A Fairness Tool for Bias in FR

Alice Loukinova, Dylan Dasgupta, Rohan Krishnamurthi, William Cutler, Joseph Robinson

Abstract

Bias in FR is an issue preventing users from accessing technology to the same extent as their equal counterparts. The goal in creating FR software was to allow anyone and everyone to have facilitated, hands-free access to their devices. As this issue has been prevalent since FR has been developed, it is not a minute problem to ignore. We have studied this bias, and looked to understand the source, as well as the methods to create a completely fair way of assessing users. As FR becomes more advanced and widespread, the issue follows in the same manner. It violates social justice values and conducts an unintentionally biased evaluation of users. The issue cannot be taken lightly and needs a solid resolution.

I. Introduction

Facial recognition (FR) is a rapidly expanding form of cybersecurity that allows a user to access their device completely hands-free. By assessing one's facial features, an identity can be created and used to unlock phones, computers, and more. FR requires a complex software backend, along with a simple front-end GUI so users of all ages and capacity can use the application equally. Bias through ethnicity and gender has been a major issue presented to software teams working with FR, as certain demographics are assessed unfairly compared to others.

By taking a balanced dataset of people from various backgrounds, tests can be run to see where bias and error occurs most. It is unfair that some users are not subject to a nondiscriminatory evaluation, while others do not have to worry about experiencing any error. The sources and methods to eliminate this bias will be explored here.

TABLE I
ABBREVIATIONS AND ACRONYMS

Label	Meaning	Purpose
M	Male	
F	Female	
A	Asian	
B	Black	
I	Indian	
W	White	
AM, BF…	Asian male, black female…	(M – AM, BF) These all serve to discuss the various subgroups under analysis
BFW	Balanced Faces in the Wild Dataset	Database under consideration
CNN	Convolutional Neural Network	Allows for higher quality image extraction
DET	Detection Error Trade-off	Plots FPR and FNR for facilitated comparison
FNR	False negative rate	Test result fails to correctly identify something as correct
FPR	False positive rate	Test result incorrectly identifies something as correct
FR	Facial recognition	Technology capable of identifying an individual through their unique facial features

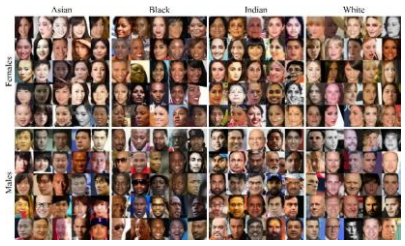

Figure 1 BFW Image set [1]. Various faces from the Balanced Faces in the Wild (BFW) database used to establish an unbiased identification method by representing each race and gender to the same extent.

A. Motivation

FR technology is used in a wide variety of fields and locales, including airport security, social media, and surveillance. It is pervasive in areas very sensitive to bias, with estimates indicating as many as 117 million Americans are affected by its use in law enforcement [2]. When used frequently in serious scenarios such as police investigations, bias in FR represents more than an inconvenience; it's a legitimate civil rights concern.

Because FR is a developing field, regulations and evaluation metrics have not been extensively developed. A complete solution does not currently exist, so the proliferation of tools for measuring bias furthers overall progress to ensuring equity in AI in practice.

Evaluating bias in FR algorithms is a monumental task in part due to two significant obstacles that we have identified. The first is the lack of a publicly available dataset that is balanced among ethnicity, sex, and the combination of the two, against which FR models could be tested. The second is that every FR algorithm is biased in its own unique ways, so there is intrinsic need for an evaluation that is automatic and custom-tailored to a given algorithm.

By implementing an easy-to-use dashboard that analyzes the bias of a given FR algorithm against our balanced dataset, we work to solve both problems simultaneously.

