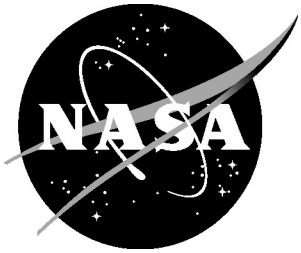
NASA/TM–20210010446



Social Bias in AI and its Implications

Sribava Sharma
Langley Research Center, Hampton, Virginia

Mallory Suzanne Graydon
Langley Research Center, Hampton, Virginia

March 2021

NASA STI Program Report Series

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

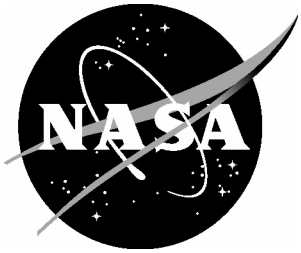
- Access the NASA STI program home page at http://www.sti.nasa.gov

- Help desk contact information:

https://www.sti.nasa.gov/sti-contact-form/

and select the "General" help request type.

NASA/TM–20210010446



Social Bias in AI and its Implications

Sribava Sharma
Langley Research Center, Hampton, Virginia

Mallory Suzanne Graydon
Langley Research Center, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia
23681-2199

March 2021

Available from:

NASA STI Program / Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199
Fax: 757-864-6500

Table of Contents

Abstract

Background. Previous studies have documented many different types of biases that exist in artificial intelligence (AI) and machine learning (ML) systems. However, these studies either do not examine the societal implications of said biases or are not easily accessible to readers. This creates a gap where software operators and developers may not be aware of the potential issues that users may face.

Objective. To review the literature on AI and ML bias with a focus on social implications and to document our findings for the purposes of education and reference. Our goals are not to identify faults within specific systems but to (a) raise awareness to the kinds of issues that have occurred in systems using technology that might be used at NASA and elsewhere and (b) to provide interested parties with a gateway into existing work on social bias in AI and ML systems.

Methods. We conducted a literature review of publications related to AI, ML, and the systems comparable to those that are or might be used at NASA. We held interviews and conversations with colleagues at NASA Langley to gain a deeper understanding of the software used in their project(s) to ensure that our review is relevant to NASA. We focused our investigation specifically on the social effects of AI and ML bias.

Results. Review of the literature reveals that bias in AI and ML can potentially have harmful social impacts on individuals and/or groups of people within our society. By affecting people differently according to characteristics such as race, gender, or sexual orientation, AI and ML systems may exacerbate existing social inequities.

Conclusion. AI and ML systems have been shown to exhibit bias leading to social harm. Those who develop and deploy software must be aware of this phenomenon so that such harmful effects can be detected and appropriately mitigated. These biases have been documented in a wide range of AI/ML systems demonstrating that it can be present in any system that functions on human datasets or interacts with people. Being versed in the downfalls that other systems have encountered is one certain method to prevent these social biases from arising in systems such as those NASA builds or buys.

1. Introduction

Today, the use of automated technology is expanding with no end in sight. We face a daily barrage of artificial intelligence (AI) and machine learning (ML) systems designed to ease our lives and to learn[1] from our behaviors. These include face-recognition software to unlock mobile phones, filtered news feeds on social media, targeted advertisements based on search history, and many more [1]. The AI and ML field is ever-growing with a plethora of possible applications. In 2017, Andrew Ng called AI "the new electricity," predicting that deep learning will transform our lives [2]. Three years later, it is difficult to think of a field of work that AI and ML have not touched.

AI technology was created with the aim of simplifying our decision-making process and creating opportunities that otherwise would not be possible, and it has made rapid progress in recent years. Less than a decade ago, Apple had just unveiled Siri [3], Google developed a neural network to recognize images of cats on YouTube videos [4], and Ian Goodfellow designed the groundbreaking machine learning framework generative adversarial network (GAN) [5]. Today, there are more virtual personal assistants than you can shake a stick at, deep learning can identify and predict the development of an age-related macular degeneration more accurately than most eye care professionals [6], and GAN can paint a picture of a cat to your specification [7, 8].

AI systems are available to use by all people in the world regardless of race, gender, or any other differentiating attributes, provided they possess the economic means to access these systems. While it is created with the intention of functioning equally for everyone, it sometimes does not. AI systems have been shown to exhibit discriminatory biases toward individuals based on race, skin color, and gender among other factors. Photos of African American individuals have been tagged as 'apes' [9, 10], facial recognition technology functions poorly on dark-skinned individuals [11], and job candidate screeners have screened out females [12]. Events such as these have caused AI to gain notoriety as users realize that machines might produce results that contribute to racial inequities.

While AI and ML systems do not have malice, they function on data that reflect, or fail to reflect, the societies and environments described by those data. If that data reflects a social inequity, machines built from that data might reproduce that inequity when performing their function. As a result, they might disproportionately disadvantage groups of people, even if their creators had no such intention [13]. These instances can be described as the result of bias in AI systems.

Philosophers David Danks and Alex London define bias as a deviation from a standard in their discussion of algorithmic bias and autonomous systems [14]. Bias can be said to be a function of two components—a standard and a measure of deviation from this standard. Therefore, any instance where an autonomous system produces results that are significantly different from its intended purpose (i.e., the standard for the system) is an occurrence of bias [14]. Although there are many flavors of bias, the biases of interest to us are the types of bias that have a negative impact on society via amplification of existing social inequities along categories such as race, gender, and sexuality. We refer to these biases in AI as social biases. Many issues surrounding the use of AI have come to light including ethics of AI, bias in AI and ML, and policies and regulations [15, 16, 17]. While each individual topic demands an in-depth analysis, our focus is on social biases exhibited in AI and ML systems.

[1] Some readers may object that anthropomorphisms like 'learn' imply a closer-than-warranted analogy between what humans do and what machine learning systems do. We sympathize, but use them anyway for consistency with the majority of the literature in this field.

Bias can be expressed in AI and ML systems in numerous ways: sample bias, historical bias, measurement bias, aggregation bias, and temporal bias.

Sample bias. Sample bias occurs when the dataset used to train the AI system is not representative of the population who interact with the AI. This type of bias is also known as unrepresentative bias: the AI development population is not representative of the user population. This occurs due to biased or skewed datasets and can be avoided by developers thoroughly inspecting training datasets [13, 16, 18]. ImageNet is a popular image dataset used to train AI and ML systems. However, a majority of the images are taken in North America and Western Europe. As a result, systems trained on ImageNet datasets perform poorly in parts of the world that are poorly represented in the dataset. For example, a search for 'bridegrooms' reveals classification of images with high confidence in the US and Australia, but poor classification in underrepresented countries such as Pakistan and Ethiopia [19].

Data collection methods may also cause sample bias [13, 16, 18]. The city of Boston released an app, StreetBump, to detect, record, and report potholes allowing the city to prioritize and repair the roads that require the most attention. However, it fast became evident that poor neighborhoods were not reporting potholes as often as affluent neighborhoods were. This observation was attributed to the fact that lower income people were less likely to have smartphones, older populations were less likely to have smartphones, and lower income people were less likely to own vehicles and more likely to use public transportation: factors that affected the apps ability to record potholes. Given these factors, the app was biased against poor neighborhoods and negatively impacted neighborhoods that were already disadvantaged [20]. A dataset collected from social media platforms may be skewed if the data are limited. User demographics differ across platforms. Women are somewhat more likely than men to use Facebook and Instagram and much more likely to use Pinterest, while men are somewhat more likely to use Reddit and Twitter [21, 22]. The userbase also differs by age, race, ethnicity, and parental educational background [23]. These factors must be taken into account when gathering data from social media platforms.

