**Is Your AI Racist? Overcoming Bias When Developing Artificial Intelligence**

Jana Infanger **|** Ariel OConnall **|** Karolyn Wooley **|** Henock Woldegeorgies

**Introduction to Artificial Intelligence**

**Definition**

Different articles have varieties of definition about Artificial Intelligent but for this document we have selected the following two definitions to give readers general idea. Techopedia.com defines Artificial intelligence (AI) is an area of computer science that emphasizes the creation of intelligent machines that work and react like humans. Some of the activities computers with artificial intelligence are designed for include:

* Speech recognition
* Learning
* Planning
* Problem solving

As per the definition of wikipedia.org, Artificial intelligence (AI) is the ability of computer program or a machine to think and learn. It is also a field of study which tries to make computers "smart". They work on their own without being encoded with commands.

**History** The term artificial intelligence was coined in 1956, but AI has become more popular today thanks to increased data volumes, advanced algorithms, and improvements in computing power and storage.

Early AI research in the 1950s explored topics like problem solving and symbolic methods. In the 1960s, the US Department of Defense took interest in this type of work and began training computers to mimic basic human reasoning. For example, the Defense Advanced Research Projects Agency (DARPA) completed street mapping projects in the 1970s. And DARPA produced intelligent personal assistants in 2003, long before Siri, Alexa or Cortana were household names.

This early work paved the way for the automation and formal reasoning that we see in computers today, including decision support systems and smart search systems that can be designed to complement and augment human abilities.

The research about Artificial Intelligence was started long time ago even before 1950’s. However, the term AI is introduced in 1950’s. The following list of events are presented to show how the evolution of AI is happened.

* 1950 Alan Turing proposed the Turing test that same year Isaac Asimov proposed the Three Laws of Robotics
* 1951 the first AI based program was written
* 1955 the first self-learning game playing program was created
* 1959 the MIT AI lab is set up
* 1961 the first robot is introduced into GM's assembly line
* 1964 saw the first demo of an AI program which understands natural language
* 1965 the first chatbot Eliza was invented
* 1974 the first autonomous vehicle is created at Stanford AI lab
* 1989 Carnegie Mellon creates the first autonomous vehicle using a neural network
* 1997 IBM deep blue beats Garry Kasparov at chess
* 1999 MIT AI labs first emotional AI is demonstrated
* 2004 DARPA introduces the first autonomous vehicle challenge
* 2009 Google starts building a self-driving car
* 2010 narrative Sciences a I am straight stability to write reports
* 2011 IBM Watson beats jeopardy champions that same year Siri Google now and Cortana become mainstream
* 2015 Elon Musk and others announced a billion-dollar donation to open AI
* 2016 Google’s deepmind defeats Korean AlphaGo champion
* 2016 Stanford issues the AI 100 report in
* 2016 UC Berkeley launches the center for human compatible artificial intelligence.
* recently we heard the well-known robot sofia Tesla introduced full self-driving car, amazon can made delivery with self-driving car and drone, uber also tested ride without driver.

**How Artificial Intelligence is being used**

Now days every industry has a high demand for AI capabilities. AI is helping human being in almost every part of our lives. To mention some of uses includes:

**Health Care**

AI applications can provide personalized medicine and X-ray readings. Personal health care assistants can act as life coaches, reminding you to take your pills, exercise or eat healthier.

**Retail**

AI provides virtual shopping capabilities that offer personalized recommendations and discuss purchase options with the consumer. Stock management and site layout technologies will also be improved with AI.

**Manufacturing**

AI can analyze factory IoT data as it streams from connected equipment to forecast expected load and demand using recurrent networks, a specific type of deep learning network used with sequence data.

**Banking**

Artificial Intelligence enhances the speed, precision and effectiveness of human efforts. In financial institutions, AI techniques can be used to identify which transactions are likely to be fraudulent, adopt fast and accurate credit scoring, as well as automate manually intense data management tasks.

**Transport**

Self-driving train, metro and cars are now available.

**Helping difficult tasks**

Some robotic systems are in progress to perform services like difficult surgeries, rescue during disaster, fire-fighting, pipe inspection, bomb disposal, exploration of dangerous/unknown environment.

**Forms of bias: how unintentional bias happens**

Algorithmic bias, as defined by Wikipediaoccurs when a computer system reflects the implicit values of the humans who are involved in coding, collecting, selecting, or using data to train the algorithm.