Table 2. Gender and Race Database Statistics: *Statistics of the Balanced Faces in the Wild (BFW) database are grouped here by subgroup and a specific value. There are a million pairs total under analysis, with a constant 30,000 positive pairs being assessed for each gender under said subgroup. Overall, F performs inferior to M for I and W, while M performs inferior to W for A and B.*

Ethnicity	Asian		Black		Indian		White		Aggregated
Gender	F	M	F	M	F	M	F	M	
Faces	2500	2500	2500	2500	2500	2500	2500	2500	20,000
Subjects	100	100	100	100	100	100	100	100	800
Faces/Subjects	25	25	25	25	25	25	25	25	25
Positive Pairs	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	240,000
Negative Pairs	85,135	85,232	85,016	85,141	85,287	85,152	85,223	85,193	681,379
Pairs Total	115,135	115,232	115,016	115,141	115,287	115,152	115,223	115,193	921,379

B. Contributions

This tool was designed due to a team effort and everyone's contributions were immensely significant. There was significant exploratory analysis done for age data, experimentation with dashboard logic and design, optimization of plots, and much organizing/documentation. The development of the dashboard was reliant on everyone completing their parts whether that be data analysis, plot generation, or data validation.

II. Prerequisite Information

FR software is a method of analyzing a person through facial features. By using biometrics to map an individual's face, it can create an identity and find a match within its database to confirm to whom the face belongs. FR has a massive range of applications and usage, from security to marketing and far more that we may not even be aware of. Police and FBI agents agree that our lives are safer because of FR technology [3], without which, could leave many crimes unsolved and suspects roaming free. This technology has been around for nearly a century by now but has only become increasingly popular over the last several years as crime rates have evolved and more severe security measures have become necessary for the population's well-being. It is fair to say that FR transformed the world of security and identification. It is another level in the biometric identification world, but to progress further we need to resolve the bias present in our current identification method. This bias will continue to hinder us from perfecting FR technology unless a fairness tool can be created and implemented. A nondiscriminatory evaluation is essential to the advancement of this software.

A. Facial Recognition (FR)

The FR industry is expected to grow immensely over the next half decade (2017-2022) and nearly double in industry value [4]. FR works similar to how we identify people in our lives. By looking at someone we know, we can give them a name and identity based on their physical appearance and unique attributes. The FR software recreates this in a more systematic and algorithmic manner, to ensure the match in their database is perfect. Refer to *figure 1* for a visualization of an FR database concept. There are certain key features on one's face that can be used to identify a person. Distances between someone's eyes, forehead and chin, and nose to chin are all valuable measurements when it comes to FR.

B. Bias in FR

FR has been in use for many decades now, since the 1960s in the United States. It has only progressed in terms of accuracy and feasibility, but there remain unsolved issues that violate social justice values and prevent the software from being perfect. One of the major problems discovered with FR technology is the bias in identifying a user [5]. Certain users are not subject to the same assessment as other users, due to their demographics. Overall, females experience more error with FR than males, and Asians as a subgroup experience the most bias out of all ethnicities. *Table* 2 supports this by displaying the unequal distribution among all races and gender. By categorizing users into four subgroups, based on ethnicity, the bias becomes far easier to target and understand. The four subgroups we have looked at are: white, Asian, black, and Indian. While there is opportunity to add more detail to these subgroups, these four best categorize all users in a more general sense. The confusion matrix in *figure 2* displays the percent error encountered between each ethnicity and gender, numerically explaining that Asian females will experience the most bias, while white males will encounter the least.

C. Similar Technology

We are not the first group to investigate this bias, as many frontier companies specialize in FR. There is countless software made in attempt to perfect FR, and an endless consumer supply. Nearly half of U.S residents with a smartphone own an iPhone, which utilizes FR software to allow access into the device. FR is found everywhere in everyday life, without us even being aware. In airports, public venues, stores, and so many more locations [6]. It has become a part of basic technology to assist groups with various tasks that all concern identification. There is no need for tedious database searching or matching pictures anymore, FR facilitates all these processes in different ways. FR is not the first method of biometric security, and certainly not the last. It is a steppingstone in the advancement of cybersecurity measures. Fingerprint identification, commonly known as "touch ID" is another example of biometric identification, it is a more out-of-date method, but nonetheless works in a similar fashion to FR. Technology similar to FR usually deals with security of some sort, but there are many programs that use

similar features despite being in a completely different field. Any software that utilizes a large database of humans/living organisms or has some process of identification through biometrics can be considered similar to FR. The software and processes behind FR technology are very complex and involved, it is hard to believe most of these interfaces can be so simple.