Historical bias. Historical bias occurs when bias that already exists in our society is reflected in the dataset leading to a biased AI. This type of bias can occur even if the dataset is truly representative of the population [13, 18]. Amazon's job application screener filtered out female applicants because the dataset used to train it revealed a world where men dominate the STEM fields and the AI mirrored this bias that exists in our society [12]. Here, the developers need to consider the biases that exist in our world and if they wish for them to be expressed in the AI. Historical bias is not always bad. African American women in Hampton Roads are fifty percent more likely to die due to breast cancer than white women are [24, 25, 26] and this data must be reflected in healthcare AI systems. However, it is harmful when healthcare AI discriminates against African American in a way that exacerbates health disparities [27]. Bias can be used for the benefit of marginalized population if it is properly utilized by AI and ML developers.

Measurement bias. Measurement bias occurs when a measurement technique does not perform equally well for all populations. The group of people that it does not perform well on will be negatively impacted. Inadequate measurement techniques can also skew datasets causing sample bias [13, 18]. Pulse oximeters, devices used to measure an individual's oxygen levels, perform less well on African American patients than their white counterparts, leading to nearly three times the frequency of undetected occult hypoxia [28]. Measurement bias is also seen when using proxy measurements to predict outcomes. For example, arrest records are used as indicators for crime rates. Using this measurement will show that poor neighborhoods have higher crime rates and African American individuals are more likely to be criminals than Caucasians. AI systems utilizing this information to predict crime rates and recommend punishments will be biased against these groups of people [29, 30].

Aggregation bias. Aggregation bias occurs when an umbrella definition is used to cover all groups of people without accounting for differences between them. This can cause minority groups to be discriminated against and a system that works well only for the group the characterization applies to, usually the majority groups. This problem can be overcome by accounting for the differences and incorporating information that correctly reflect this. In diabetes patients, HbA1c levels are universally used as a diagnostic tool and to monitor patients. However, it has been described that the relationship between HbA1c and blood glucose levels differ based on race and ethnicity [31].

Temporal bias. Temporal bias occurs over time when the population's behavior changes over time and the AI does not adapt to the changes. Populations and systems change over time due to changes in trends or system updates. The AI should mirror the population, but if it does not due to changes in one without an analogous change in the other then this dissonance can lead to a biased AI. In humans, changes can occur periodically (e.g., seasonally) or due to sudden dramatic events. An AI system should be adaptable to respond and reflect these changes in its functions. Research has shown how introduction of new features on social media platforms can affect user activity [32] and how major disasters, such as hurricanes and earthquakes, can temporarily change social datasets [33].

For more examples of bias in AI and ML systems, refer to other articles surveying the field [13, 18, 16].

These biases can enter an AI system through various avenues. To rectify the problem, it is important to identify how bias has entered a system. Below we discuss the main sources of bias in AI and ML technology: data collection, data processing, and training models.

Data collection. Bias may enter the system during the data collection process due to the manner the data was collected or if the dataset itself is skewed. Sample, historical, and measurement bias are ways bias can be expressed if the data collected for AI training is not representative of the entire population. AI systems that collect data automatically for the purposes of continued training and refinement of its function are also susceptible during the collection process. To avoid this, developers must vet their datasets to ensure accuracy and proper representation. Faulty datasets are a major cause for bias in AI. Additionally, researchers have developed algorithms to analyze datasets for bias providing software developers with the necessary tools to mitigate bias [16].

Data processing. Bias can be introduced into AI systems via operations that process the data. These include actions such as data cleaning, enrichment, and aggregation. Data cleaning is the process of correcting and fixing any errors in the dataset prior to training the AI. While it is meant to prevent bias in data, it can be counterproductive if key data components are removed, or if missing and incorrect values are filled in using assumptions. Data enrichment is the process of enhancing the data collected via annotations by machines or humans. The annotations are meant to categorize the data and make it more accessible to use by the AI. Annotations may contain increased number of errors, may not be uniform when performed by different people and may be subjective when performed by humans. Inaccurate enrichment by ML methods can lead to introduction of bias via machine data processing. Developers must ascertain if the data modification processes they employ are the best fit for the AI and the datasets used [16].

AI and ML training model. The data collected and processed is used to train the AI system utilizing a training model. This can be a source of bias. Every AI and ML system requires a training model that best fits its needs based on the systems' purpose and the training dataset. Different models are tested to identify the best fit prior to system training. Models must be validated using the dataset and benchmark datasets to examine their performance and compared against each other. Postprocessing modifications

and continued adjustments during deployment are necessary to allow the trained system to adapt and function correctly over time. If the incorrect model is chosen or if any of the downstream processes are erroneous, then it can lead to a malfunctioning system [13].

2. AI at NASA

NASA has employed AI and ML systems in diverse fields for various purposes, and is considering many more applications. In this section, we briefly characterize some of these applications. We have not exhaustively catalogued current or proposed AI or ML systems at NASA, nor have we examined any of these systems for evidence of problematic bias. We list these systems to illustrate both the diversity of AI and ML systems at NASA and the ways in which these AI and ML systems relate to humans. Later sections will present examples illustrating how non-NASA AI and ML systems—some with parallels to current or proposed NASA AI and ML systems—have exhibited social bias.

Many current and proposed AI and ML projects at NASA have no obvious relationship to social data or decisions that might have the potential to exhibit social bias. For example:

- **Early warning of storms and other disasters.** We spoke to one researcher who is investigating the use of machine learning tools to identify severe storms by analyzing imagery from geostationary satellites to identify tropopause-penetrating updrafts [34]. Researchers have also sought funding to curate datasets for use in training future early warning systems.
- **Assessing the quality of manufactured components.** We also spoke to a researcher who is using convolutional neural networks to examine components produced by an additive manufacturing process for defects [35, 36, 37, 37].
- **Optimizing communications for aerospace applications.** NASA is co-sponsoring a workshop on the use of AI and ML to develop cognitive telecommunication systems [38]—systems that 'intelligently' route signals and allocate radio bandwidth—for aerospace applications [39].

There are also current and proposed AI and ML projects at NASA that have more obvious relation to data derived from, or decisions that impact, individual human beings. For example:

- **Digital assistants.** The Intelligent Response and Interaction System (IRIS) project aims to develop dialogue-based voice assistants to support human space exploration missions [40]. Among other things, automatic speech recognition technology would allow crew members performing extravehicular or maintenance procedures to command or communicate with automated systems without using their hands. Similar digital assistants may one day assist crews with complex tasks such as diagnosing and treating medical emergencies when communications blackouts or delays limit the availability or timeliness of Earth-based expert support.
- **Crew state monitoring (physiological sensing).** There are efforts underway to model pilot attributes that are difficult to measure in terms of values that are more easily obtained in a cockpit. Some of these efforts are aimed at assessing the physiological state of pilots, e.g., the concentration of certain gasses in pilots' bloodstreams [41]. Others are aimed at assessing pilots' mental states, e.g., their attention level [42, 43]. Should these efforts find strong correlations, the models could be used to power ML systems to monitor pilot health or maintain focus.

- **Assessing employees and job applicants.** There is interest in using AI and ML technology to assess human performance and assign tasks accordingly [44]. This is intended as a means of better matching work to people and making more objective personnel decisions.
- **Building 'smart campuses'.** There are efforts to transform NASA facilities into 'smart campuses' fitted with an array of sensors and software to monitor spaces and activities [45]. These systems are intended to improve the campuses by, e.g., reducing energy usage. But it is not difficult to imagine that some would propose using the same or similar sensors, perhaps augmented by autonomous vehicles or facial recognition technology, for security purposes [46].