Bias can enter into Artificial Intelligence as a result of pre-existing cultural, social, or institutional expectations; because of technical limitations of their design; or by being used in unanticipated contexts or by audiences who are not considered in the software's initial design.

And thus, the fundamental software principle of ‘garbage in, garbage out’ is born. Today, however, artificial intelligence (AI) has raised the stakes on the old conundrum, as the ‘garbage out’ from AI leads to appalling examples of bias.

For example, if you want to use AI to make recommendations on who best to hire, feed the algorithm data about successful candidates in the past, and it will compare those to current candidates and spit out its recommendations.

Just one problem. If the input data are biased – say, consisting of mostly young white males, then who will the AI recommend? You guessed it: mostly young white males.

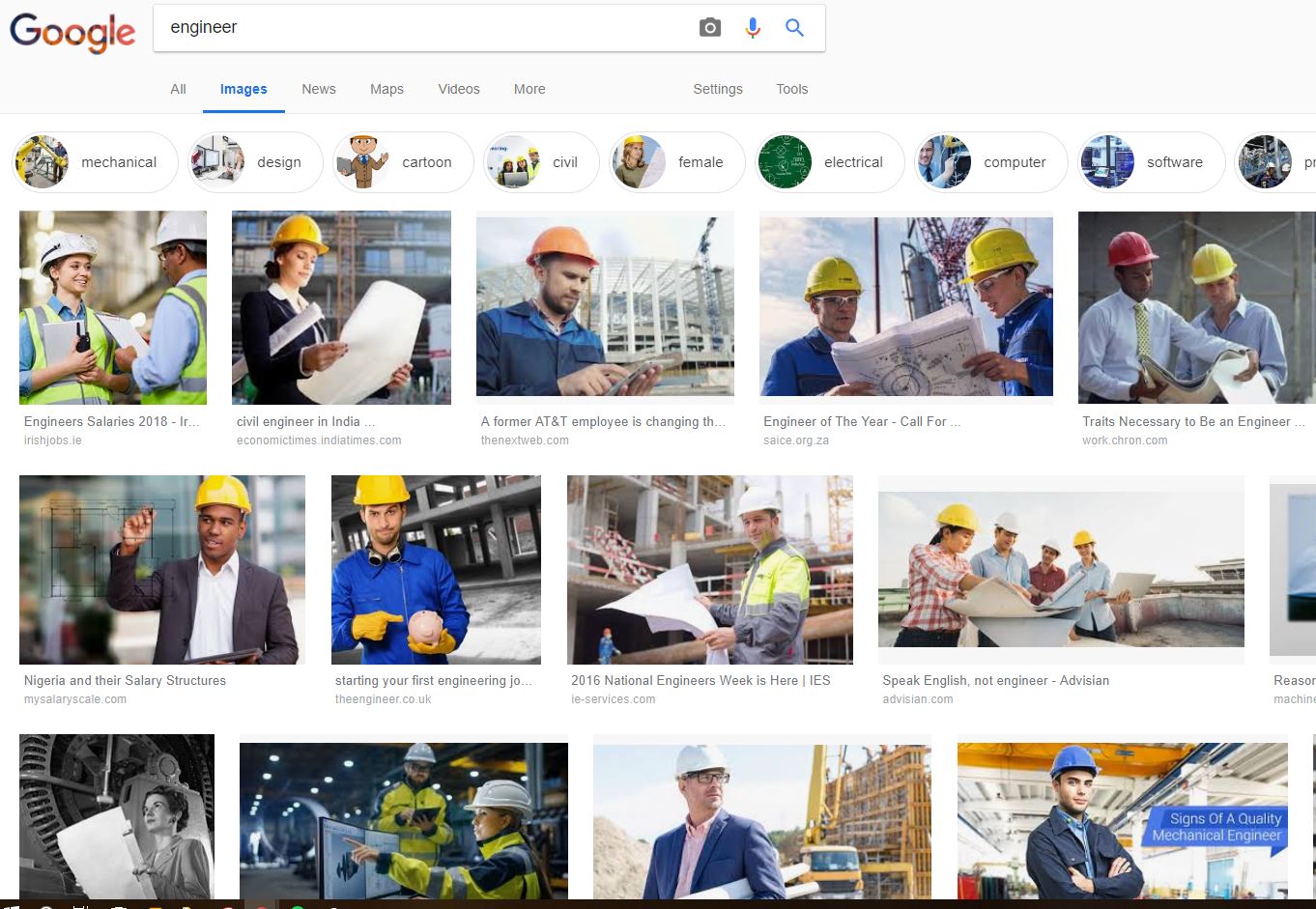
Whether the AI algorithms are themselves biased is also an open question. “[Machine-learning algorithms] haven’t been optimized for any definition of fairness,” [says](http://www.fortune.com/longform/ai-bias-problem/) Deirdre Mulligan, associate professor, UC Berkeley School of Information. “They have been optimized to do a task.”

