alone resource-intensive and impractical. Moreover, AI's capacity to process extensive data volumes reduces its reliance on human shortcuts that often lead to prejudices. Involving AI assistants in areas like job posting or application evaluation can actively prevent or identify biased human decisions, promoting fairness for minorities and socially-disadvantaged groups in the employment process (Zhang et al., 2019).

Using AI to Mitigate Bias

Advocates of artificial intelligence (AI) contend that it can serve as a valuable tool in situations requiring impartial judgments, such as fair employment decisions during hiring processes. An article in Harvard Business Review highlights that machine learning systems, unlike humans, disregard variables that do not accurately predict outcomes based on available data (Manyika, Silberg, and Presten, 2019). Despite these assertions, using AI for such purposes has its downsides. For instance, in 2018, Amazon encountered unexpected bias in its AI recruiting tool, which exhibited a gender bias against women applying for technical positions, reflecting the existing male dominance in the technology sector (Dastin, 2018).

Contrary to leveraging AI for making decisions, our team was intrigued by the idea of harnessing AI's adeptness at detecting biases and prejudices. The focus shifted toward using AI to uncover human biases that might have otherwise gone unnoticed. In essence, our proposed solution advocates for employing AI to identify biases in both human and machine learning decisions. To assess the viability of this approach, we endeavored to apply existing tools designed to check bias on sample data and models, including Google's What-If Tool and IBM's Fairness 360 toolkit.

Using AI to Mitigate Bias

Google's What-If Tool

Initially, we employed Google's What-If Tool (WIT) to visually explore the behavior of a sample machine learning model across various inputs and diverse machine learning fairness metrics. Developed by Wexler in 2018, WIT serves as a visual interface for investigating machine learning model behavior. In our study, we applied WIT to analyze a publicly available dataset[1] encompassing criminal history, demographics, and COMPAS risk scores for defendants in Broward County. COMPAS, a contentious algorithm utilized by US courts to predict recidivism in criminal defendants, exhibited evident racial bias when analyzed through a basic machine learning algorithm trained on the COMPAS data.

Comparisons of inference scores highlighted the impact of features such as race on predicted recidivism scores (where 1 signifies low risk). Segregating the data based on racial features uncovered a disproportionate prediction of low-risk scores (depicted in blue) for Caucasians compared to African-Americans. This observation suggests the potential use of AI in detecting biases in both human and machine decision-making processes. Moreover, we found WIT to be user-friendly for machine learning beginners, given its integration with Google Colaboratory and comprehensive documentation.

The tool's flexible axis options enabled us to visually explore relationships between different factors, providing insights into potential instances of racial bias. Through additional experimentation, we determined that we could manually edit the features of a datapoint to observe shifts in the recidivism prediction score. Additionally, we could generate partial dependence plots illustrating the marginal effect of a feature on the model's predictions.

The tool suggested adjustments to threshold values to align with various fairness definitions, such as demographic parity and equal opportunity. Consequently, the What-If Tool appears to be a viable resource for beginners to scrutinize existing biases in algorithms.

IBM's AI Fairness 360

Furthermore, we explored IBM's AI Fairness 360 tool to assess various bias mitigation algorithms and metrics using a sample dataset. AI Fairness 360, an open-source toolkit designed for mitigating discrimination in machine learning models across the AI application lifecycle (Bellamy et al., 2019), offers three approaches to address bias in AI algorithms: pre-processing, in-processing, and post-processing of data. In this solution, our emphasis was on employing a pre-processing algorithm called reweighing. Reweighing involves assigning weights to each feature in the data before training a model, preventing bias in the model's output.

After applying reweighing and using the disparate impact metric to measure the disparity before and after the application of weights, we observed that reweighing successfully eliminated bias in the data. Our experimentation with this tool focused on assessing racial bias in the COMPAS dataset. By implementing Learning Fair Representations (LFR) and reweighing algorithms, we compared the impact of these pre-processing techniques on recidivism predictions and disparate impact—the ratio of favorable outcomes (low COMPAS scores below 12) for an unprivileged group (other races) to a privileged group (African Americans)[1].

Prior to the application of AI Fairness 360 packages, the disparate impact stood at 0.89420 (to 5 decimal places). Subsequently, after applying LFR and reweighing, there was a substantial improvement, with the disparate impact reaching 1.00000 (to 5 decimal places). It is worth acknowledging that numerous contributing factors exist within this dataset; nevertheless, our data unequivocally indicates that AI Fairness 360 serves as a potent tool for mitigating potential sources of bias.