III. EXPLORATORY DATA ANALYSIS

In this section some methods for measuring bias in facial recognition will be explored.

A confusion matrix is a very useful tool for measuring recall, precision, specificity, and accuracy in a model. Recall measures how much was correctly predicted out of all the positive cases.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (1)$$

Precision looks at how many cases are truly positive out of the positive ones predicted correctly.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

Specificity is a measure of how correct the true negative rate is.

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (3)$$

Accuracy looks at how much was predicted correctly out of all the classes.

$$Acc = \frac{True\ Positive + True\ Negative}{True/False\ Positive + True/False\ Negative} \quad (4)$$

The value of a given cell corresponds to the extent to which an image belonging to the subgroup in that row was matched incorrectly with an image from the subgroup in the corresponding column. Naturally, these errors coalesce at the diagonals since it is more likely to confuse an image with another from the same subgroup. Since the model displayed below measures errors, lower diagonal values would indicate a model that is better at making predictions. The model tends to discriminate the best between white male (WM) faces and the worst between Asian female (AF) faces with the highest number being in the top left of the matrix and the lowest in the bottom right. This really emphasizes that the subgroups are meaningful and distinct.

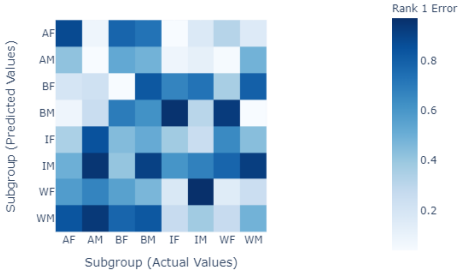

*Figure 2: **Confusion matrix.** The error (%) for the various faces of BFW vs. all others. AF performs the worst as WM performs the best, meaning AF is confused most often. This shows that subgroups are useful for FR because we can break each category/group down to analyze which genders and which races are being confused the most*

Another effective plot is the Detection Error Trade-off (DET) Curve which is a plot of measured error rates and shows the false negative rate (FNR) as a function of the false positive rate (FPR). This shows the tradeoff between sensitivity with the FPR and specificity with the FNR. A false positive is when the prediction is positive and it's false, this is a Type 1 Error. A false negative is when the prediction is negative and it's false, this is a Type 2 Error. The three DET plots below show a comparison between genders, ethnicities, and the subgroups. These plots are all based on the assumption of a global threshold meaning it's constant for each subgroup.

In Figure 3, the curve representing male faces is lower meaning that it performs better. In Figure 4, the lowest curve is for a white ethnicity and the highest represents an Asian ethnicity, indicating the model performs the best on white faces and the worst on Asian faces. Similarly, in Figure 5, the lowest curve belongs to white males and the highest belongs to Asian females. The main takeaway is that a threshold varying across different subgroups would perform better than a global one and can yield a constant FPR.