There are also proposals whose relationship to data or decisions that might exhibit social bias is not clear. For example, there have been proposals to use AI and ML technologies to enable air traffic control at much greater scale than is currently provided. Social factors might influence trust in such automation, leading to disparate impacts across social categories.

We have not examined any of these systems for their potential to contribute to social bias. But as we will show in the next section, non-NASA AI and ML systems that have been trained using social data or used to make decisions impacting individual people have caused social harm.

NASA is aware that using AI and ML responsibly requires identifying and addressing the ethical issues arising from that use. Accordingly, there is an ongoing effort to draft guidance to ensure that NASA's use of AI and ML is fair, mitigates discrimination and bias, supports diversity and inclusion, is explainable and transparent, is accountable, is secure and safe, provides benefits to society, and is scientifically and technically robust. We provide these examples in part to underscore the importance of such efforts.

3. Examples of bias in AI and ML systems

In this section, we present examples from the literature of social bias in AI and ML systems. We do not claim that the examples we present here are a complete or proportional depiction of the ways AI and ML might contribute to social inequity. We do not claim to have deep or personal knowledge of any of these systems. We merely present examples from publicly available literature, organized into categories chosen for narrative convenience (rather than, e.g., use of a specific algorithm or data set, or even a specific function or purpose). We contend only that the sheer number of available examples, and their distribution over diverse kinds of systems, suggests that the potential for harm might exist broadly across AI and ML systems that interface with, or are built on data derived from, members of the public. The nature of the harms shown in the examples itself suggests the importance of preventing or mitigating it.

3.1. Speech recognition

Automated speech recognition (ASR) systems are AI software that employ algorithms to recognize spoken words and execute actions based on the voice commands. These systems have a plethora of applications in an array of fields. Examples include in-car systems, virtual assistants, medical records dictation, and automated caption generating systems.

Examination of five popular ASR systems by researchers at Stanford University revealed that they exhibited roughly twice as many errors while recognizing the speech of African Americans than age- and gender-matched white Americans [47]. While the ASR systems correctly detected African Americans'

sentence structures, they struggled to identify the rhythm and intonation of African American Vernacular English. This led the researchers to conclude that the limitation of the ASR systems is most likely due to inadequate training dataset [47]. The performance gap documented in this study can have widespread impact on African American individuals as ASR system become increasingly prevalent. African Americans might find it more difficult than their white counterparts to operate devices that incorporate ASR software, such as mobile phones, in-car voice command systems, virtual assistants, and hands-free computer systems for disabled individuals. In addition, the authors extrapolate that individuals may face unfair disadvantage in processes that involve ASR systems [47]. ASR systems have been used in hiring processes to screen applicants and to transcribe court proceedings in the criminal justice system where inaccurate speech-to-text transcription can lead to negative outcomes for the individuals involved. While the focus of this study is on bias against African American individuals, it is important to highlight that these findings can be generalized for any individual whose speech is not recognized by ASR systems due to a lack of representation in the training dataset.

Another instance of bias in ASR systems was identified by researchers studying YouTube's automated caption generating system [48, 49]. This AI, Google's speech recognition software [50], is designed to recognize the words spoken in a video on YouTube's video hosting platform and generate accurate captions in various languages for the audience viewing the video. Five distinct dialects from different geographical regions—California, Georgia, New England, New Zealand, and Scotland—were chosen to evaluate the accuracy of the automatic caption generating system. This ASR system exhibited lower accuracy for women than for men, and lower accuracy for the Scottish dialect than any of the other four tested, with the first quartile error rate for Scotland exceeding the third quartile error rate for California [48]. A follow-up study supported these findings by showing that YouTube's ASR had a slightly higher word error rate for non-white English speakers than for Caucasian American English speakers [49]. The bias exhibited here, against women, speakers of unique dialects, and people of color, unfairly discriminates individuals of these populations. These users cannot rely on automatic transcription making videos of themselves accessible to users requiring captions as much as white English-speaking Americans. This can impact the quality of their content, limit their audience, and diminish monetary returns due to reduced exposure. Furthermore, the inaccurate ASR system invalidates the captions generated and disadvantages individuals who rely on these captions for enjoying YouTube videos. The Deaf community is one such population who rely on the captions to understand the content of the videos and the unreliability of these captions has led them to be referred as 'craptions' [51, 52]. As one of the largest video-hosting platforms, this can severely limit the ability of hearing-impaired people from accessing YouTube videos for educational and entertainment purposes.

In addition to these two studies, further instances of bias have been reported in speech recognition software highlighting their failure to cope with variations in dialect and accents [53, 54, 36, 55]. These studies highlight how speech recognition systems have difficulty with sociolinguistic variations across both gender and dialects. Consequently, individuals from populations whose speech is less well represented in the systems' training data face disproportionate difficulty in using systems such as virtual assistants, in-car voice recognition systems, and hands-free computer systems. This results in lost time, failure to complete tasks, and, ultimately, social harm and inequity. These individuals cannot fully realize the benefits of the products and services they have paid for, and they may be subjected to unfair processes if ASR systems are employed for decision making processes such as in hiring or in courts, resulting in undesirable outcomes and exacerbation of social inequity.

3.2. Facial recognition

Facial recognition software systems are AI systems designed to analyze images of faces. Some facial recognition software identifies individuals, either determining whether two images are of the same person or identifying an individual from a database of images of many individuals. Some software categorizes the race or gender of individuals based on an image of their face. Other software determines whether images show a face in a specific configuration, e.g., smiling or blinking. Ideally, all of this software should function equally well for images of subjects of all races and genders. Uses for this AI include photography, ID verification, biometrics, and security services.

Numerous instances of facial recognition going awry have been recorded in press articles. For example, Nikon cameras used face recognition AI to warn users when an individual being photographed is blinking so as to avoid capturing images where the subject(s) eyes are closed. But the cameras sometimes reported that subjects of East Asian were blinking when they were not [56]. In a similar case, one Chinese user alleged that her iPhone X's face recognition feature unlocked her phone for a Chinese colleague [57]. The phone's maker, Apple, Inc., disputes this account but admits that facial recognition software may not accurately differentiate between twins, siblings, or individuals under the age of 13 [58]. In both these cases, the products' poor performance for members of some racial groups prevented those individuals from enjoying the products they purchased to the same extent as other users are able to.