Adversarial Debiasing

Adversarial debiasing operates as an In-Processing Algorithm, involving the construction of two distinct models. The first model predicts the target based on prior feature engineering and pre-processing steps applied to the training data. Simultaneously, the second model serves as an adversary, attempting to predict the sensitive attribute based on the predictions generated by the first model. In an unbiased scenario, the adversarial model should struggle to accurately predict the sensitive attribute. The adversarial model plays a crucial role in guiding modifications to the original model, weakening its predictive capability until it no longer accurately predicts the protected attributes based on the outcomes.

This iterative process enhances the accuracy of each model's representation, resulting in improved performance over repetitions. The practical application of this method is particularly beneficial in decision-making systems such as insurance, loan/banking, and judicial/law enforcement systems. The initial model in the sequence is exclusively trained using non-discriminatory data. The second model operates as the first adversary, attempting to identify the sensitive attribute with access to all available data. Subsequently, the original model undergoes adjustments through weighing to diminish bias.

In this chain, each subsequent model functions as an adversary, aiding in the adjustment of its predecessor to reduce bias. This sequential refinement ensures that the identification of bias becomes progressively less biased in accordance with the dataset. Notably, our approach extends the principles outlined in "Mitigating Unwanted Biases with Adversarial Learning" by BH Zhang, B Lemoine, M Mitchell.

Adversarial Debiasing

Addressing bias in AI presents a significant challenge, primarily stemming from underrepresentation in datasets and the oversight of certain nuances during development. Achieving true AI bias-free status is complicated as it relies heavily on the datasets used. Developers bear the responsibility of acknowledging and navigating potential skewness or bias within the data. To mitigate bias's impact on both data and AI systems, we propose incorporating counterfactual fairness (Wu et al., 2019) and leveraging generative modeling, as seen in Generative Adversarial Nets (GANs). This involves generating a counterfactual world resembling the original data while ensuring that specific attributes do not drive patterns in the data (Goodfellow et al., 2014) within the proposed bias-cleaning algorithm.

An inherent challenge in adversarial debiasing lies in its speed, with randomized unweighting, or adjusting parameters, proving insufficient for substantial change. Particularly for intersectional characteristics affecting multiple parameters simultaneously, teaching the algorithm to effectively address them can be problematic, leading to diminishing returns rapidly. Moreover, adversarial debiasing faces limitations in addressing biases stored in image recognition.

Despite these challenges, the universality of this solution cannot be overlooked. A suggested implementation involves users uploading resumes to job-hunting sites like LinkedIn, Indeed, or ZipRecruiter. On these platforms, the data undergoes processing and debiasing, streamlining the information provided to companies. This pre-processing significantly reduces the workload for companies and facilitates seamless integration into the job application workflow. Ultimately, this approach ensures that the debiasing algorithm is implemented efficiently, demonstrating its potential for widespread applicability (H.R., 2020).