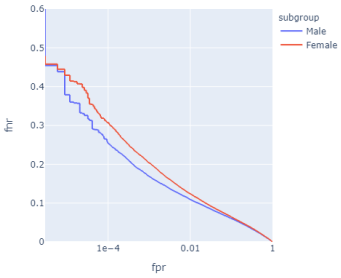

Figure 3: Detection Error Tradeoff (DET) Curve between genders

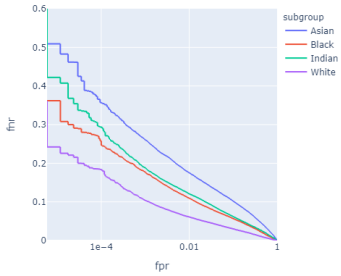

Figure 4: Detection Error Tradeoff Curve between ethnicities

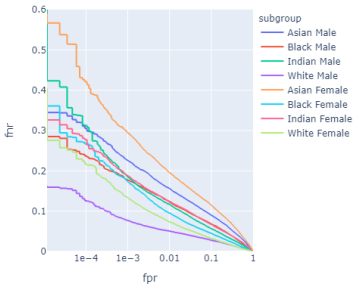

Figure 5: Detection Error Tradeoff Curve between subgroups

Another visualization is a signal detection model (SDM). In Figure 6, the SDM curves across subgroups show a discrete distribution of scores. The imposter scores in orange have a median at 0 and follow a gaussian pattern with most of the variation across subgroups occurring in the upper percentiles, while genuine pairs in blue vary in both the lower and upper percentiles. Generally, less overlap between genuine scores and imposter scores is better. In Figure 6, the highest area of overlap between the genuine and the imposter pairs occurs for the Asian female subgroup. In comparison, there is virtually no overlap for the white male subgroup, which consistently has the best performance.

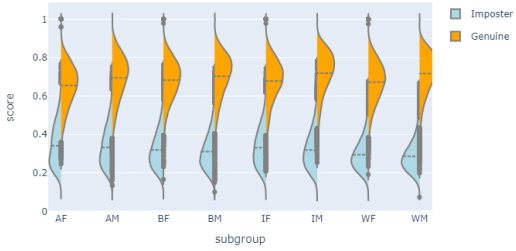

Figure 6: Signal detection model (SDM) across subgroups

The last tool for performance measurement to be discussed is the Receiver Operating Characteristic (ROC) curve. The ROC is a probability curve with true positive rate (TPR) plotted against the false positive rate (FPR).

$$TPR(Sensitivity) = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

$$FPR(1 - Specificity) = \frac{False\ Positive}{True\ Negative + False\ Positive} \quad (6)$$

This means that the top left corner where the false positive rate would be zero and the true positive rate would be one is the ideal spot for a perfect model. This is very idealistic and not usually the case, however other indicators of a good model are a larger area under the curve and a steeper curve. The steeper the curve the more the TPR is maximized and the FPR is minimized. The ROC curve clearly displays the inverse relationship between sensitivity specificity, as sensitivity increases the specificity decreases. In Figure 7 and Figure 8, the ROC curves for gender and subgroups can be seen. In Figure 7, the curve representing male faces is steeper and has a larger area under the curve indicating better performance. Similarly, in Figure 8, the steepest curve with the most area under the curve belongs to the white male subgroup.