Facial recognition software has been trialed for use in the legal system for the purposes of facial detection with the goal of accurately identifying and arresting suspected criminals [59, 60, 61, 62]. However, a high error rate coupled with an inherent racial bias within the AI has led to misidentification and wrongful arrests [11, 63, 64, 62, 65]. In the summer of 2020, amid the pandemic, Robert Julian-Borchak Williams was arrested at his home in front of his wife and two daughters after facial recognition software employed by the Detroit Police identified him as the individual seen committing larceny in security camera footage. All charges against Mr. Williams were dropped due to the lack of evidence and he was released, but not before he had spent time in jail [62, 66]. Subsequently, the Detroit Police Chief admitted to a 96% error in rate in facial detection when attempting to identify suspects and changes were made to the way facial detection technology is used in the criminal justice process [62, 65]. Bias has been shown in facial recognition software used in other US law enforcement agencies as well. Analysis showed that one such system, Amazon's Rekognition, is biased against African Americans [67]. The test revealed that the software mistakenly identified 28 Congress members as criminals, with people of color misidentified at a higher rate than white individuals were. Given the many organizations that employ Rekognition, African Americans face an unfair disadvantage with the use of facial recognition technology by law enforcement. Revelation of bias within Rekognition prompted harsh criticism from Amazon's shareholders, members of Congress, and researchers to halt use of Rekognition pending correction of the bias [67, 68]. Currently, Amazon has placed a one-year long moratorium on the use of Rekognition by law enforcement to allow for policy making to monitor and regulate the use of facial recognition software in criminal justice [69]. The bias observed here results in a lopsided number of African Americans misidentified as criminals can further exacerbate the social inequality in today's society where African Americans are nearly five times more likely to be arrested than whites are [70, 71, 72] . The stark issue of discriminatory policing, arresting, and prosecution against African Americans has come to a head with the mounting incidents of police brutality. These social injustices are not lessened, but worsened instead, by facial recognition systems due to their bias even though the intention is to improve the justice system.

Joy Buolamwini, a researcher at MIT, investigated commercially used facial recognition systems for occurrences of social bias [11]. Her work revealed that these systems exhibit bias against darker-skinned

individuals after experiencing a facial recognition system that failed to identify her due to the color of her skin. As per the analysis, facial recognition had 34.7% error rate for darker-skinned females while light-skinned males only had 0.8% error rate. The study also showed that dark-skinned individuals were more likely to be misidentified and their gender more likely to be misclassified than light-skinned individuals'. The researchers concluded that dark-skinned females were the most misclassified group while facial recognition systems performed best on light-skinned males and that facial recognition systems performed better on men than on women. These findings support the instances of social bias experienced by individuals mentioned earlier and demonstrates that there is indeed social bias innately present in facial recognition systems. The individual experiences of bias are not mere anecdotal tales; there is evidence of discrimination by face recognition AI. These findings highlight the disproportionate disadvantage faced by African Americans and the ramifications it can have on individuals and their families, as in the example of Mr. Williams mentioned above. He and his family lost time and money, he was jailed, he was humiliated in public and his children may need therapy [62, 66]. Dark-skinned individuals may have limited access to services and products due to shortcoming of face recognition AI. For example, in the UK, dark-skinned women were almost four times as likely to have their passport photos rejected by the face recognition system than for light-skinned men. The women are told by the AI that their mouth looks open when, in fact, they were not. This has resulted in a difficult and time-consuming process for certain individuals to obtain their passport—all due to a biased AI [73].

3.3. Image classifiers

Image classifiers are AI systems capable of identifying and categorizing images or aspects of an image. These systems are trained using Deep Neural Networks (DNN) to perform intended tasks. Applications of image classifiers include photo tagging on social media and even safety-related applications such as perceiving the world around self-driving cars and diagnosing medical conditions based on medical imaging [74, 75]. The success of these systems rests on the AI accurately categorizing images and correctly labelling them; failure might result in harm.

Recently, Google Photos tagged an African American user and his friend as 'gorillas' [9, 10]. The AI was also confusing white users with dogs and seals [10]. Flickr, another popular image hosting platform, introduced an auto-tagging software which would categorize and tag images with appropriate labels to streamline search results. However, dark-skinned individuals were labelled as 'animals' and 'apes'. Pictures from Dachau and Auschwitz concentration camps were labelled as 'sport', 'jungle gyms', and 'trellis'. Images of Native Americans were labelled as 'costume'. Failure of these AI systems to function properly resulted in certain groups to be singled out unfairly causing anger and insult [10]. Image classifier have also been shown to misclassify images if the images are slightly altered [74, 37]. Research revealed that modifying an image by transplanting objects onto the image or by altering the position of objects in the image affects the classifiers' ability to detect the object itself as well as other objects in the image [74]. In a real-world example, targeting road sign classification in AI systems led to misclassification of the signs in question. Here, the researchers developed an algorithm, Robust Physical Perturbations (RP2), capable of generating physical alterations to modify 'STOP' signs and right turn signs mimicking real world scenarios such as graffiti and art that can impede an image classifier. They were able to successfully cause image classifiers to fail. The results showed that interferences that cause distortions of an image cannot be successfully identified by the AI. This can have drastic consequences in vehicles, UAVs, drones, and robots that rely of accurate classification of real-world images for mission safety and success [37].

Researchers at Columbia University investigated bias in DNN based image classification system by developing a technique called DeepInspect to detect bias errors [75]. They separated errors into two categories: (1) confusion, where a system cannot differentiate objects within a single image, and (2) bias, where a whole class of images are misclassified (e.g., dark-skinned people misclassified). Using their DeepInspect technique, the researchers discovered both confusion and bias errors in widely used DNN image classifying models. They recommend employing this technique to test image classifier systems prior to use to validate their accuracy. Systems in use can also be tested to determine if bias is present. If it is, the AI needs to be retrained with a more appropriate training model to reduce errors [75]. In addition, targeting DNN based image classifiers with object manipulation led to misclassification of images and, essentially, fooling the AI system.

3.4. Bias via proxy measurements

Proxy data are measurements that are used as stand-ins for other immeasurable values [76]. A value maybe immeasurable due to legal purposes, lack of measuring tools, or protection of individuals' privacy. Some examples of proxy measurements are a country's GDP as a measure of the quality of life of the citizens in the country, a student's exam scores as a measure of their intelligence, and homicide rates as a measure of public safety. Essentially, the proxy data are indirectly measuring the desired outcome and can be a less costly alternative to measuring the fields of interest [76, 77]. However, it is vital that the proxy and the data of interest are strongly related for one to be an accurate predictor of the other. Bias occurs when the link is not as strong as it should be or if the proxy measure is unintentionally measuring and predicting a secondary output that was not meant to be captured [76, 78, 18, 77].

Various AI systems employ proxy measurements to predict outcomes and aid with decision making processes. Some of these systems have inadvertently disadvantaged individuals by exhibiting social bias. In the US, an AI system used to determine health care resources necessary to treat a patient was found to be discriminatory against Black individuals [27]. The algorithm used total cost of treatment per year as a proxy for the patient's need for healthcare. On surface, this appeared to be a fair assumption since a patient with greater needs for care will acquire a greater cost for treatment. However, the algorithm was based on data that reflected both (a) higher incidents of chronic health conditions among Black Americans than among their white counterparts and (b) that an average $1,800 less per year had been spent treating Black individual than to treat a white individual with the equal health conditions. This meant that the algorithm assigned similar risk scores to a white patient and a much sicker Black patient. Reliance on this risk score thus meant that Black individuals were being disproportionately denied equal health care services by the AI due to the fact that the proxy measurement was not an accurate predictive tool [27].

Bias via proxy measurements have been revealed in facial recognition technology. A paper published in 2016 claimed to have developed an AI system that could determine whether an individual is a criminal based on a photograph of that individual's head [79]. The authors stated that their program can discern a criminal vs a non-criminal with 90% accuracy and that an individual's facial structure revealed criminal predisposition. The AI was measuring facial features to predict his/her criminality. The study was shown to have had a biased training dataset where the AI was measuring the individuals' facial expressions to detect criminality rather than facial features [80]. Since the images of 'criminals' all have frowns or scowls and the images of 'non-criminals' have a smile or relaxed facial expression, the AI learned to classify smiling people as non-criminals and non-smiling people as criminals [80].

