Final Project: The Impact of Bias in AI Technology: An AI Ethical Challenges and Analysis

EAI 6400： Data Govern & Responsible AI

Course CRN: 202425

Professor: Dr. Mimoza Dimodugno

Xingyu Lian

Zirui Qiao

Zhenyu Zhang

Lian.xin@northeastern.edu

[qiao.zir@northeastern.edu](mailto:qiao.zir@northeastern.edu)

zhang.zhenyu6@northeastern.edu

Date: 03/28/2024

**Introduction:**

The facial recognition technology has been widely used in every part of our life, however, the problem of bias in data privacy has been raised as well. We want to what is the main reason that cause this bias in facial recognition technology, the regulation, or guidelines now to solve this problem, and what potential solutions could be used to improve this current solution.

After extensive research about the bias in Facial Recognition Technology, it has become increasingly clear that the current legislations are weakness in mitigating the ethical concerns and risks associated with facial recognition technology. Thus, we think there should have more accurate and strict regulation to solve the bias in Facial Recognition Technology.

**Bias in Facial Recognition:**

First of all, the inaccurate facial recognition models often cause bias due to imbalance in training data, particularly in the domain of policing. As we have discussed in the lecture, an imbalance in training data will most definitely result in bias in the model performance. Such bias may mostly affect the rights of the underrepresented social groups, due to the high likelihood of low representation in the training data, given their minority in the social structure, and the bias in policing and justice system can be especially damaging.

For instance, the Robert Williams case of 2020, where a man was arrested after Detroit Police’s facial recognition application wrongfully identified him as the mugger they were searching for. The Robert Williams case and many like it highlight the pressing concern with facial recognition technology in criminal investigations, where a wrongful identification can severely violate the personal freedom of the innocent public.

It is particularly troublesome that new, unstable technologies like FRT are being incorporated within systems of policing and surveillance in the US where histories of racial disparity and inequality already exist (Perkowitz, 2021). Detailed studies by researchers at MIT and Microsoft Research and at the US National Institute of Standards and Technology (NIST) identified persistent inaccuracies in algorithms that were designed to detect and/or identify faces when applied to people of color. The algorithms were also less accurate for women than for men, with the largest errors, up to 35 percent, arising for female faces of color, according to the MIT/Microsoft study (Perkowitz, 2021). In other words, the legislations should emphasize the demographic diversity in the development of these algorithms, as historically, datasets have lacked representation from nonwhite populations, leading to inherent biases.

These recent ethical studies and legal disputes have led to a temporary pitstop of application of facial recognition technology in policing and justice system. Policymakers have begun to call for legislation to establish standards and regulation for facial recognition technologies, as in a series of hearings held in 2019, the US House Oversight Committee achieved bipartisan support in favor of regulation. Meanwhile, local governments of Somerville, Massachusetts, and of Oakland and San Francisco, California, have prohibited the use of FRT in the police departments (Perkowitz, 2021). However, these local actions can not address the existing bias in FRT in the domain of policing nationwide.

**Proof with Biased and Unbiased Dataset:**

To prove the affection of bias, we use the unbalanced datasets -- LFW to do an experiment (Labelled Faces in the Wild (LFW) Dataset, 2018).

The LFW (Labeled Faces in the Wild) datasets is a key resource for those working on face recognition technology, specifically designed to tackle the challenges of recognizing faces in "unconstrained" environments, which means the faces aren't perfectly posed and the lighting conditions can vary widely. Originating from the University of Massachusetts, Amherst, the datasets comprises thousands of face images collected from the web, each meticulously processed to ensure the faces are centered.

The component we focused on, the lfw\_allnames.csv file, is particularly useful as it lists the names associated with each face in the datasets, along with the number of images available for each individual. This makes it an essential tool for organizing the data and selecting subsets for training and testing face recognition models.

# Exploratory Data Analysis:

After load the dataset as a dataframe, we first explore it use describe() function, the result is shown in Figure 1.0:

屏幕上有字

描述已自动生成

Figure 1.0: lfw\_allnames describe

From Figure 1.0, we can see that there are 5749 images in total and the mean is 2.3, percentile 25, 50 and 75 are 1, 1 and 2 respectively which means most people have only 1 or 2 images, while the max number of image provided is 530. Therefore, this dataset is very unbalanced on the number of image provided.

By showing the distribution of people by number of images, we explore this trend more visually.

图表, 直方图

描述已自动生成

Figure 1.1: Distribution of People by Number of Images

图形用户界面

低可信度描述已自动生成

Figure 1.2: Distribution of Exact Image Percentage

After calculation, 70.8% of people have only one image and 13.55% people have only 2 images. The percentile of 3 images is 89.4%, almost 90%.

