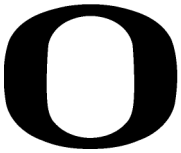
Presented to the Interdisciplinary Studies Program:



UNIVERSITY OF OREGON
**APPLIED INFORMATION MANAGEMENT**

Applied Information Management
and the Graduate School of the
University of Oregon
in partial fulfillment of the
requirement for the degree of
Master of Science

# Identifying and Mitigating Bias in Machine Learning Applications

CAPSTONE REPORT


**Laura Bald**
**Admissions and System Manager**
**Portland State University**

University of Oregon
Applied Information
Management
Program


**Spring 2019**


Continuing and Professional
Education
1277 University of Oregon
Eugene, OR 97403-1277
(800) 824-2714

Approved by

_____

Dr. Kara McFall
Director, AIM Program

# Identifying and Mitigating Bias in Machine Learning Applications

Laura Bald

Portland State University

**Abstract**

This study addresses the existence of bias in machine learning applications and examines techniques for identifying and mitigating bias using scholarly literature published between 2012 and 2019. The intended audience is machine learning engineers, system analysts, and data analysts of any industry. This study is significant because there may be considerable ethical implications caused by machine learning bias; identifying and mitigating these biases is key to the development and deployment of effective machine learning algorithms.

*Keywords: machine learning, artificial intelligence, bias, variance, machine bias, ethics, algorithmic bias, machine learning models*

## Table of Contents

Abstract ..... 3

Introduction to the Annotated Bibliography ..... 7

Problem Statement ..... 7

Purpose Statement ..... 10

Research Questions ..... 11

Audience Profile ..... 11

Search Report ..... 12

Reference Evaluation ..... 14

Annotated Bibliography ..... 17

Introduction to the Annotated Bibliography ..... 17

Understanding Machine Learning ..... 17

Explanations for Bias in Machine Learning ..... 21

Solutions for Mitigating Bias in Machine Learning ..... 34

Conclusion ..... 48

Understanding Machine Learning ..... 48

Explanations for Bias in Machine Learning ..... 50

Solutions for Mitigating Bias in Machine Learning ..... 53

Final Thoughts ..... 56

References ..... 58

## List of Tables and Figures

*Figure 1.* Graphical illustration of bias and variance..... 22

## Introduction to the Annotated Bibliography

### Problem Statement

Artificial intelligence (AI) is defined for the purposes of this study as "intelligence exhibited by an artificial entity to solve complex problems" (Borana, 2016, p. 64); Borana (2016) notes that "such a system is generally assumed to be a computer or machine" (Borana, 2016, p. 64). The related term machine learning (ML) is defined as a type of AI in which the "systems automatically learn programs from data" (Domingos, 2012, p. 78). Whereas AI produces output based on human input and human adjustments of algorithms, ML autonomously learns from inputs and outputs and adjusts its own algorithms (Domingos, 2012). Some authors, such as Li (2019), use the two terms interchangeably.

The concept of AI stretches back as far as Greek antiquity with the animalistic automatons of Hephaestus (Yapo & Weiss, 2018). Humanoid automatons were first attempted in third century China; built again during the twelfth century by a Muslim scholar, engineer, and inventor named al-Jazari; and rendered in detailed sketches during the Renaissance period by Leonardo da Vinci (Hamet & Tremblay, 2017). A long history of automatons and experimentation led to William Gray Water developing the first electronic, autonomous robot in 1948 that used electronic connections to mimic brain cells and brain functionality (Hamet & Tremblay, 2017). In 1955, John McCarthy coined the term *artificial intelligence* and founded the field of AI with his colleagues in 1956 at a Dartmouth College conference on artificial intelligence (Hamet & Tremblay, 2017).

Even before the term *artificial intelligence* was coined, scientists were researching and conducting innovative ML algorithms (Buchanan, 2005). In 1947, Alan Turing gave a lecture to the London Mathematical Society in which he proposed a machine that would learn from its own

experiences (Turing, 1995). In the early 1950s, Arthur Samuel created a program that plays checkers against humans and through learning schemes, was designed to outperform a person of average intelligence (Samuel, 1959). In 1961, Samuel's program beat the Connecticut state checker champion (McCarthy & Feigenbaum, 1990). Scientists have continued to research, refine, and experiment with ML throughout the twentieth century (Buchanan, 2005).