Basically, if the data that the machine is fed contains unconscious bias, then the outcome of the machine learning will also contain that bias.

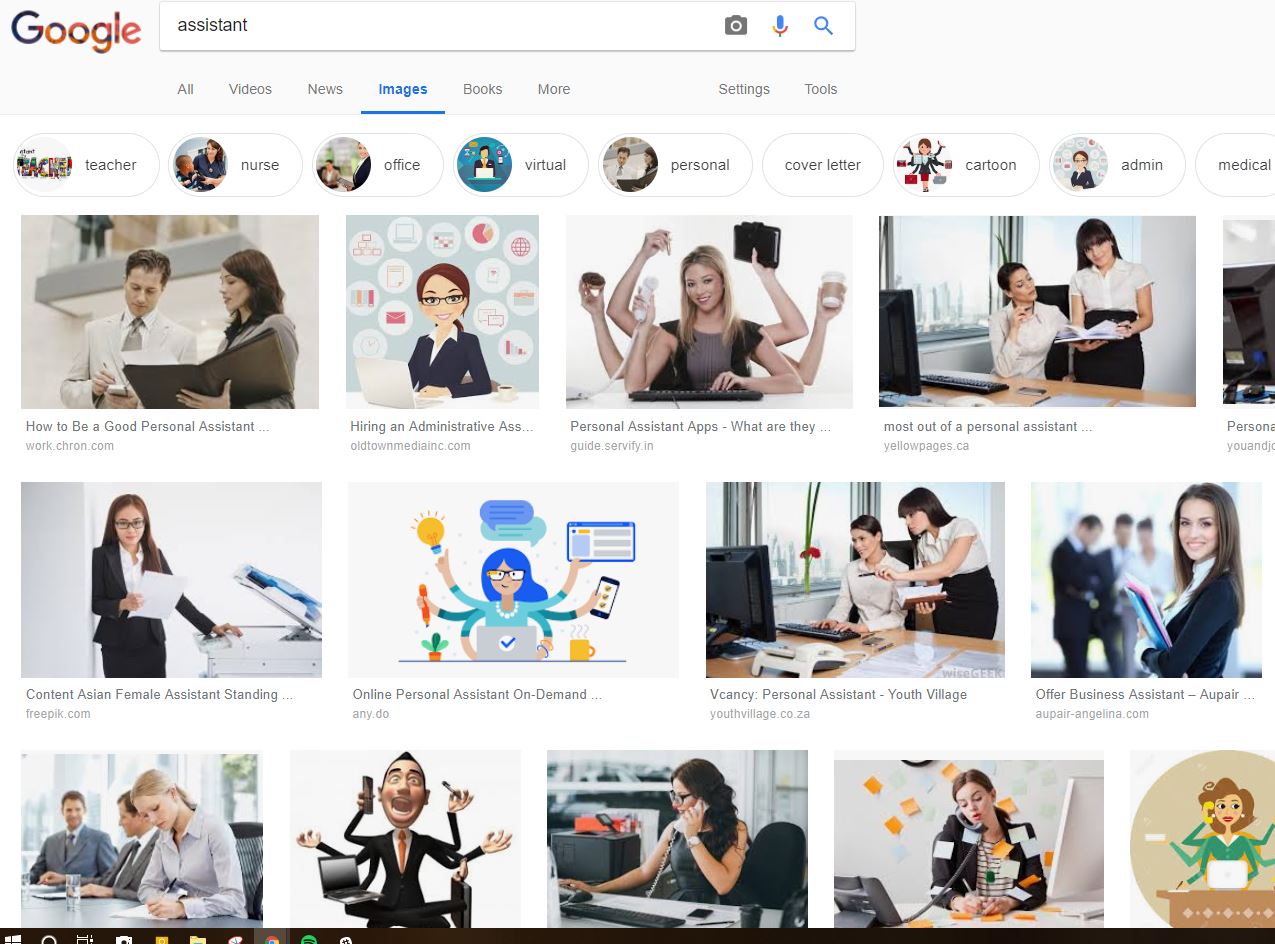
“It’s not that the programmer sits down and does this on purpose, it’s something that he or she just doesn’t know and that way, the algorithms of these outcomes become flawed.”