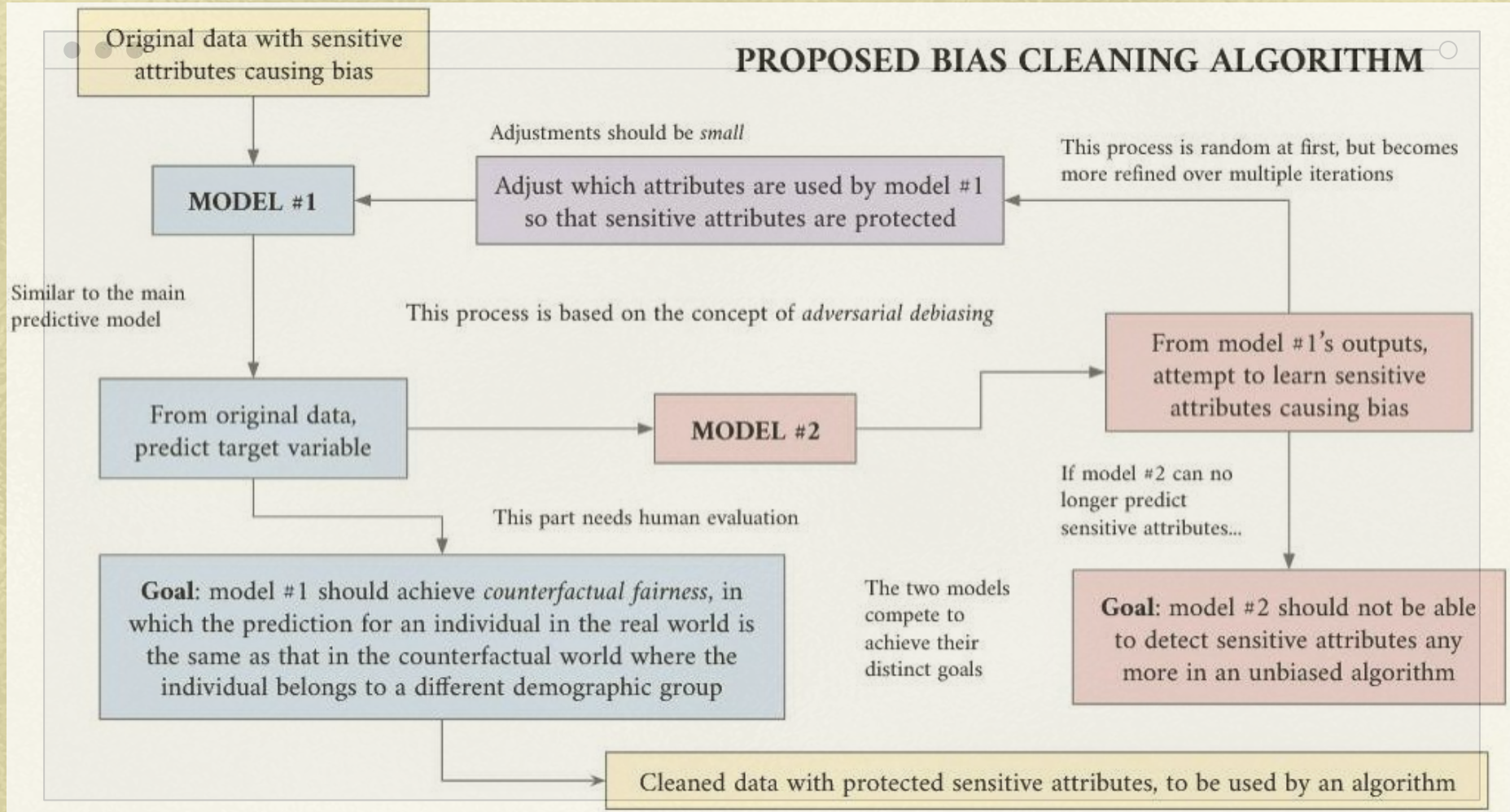
Proposed Bias-Cleaning Algorithm

Adversarial Debiasing

The implementation of solutions to address digitized bigotry poses a significant challenge in various domains, including sentencing, justice reform, and healthcare, where algorithms are increasingly prevalent. Recognizing the diverse manifestations of bias in different contexts, our team realized the impracticality of devising personalized solutions for each situation. The inherent specialization and personal nature of these issues rendered modifying existing algorithms ineffective on a large scale. Consequently, our team shifted focus to explore the possibility of creating a comprehensive algorithm capable of reading code and evaluating associated issues.

However, a notable obstacle emerged concerning the algorithm's ability to deduce finality or completeness, encapsulated in the Entscheidungsproblem, or the decision paradox (Turing, 1937). It became evident that the bias introduced by algorithms is not inherent in the algorithms themselves but rather stems from the underlying data. In an ideally unbiased world, algorithms would exhibit no bias, as there would be no bias in the data. In such a scenario, societal structures would not indirectly influence various person-related statistics, eliminating patterns for algorithms to observe.

Our proposed algorithmic solution is adversarial debiasing, designed to eliminate structural bias embedded in individual data. In this approach, two algorithms are employed. The first algorithm (A) operates on a given dataset for an individual, while the second algorithm (B) attempts to predict the person's protected identity from the data. The primary objective of the first algorithm is to iteratively adjust the data to diminish the predictive power of the second, while the second algorithm continuously trains to identify more patterns, simulating the long-term development of algorithmic biases (Lemoine, 2018).


PROPOSED BIAS CLEANING ALGORITHM
Original data with sensitive attributes causing bias
MODEL #1
Adjustments should be *small*
Adjust which attributes are used by model #1 so that sensitive attributes are protected
This process is random at first, but becomes more refined over multiple iterations
Similar to the main predictive model
This process is based on the concept of *adversarial debiasing*
From original data, predict target variable
MODEL #2
From model #1's outputs, attempt to learn sensitive attributes causing bias
If model #2 can no longer predict sensitive attributes...
This part needs human evaluation
Goal: model #1 should achieve *counterfactual fairness*, in which the prediction for an individual in the real world is the same as that in the counterfactual world where the individual belongs to a different demographic group
The two models compete to achieve their distinct goals
Goal: model #2 should not be able to detect sensitive attributes any more in an unbiased algorithm
Cleaned data with protected sensitive attributes, to be used by an algorithm

Solution #3

Based on our problem statement and hypotheses, we have developed a tool called Authentic Knowledge or AK, which can be used by AI developers to identify bias in their models and to mitigate the bias by training their model on new data provided by AK.

Within AK's platform, developers can grant access to their model's API along with sample data. AK utilizes Generative Adversarial Networks to produce random data matching the format and constraints of the original. Once a substantial repository of generated data is prepared, AK initiates bias analysis on the model.