People with only one image cannot be used for both training and testing (otherwise, the model will train itself using test data, cause overfitting). Therefore, we collect people with at least 2 images.

文本

描述已自动生成

Figure 1.3: Save People with 2 or more Images to CSV File

Explore the data with cumulative distribution plot.

图表, 散点图

描述已自动生成

Figure 1.4: Cumulative Distribution of Number of Images

The closest percentile to 90% is 9 (90.6%). Now, visualize the top 20 people by number of images.

图表

描述已自动生成

Figure 1.5: Top 20 People by Number of Images

# Model Training:

After the simple data processing and exploration, we try to classify 1680 classes using a custom CNN model and a pretrained ResNet-34 model with unbalanced original data and balanced preprocessed data.

We use Oversampling to balance the dataset for the reason that we want to control the variables (so that the balanced and unbalanced datasets use the same amount of data to train the model), which better highlights the impact of data bias.

The following image shows the Custom CNN model structure:

文本

描述已自动生成

Figure 2.0: Custom CNN Model Structure

Since this experiment was used to analyze the effect of data bias, the structure of the CNN model was not carefully designed, which inevitably led to over-fitting. Therefore, we add “Learning Rate Decay” in the training loop which decrease learning rate when validation loss increase. There are 4 combination in total and each is trained for 10 epochs.

## Custom CNN Model:

### Custom CNN with Unbalance Datasets:

图表, 折线图

描述已自动生成

Figure 2.1: Custom CNN with Unbalance Datasets Loss Curve

文本

描述已自动生成

Figure 2.2: Custom CNN with Unbalance Datasets Test Accuracy

The validation accuracy reached 24.8% after 10 epochs. Although applying “Learning Rate Decay”, the validation loss still keep increasing which shows a clear trend of over-fitting. The model achieve 0.06% accuracy on test datasets. 24.8% vs 0.06% shows a huge over-fit trend.

### Custom CNN with Balance Datasets:

图表, 折线图

描述已自动生成

Figure 2.3: Custom CNN with Balance Datasets Loss Curve

文本

描述已自动生成

Figure 2.4: Custom CNN with Balance Datasets Test Accuracy

The validation accuracy reached 77.32% after 10 epochs. The validation loss was decreasing at the beginning but soon increases which also shows a trend of over-fitting. The model achieve 2.20% accuracy on test datasets. 77.32% vs 2.20% shows a larger trend of over-fitting compare to using unbalance datasets.

### Section Conclusion:

Although both balanced and unbalanced datasets can lead to overfitting, the balanced dataset was tested with greater accuracy but with greater overfitting. This could be the result of oversampling. It was calculated that in the unbalanced dataset, the average person had 4.45 images, but 46.4% had only 2 images and 63.7% had less than 4 images. Therefore, for these 63.7% of people, the model using the balanced dataset must learn from only 1 or 2 images multiple times, resulting in severe overfitting.

## Pretrained ResNet-34 Model:

### Pretrained ResNet-34 with Unbalance Datasets:

图表, 折线图

描述已自动生成

Figure 2.5: Pretrained ResNet-34 with Unbalance Datasets Loss Curve

文本

描述已自动生成

Figure 2.6: Pretrained ResNet-34 with Unbalance Datasets Test Accuracy

The validation accuracy reached 30.88% after 10 epochs. The validation loss keep decreases except epoch 8 shows that the model is learning well. The model achieve 4.52% accuracy on test datasets. 30.88% vs 4.52% still shows a large trend of over-fitting compare to using unbalance datasets.

### Pretrained ResNet-34 with Balance Datasets:

图表, 折线图

描述已自动生成

Figure 2.7: Pretrained ResNet-34 with Balance Datasets Loss Curve

文本

描述已自动生成

Figure 2.8: Pretrained ResNet-34 with Balance Datasets Test Accuracy

The validation accuracy reached 77.86% after 10 epochs. Validation losses are always decreasing and are the best performing of all the model and dataset combination. The model achieve 8.63% accuracy on test datasets. 77.86% vs 8.63% still shows a large trend of over-fitting compare to using unbalance datasets.

### Section Conclusion:

Both pretrained models show good learning trends, and it is reasonable to believe that more epochs can make the models more accurate on the test dataset. Similar to custom CNN model, the balanced dataset was tested with greater accuracy but with greater overfitting. The reason for this may also be the same as mentioned in the custom CNN section.

In both custom CNN model and pretrained ResNet-34 model, balanced dataset performs better than unbalanced dataset.