In the 1970s and 1980s, AI and ML experienced a drop in funding after a report by James Lighthill (1973) claimed significant disappointment in the field of AI after initial high hopes. Lighthill (1973) claimed that innovations in AI, particularly in robotics and language processing, had been promised and hyped by the scientific community, but progress was much slower than anticipated, with few positive results. The period of time following the Lighthill report was designated the *AI Winter* and lasted until the 1990s (Brachman, 2006). There are varying opinions as to the genesis of the next rise of AI; Hamet and Tremblay (2017) claim it occurred due to the use of AI in medical diagnosis and Deng (2018) attributes the rise to speech recognition applications.

Today, ML is commonly used in a multitude of industries, such as health care, manufacturing, education, and marketing (Jordan & Mitchell, 2015). Examples of ML uses include automatic speech recognition, spam filters, driverless cars, digital assistants, advanced cybersecurity (Brundage et al., 2018), predictions of student success (Daud et al., 2017), predictions of recidivism (Brennan, Dieterich, & Ehret, 2008), and the use of predictive text that most smartphone users experience any time they start typing a text or email on their phones (Indrajith & Vijayakumar, 2016). Artificial intelligence and machine learning have the potential to make most tasks, from the mundane to the complex, more efficient and safer from human error (Borana, 2016).

While the possibilities and benefits of AI and ML are widely recognized, experts warn that the technologies pose risks that must be considered and safeguarded against before full adoption (Li, 2019). For example, Amazon used an AI technology that was trained to scan and recognize patterns in resumes of historically successful employees, then make hiring recommendations from a database of new applicant resumes (West, Whittaker, & Crawford, 2019). However, Amazon's past hiring practices were alleged to have been implicitly discriminatory against women, and it was reported that the AI hiring tool was downgrading resumes from female candidates, as it had also learned these biases from the historical data inputs (West et al., 2019).

Even seemingly innocuous tasks, like predictive text suggesting the completion of a sentence in an email or predicting the use of an emoji, can result in biases (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). Word-embedding algorithms that contain biases reflect gender stereotypes; for example, predictive text may suggest *man* in relation to *computer programmer* or *woman* in relation to *homemaker* (Bolukbasi et al., 2016). Biases may not always enter into ML systems accidentally; within 24 hours of release, a Twitter chat bot using AI technologies was trained by internet users to repeat racist and misogynistic sentiments (Li, 2019). This and other types of biases have been found to be integrated into ML algorithms and then perpetuated by these algorithms (Yapo & Weiss, 2018).

Michael Li (2019) argues that "all human data is fundamentally biased in some way" (para. 11). Human beings, who are empathetic, conscious, and complex, have the capacity to understand and combat bias through cognitive debiasing techniques (Croskerry et al., 2013). Machines do not hold these traits and thus, must rely on humans to actively combat the bias by training the machines to do so (Li, 2019).

In addition to incorporating anti-bias training into the algorithms, humans must continue to carefully monitor AI outputs and correct biased results (Li, 201). Karen Hao (2019) explains that bias may be incorporated into many stages of the ML process, including the stage before implementation in which data scientists are framing the problems they want to address. For example, if a credit card company wants to predict a customer's *creditworthiness*, the definition of *creditworthiness* may be defined in the ML algorithm based on the company's ultimate business goal rather than fairness to the customer (Hao, 2019). Susceptibility to bias means that AI and ML require constant vigilance at all stages of implementation; Li (2019) contends that in order to recognize and mitigate biased results, machine learning engineers and business leaders who use this technology must have specialized training in understanding both the possibilities and risks of AI.

The focus of this annotated bibliography will be on best practices in recognizing and mitigating biases that arise during the development and adoption of ML algorithms and technology. The research is significant in addressing inherent flaws in ML; Ribeiro, Singh, and Guestrin (2016) declare that knowing and trusting a model or a prediction is vital to the use of machine learning.