A text classifier AI employed by Amazon to screen job applications and select the best candidates for various positions was discovered to be discriminatory against women [12]. Text classifiers are programs capable of identifying, characterizing, and grouping text for purposes as intended by the developer. Amazon's AI scanned applications for words and phrases and rated applicants on a scale from one to five stars. The presence or absence of particular words or phrases was thus used a proxy measurement for the qualification of an applicant. The scoring weights were derived from a decade of hiring decisions and included words that were directly or indirectly correlated to the sex of the applicant. For example, a reference to a "woman's" club or activity was scored negatively. Terms such as 'executed' or 'captured', which are more likely to be used by male applicants, were scored positively [12]. As a result, the system inadvertently favored male applicants. Further studies have also shown proxy bias in text classifiers where certain professions and adjectives are associated with one gender over the other [81, 82]. Use of such AI in hiring practices results in unfair treatment towards women and exacerbates the gender divide in areas where women are already underrepresented, such as in the STEM fields. Men/women will be less likely to be recruited and offered jobs in certain fields due to stereotypes associated with words.

Further examples include an algorithmic model capable of predicting if an individual would repay their loans by observing if that individual's email contained their name [83] when it was shown that individuals with African American names are more likely to face discrimination compared to individuals with white American names [84]. Tenant screening software employed by landlords to predict trustworthiness of tenants have been shown to be discriminatory [85, 86, 87]. Proxy measures such as credit scores, social media posts, and the frequency one visits bars have been used to predict if a tenant is more or less likely to pay rent. These practices only succeed in propelling and exacerbating the bias and prejudice that exists in our society to unfairly discriminate individuals from accessing services [86].

4. Detecting and mitigating bias

The issue of social bias in AI systems cannot be addressed if we do not accept that the systems we create can be biased. Our society is filled with biases and stereotypes that harm people. The datasets collected to train AI systems will reflect (to a lesser or greater extent) the world we live in and will cause the AI to mimic our societal environment [14, 16, 88]. Therefore, it does not come as a surprise that today's AI systems built on historically biased data exhibit similar biases. Once we accept this, it becomes possible to detect social bias in existing systems, mitigate its harms, and prevent the occurrence of social bias in future systems.

As mentioned before, a major source of bias in AI systems is the training dataset. It is vital to ensure that the dataset accurately represents the entire population to prevent bias entering the systems due to unrepresented data. If the training dataset is more representative of one group of people than another, then it may perform poorly and harm members of the poorly represented groups [14, 16]. This is seen in commercial facial recognition systems' failure to recognize dark-skinned individuals due to a lack of representation of these groups in the training dataset. However, even if the training dataset is all encompassing, it may exhibit historical social bias. Prejudices and stereotypes from the world are reproduced and amplified by the AI exacerbating the social bias experienced by groups of people [89, 16]. This is evident in the previous example of Amazon's job application screener that screened out women because the dataset used to train the AI reflected a work environment dominated by men. In addition, social bias can occur in an AI that continuously collects information for the purpose of continued training and refinement of its process [90]. To mitigate these effects, researchers have developed algorithms to detect biased datasets,

that can be employed by software developers to assess their datasets prior to training an AI, and tests to detect bias in AI systems [91, 89, 92, 93, 94, 95]. Furthermore, leading researchers in the field of AI and ethics have developed a comprehensive list of questions that developers can use to assess the fairness of their AI system in order to reduce potential harm [96, 97, 98]. Some of these tools also have the capability of scrubbing datasets and debiasing them, as in removing gender stereotypes in word embeddings used by AI systems [82]. These are a few of the techniques available to prevent the training of a biased AI. Although such tools may not be available for every type of AI system, it is the stakeholders' responsibility to investigate, identify, and utilize all the resources at their disposal.

Building a diverse team with an array of expertise will aid with bias reduction in AI systems. Individuals from various unique backgrounds and knowledgebases can test and question the AI in an all-around robust manner that otherwise would not be possible. Assessing the AI from different angles will ensure that it can withstand a userbase consisting of individuals of diverse backgrounds without discriminating against them [99]. Google Photos' mistagging of African American individuals as 'gorillas' was not expected by the developers, which may have been anticipated and assessed with a more diverse team [100, 101]. Additionally, vigorous alpha and beta testing of software is essential to identify and remove bugs and glitches that may otherwise be present in the AI leading to unintentional discrimination and harm against individuals [88]. However, this may not be viable for smaller organizations or projects limited by financial constraints.

Once a system is released, organizations must create a feedback mechanism where users can report instances of social bias with dedicated teams on the receiving end responsible for handling these cases promptly to reduce the impact on disadvantaged populations. A clear line of communication between the AI, the stakeholders, and the users can help finetune and improve the system while building a triangle of trust. Finally, for transparency, a centralized database of social bias cases and reports—both those that occur nationally and those that occur globally—needs to be created [102]. We fear that the reported cases of social bias in AI represent only a fraction of the true number of incidents. Due to lack of transparency, we do not know what we do not know. This phenomenon, known as silent failures—failures that exist but are not known about—creates a circle of failure where different companies creating comparable AI systems make similar mistakes that disadvantage the same groups of people [103]. This is observed in ASR and facial recognition system examples discussed previously. Instead, a centralized database can allow for a united progression of technology benefitting the users. Sharing of information will allow for better policy making and research reproducibility [102, 16]. Such systems exist for reporting cybersecurity breaches [104] and aviation failures [105] and have been successful in preventing the manifestation of repeated offenses.

While we have focused on systems that contributed to social inequities, there are also examples of systems reducing existing inequities. For example, an ML system built to predict patients' experience of knee pain showed much less racial disparity in the accuracy of its predictions than standard radiographic measures [106]. Since such predictions inform decisions about which treatments to offer to patients, reducing inequity in the predictions would reduce inequities in patients' quality of life. Examples like this suggest that, with attention, it might be possible for ML to reduce, rather than exacerbate, social inequity.

5. Relevant communities of practice

With the rapid advancement of AI technology, a parallel research community has evolved to delve into the ethical practices and legal policies surrounding the use of AI. Below, we list some of the organizations, researchers, and reference models that featured prominently in our literature search.

Organizations:

- **Partnership on AI.** The Partnership on AI is a nonprofit organization formed by leading companies including Facebook, Microsoft, IBM, Google, and Amazon [107].
- **OpenAI.** OpenAI is a research consortium of for-profit organizations initially formed by Elon Musk in 2015 with a focus on artificial general intelligence [108]. Their goal is to develop "highly autonomous systems that outperform humans at most economically valuable work" for the benefit all of humanity. They have released numerous products and applications with a wide range of uses.
- **OpenCog.** OpenCog is a nonprofit organization focused on creating an open-source AI framework for the creation of an advanced AI system with cognitive capabilities comparable to human intelligence [109]. SingularityNET, one of their many projects, is a decentralized AI marketplace, the first of its kind, that allows users to share, advertise, and purchase AI projects [110].
- **DeepMind.** DeepMind is a British company, acquired by Alphabet Inc., with the goal of advancing AI systems [111]. They have published articles on AI safety and formed a DeepMind Ethics & Society division to understand the ethical implications of AI use. Their research covers a wide range of AI uses, including protein folding, macular degeneration, and efficient cooling of data centers to save energy.
- **Brookings Institution Artificial Intelligence and Emerging Technology.** The Brookings Institution is a nonprofit American research organization based in Washington D.C. They conduct research on a wide range of public policy topics and influence policy making. One of the fields of research is AI and emerging technology with a focus on governance, bias, and national security [112]. Their publications have identified best approaches AI governance and AI practices to benefit the public.
- **NIST.** The National Institute of Standards and Technology has a dedicated division to conduct research on AI applications, bias, and security [113]. They are committed to establishing appropriate standards for AI use and practice.