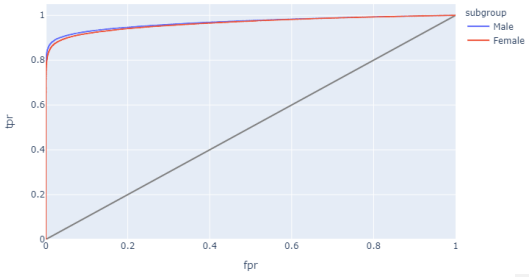

Figure 7: Receiver Operating Characteristic (ROC) curve between genders

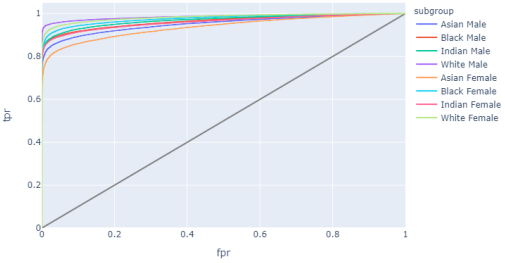

Figure 8: Receiver Operating Characteristic (ROC) curve across subgroups

A. Who is the dashboard for?

The dashboard is created for those looking to fully understand the back end of our fairness tool. Each of the plots created here can be recreated with a different set of data through the dashboard. It allows for user input and can be used to display user-inputted data in the selected plots. Researchers interested in FR and attempting to create their own model can utilize the dashboard to evaluate their own models. Rather than tediously testing it on their own, the dashboard allows for a far quicker and automatic analysis of one's FR model. Students involved in FR whether it is through research, their coursework, or other motivation can use dashboard as a great learning tool [7] For someone who wants clean visual aids, as well as an in-depth explanation of how we can arrive at these plots can look to the dashboard for more technical related questions.

B. Tools Used

- Plotly
- Dash
- Pandas
- Dash DataTable
- Dash Core Components

How should I cite Python libraries? Should I even keep this section or just introduce and cite the tools as they come up?

C. Implementing the Data

Figure 9: Dashboard Data Flow: This diagram shows how data is read and stored by the dashboard throughout its processes.

The most important step in creating the dashboard was developing a flow for the data and understanding how it would interact with the components as the user interacts with the dashboard GUI. With time and processing efficiency in mind, the following data flow, shown in *Figure 9* was created.

When the dashboard is first opened, it loads and presents the default dataset – bfw-v0.1.5-datatable. The default data is stored as a pickle in the app's directory. The user is also given the option to upload their own dataset. If he/she chooses to do so, the dataset is first checked to see if it conforms to the requirements of the dashboard and if so, it is relabeled and converted to a pickle file, which is then stored in the *cache-directory* folder of the app's directory. From there, either the default or newly cached data's pickle file is read by three components of dashboard - the data preview, score distribution, and error evaluation sections. While being read from the cache, the filters specified by the user are applied to the dataset. The filtered data is then displayed in the tables presented in the data preview section and the charts shown in both the score distribution and error evaluation sections.

D. Frontend-Backend Interaction

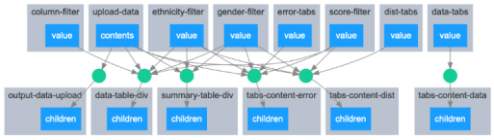

Figure 10: Callback Graph

While the backend of the dashboard follows the flow shown in *Figure 9*, the frontend of the dashboard had to be designed to keep up with the many user interactions while communicating them to the backend.

The primary component with the most critical connections is the 'upload-data' component. This component triggers the callback that contains first three steps of the data flow shown in *Figure 9*. When the dashboard is initialized, the callback reads the default data to be shared with the rest of the dashboard. If the user uploads his/her own data via the 'upload-data' component, the callback is triggered again and the data is validated, cached, and read by the rest of the dashboard. It connects to the 'output-data-upload', 'data-table-div', 'summary-table-div', 'tabs-content-dist', and 'tabs-content-error'. The connection denotes that any time 'upload-data' is triggered, the connected components are all updated. It's this connection that automatically updates all the charts and plots whenever the data is updated.

The 'ethnicity-filter', 'gender-filter', and 'score-filter' components represent the three global filters that are applied to all the charts and plots in the dashboard. Whenever the filters are updated, the data is read from the cache with the filters applied and sent to the 'data-table-div', 'summary-table-div', 'tabs-content-dist', and 'tabs-content-error' components. The 'column-filter' component only affects the layout of the data preview, so it's only connection is to 'data-table-div'.

The tabs are represented by the 'error-tabs', 'dist-tabs' and 'data-tabs' components, which are connected to 'tabs-content-error', 'tabs-content-dist', and 'tabs-content-data', respectively. Whenever a tab is selected, the value is sent to the child and the contents of the tab are updated.

Commented [DD1]: What should I refer to it as? Who should I credit it to? Cite Github?

Commented [DD2]: these names only really make sense to me so should I create a table giving more understandable aliases to the callbacks and html divs?