# Model Evaluation:

Now it's time to see how much of impact an unbalanced datasets can have. By exploring the distribution of correctly predicted test data for each combination of model datasets, we can visualize the impact easily. In this section, we need to compare the cumulative distribution graphs with Figure 1.4. For ease of reference, Figure 1.4 has been replayed here.

图表, 散点图

描述已自动生成

Figure 1.4: Cumulative Distribution of Number of Images

## Custom CNN Model:

### Custom CNN with Unbalance Datasets:

图表, 散点图

描述已自动生成

Figure 3.0: Custom CNN with Unbalance Datasets Cumulative Distribution Graph

Custom CNN with Unbalance Datasets only has one correct result which is “George\_W\_Bush” with 530 images (the most image provider). The model has been biased beyond belief.

### Custom CNN with Balance Datasets:

图表, 折线图

描述已自动生成

Figure 3.1: Custom CNN with Balance Datasets Cumulative Distribution Graph

The custom CNN using the balanced dataset also shows lower accuracy (37 out of 1680 tests), but its data distribution is closer to that of the original dataset (Figure 1.4), which suggests that its bias is not very large. This model dataset combination is overall much better than the custom CNN model for the unbalanced dataset.

## Pretrained ResNet-34 Model:

### Pretrained ResNet-34 with Unbalance Datasets:

图表

描述已自动生成

Figure 3.2: Pretrained ResNet-34 with Unbalance Datasets Cumulative Distribution Graph

Pretrained ResNet-34 with Unbalance Datasets made 76 correct prediction out of 1680. Its data distribution is close to that of the original dataset (Figure 1.4). However, there are still a huge distance between them. In Figure 1.4, The closest percentile to 90% is 9 (90.6%); but in Figure 3.2, the closest percentile to 90% is 44 (89.5%).

### Pretrained ResNet-34 with Balance Datasets:

图形用户界面, 图表

中度可信度描述已自动生成

Figure 3.3: Pretrained ResNet-34 with Balance Datasets Cumulative Distribution Graph

Pretrained ResNet-34 with Balance Datasets made 145 correct prediction out of 1680, which is the highest among all model dataset combinations. Its data distribution is very close to that of the original dataset (Figure 1.4). In Figure 1.4, The closest percentile to 90% is 9 (90.6%); and in Figure 3.3, the closest percentile to 90% is also 9 (89.0%).

图表, 折线图, 散点图

描述已自动生成

Figure 3.4: Pretrained ResNet-34 with Unbalance and Balance Datasets Cumulative Distribution Graph

### Conclusion:

Both Custom CNN mode and Pretrained ResNet-34 shows that model training with balanced (smaller bias) dataset result in more accurate prediction with smaller bias. Although we can use data preprocessing to balance the classes, this also leads to severe model over-fitting (77.32% vs 2.20% and 77.86% vs 8.63%). This suggests that simply using data preprocessing to reduce data bias is not enough, and that the best case scenario is still to collect a dataset that is inherently biased with little bias in order to achieve the best accuracy.

In addition, the imbalance of data classes is only one of the manifestations of bias, which also includes sampling bias, measurement bias, selection bias, confirmation bias, etc. This experiment is mainly used to show the effect of Class Imbalance Bias of the data, from which analogies are made to the effect of other bias. Therefore, it is very challenging to collect a dataset with a small bias in real life.

**Regulation and Guidelines of problem:**

Many Countries has built their many regulations to protect the data privacy like European Union’s General Data Protection Regulation (GDPR) which focus on give people right to protect their own data. America’s California Consumer Privacy Act (CCPA) which focuses on give customers right to know where their data was used. And China’s Personal Information Protection Act (PIPL) which focus on who to processing the personal information been collected. We see that these regulation gives clear guidelines of how to protect the data on security, accountability, transparency and so on. However, there is a few things about how to deal with bias in dataset and even when they mention the fairness of data, there isn’t exactly standard to put this regulation to each dataset that will be used.

**EU's General Data Protection Regulation (GDPR):**

The scope of GDPR jurisdiction covers all companies that process the data of EU residents. Companies in the EU must comply with the GDPR. Companies outside the EU also need to respect the GDPR if they process the data of EU residents. This law is the most stringent for user protection and will increase the compliance costs of enterprises to a certain extent. The spirit is commendable, but if it is really implemented by enterprises, the cost will be extremely high.

* Three principles of legality, fairness, and transparency: Data information related to the individual data subject should be processed in a legal, fair, and transparent manner.
* Data collection should have a clear purpose: Personal information collection should have a specific, clear, and legal purpose, and any method that is inconsistent with the above purposes will no longer allow the data to be processed.
* Minimization principle of data collection: Personal data collection should be limited to all necessary data related to the purpose of data processing.
* Accuracy: Personal data should be accurate and, if necessary, kept as up to date as possible.