Researchers:

Name	Affiliation	Details
Bryson, Joanna J.	University of Bath, UK	http://www.cs.bath.ac.uk/~jjb/
Buolamwini, Joy	MIT Media Lab, USA	https://www.media.mit.edu/people/joyab/
Caliskan, Aylin	George Washington University, USA	https://www2.seas.gwu.edu/~aylin/
Chouldechova, Alexandra	Carnegie Mellon University, USA	http://www.andrew.cmu.edu/user/achoulde/
Gebru, Timnit	Unaffiliated	@timnitGebru on Twitter
Hinton, Geoffrey	University of Toronto, Canada	https://www.cs.toronto.edu/~hinton/
Obermeyer, Ziad	University of California, Berkeley, USA	https://publichealth.berkeley.edu/people/ziad-obermeyer/
Raji, Inioluwa Deborah	Mozilla Foundation	@rajiinio on Twitter
Turner Lee, Nicol	Brookings Institution, USA	https://www.brookings.edu/experts/nicol-turner-lee/

Reference models for bias and AI ethics:

- Harini Suresh and John V. Guttag propose a framework for understanding how bias enters ML systems [13].
- Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman identify ways in which social data—data about human users—is misused and ways to prevent such misuse [16].
- David Leslie provides a guide for the responsible design and implementation of AI systems [15].

6. Conclusions

The purpose of this article is to compile information on the topic of social bias in AI, raise awareness of its potential harm, and suggest that those constructing and deploying such systems consider the possible presence of social bias and take adequate measures to prevent or mitigate the harms that might result. We conducted a literature survey on the phenomenon of biased AI and have categorically demonstrated how it can disproportionally disadvantage individuals along racial, gender, or other social category lines. We presume that AI systems are not being created with malicious intent. Nevertheless, these systems may not function equally for all groups of people. Unequal performance across social categories can lead to unfair treatment of vulnerable populations.

We have highlighted numerous ways social bias can manifest in different AI systems through examples in commercially employed AI systems and summarized current research in the field. We have no reason to believe these systems are uniquely susceptible to inadvertent social bias. Moreover, we feel it is the responsibility of those who develop and deploy systems to create fair and equitable systems that do not exacerbate existing social inequities. Accordingly, we suggest that those who develop and deploy systems that interact similarly with human beings (a) consider the potential for their systems to cause social harm and (b) take efforts, commensurate with that potential risk, to eliminate or mitigate such harms. We hope that use of resources listed in this article will aid in that endeavor.

Works Cited

[1] O. Li, "Artificial Intelligence is the New Electricity: Andrew Ng," *Medium,* 28 April 2017.

[2] C. Jewell, "Artificial Intelligence: The New Electricity," *WIPO Magazine,* June 2019.

[3] J. Golson, "Siri Voice Recognition Arrives on the iPhone 4S," *MacRumors,* 4 October 2011.

[4] Q. V. Le, "Building High-level Features Using Large Scale Unsupervised Learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, 2014.

[6] J. Yim, R. Chopra, T. Spitz, J. Winkens, A. Obika, C. Kelly, H. Askham, M. Lukic, J. Huemer, K. Fasler, G. Moraes, C. Meyer, M. Wilson, J. Dixon, C. Hughes, G. Rees and P. Khaw, "Predicting Conversion to Wet Age-Related Macular Degeneration Using Deep Learning," *Nature Medicine,* vol. 26, no. 6, p. 892–9, June 2020.

[7] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros and B. A. Research, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[8] M. Byrne, "Build Your Own Terrifying Cat-Blob with Machine Learning," *Vice,* 23 February 2017.

[9] L. Grush, "Google Engineer Apologizes After Photos App Tags Two Black People as Gorillas," *The Verge,* 1 July 2015.

[10] M. Zhang, "Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software," *Forbes,* 1 July 2015.

[11] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research: Conference on Fairness, Accountability, and Transparency,* p. 1–15, February 2018.

[12] J. Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," *Reuters,* 9 October 2018.

[13] H. Suresh and J. V. Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning," *arXiv,* 17 February 2020.

[14] D. Danks and A. London, "Algorithmic Bias in Autonomous Systems," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[15] D. Leslie, "Understanding Artificial Intelligence Ethics and Safety," The Alan Turing Institute, 2019.

[16] A. Olteanu, C. Castillo, F. Diaz and E. Kıcıman, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *Frontiers in Big Data: Data Mining and Management,* July 2019.

[17] A. Gupta, M. Ganapini, R. Butalid, C. Lanteigne, A. Cohen, M. Akif, T. De Gasperis, V. Heath and E. Galinkin, "The State of AI Ethics," Montreal AI Ethics Institute, 2020.

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *arXiv,* 17 September 2019.

[19] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson and D. Sculley, "No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World," in *Proceedings of the Workshop on Machine Learning for the Developing World*, Long Beach, CA, USA, 2017.

[20] K. Crawford, "The Hidden Biases in Big Data," *Harvard Business Review,* 1 April 2013.

[21] A. Perrin and M. Anderson, "Share of U.S. Adults Using Social Media, Including Facebook, Is Mostly Unchanged Since 2018," Pew Research Center, 2019.

[22] M. Anderson, "Men Catch Up with Women on Overall Social Media Use," Pew Research Center, 18 August 2015. [Online]. Available: https://www.pewresearch.org/fact-tank/2015/08/28/men-catch-up-with-women-on-overall-social-media-use/. [Accessed 17 February 2021].

[23] E. Hargittai, "Whose Space? Differences Among Users and Non-Users of Social Network Sites," *Journal of Computer-Mediated Communication,* vol. 13, no. 1, p. 276–97, October 2007.

[24] Virginia Department of Health, "Cancer in Virginia: Overview and Selected Statistics," Virginia Department of Health, Richmond, VA, USA, 2016.

[25] S. Harris, "Black Women 50 Percent More Likely to Die of Breast Cancer than White Women in Hampton Roads," *10 On Your Side WAVY TV,* 23 May 2018.

[26] K. Hafner, "Breast Cancer Rates are Especially High in Hampton Roads. There's a Patchwork of Problems Behind the Numbers," *The Virginian Pilot,* 28 September 2018.

[27] Z. Obermeyer, B. Powers, C. Vogeli and S. Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science,* vol. 366, no. 6464, p. 447–453, 2019.

[28] M. W. Sjouing, R. P. Dickson, T. J. Iwashyna, S. E. Gay and T. S. Valley, "Racial Bias in Pulse Oximetry Measurement," *New England Journal of Medicine,* vol. 383, no. 25, p. 2477–8, December 2020.

[29] J. Larson, S. Mattu, L. Kirchner and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica,* 2016.

[30] J. Angwin, J. Larson, S. Mattu and L. Kirchner, "Machine Bias," *ProPublica,* 23 May 2016.

[31] W. H. Herman and R. M. Cohen, "Racial and Ethnic Differences in the Relationship Between HbA1c and Blood Glucose: Implications for the Diagnosis of Diabetes," *Journal of Clinical Endocrinology and Metabolism,* vol. 97, no. 4, p. 1067–72, April 2012.

[32] M. M. Malik and J. Pfeffer, "Identifying Platform Effects in Social Media Data," in *Proceedings of the 10th International Conference on Web and Social Media (ICWSM)*, 2016.