**Examples of Unintentional Bias:**

Our first example kicks off with a few simple google searches. Here is a search result for the word “engineer.” What do you see?



Here is another example: an image search of the word “assistant.” As you can see the majority of the pictures are all of women, or women assisting men.



This gender bias extends beyond just image searches, a recent paper shows that some concerning biases seen in human psychology experiments are also readily acquired by algorithms. The words “female” and “woman” were more closely associated with arts and humanities occupations and with the home, while “male” and “man” were closer to maths and engineering professions.

And the AI system was more likely to associate European American names with pleasant words such as “gift” or “happy”, while African American names were more commonly associated with unpleasant words.

One previous study showed that an identical CV is 50% more likely to result in an interview invitation if the candidate’s name is European American than if it is African American.

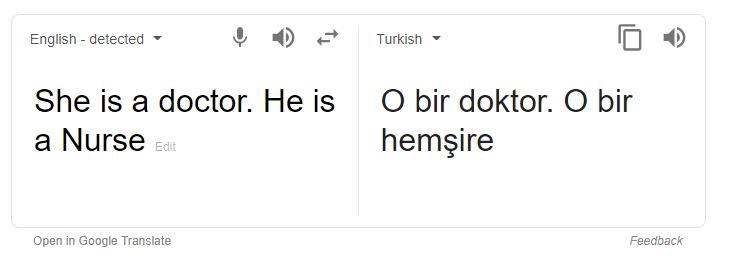
“A lot of people are saying this is showing that AI is prejudiced. No. This is showing we’re prejudiced and that AI is learning it.”

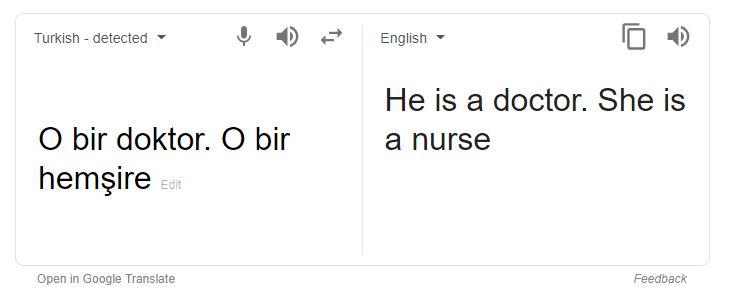
Computer vision is often bad at recognizing people of color. One of the most infamous examples comes from 2015. Google Photos, which automatically labels photos, usefully categorized graduation photos and images of buildings. However, it also labeled black people as gorillas.

In 2016, the [Beauty.AI](http://beauty.ai/) website was using AI robots as judges for beauty contests. It found that people with light skin were judged much more attractive than people with dark skin. And in 2017, [FaceApp](https://www.faceapp.com/), which uses neural networks to create filters for photographs, created a hotness filter that lightened people's skin and gave them more European features.



Another example: take a pair of sentences like “She is a doctor. He is a nurse.” Use Google Translate to translate them into Turkish and then translate them back into English. The genders become flipped to so that the sentences now say, “He is a doctor. She is a nurse.” Turkish has a gender-neutral singular pronoun that becomes translated into a stereotype in English. This happens with other languages that have gender-neutral singular pronouns. It’s been documented for a variety of words that translation stereotypes hold that women are lazy, women are unhappy, and many more characterizations.





Machine learning can actually amplify bias. An example of this is discussed in “[Men also like shopping: Reducing gender bias amplification using corpus-level constraints](http://ai2-website.s3.amazonaws.com/publications/yatskar_bias.pdf)”, which looked at visual semantic role labeling of images in a dataset.The researchers found that 67% of images of people cooking were women but the algorithm labeled 84% of the cooks as women. There is a risk of machine learning algorithms amplifying what we see in the real world.