To achieve this, AK employs permutation (or scattering) and feature interactions to generate multiple datasets. These datasets are sequentially sent to the model's API, and the resulting outputs are collected. AK then compares the results across all datasets, scrutinizing the impact of each field and field combination on the model's outcomes through permutations and feature interactions. These metrics are shared with AI developers, offering insights into potential biases.

If biases are identified, AK supplies relevant datasets generated earlier, focusing on the features responsible for the bias. These datasets exhibit ample diversity and randomness to counteract bias when used for model retraining.

Users can customize batch sizes and the extent of permutation or feature interaction during the bias detection phase. They also retain the freedom to handpick datasets for download and subsequent retraining.

Solution #3

FOR STRUCTURED-DATA DRIVEN MODELS

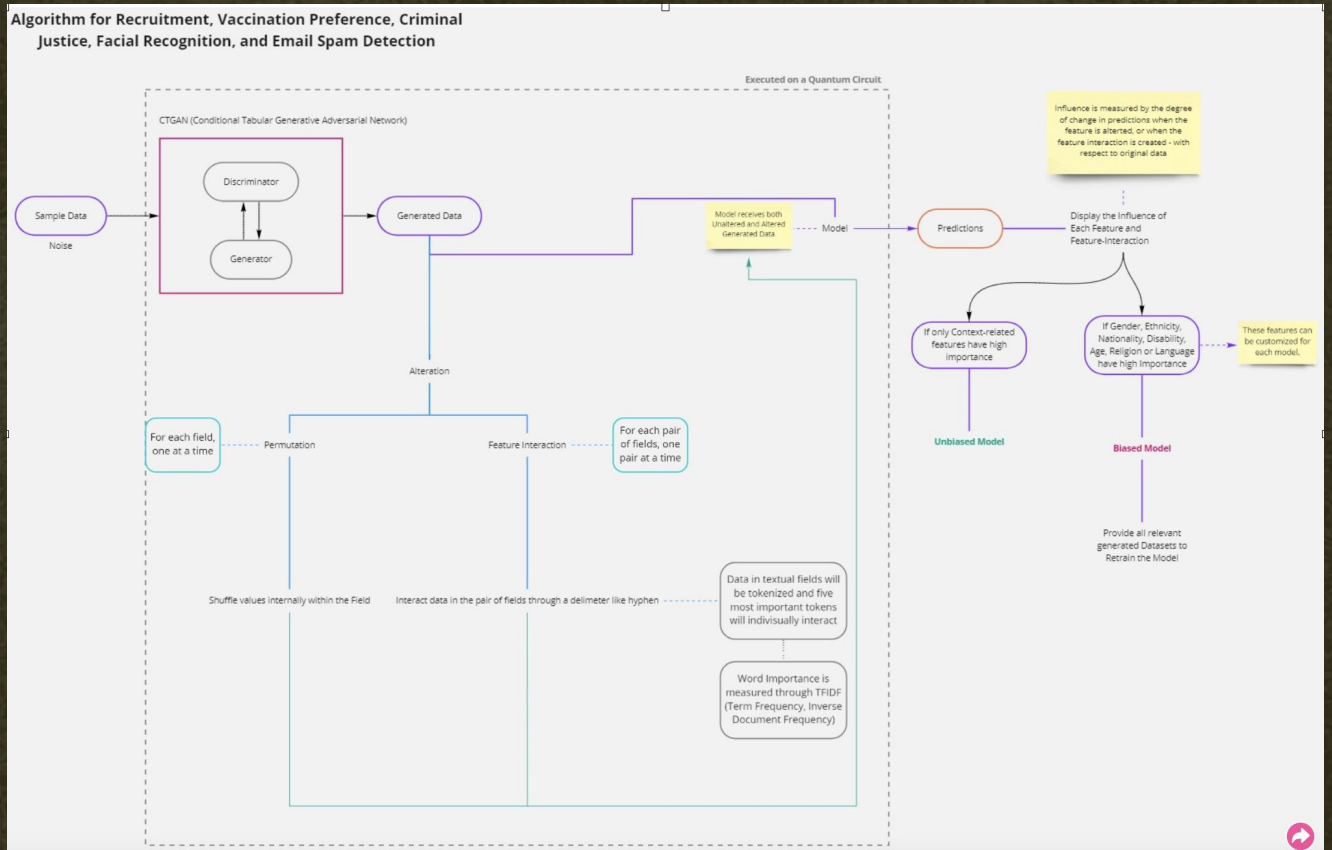
AK operates similarly for models running on structured data, encompassing both numeric and textual information. These models span various domains such as Recruitment, Vaccination Preference, Criminal Justice, Facial Recognition, and Spam Detection. Utilizing sample data as a form of noise, AK employs a Conditional Tabular Generative Adversarial Network (CTGAN) to generate novel data. Subsequently, this data undergoes a systematic alteration process involving permutations and feature interactions.