**Purpose Statement**

The purpose of this qualitative study is to explore the implications of bias in machine learning and to identify strategies for recognizing and reducing machine learning bias. The design of the study is a literature review and the method of inquiry is collection, sorting, review, annotation, and analysis of selected research sources. The primary audience for this study is machine learning engineers and system analysts and data analysts in any given industry who play a role in designing, testing, implementing, and using machine learning technologies.

## Research Questions

**Main question.** What are the best practices for identifying and mitigating bias in machine learning?

**Sub-questions.**

- Can some bias be beneficial to machine learning?
- What are the ethical implications of bias in machine learning?
- Is it more effective to create an algorithm without bias or to teach a machine to remove bias through learning?

## Audience Profile

The intended audience for this study is the machine learning engineer, system analyst, and data analyst of an organization in any given industry. The ML developer is the individual programming the algorithms used in machine learning ("Machine Learning Engineer Job Description," n.d.). The system analysts are individuals who are instrumental in developing technological strategies and implementing the tools for use in organizations ("System Analyst Job Description," n.d.). The data analysts are responsible for developing and implementing data analyses and strategies, as well as interpreting analysis results ("Data Analyst Job Description," n.d.).

The ML developers will benefit from this study because their biases will be subject to integration into their ML algorithms from their very inceptions. They will be the first defenders against bias and this study may provide suggestions for mitigating the potential bias issues. The systems and data analysts will benefit from this study because they are responsible for recommending strategies to non-technical colleagues and if they are not aware of all of the

possible consequences of ML algorithms, their recommendations on the use of ML may be flawed.

This audience will benefit by developing awareness of unexpected outcomes from the use of ML algorithms and confidence in asking and examining what is *under the hood* of these technologies. One goal of this study is to provide the audience with insights about the questions they should ask vendors, data scientists, and programmers regarding the use of ML and the potential warnings to look for during the research and implementation of systems that make use of ML.

## Search Report

**Search strategy.** I began my search strategy by using Google to narrow down the common vernacular for this topic. I accomplished this focus by searching for the phrases *machine learning bias, artificial intelligence and machine learning, human bias in machine learning,* and *human bias in artificial intelligence* in the Google search engine. The results included common words and phrases that I used as the keywords for the scholarly databases.

My initial searches returned several articles that were fully dedicated to the topic of bias in machine learning. Many articles that I found referenced this topic but were primarily focused on other subjects. Using the reference lists in those articles, I was able to find more relevant scholarly resources that pertained specifically to this topic. To cross reference articles, I used Google Scholar or the Portland State University (PSU) library search function to search for the article title. These search engines returned the original source where I found all metadata.

**Keywords.** My initial research indicated that the topic of bias concerns with machine learning can be covered under many interchangeable terms and keywords, such as:

- machine learning,

- artificial intelligence,
- bias,
- machine bias,
- algorithmic bias,
- neural networks,
- deep learning, and
- ethical considerations.

I found that terms like *machine bias* and *machine learning bias* are not as commonly used in scholarly articles, although they have been used as keywords in more recent white papers and non-scholarly technical articles. However, these searches often produced cited, scholarly articles that are either in line with this topic or cite other relevant articles.

**Libraries and Search Engines.** I accomplished research for this annotated bibliography, focusing on cited, scholarly literature, using the following libraries and search engines:

- Google Scholar – Scholarly articles and books directly related to the keywords.
- Web of Science – Scholarly articles and journals in the arts and humanities, social sciences, and sciences.
- Portland State University library subject guides – Research guide for Computer Science.
- University of Oregon library research guides – Research guide for Computer Information Science.

**Databases.** The library subject and research guides suggested the following databases, which I used to find reference sources for this annotated bibliography:

- Inspec,
- arXiv,

- ACM Digital Library,
- IEEE Xplore,
- CiteSeer,
- Academic Search Premier, and
- MathSciNet.