In research performed by Zeynep Tufekci, who has provided insights into the intersection of technology and society. It was found that “the number of people telling me that YouTube autoplay ends up with white supremacist videos from all sorts of starting points is pretty staggering.” Examples include:

* “I was watching a leaf blower video and three videos later, it was white supremacy.”
* “I was watching an academic discussion of the origins of plantation slavery and the next video was from holocaust deniers.”
* “I was watching a video with my daughters on Nelson Mandela and the next video was something saying that the black people in South Africa are the true racist and criminals.” It's scary.

**Intentional Bias**

Intentional Bias is defined as “when a person or group deliberately alters data in order to change the results of an experiment or study. This type of bias influences the information gathered to go in a certain and predetermined direction. To be an intentional bias the alteration of data and results must be deliberate.”

Nefarious actors could deliberately introduce bias into AI systems. Some examples of how this could negatively impact future society are:

* Hate groups using AI to target or exclude people on the basis of race, gender, religion, or other characteristics. Biased Algorithms could either give hate groups either justification or more advanced means to directly do so.
* Biased data could serve as bait. Corporations could release biased data with the hopes that other corporations use it to train their artificial intelligence systems, causing them to diminish the quality of their own products.
* Biased algorithms could make it easier to mask discriminatory lending, hiring or other bad business practices.

Therefore, fixing intentional bias requires diversified training data. Academic and industry workers have called for legislative oversight that addressed intentional bias. As a society we must observe how algorithms are implemented and what outcome they produce. As written by Douglas Yeung, “Identifying and addressing bias in those who develop algorithms, and the data used to train them, will go a long way to ensuring that artificial intelligence systems benefit us all, not just those who would exploit them.”

**Examples of Intentional Bias**

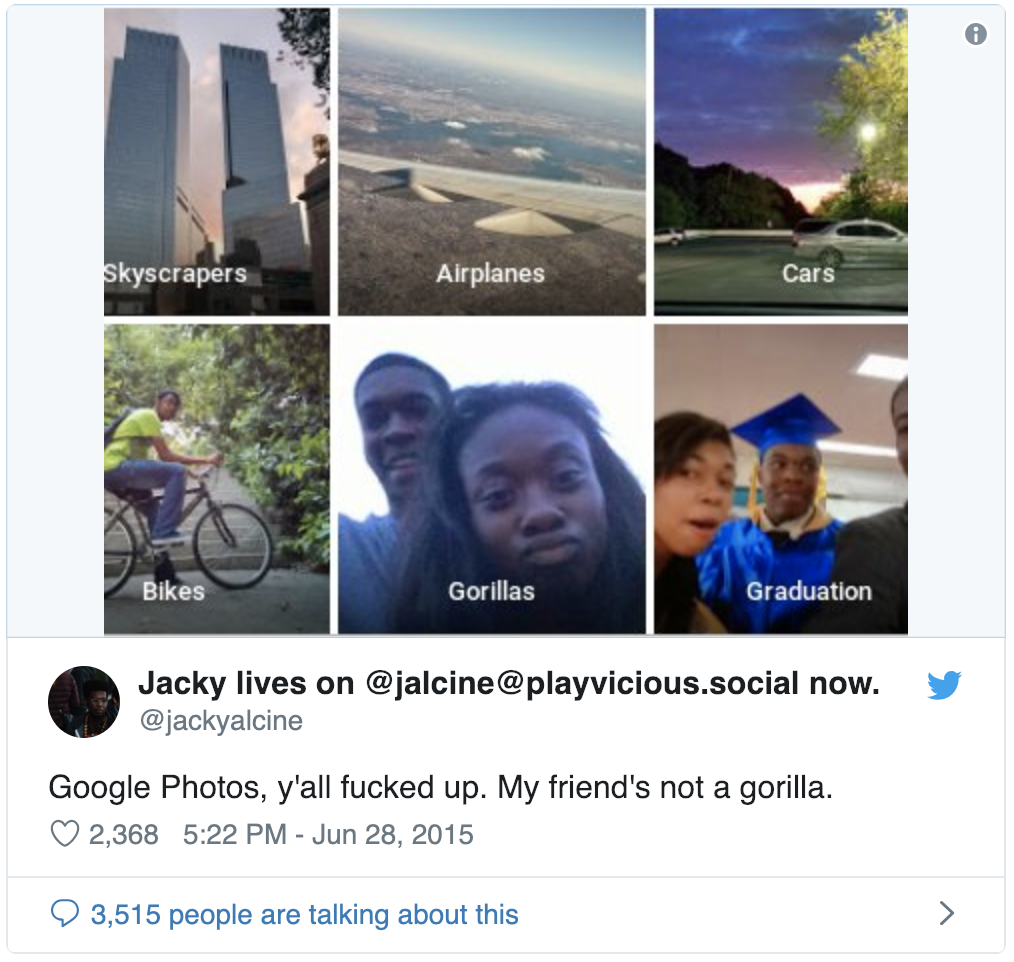
An example of intentional bias being caused by users is the 2016 twitter AI named Tay. Tay, also known as TayTweets, was created by microsoft to try and see if an AI can make tweets similar to a teenager as an experiment in “conversational understanding.” “Unfortunately, the conversations didn't stay playful for long. Pretty soon after Tay launched, people starting tweeting the bot with all sorts of misogynistic, racist, and Donald Trumpist remarks. And Tay — being essentially a robot parrot with an internet connection — started repeating these sentiments back to users, proving correct that old programming adage: flaming garbage pile in, flaming garbage pile out.”