The modification of datasets occurs by altering one field or a group of fields at a time. AK then submits these modified datasets to the model, which has been previously tested with unaltered data, recording each output. For a model dealing with N fields, this translates to a minimum of N(N+1)/2 iterations. In the case of textual fields, up to 5 interactions are generated from the five most significant words or tokens, determined by TFIDF (Term Frequency & Inverse Document Frequency).

The computational demands of Generative Adversarial Networks (GANs) can be significant, leading to extended data generation times. For intricate models, the alteration process may also consume a substantial amount of execution time. To address these challenges, AK optimizes the process by implementing it on Quantum Circuits, resulting in a drastic reduction in computational requirements by orders of magnitude.

Solution #3

FOR STRUCTURED-DATA DRIVEN MODELS



Solution #3

FOR IMAGE-DATA DRIVEN MODELS

In the realm of image-based models, particularly those focused on Facial Recognition, a prevalent form of bias often surfaces in the form of diminished accuracy when identifying faces of individuals with diverse skin tones. While many commercially available facial recognition models excel at recognizing faces under optimal conditions, they struggle when faced with challenges such as small, poorly-lit, or blurry facial features. Notably, there is a tendency for these models to exhibit higher recognition rates for faces of individuals with lighter skin tones compared to others.

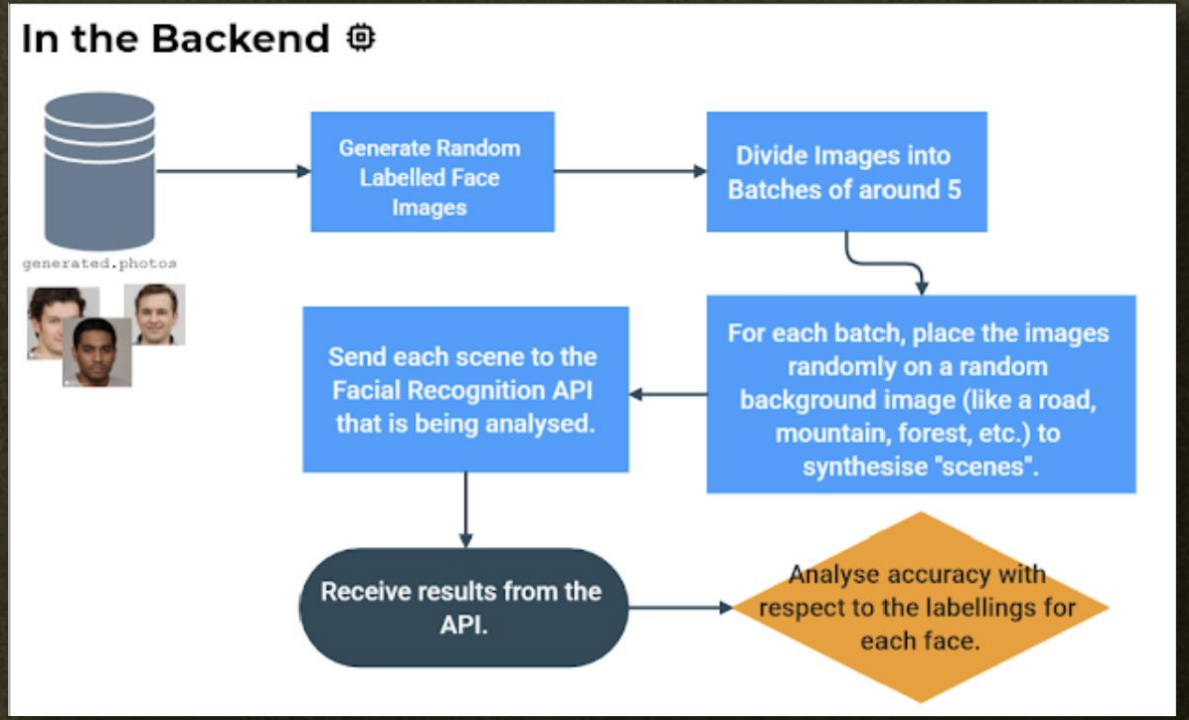
Addressing this issue, AK employs a Generative Adversarial Network (GAN) to generate a variety of facial images. These images, organized into groups, are then amalgamated with diverse backgrounds, including random landscapes, in various sizes and positions. The resulting images undergo analysis by the model in question, and the outcomes are meticulously documented. AK's assessment goes beyond mere face detection, incorporating confidence values associated with each prediction. Through repeated iterations, AK furnishes developers with metrics detailing the model's accuracy across different age groups, genders, and racial identities. The images crafted in the preceding steps are provided to developers for the purpose of retraining the model and alleviating bias.