[33] K. Crawford and M. Finn, "The Limits of Crisis Data: Analytical and Ethical Challenges of Using Social and Mobile Data to Understand Disasters," *GeoJournal,* vol. 80, no. 4, p. 491–502, August 2015.

[34] K. M. Bedka, Interviewee, *Personal communication.* [Interview]. November 2020.

[35] R. Ledesma and A. Ramlatchan, "Convolutional Neural Networks for Image Classification in Metal Selective Laser Melting Additive Manufacturing," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021.

[36] G. Droua-Hamdani, S.-A. Selouani and M. Boudraa, "Speaker-Independent ASR for Modern Standard Arabic: Effect of Regional Accents," *International Journal of Speech Technology,* vol. 15, no. 4, p. 487–493, December 2012.

[37] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno and D. Song, "Robust Physical-World Attacks on Deep Learning Visual Classification," *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR),* p. 1625–1634, June 2018.

[38] Cognitive Communications, "Background," Worldwide Universities Network Initiative, 16 January 2012. [Online]. Available: https://web.archive.org/web/20120116043349/http://www.wun-cogcom.org/background.html. [Accessed 17 February 2021].

[39] NASA; IEEE, "Call for Papers," 17 January 2021. [Online]. Available: https://ieee-ccaa.com/call-for-papers/. [Accessed 17 January 2021].

[40] JSC Office of Chief Technologist, "FY21 CIF IRAD Project Awards Announced!," [Online]. Available: https://www.nasa.gov/sites/default/files/atoms/files/roundup_announcement_of_awardees.pdf. [Accessed 20 January 2021].

[41] K. D. Kennedy, Interviewee, *Personal communication.* [Interview]. September 2020.

[42] N. J. Napoli, M. Paliwal, V. R. Rodrigues, A. Harrivel, K. D. Kennedy and C. L. Stephens, "Comparing Deep Neural Network Architectures using Transfer Learning to predict Cognitive States," in *Proceedings of the AIAA SciTech Form and Exhibition*, 2021.

[43] C. Stephens, A. Harrivel, L. Prinzel, R. Comstock, N. Abraham, A. Pope, J. Wilkerson and D. Kiggins, "Crew State Monitoring and Line-Oriented Flight Training for Attention Management," in *Proceedings of the 19th International Symposium on Aviation Psychology*, Dayton, OH, USA, 2017.

[44] N. Skytland, "The Future of Work Framework," *NASA Blogs,* 15 November 2018.

[45] T. Soderstrom, "A Recipe for Innovation," *IT Talk,* p. 10, October–December 2019.

[46] X. Zhou, "Application Research of Face Recognition Technology in Smart Campus," *Journal of Physics: Conference Series,* vol. 1437, no. 012130, 2020.

[47] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky and S. Goel, "Racial Disparities in Automated Speech Recognition," *Proceedings of the National Academy of Sciences of the United States of America (PNAS),* vol. 117, no. 14, p. 7684–9, 7 April 2020.

[48] R. Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (EthNLP)*, 2017.

[49] R. Tatman and C. Kasten, "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions," in *Proceedings of Interspeech*, Stockholm, Sweden, 2017.

[50] H. Liao, E. McDermott and A. Senior, "Large Scale Deep Neural Network Acoustic Modeling with Semi-Supervised Training Data for YouTube Video Transcription," in *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013.

[51] M. Lockrey, "YouTube Automatic Craptions Score an Incredible 95% Accuracy Rate!," *Medium,* 25 July 2015.

[52] L. Besner, "When Is a Caption Close Enough?," *The Atlantic,* 9 August 2019.

[53] B. Wheatley and J. Picone, "Voice Across America: Toward Robust Speaker-Independent Speech Recognition for Telecommunications," *Digital Signal Processing,* vol. 1, no. 2, p. 45–63, April 1991.

[54] M. Benzeguiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi and C. Wellekens, "Automatic Speech Recognition and Speech Variability: A Review," *Speech Communication,* vol. 49, no. 10–11, p. 763–786, October–November 2007.

[55] L. Plunkett, "Kinect Doesn't Speak Spanish (It Speaks Mexican)," *Kotaku,* 1 September 2010.

[56] A. Rose, "Are Face-Detection Cameras Racist?," *Time,* 22 January 2010.

[57] M. Papenfuss, "Woman in China Says Colleague's Face Was Able to Unlock Her iPhone X," *Huffpost,* 14 December 2017.

[58] Apple, Inc., "Face ID Security," 2017.

[59] M. Day, "Amazon Officials Pitched Their Facial Recognition Software to ICE," *The Seattle Times,* 23 October 2018.

[60] N. Statt, "Amazon Told Employees it Would Continue to Sell Facial Recognition Software to Law Enforcement," *The Verge,* 8 November 2018.

[61] D. Harwell, "Amazon's Facial-Recognition AI is Supercharging Police in Oregon. But What if Rekognition Gets it Wrong?," *The Washington Post,* 30 April 2019.

[62] B. Allyn, "'The Computer Got it Wrong': How Facial Recognition Led to False Arrest of Black Man," *NPR,* 24 June 2020.

[63] I. D. Raji and J. Buolamwini, "Actionable Auditing: Auditing the Impact of Publicly Naming Biased Performance Results of Commercial AI Products," *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society,* 2019.

[64] K. Wiggers, "MIT Researchers: Amazon's Rekognition Shows Gender and Ethnic Bias," *Venture Beat,* 24 January 2019.

[65] T. B. Lee, "Detroit Police Chief Cops to 96-Percent Facial Recognition Error Rate," *Ars Technica,* 30 June 2020.

[66] K. Hill, "Wrongfully Accused by an Algorithm," *The New York Times,* 24 June 2020.

[67] J. Snow, "Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots," American Civil Liberties Union, 2018.

[68] Concerned Researchers, "On Recent Research Auditing Commercial Facial Analysis Technology," *Medium,* 26 March 2019.

[69] Amazon, "We Are Implementing a One-Year Moratorium on Police Use of Rekognition," 10 June 2020. [Online]. Available: https://www.aboutamazon.com/news/policy-news-views/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition. [Accessed 5 January 2021].

[70] T. Williams, "Black People Are Charged at a Higher Rate Than Whites. What if Prosecutors Didn't Know Their Race?," *The New York Times,* 12 June 2019.

[71] J. Kaplan and B. Hardy, "Early Data Shows Black People Are Being Disproportionally Arrested for Social Distancing Violations," *ProPublica,* 8 May 2020.

[72] A. Srikanth, "Black People 5 Times More Likely to Be Arrested than Whites, According to New Analysis," *The Hill,* 11 June 2020.

[73] M. Ahmed, "UK Passport Photo Checker Shows Bias Against Dark-Skinned Women," *BBC News,* 7 October 2020.

[74] A. Rosenfeld, R. Zemel and J. K. and Tsotsos, "The Elephant in the Room," *arXiv,* 9 August 2018.

[75] Y. Tian, Z. Zhong, V. Ordonez, G. Kaiser and B. Ray, "Testing DNN Image Classifiers for Confusion & Bias Errors," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (ICSE)*, 2020.

[76] J. D. Mahnken, X. Chen, A. R. Brown, E. D. Vidoni, S. A. Billinger and B. J. Gajewski, "Evaluating Variables as Unbiased Proxies for Other Measures: Assessing the Step Test Exercise Prescription as a Proxy for the Maximal, High-Intensity Peak Oxygen Consumption in Older Adults," *International Journal of Statistics and Probability,* vol. 3, no. 4, p. 25, 2014.