Now, while these screenshots make it seem like Tay has become a product of the internet’s worst tendencies, it’s not as straightforward as that. Once the bot’s tweets were looked through, it was noticeable that the worst tweets were the result of copying users. One function of the AI was that if you told Tay to “repeat after me” it would tweet it soon after.

The result of Microsoft’s Tay experiment raised some serious questions like, how are we going to teach AI using public data without incorporating the worst traits of humanity? And if we create bots that mirror their users, how do we prevent them from becoming this controversial? In an emailed statement given later to [*Business Insider*](https://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3?r=UK&IR=T), Microsoft said: "The AI chatbot Tay is a machine learning project, designed for human engagement. As it learns, some of its responses are inappropriate and indicative of the types of interactions some people are having with it. We're making some adjustments to Tay." ***Source:*** *“Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day” - The Verge*

Another example of intentional bias showed when a user google image searched the word “gorilla” back in 2015. A software engineer pointed out that the image recognition technology in google photos was classifying black people as “gorillas.”



A spokesperson for Google confirmed that the image category “gorilla”, “chimp”, “chimpanzee”, and “monkey” were blocked on google photos after the tweet went viral. “Image labeling technology is still early and unfortunately it’s nowhere near perfect,” said the rep. It may seem strange that Google didn’t fix the problem in the AI instead of taking those categories down. I think it shows how difficult it is to come up with a complete solution to this error and how complex AI programming is. It’s also a good reminder of how software needs to be tested and trained by a diverse group of people.

**Ways developers are working to overcome bias in artificial intelligence**

It was previously thought that the purity of math would be the quick solution to easily overcoming bias. After all, a computer’s responses are just a physical process, and therefore should be unable unable to display bias. While the previous assumption was that AI would provide neutral data as a machine is incapable of developing a personal position on morality or ethical platforms, developers soon experienced the flaws in machine learning, and the pitfalls of algorithmic bias. Developers began searching for solutions to these various forms of bias, resulting in a series of changes, with some answers being simple, and others more complex.

One of the most simple ways developers are working to fight algorithmic bias is by overcoming the lack of diversity in the development community. A more diverse group of contributors can help fight unintentional bias by providing more varied perspectives when creating artificial intelligence. More diverse creators can lead to more diverse and inclusive data. An example of this would be the AI who when given the search word “engineer” displayed mostly caucasian males. Had more women been involved in the creating environment for this tech, the issue may have been caught quicker or may not occured at all.

Another solution for writing unbiased algorithms is that developers are simply writing better code. As artificial intelligence gains popularity, developers are getting more experience in writing for AI and improving on anticipating common stumbling blocks based on previous issues and experiences. Developers now have higher standards, and better awareness when creating AI. The data is only as unbiased as the person creating it, and developers are now actively looking to make sure the data supplied is more diverse to match the diverse needs of the consumer. One of the most common ways this is being implemented is by improving testing systems.