Recognizing the time-intensive nature of GANs, we are exploring innovative approaches to streamline this process. One avenue being investigated involves leveraging a Quantum Approach utilizing Google's Cirq. Additionally, we have integrated the use of the Generated.Photos API to access labeled facial images, aiming to enhance efficiency and reduce execution times.

Solution #3

FOR IMAGE-DATA DRIVEN MODELS

An Oversimplified Representation of the Algorithm.



Solution #3

FOR NATURAL LANGUAGE DRIVEN MODELS

While we're in the process of exploring the development of a chatbot capable of engaging with other chatbots in various English accents and dialects (as well as other languages) to identify biases in text comprehension, we've devised a method for generating effective textual datasets to train chatbots. Our NLP program at AK can analyze existing conversational datasets (provided by AI developers who share the datasets their models were trained on) and craft new, diverse conversations. This involves substituting words with synonyms or slang expressions from different cultures, while preserving the original intents or semantics of the chatbot. Synonyms are sourced from a comprehensive dictionary database. Additionally, we're actively developing a technique to generate conversations with unconventional grammar and sentence structures, aiming to enhance the experience for non-native speakers. The resulting datasets, complete with labeled tokens, empower AI developers to retrain their chatbots, making them more compatible with diverse conversational styles across various cultural and geographical backgrounds.

Solution #3

FOR NATURAL LANGUAGE DRIVEN MODELS

Solution #3

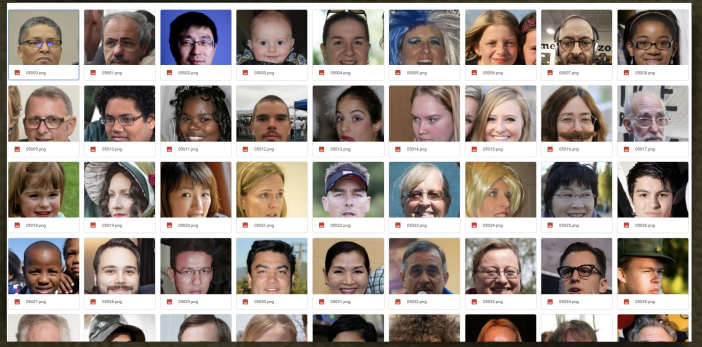
FOR NATURAL LANGUAGE DRIVEN MODELS

Before solidifying our understanding of Authentic Knowledge, we delved into researching existing datasets to identify potential bias and undersampling. This exploration substantiated our assertion that the predominant bias often originates from the data itself. To validate this, we randomly selected over 100 images from the Flicker Faces HQ (FFHQ) Dataset, which stands as one of the largest repositories of facial images globally, extensively used for training various models. The deliberate randomness in our image selection, coupled with the substantial sample size, ensured that our observations accurately represented the entire dataset. Our analysis revealed that, despite an equal distribution of male and female images, the dataset exhibited limited ethnic diversity, with a noticeable skew toward specific ethnic groups.

Solution #3

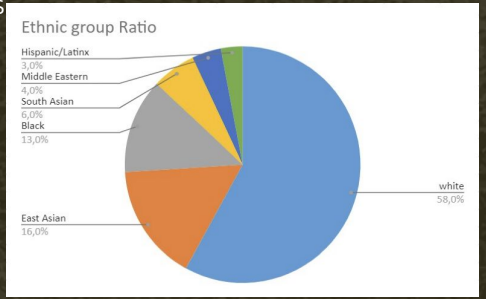
EXPERIMENT - BIAS IN IMAGE TRAINING-DATA

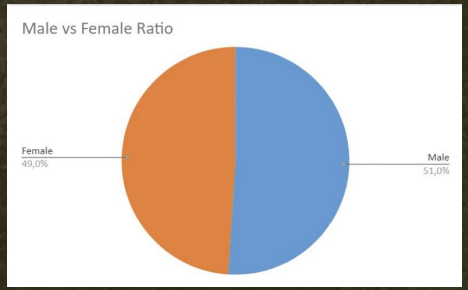
Prior to solidifying our understanding of Authentic Knowledge, we delved into an examination of existing datasets, specifically focusing on identifying biases and instances of undersampling. This investigative process was instrumental in validating our assertion that the primary source of bias resides within the data itself. To substantiate our findings, we randomly selected over 100 images from the Flicker Faces HQ (FFHQ) Dataset, a globally recognized repository of facial images extensively used in model training.