[77] A. E. R. Prince and D. Schwarcz, "Proxy Discrimination in the Age of Artificial Intelligence and Big Data," *Iowa Law Review,* vol. 105, p. 1257–1318, 2020.

[78] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing and B. Schölkopf, "Avoiding Discrimination through Causal Reasoning," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017.

[79] X. Wu and X. Zhang, "Automated Inference on Criminality using Face Images," *arXiv.org,* vol. 1611.04135v3, 26 May 2017.

[80] C. Bergstrom and J. West, "Case Study: Criminal Machine Learning," 2017. [Online]. Available: https://www.callingbullshit.org/case_studies/case_study_criminal_machine_learning.html. [Accessed January 6, 2021].

[81] M. Prates, P. Avelar and L. C. Lamb, "Assessing Gender Bias in Machine Translation: A Case Study with Google Translate," *Neural Computing and Applications,* vol. 32, p. 6363–6381, 2020.

[82] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama and A. T. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," in *Proceedings of the 30th International Conference on Neural and Information Processing Systems (NIPS)*, Barcelona, Spain, 2016.

[83] T. Berg, V. Burg, A. Gombović and M. Puri, "On the Rise of Fintechs: Credit Scoring using Digital Footprints," National Bureau of Economic Research (NBER), Cambridge, 2018.

[84] M. Bertrand and S. Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," National Bureau of Economic Research (NBER), Cambridge, 2003.

[85] C. Dewey, "Creepy Startup Will Help Landlords, Employers, and Online Dates Strip-Mine Intimate Data from Your Facebook Page," *The Washington Post,* 9 June 2016.

[86] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *California Law Review,* vol. 104, p. 671–732, June 2016.

[87] L. Kirchner and M. Goldstein, "Access Denied: Faulty Automated Background Checks Freeze Out Renters," *The Markup,* 28 May 2020.

[88] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda and C. Wagner, "Bias in Data-driven AI Systems-An Introductory Survey," *WIREs: Data Mining and Knowledge Discovery,* vol. 10, no. e1356, 2020.

[89] A. Caliskan, J. J. Bryson and A. Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases," *Science,* vol. 356, no. 6334, p. 183–186, 2017.

[90] J. Vincent, "Twitter Taught Microsoft's AI Chatbot To Be a Racist Asshole in Less Than a Day," *The Verge,* 24 March 2016.

[91] A. Datta, M. C. Tschantz and A. Datta, "Automated Experiments on Ad Privacy Settings," *Proceedings on Privacy Enhancing Technologies,* vol. 2015, no. 1, p. 92–112, April 2015.

[92] M. Cisse, Y. Adi, N. Neverova and J. Keshet, "Houdini: Fooling Deep Structured Prediction Models," in *Advances in Neural Information Processing Systems: Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2017.

[93] W. Knight, "Forget Killer Robots: Bias Is the Real AI Danger," *MIT Technology Review,* 3 October 2017.

[94] S. Komkov and A. Petiushko, "AdvHat: Real-World Adversarial Attack on ArcFace Face ID System," *arXiv.org,* 23 August 2019.

[95] D. McDuff, S. Ma, Y. Song and A. Kapoor, "Characterizing Bias in Classifiers using Generative Models," in *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[96] N. Turner Lee, P. Resnick and G. Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," 2019.

[97] M. A. Madaio, L. Stark, J. Wortman Vaughan and H. Wallach, "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI," in *Proceedings of the 2020 CHI Conference on Human Factors in Computer systems*, 2020.

[98] Partnership on AI, The, "Understanding Facial Recognition Systems," The Partnership on AI, San Francisco, CA, USA, 2020.

[99] S. West, M. Whittaker and K. Crawford, "Discriminating Systems: Gender, Race, and Power in AI," AI Now Institute, 2019.

[100] N. Turner Lee, "Detecting Racial Bias in Algorithms and Machine Learning," *Journal of Information, Communication and Ethics in Society,* vol. 16, no. 3, p. 252–260, August 2018.

[101] D. Miller, "Design Biases in Silicon Valley Are Making the Tech We Use Toxic, Expert Says," *ABC News,* 22 October 2017.

[102] K. Lloyd and A. Hamilton, "Bias Amplification in Artificial Intelligence Systems," in *Proceedings of the Conference on Artificial Intelligence in Government and Public Sector (AAAI FSS)*, 2018.

[103] A. Orgla, "No More Silent Failures!," *Medium,* 2019.

[104] The MITRE Corporation, "Common Vulnerabilities and Exposures (CVE)," 17 February 2021. [Online]. Available: https://cve.mitre.org/. [Accessed 17 February 2021].

[105] National Aeronautics and Space Administration, "Aviation Safety Reporting System," 17 February 2021. [Online]. Available: https://asrs.arc.nasa.gov/. [Accessed 17 February 2021].

[106] E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan and Z. Obermeyer, "An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations," *Nature Medicine,* vol. 27, p. 136–40, January 2021.

[107] The Partnership on AI, "The Partnership on AI: About," [Online]. Available: https://www.partnershiponai.org/about/. [Accessed 13 January 2021].

[108] OpenAI, "OpenAI Charter," 9 April 2018. [Online]. Available: https://openai.com/charter. [Accessed 13 January 2021].

[109] OpenCog, "The Open Cognition Project," [Online]. Available: https://wiki.opencog.org/w/The_Open_Cognition_Project. [Accessed 14 January 2021].

[110] S. Rogers, "SingularityNET Talks Collaborative AI as its Token Sale Hits 400% Oversubscription," *VentureBeat,* 7 December 2017.

[111] DeepMind, "DeepMind: About," [Online]. Available: https://deepmind.com/about. [Accessed 13 January 2021].

[112] Brookings Institution, The, "Artificial Intelligence and Emerging Technology Initiative," [Online]. Available: https://www.brookings.edu/project/artificial-intelligence-and-emerging-technology-initiative/. [Accessed 13 January 2021].

[113] National Institute of Standards and Technology, "NIST: Artificial Intelligence," [Online]. Available: https://www.nist.gov/artificial-intelligence. [Accessed 13 January 2021].

[114] Amara, "YouTube Automatic Captions Need Work: A Chat with Michael Lockray," *Amara Blog,* 1 May 2015.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01/03/2021	TECHNICAL MEMORANDUM	August 2020 – February 2021

4. TITLE AND SUBTITLE
Social Bias in AI and its Implications

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Sribava Sharma
Mallory Graydon

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
340428.02.40.07.01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA Langley Research Center
Hampton, VA 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-001

10. SPONSOR/MONITOR'S ACRONYM(S)
NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
NASA/TM-20210010446

12. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified - Unlimited
Subject Category
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES
An electronic version of this document can be found at https://ntrs.nasa.gov.

14. ABSTRACT

Previous studies have documented many types of biases in artificial intelligence (AI) and machine learning (ML) systems. We reviewed the literature on AI and ML bias with a focus on social implications and found that such biases might harm individuals and/or groups of people. By affecting people differently across categories such as race, gender, or sexual orientation, AI and ML systems might exacerbate social inequities. We recount examples of issues in systems that use technologies that might be used at NASA and elsewhere so that similar issues may be mitigated in future systems. We also provide readers with a gateway into the relevant literature and practice.

15. SUBJECT TERMS
Artificial intelligence (AI), machine learning (ML), social bias, social inequity, engineering

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			HQ - STI-infodesk@mail.nasa.gov
U	U	U	UU	26	19b. TELEPHONE NUMBER *(Include area code)* 757-864-9658