While a large part of bias in AI creation happens without premeditation, intentional bias can occur, both in the creative environment, and with the consumer. To counteract intentional bias forming in the production of AI, the programming community has been calling for a standard of quality control auditing to occur for AI software both prior to and post production. It is hoped that requiring all created AI software to be subject to the scrutiny of peers within the tech community may help deter any intentional bias that may occur as well as catch any potential overlooked algorithmic bias.

Overcoming the bias of the consumer has proven to be a bit more tricky. User input affects machine learning dramatically, just ask anyone who used Twitter the day “TayTweets” was released. Developers are trying to creatively come up with solutions to account for and limit biased user data without overstepping the whole purpose, which is allowing the user to supply their own input. One of the ways this is being done is by compensating for cultural bias when writing code for AI. Artificially created systems reflect our society’s values. If you have an imperfect society providing data for machine learning, it’s no surprise the results of machine learning will ultimately reflect that culture’s various inherent forms of bias. Making sure AI systems are well trained on balanced data prior to being released for use helps combat the machine learning of cultural bias. Another option being considered and implemented is the policing and monitoring of AI software after it has been released. Observing AI in action and stepping in to take corrective measures when AI becomes too deviant from its intended purpose may help safeguard the integrity of the software.

In some ways, ultimately it is up to society to determine where the future of AI standards lie. Companies follow the profits, and with it the expectations of the consumer.

Report issues, help test software, become a society that expects developers to audit AI software, engage!

**Final Points:**

AI has the potential to reinforce existing biases because, unlike humans, algorithms may be unequipped to consciously counteract learned biases. “A danger would be if you had an AI system that didn’t have an explicit part that was driven by moral ideas, that would be bad,” Joanna Bryson said.

**Auditing:**

* Should it be standard?
* Who should do it, the companies or an external source?
* Should the government be involved
* Should consumer input be under scrutiny

**Education:**

* Awareness of bias in AI should be required prior to development

**Accountability:**

* Fiduciary Responsibility, should technology have its own version
* Should there be repercussions
* Should there be regulations in place
* The burden of responsibility, can it fall on the consume

**Developers:**

* should diversity be required?
* better testing, better data, better AI training

**Work cited**

“Artificial Intelligence – What It Is and Why It Matters.” *SAS*, [www.sas.com/en\_us/insights/analytics/what-is-artificial-intelligence.html](http://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html).

“Artificial Intelligence.” *Wikipedia*, Wikimedia Foundation, 7 Dec. 2018, simple.wikipedia.org/wiki/Artificial\_intelligence.

- “Algorithmic Bias,” Wikipedia, <https://en.wikipedia.org/wiki/Algorithmic_bias>

Bloomberg, Jason. “Bias Is AI's Achilles Heel. Here's How To Fix It.” Bias Is AI's Achilles Heel. Here's How To Fix It, Forbes.com, 13 Aug. 2018, 9:40am, [www.forbes.com/sites/jasonbloomberg/2018/08/13/bias-is-ais-achilles-heel-heres-how-to-fix-it/#1c3404ed6e68](http://www.forbes.com/sites/jasonbloomberg/2018/08/13/bias-is-ais-achilles-heel-heres-how-to-fix-it/#1c3404ed6e68).

Bryson, Joanna. “Three Very Different Sources of Bias in AI, and How to Fix Them.” Three Very Different Sources of Bias in AI, and How to Fix Them, 13 July 2017, [www.joanna-bryson.blogspot.com/2017/07/three-very-different-sources-of-bias-in.html](http://www.joanna-bryson.blogspot.com/2017/07/three-very-different-sources-of-bias-in.html).

Darmody, Jenny. “AI can cause unconscious bias – but it could also be the solution.“, Silicon Republic, 20 July, 2018, <https://www.siliconrepublic.com/machines/alexa-gorman-inspirefest-ai-diversity>

Devlin, Hannah. “AI programs exhibit racial and gender biases, research reveals” The Guardian, 13 April, 2017, <https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals>

“Intentional Bias.” *Dunning-Kruger Effect Definition | Psychology Glossary*, www.alleydog.com/glossary/definition.php?term=Intentional Bias.