The key to law is to collect minimum, clear and accurate user data which make sure that the dataset collect is fair. This law prevents the happening of bias in vicious circle stop from the root as there is no bias data, algorithm will not give bias result, and people will not use bias result to do next data collection and prediction.

**America’s California Consumer Privacy Act (CCPA):**

The CCPA only applies to the personal information of California residents, where personal information means information that directly or indirectly associated with a particular consumer or household. Also, if data is collected from California residents, the any organization need to follow CCPA regardless of where the organization is headquartered. Moreover, CCPA gives customers four right to protect their data that is to know, to ensure, to opt-out and not to discriminate.

* Right to know: Consumers should be informed about what personal information an organization collects about them and how it is used.
* Right to erasure: With some exceptions, consumers can have information collected about them deleted.
* Right to Opt-Out: Consumers can prevent the sale of their information to third parties.
* Right not to discriminate: Organizations cannot treat users differently for exercising their CCPA rights.

Thus, we can see that the goal of this law is to protect the data privacy of customers, different with GDPR which focus on limit the data that companies could collect, CCPA focus more on let customers know and choose whether to sell their data or not.

**China’s** **Personal Information Protection Act (PIPL):**

For PIPL, personal information refers to any type of information, whether electronic or other records, relating to an identified or identifiable natural person within the People's Republic of China. The processing of PI includes collection, storage, use, change, transmission, provision, disclosure, deletion, etc. The processing of PI is allowed only done these several condition:

* Obtain individual consent.
* Necessary for the performance of a contract to which the individual is a party or for the implementation of human resources management.
* Necessary to perform legal duties or obligations.
* As necessary to respond to a public health emergency or to protect a person's life, health, or property during an emergency.
* For news reporting and other activities in the public interest.
* To process PI that the individual has disclosed to himself or otherwise lawfully disclosed.
* Laws and regulations provide otherwise.

Even people who processing data are get allowed, they need to follow several requirements when processing data:

* Legal, fair, necessary, and good faith. PIPL Article 5.
* Purpose limitation and data minimization. PIPL Article 6.
* Open and transparent. PIPL Article 7.
* Accuracy and completeness. PIPL Article 8.
* Safety and Accountability. PIPL Article 9.
* Limited Data Retention. PIPL Article 19.

These requirements are in order to make sure the processing of data protect the PI by regulation.

**Problem in Strategy:**

By comparing the laws mentioned above, we can see that GDPR focus on the data collect regulation, CCPA only regulation big organization that get benefit on data, and PIPL focus on how to processing personal information.

The facial recognition technology needs facial picture which belong personal information should be regulated under these laws. First, GDPR clearly define what kind of data should or should not be collect. Thus, the facial picture that been collect into dataset under the regulation of GDPR will not have biased problem in most case. However, this regulation can’t make sure that the bias problem for facial recognition will never appear because there don’t have an exact data description about fairness, necessary, accurate and so on. Thus, different companies have different ideas about these definitions and there might exist bias of data collection with these flexible definitions.

Then, when talking about CCPA, it only regulates company which not only rely on the data to get benefit but also have a large amount of benefit than average company. Thus, this law will only give customer to know the use of their data but less regulation if small size company have bias on the facial picture they collect. Lastly, PIPL care about the collect and process of personal information seriously, however, it has similar problem like GDPR that the definition of standard doesn’t have exactly number to make it as clear as possible. Also, the punishment of PIPL is also not enough, when people don’t collect or process these facial picture collect, the most common punishment is don’t give them right to these data in the future or let customers choice whether give their data to this data processing group in the future or not.

**Current situation:**

Currently, citizens may combat the misuse of FRT with lawsuits under the current legislations of consumer right protection, and there have been successful cases. In 2020, ACLU and other entities filed class-action lawsuit against Clearview AI for scraping three billion facial images from social media without the data owner’s consent. This clear violation of Illinois Biometric Information Privacy Act (IBIOA) and other consumer privacy legislations has led to a settlement with the additional condition of Clearview AI being permanently forbidden from making its faceprint database available to most businesses and other private entities in the US (ACLU, 2022). However, one must realize that current legislations and lawsuits under them only address the issue of consent, not the bias in the FRT models.