Commented [DD3R2]:

E. Dashboard Components

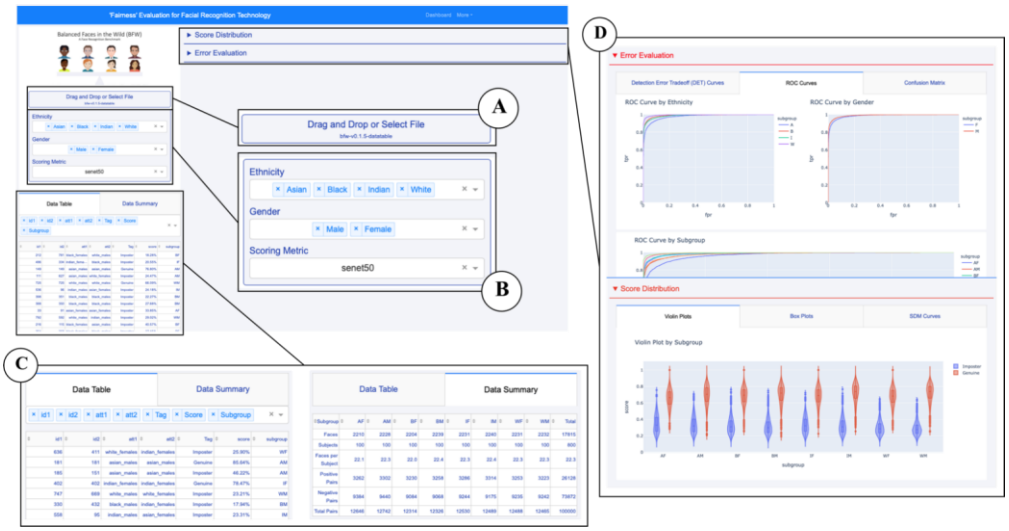

Figure 11: Dashboard Layout

(A) Data Upload

The data upload component is where the user can upload his/her own dataset. The dataset can either be dragged and dropped directly onto the component, or the component can be clicked, which will prompt a file directory to appear, allowing the user to select a dataset. This component was built using the *dash_core_components* library's *Upload* component **[reference]**. The *Upload* component allows users to upload CSV files into the dashboard, which is then read as a base64 encoded string that contains the filename and location.

(B) Global Filters

The filters located below the data upload component are dropdown filters that apply to the entire dashboard. By default, all genders and ethnicities are selected, and the scoring metric is set to 'senet50'. Whenever the user changes any of the filters, the data is re-read with the filters applied. These filters were created using the *dash_core_components* library's *Dropdown* component **[reference]**. The *Dropdown* component takes either multiple values as a list or single values as a string, which are read by each of the visualization components and applied to the dataset.

(C) Data Preview and Data Summary

The Data Table and Data Summary tabs on the dashboard enable the user to preview the dataset and the summary statistics of it. When the dataset is read, a random subset of the data is taken and displayed in the Data Table tab. This data table is reactive to the global filters and has an additional column filter, which allows the user to show and hide specific columns in the data preview. The Data Summary tab provides the user with a numerical breakdown of the dataset grouped by subgroup. The counts shown in this table represent the entire dataset and are also reactive to the global filters. Both tables were created using a combination of *Pandas***[reference]** and the *Dash DataTable* **[reference]** component. Pandas is used to create the tables and calculate the dataset summary, while Dash DataTable is used to convert the Pandas tables to the interactive HTML tables shown on the dashboard.

(D) Score Distribution & Error Evaluation

The score distribution and error evaluation components are dropdown sections created using HTML's Summary and Details elements. When either heading is clicked, a series of tabs containing various visualizations are revealed. The tabs were created using the was built using the *dash_core_components* library's *Tabs* and *Tab* components **[reference]**. The *Tab* component controls the style, value, and contents of each individual tab, while the *Tabs* component holds multiple *Tab* components together. At both the score distribution and error evaluation components the global filters are taken in and applied to the data being read, which is then passed to *Plotly* **[reference]** to create the visualization denoted by the selected tab.