<https://www.alleydog.com/glossary/definition.php?term=Intentional+Bias>

Kirkpatrick, Keith. “Battling Algorithmic Bias.” cacm.acm.org, Oct. 2016  
[Https://Cacm.acm.org/Magazines/2016/10/207759-Battling-Algorithmic-Bias/Abstract](https://cacm.acm.org/Magazines/2016/10/207759-Battling-Algorithmic-Bias/Abstract).

Penchikala, Srini “Analyzing and Preventing Unconscious Bias in Machine Learning” 14 August, 2018, InfoQ, <https://www.infoq.com/articles/machine-learning-unconscious-bias>

Price, Rob. “Microsoft Is Deleting Its AI Chatbot's Incredibly Racist Tweets.” *Business Insider*, Business Insider, 24 Mar. 2016, [www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3?r=UK&IR=T](http://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3?r=UK&IR=T)

“What Is Artificial Intelligence (AI)? - Definition from Techopedia.” *Techopedia.com*, [www.techopedia.com/definition/190/artificial-intelligence-ai](http://www.techopedia.com/definition/190/artificial-intelligence-ai).

*Where Can Artificial Intelligence Be Used? - Quora*. [www.quora.com/Where-can-artificial-Intelligence-be-used](http://www.quora.com/Where-can-artificial-Intelligence-be-used).

Vincent, James. “Twitter Taught Microsoft's Friendly AI Chatbot to Be a Racist Asshole in Less than a Day.” *The Verge*, The Verge, 24 Mar. 2016, [www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist](http://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist)

Vincent, James. “Google 'Fixed' Its Racist Algorithm by Removing Gorillas from Its Image-Labeling Tech.” The Verge, The Verge, 12 Jan. 2018, [www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai](http://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai).

Yeung, Douglas. “Intentional Bias Is Another Way Artificial Intelligence Could Hurt Us.” RAND Corporation, 22 Oct. 2018, [www.rand.org/blog/2018/10/intentional-bias-is-another-way-artificial-intelligence.html](http://www.rand.org/blog/2018/10/intentional-bias-is-another-way-artificial-intelligence.html).

Work Log:

Ariel: time: approx 6-9 hrs  
1/10/ 1/15 1/17 1/21 1/22   
Updated running Google doc with Henock’s google doc, alphabetized the citations and tried (unsuccessfully) to merge the formatting, created the section titles, cleaned up the spaces between contributions, added the citations of: Battling Algorithmic Bias, Three Very Different Sources of Bias in AI, and How to Fix Them, & Bias Is AI's Achilles Heel. Here's How To Fix It, and wrote the section for **Ways developers are working to overcome bias in artificial intelligence**. I started the powerpoint presentation, set up the basic slide theme, input my section for ways to overcome bias and set up the general layout of the rest of the slides to be filled in. I updated the powerpoints to include slides with quotes and a final slide on 1/22/19.

01/13/19 Jana- Researched and cited 4 articles, added to group document: Bias Is AI's Achilles Heel. Here's How To Fix It, Forbes; “AI programs exhibit racial and gender biases, research reveals“- The Guardian; Silicon Republic, by Jenny Darmody: “AI can cause unconscious bias – but it could also be the solution“; InfoQ, by Srini Penchikala, Analyzing and Preventing Unconscious Bias in Machine Learning; (2 hours)

01/18/19 Jana- editing assigned section of quotes, working on flow. (.5 hour)

01/21/19 Jana Wrapped up editing and flow of assigned section. (1 hour)

01/21/19 Jana- added pictures & slides to presentation (.5 hour)

Karolyn:

1/10/19: *Researched resources for paper. Found Twitter Taught Microsoft’s AI chatbot to be a racist asshole Article (1 hour).*

1/15/19: *Made changes to the paper so that it was all similar and grammatically correct (.5 hours).*

1/15/19: *Researched resources for paper and found Microsoft is Deleting AI Chatbot’s incredibly racist tweets article (1 hour).*

1/15/19 : *wrote part of intentional bias section (.5 hours).*

*1/25/19: added description of intentional bias and “gorilla” example (2 hours).*

*1/28/19: added final slides and pictures in Intentional bias section (.5 hours).*

Henock: created and filled in the sections for **How Artificial Intelligence is being used, History, & Definition**, and provided the citations of: Artificial Intelligence – What It Is and Why It Matters, Artificial Intelligence, What Is Artificial Intelligence (AI), & *Where Can Artificial Intelligence Be Used*.