**Potential Solution:**

To address bias in FRT, we must focus on both responsible data collection and utilization. In future legislations, user consent should be mandatory for data collection, and the resulting dataset must be fair and representative. This means treating data from minority groups with particular care to avoid skewing the results. Furthermore, users of algorithmic predictions should critically evaluate the outputs, considering potential bias that may arise from the algorithms themselves.

For instance, we can establish a precise definition of accuracy. By developing multiple models for the dataset in question, if any model demonstrates an accuracy below 80%, it indicates the presence of an accuracy issue. To address the fairness issue,

we observe that if the target image in the dataset comprises less than 10% of the total images, there exists a fairness problem as the minority target might not be adequately represented to address the necessity issue, if duplicate data in the dataset exceeds 20%, it suggests that unnecessary data has been collected, which is not permissible.

Other than including clear definition and performance standards, future legislations may increase its range of cover to all personal information as we discussed in PIPL section.

Furthermore, the penalties outlined in these laws should be more substantial than the direct harm caused to the organizations' interests if they collect or process biased facial images, as demonstrated in GDPR or CCPA.

Lastly, future legislations should include measures to effectively and proactively address issue of misuse and bias, such as a channel for reporting suspected bias in commercially applied FRT and a board of legislators and industry professionals to determine accountability of wrongfully applied FRT.

Ultimately, we’d like to advise legislators, for the purpose of oversight and regulation of FRT in future legislations, to consider the following questions:

1. Who is responsible for overseeing the development, procurement, and testing of FRT systems, to ensure robust management and procedures to address bias effectively?
2. Under what circumstances and for what purposes is the use of FRT to capture individuals' images deemed appropriate?
3. What specific measures regarding consent, notification, and oversight should be implemented to ensure fairness and transparency in these applications?
4. What criteria should guide the establishment and utilization of facial database, and for what purposes should they be employed?
5. What mechanisms of accountability should be instituted for different applications of FRT?
6. How are procedures for lodging complaints and raising challenges made accessible and equitable to all individuals?

**Conclusion:**

In summary, we can determine that the unbalanced dataset is one of the main reasons that bias problem exist in AI risk. Although regulations like GDPR, CCPA, PIPL have made efforts to people’s right to know the use of their data, to protect personal data, and to process data, there should be extra mechanisms to regulate the bias in the data. We discussed several proposed solutions, such as establishing clear standard for model performance as part of legislation and prohibited the use of facial recognition technology when bias event happened, as well as extending the range of legislations to cover all personal information, increase the penalty of breaching individual privacy, and establishing channels of reporting and addressing bias in FRT.

References:

ACLU. (2022, May 9). In Big Win, Settlement Ensures Clearview AI Complies with Groundbreaking Illinois Biometric Privacy Law. Retrieved from <https://www.aclu.org/press-releases/big-win-settlement-ensures-clearview-ai-complies-with-groundbreaking-illinois>

Agnė Limantė. 2023. Bias in Facial Recognition Technologies Used by Law Enforcement: Understanding the Causes and Searching for a Way Out, Nordic Journal of Human Rights, <https://doi.org/10.1080/18918131.2023.2277581>

Almeida, D., Shmarko, K., & Lomas, E. (2022). The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: A comparative analysis of US, EU, and UK regulatory frameworks. *AI and Ethics, 2*, 377–387. <https://doi.org/10.1007/s43681-021-00077-w>

Baik, Jeeyun Sophia. 2020. Data Privacy Against Innovation or Against Discrimination?: The Case of the California Consumer Privacy Act (CCPA). *Telematics and Informatics*, 52. <https://doi.10.1016/j.tele.2020.101431>, Available at SSRN: <https://ssrn.com/abstract=3624850>

California Consumer Privacy Act (CCPA). 2018. Retrieved from <https://oag.ca.gov/privacy/ccpa#sectiong>

Calzada I. 2022. Citizens’ Data Privacy in China: The State of the Art of the Personal Information Protection Law (PIPL). *Smart Cities*. 2022; 5(3):1129-1150. <https://doi.org/10.3390/smartcities5030057>

Christina Tikkinen-Piri, Anna Rohunen, Jouni Markkula. 2018. EU General Data Protection Regulation: Changes and implications for personal data collecting companies. Computer Law & Security Review, Volume 34, Issue 1, 134-153, <https://doi.org/10.1016/j.clsr.2017.05.015>.

Labelled Faces in the Wild (LFW) datasets. (2018, May 17). <https://www.kaggle.com/datasets/jessicali9530/lfw-dataset>

Perkowitz, S. (2021). The Bias in the Machine: Facial Recognition Technology and Racial Disparities. *MIT Case Studies in Social and Ethical Responsibilities of Computing*. <https://doi.org/10.21428/2c646de5.62272586>