**Documentation method.** I used Zotero to collect the resources I found. This tool gathers Adobe portable document format (PDF) documents, links, and meta-data directly from the online source. Using Zotero, I created separate folders to categorize the sources: Primary, Non-Scholarly, Paid Books/Articles, and Unrelated. I used the Zotero note function to add a brief summary of the article and how it could be used (or not used) in the annotated bibliography. I also linked related articles to help organize the flow of data and reviews. I tagged each article with the related keywords so I could keep track of the different topics covered.

**Reference Evaluation**

**Reference evaluation criteria.** I evaluated references using the five characteristics described in the *Evaluating Information Sources* guide by University of Florida's Center for Public Issues Education (CPIE) (2014). The guide includes definitions and examples for each characteristic (Center for Public Issues Education, 2014). The characteristics listed are authority, timeliness, quality, relevancy, and bias (Center for Public Issues Education, 2014). If references did not meet these criteria, I did not use them for this annotated bibliography.

***Authority***. The author's credentials must be credible; an author tends to be more credible if the author has an advanced degree, is associated with a reputable organization, and has been cited many times (Center for Public Issues Education, 2014). The credibility of the publishing organization can also be examined to determine authority (Center for Public Issues Education,

2014). The references I chose were authored by individuals with advanced degrees in computer science and computer engineering or with significant professional experience working with AI and ML. I deemed references to be authoritative if they were peer-reviewed, published in a well-known and well-established computer science community, or were cited in at least ten other scholarly sources.

***Timeliness.*** The date of publication for a reference must be considered appropriate for the topic (Center for Public Issues Education, 2014). The topic of machine learning requires more recent publications to ensure the information is accurate and current. However, researching the history of certain technologies did not necessarily require a recent publication to be considered credible. I began searching within a date range between 2010 to 2019, but I found that keeping the date range wide was still valuable as the topic of machine learning has been discussed for decades and the technology is still relevant.

***Quality.*** I deemed references to be high quality if they were grammatically correct, free from spelling errors, and logically organized (Center for Public Issues Education, 2014).

***Relevancy.*** I deemed the references to be relevant if they were related to the topic of ML and bias in AI and ML.

***Bias.*** I deemed references to be unbiased if the author's objectives were to inform as opposed to persuade the reader towards the author's viewpoint or to sell a product or service (Center for Public Issues Education, 2014). For example, I found the article written by Scott Fortmann-Roe on his personal website, where he features modeling software. However, this software and his essays are provided at no cost, so I deemed this source to be free of the bias that exists when an author attempts to gain monetarily through the sale of related products or services. The topic of machine learning bias often overlaps with the topic of ethics, so several

references did place value statements in their writing. However, the references I chose also provided factual explanations for these ethical dilemmas and offered frameworks for mitigating biases. Due to the purpose of my study, I found these references highly valuable; the ethical value statements found in the writing served to provide context to the topic of bias in machine learning.

# Annotated Bibliography

## Introduction to the Annotated Bibliography

The following annotated bibliography contains fifteen carefully selected references that examine ML and bias in ML. The references are organized into three categories: (1) understanding machine learning, (2) explanations for bias in machine learning, and (3) solutions for mitigating bias in machine learning. These topics are included to define and describe ML and ML techniques, provide frameworks for recognizing and identifying bias in ML, and offer suggested techniques and models for mitigating bias in ML.

Each annotation is comprised of three elements: (1) the full bibliographic citation, (2) an abstract, and (3) a summary. The abstracts included were written by the author(s) unless otherwise noted. The summaries highlight the source's relevancy to bias in ML and attempt to identify techniques and best practices for mitigating bias in ML.