SCREENSHOT FROM FFHQ

The deliberate randomness in our image selection process, coupled with the substantial sample size, ensured that our conclusions accurately reflected the diversity inherent in the entire dataset. Our analysis revealed that, although the dataset exhibited an equal distribution of male and female images, it displayed limited ethnic diversity, with a notable skew towards specific ethnic groups.





Solution #3

EXPERIMENT - BIAS IN IMAGE RECOGNITION MODELS

We conducted an experiment utilizing the widely-used DeepAI facial recognition service to emulate the functionality of Authentic Knowledge. Our approach involved developing a JavaScript program that utilizes Generated.Photos's API to retrieve random facial images, which are then labeled and placed on diverse backgrounds obtained from Picsum. The program incorporates an algorithm that positions these facial images randomly in terms of size and location on the backgrounds, records these placements, and subsequently submits the composite images to DeepAI for analysis.

Upon analysis, DeepAI provides feedback by identifying face positions within the images, accompanied by a Confidence Level indicating the certainty of the predictions. We further scrutinized the results by visually delineating bounding boxes around DeepAI's predictions. Our program then correlated these predictions with various attributes such as age, ethnicity, skin color, eye color, hair color, hair length, and gender of the facial images.

Solution #3

EXPERIMENT - BIAS IN IMAGE RECOGNITION MODELS

```
      [256, 128],
      [256, 192],
      [256, 256],
    ];
    var selectedPos = getRandom(positions, 10);

    var canvas = document.getElementById('canv');
    var context = canvas.getContext('2d');
    var bgImg = new Image();
    bgImg.src = "https://picsum.photos/320";
    bgImg.crossOrigin = "anonymous";
    bgImg.onload = function() {
      context.drawImage(this, 0, 0);

      let i = 0;

      for (var x = 0; x < 10; x++) {
        var imgObj = new Image();
        imgObj.src = res["faces"][x]["urls"][1][64];
        var X, Y;
        [X, Y] = selectedPos[x];
        imgObj.X = X;
        imgObj.Y = Y;
        imgObj.crossOrigin = "anonymous";
        imgObj.onload = function() {
          let size = 32 + Math.random() * 32;
          context.drawImage(this, this.X, this.Y, size, size);
          i++;
          if (i == 10) {
            var renderedImg = new Image();
            renderedImg = canvas.toDataURL("image/png");
```

```
            renderedImg = canvas.toDataURL("image/png");

            deepai.setApiKey('22dc5752-8fa6-40ce-a382-ae5bc8295e94');

            (async function() {
              var resp = await deepai.callStandardApi("facial-recogniti
                image: renderedImg,
              });
              var c = document.getElementById("canvOut");
              var ctx = c.getContext("2d");
              var copyImg = new Image();
              copyImg.src = renderedImg;
              copyImg.crossOrigin = "anonymous";
              copyImg.onload = function() {
                ctx.drawImage(this, 0, 0);

                ctx.lineWidth = "3";
                ctx.strokeStyle = "red";
                for (var k = 0; k < resp.output.faces.length; k++) {
                  ctx.rect(...resp.output.faces[k].bounding_box);
                  ctx.stroke();
                }

              }
            })()

          }
        }
      }

    }
```

Implementation of JavaScript into the algorithm

Solution #3

EXPERIMENT - BIAS IN IMAGE RECOGNITION MODELS

Generated Image

Faces Detected

Solution #3

EXPERIMENT - BIAS IN VACCINATION PREFERENCE MODELS

Since COVID Vaccine Preference/Distributions models are not publicly available, we set up a simulation of the same and hard-coded a bias resembling the one in Stanford's Vaccination Algorithm. This was in the form of up to 25% lower preference for people who were between the ages of 20 and 35, and worked offline; and gave up to 25% higher preference to people above 60 years of age, working online. Although the exact cause of bias in Stanford's algorithm was different, our simulation created a similar effect. Besides age and nature of work, the simulation also included factors like previous health condition (on a 0 to 1 scale), number of COVID-19 contacts, and job profile (Health, Law Enforcement, or others). The logic for these factors was hard-coded as well. However, we also included less relevant factors like favorite color, favorite music, and education level. These factors were present in the data but were not used by the model.

This program was written in Python and hosted on PythonAnywhere in order to allow AK to access it like a regular API. Once this was in place, we proceeded to apply AK's process and recorded the results.