The score distribution component contains a Violin Plot, Box Plot, and SDM Curves Plot detailing the distribution of scores across each subgroup. The Violin Plots and Box Plots provide the same insight as the SDM Curves which are shown and described in *Figure 6*. The error evaluation component contains the DET Curves (*Figures 3, 4, 5)*, ROC Curves (*Figures 7,8)*, and Confusion Matrix (*Figure 2)*.

IV. Future

FR software is simply the beginning of a new wave of advanced biometric cybersecurity measures, that use human features to protect one's device. Initially, fingerprint identification was created in the late 1800s [7], then came FR shortly after, and we have only progressed from there. Creating a just method of evaluating users is key to the future of biometric security, if this issue remains present the advances made in this field are automatically limited by this prevalent issue. Concepts such as retina scanners are hindered by dilemmas such as this, because if FR cannot be perfected and fully functional: why would a more complex security measure be? FR has been used to help countless people and is far more useful than one might imagine. Police use FR to track down suspects and prevent further crimes from happening. Airports use FR to locate passengers or possible suspects to ensure their customers don't have to worry [8]. The list goes on: shopping centers, law enforcement, and more. FR is only just beginning to assist our world, by perfecting the technology and creating a completely unbiased method to assess users: the possibilities are endless.

References

[1] Robinson, Joseph P., et al. "Face Recognition: Too Bias, or Not Too Bias?." *arXiv preprint arXiv:2002.06483* (2020).

[2] Bedoya Alvaro., et al. "The Perpetual Lineup." Center on Privacy and Technology, Georgetown Law, 18 Oct. 2016, https://www.perpetuallineup.org/

[3] Valentino-devries, Jennifer. "How the Police Use Facial Recognition, and Where It Falls Short." *The New York Times*, The New York Times, 12 Jan. 2020, www.nytimes.com/2020/01/12/technology/facial-recognition-police.html.

[4] Symanovich, Steve. "How Does Facial Recognition Work?" *Official Site*, us.norton.com/internetsecurity-iot-how-facial-recognition-software-works.html.

[5] Singer, Natasha, and Cade Metz. "Many Facial-Recognition Systems Are Biased, Says U.S. Study." *The New York Times*, The New York Times, 19 Dec. 2019, www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html

[6] Marr, Bernard. "Facial Recognition Technology: Here Are The Important Pros And Cons." *Forbes*, Forbes Magazine, 19 Aug. 2019, www.forbes.com/sites/bernardmarr/2019/08/19/facial-recognition-technology-here-are-the-important-pros-and-cons/#7533dba314d1.

[7] Saravanan, Raja, and Raja Saravanan. "Facial Recognition Can Give Students Better Service (and Security)." *Ellucian*, www.ellucian.com/insights/facial-recognition-can-give-students-better-service-and-security.

[8] German, Ed. *History of Fingerprints*, onin.com/fp/fphistory.html.

[9] Martin, Nicole. "The Major Concerns Around Facial Recognition Technology." *Forbes*, Forbes Magazine, 25 Sept. 2019, www.forbes.com/sites/nicolemartin1/2019/09/25/the-major-concerns-around-facial-recognition-technology/#79bd0e704fe3.

[10]

Alice Loukinova (M'76–SM'81–F'87) and all authors may include biographies. Biographies are often not included in conference-related papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author's educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state, and country, and year the degree was earned. The author's major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author's last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (publisher name, year) similar to a reference. Current and previous research interests end the paragraph. The third paragraph begins with the author's title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the IEEE. Finally, list any awards and work for IEEE committees and publications. If a photograph is provided, it should be of good quality, and professional-looking. Following are two examples of an author's biography.

Dylan Dasgupta

Rohan Krishnamurthi.

B
William Cutler