Solution #3

EXPERIMENT - BIAS IN VACCINATION PREFERENCE MODELS

```
from flask import Flask, request
from flask.json import jsonify
from myFuncs import *
import json
from flask_cors import CORS


app = Flask(__name__)
CORS(app)

@app.route("/", methods=['GET', "POST"])

def ri():

        data = {}
        try:
                data = dict(request.get_json())
        except:
                pass
        finally:
                with app.app_context():

                        if len((data).keys()) > 0:

                                frame = data['frame']
                                R = getPreferences(frame)

```

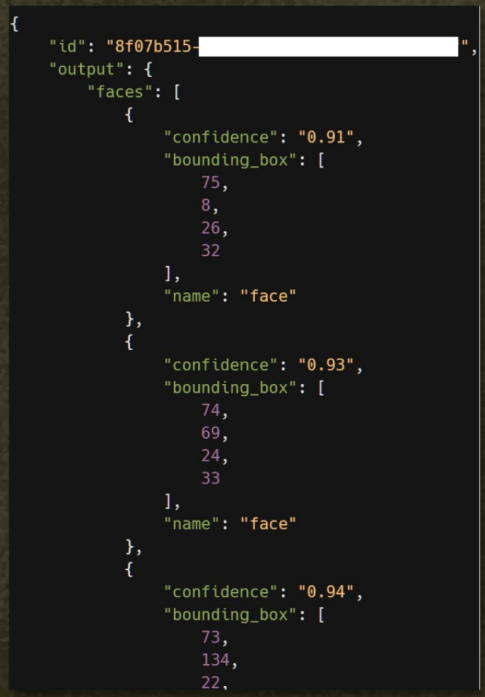
Solution #3

RESULTS

Our experimentation with existing artificial intelligence to identify and rectify bias in both human decision-making and machine learning has validated the effectiveness of our proposed solution. Specifically, the What-If Tool emerges as a suitable choice for beginners aiming to scrutinize inherent biases in algorithms. Conversely, while AI Fairness 360 is less user-friendly for those lacking coding proficiency, it stands out as a comprehensive tool for mitigating bias and deserves promotion among computer scientists involved in developing machine learning algorithms.

To promote a more sustainable and ethical AI landscape, it is crucial to advocate for the adoption of these user-friendly bias-checking systems in businesses and public organizations. These tools can be seamlessly integrated into web courses or online programs, making them accessible to the public through various educational platforms. During the prototyping phase of our bias-cleaning algorithm, it became evident that an approach inspired by adversarial debiasing could be applied effortlessly to classification and regression problems. However, concerns arise about the speed of implementation, as the randomized unweighting by model 2 may not be swift enough to induce significant changes in the fairness of model 1. Additionally, the complexities of relationships among intersectional characteristics pose challenges to the efficacy of this model. Consequently, while the principles of our proposed model are feasible for simple machine learning models, the algorithm exhibits sluggishness and ineffectiveness when confronted with more complex datasets.

To address these limitations, further research, potentially extending beyond the confines of this challenge, is essential to identify alternative bias-cleaning methods that can complement and enhance the efficacy of our proposed algorithm.



LIMITATIONS AND ANALYSIS

While our solution describes an innovative method of increasing fairness in machine learning models, there are disadvantages: These two disadvantages can have further implications when an adversarial chain is used in large scale learning systems. Nevertheless, it all depends on the priorities of the end-user.

1. Need for more computational resources
2. Higher complexity
3. Faster execution with traditional methods will demand a huge amount of computational resources, and hence there is a need to explore other options that can make the analyses quicker and still ensure their accuracy.

LIMITATIONS AND ANALYSIS

Through our experimentation with existing artificial intelligence to identify and address bias in both human decision-making and machine learning, we have substantiated the effectiveness of our suggested solution. Specifically, the What-If Tool emerges as a viable option for novices seeking to scrutinize inherent biases in algorithms. On the other hand, while less user-friendly for individuals lacking coding proficiency (such as the general public), AI Fairness 360 stands out as a comprehensive tool for mitigating bias and warrants promotion among computer scientists engaged in the development of machine learning algorithms.

Moving forward, in order to foster a more sustainable and ethical landscape for AI implementation, it is imperative to advocate for the adoption of these user-friendly bias-checking systems within businesses and public organizations. These tools can be seamlessly integrated into web courses or online programs, making them accessible to the public through various educational platforms. During the prototyping phase of our bias-cleaning algorithm, it became evident that an approach inspired by adversarial debiasing could be effortlessly applied to classification and regression problems. However, concerns arise regarding the speed of implementation, as the randomized unweighting by model 2 may not be swift enough to induce significant changes in the fairness of model 1. Additionally, the intricacies of relationships among intersectional characteristics pose challenges to the efficacy of this model. Consequently, while the principles of our proposed model prove feasible for simple machine learning models, the algorithm exhibits sluggishness and ineffectiveness when confronted with more complex datasets.

To address these limitations, further research—potentially extending beyond the confines of this challenge—is essential to identify alternative bias-cleaning methods that can complement and enhance the efficacy of our proposed algorithm.

REFERENCES AND BIBLIOGRAPHY