## Understanding Machine Learning

Bakshi, K., & Bakshi, K. (2018). Considerations for artificial intelligence and machine learning: Approaches and use cases. Paper published at the *2018 IEEE Aerospace Conference* (pp. 1-9). Big Sky, MT: Institute of Electrical and Electronics Engineers. doi:10.1109/AERO.2018.8396488

**Abstract.** As data sets grow, leveraging machines to learn valuable patterns from structured data can be extremely powerful. The volume of data is too large for comprehensive analysis, and the range of potential correlations and relationships between disparate data sources are too great for any analyst to test all hypotheses and derive all the value buried in the data. Machine learning (ML) is ideal for exploiting the opportunities hidden in big data. Machine learning is a type of artificial intelligence (AI) that allows

software applications to become more accurate in predicting outcomes without being explicitly programmed. The basics of machine learning is to build algorithms that can take input data and use statistical analysis to predict an output value within an acceptable range. This paper explores the basics of machine learning, discussing concepts and topics like supervised, unsupervised and reinforcement learning, regression, classification, model evaluation metrics, overfitting, variance versus bias, linear regression, ensemble methods, model selection, Decision Trees, Random Forests. The paper then will review several several *[sic]* use cases, where machine learning can be applied, including but not limited to Aerospace, Internet of Things (IoT) and Computer Network Analytics use cases. The applicability of AI and ML will be reviewed in these use cases. Finally, the latest trends in machine learning will be discussed.

**Summary.** This article provides an overview of machine learning by defining the purpose of ML, explaining the different categories of ML algorithms, and providing examples of ML algorithms. The authors define the purpose of ML as a way to determine insights from large and disparate datasets in a much more accurate and efficient fashion than humans can manually accomplish. They define three major categories of ML algorithms: supervised, unsupervised, and semi-supervised learning. They choose to only focus this article on supervised and unsupervised learning. Supervised learning is defined as an algorithm that receives both the inputs and the outputs of data in the form of a labeled dataset. The two major types of supervised learning algorithms are classification, in which the goal is to predict the assignment of categories to data, and regression, defined as the prediction of a continuous number, such as income. Unsupervised learning is described as a type of learning algorithm in which labels, or output, are not provided,

thus the algorithm must discover its own patterns and output. Unsupervised learning algorithms include transformation, defined as using the existing formations of data to create new formations, and clustering, defined as the partitioning of data into separate groups of similar items.

The authors explore 20 types of algorithms in each category, including k-nearest neighbors (k-NN), linear models, decision trees, kernelized support vector machines, and agglomerative clustering. In some of the descriptions, they provide advantages and disadvantages of the model; for example, k-NN is easy to understand and usually gives a good performance, but does not always perform efficiently with large datasets.

This source is relevant to this study because it provides a straightforward overview of ML and the types of algorithms used in ML. Understanding the different types of algorithms and types of learning is useful for identifying anomalies like bias in ML.

de Saint Laurent, C. (2018). In defence [*sic*] of machine learning: Debunking the myths of artificial intelligence. *Europe's Journal of Psychology, 14*(4), 734–747. doi:10.5964/ejop.v14i4.1823

**Abstract.** There has been much hype, over the past few years, about the recent progress of artificial intelligence (AI), especially through machine learning. If one is to believe many of the headlines that have proliferated in the media, as well as in an increasing number of scientific publications, it would seem that AI is now capable of creating and learning in ways that are starting to resemble what humans can do. And so that we should start to hope – or fear – that the creation of fully cognisant machine might be something we will witness in our life time. However, much of these beliefs are based on deep misconceptions about what AI can do, and how. In this paper, I start with a brief

introduction to the principles of AI, machine learning, and neural networks, primarily intended for psychologists and social scientists, who often have much to contribute to the debates surrounding AI but lack a clear understanding of what it can currently do and how it works. I then debunk four common myths associated with AI: 1) it can create, 2) it can learn, 3) it is neutral and objective, and 4) it can solve ethically and/or culturally sensitive problems. In a third and last section, I argue that these misconceptions represent four main dangers: 1) avoiding debate, 2) naturalising our biases, 3) deresponsibilising creators and users, and 4) missing out some of the potential uses of machine learning. I finally conclude on the potential benefits of using machine learning in research, and thus on the need to defend machine learning without romanticizing what it can actually do.

**Summary.** This article focuses on explaining how machine learning works and describing the myths, risks, and potentials of ML. With AI and ML advancements often featured in the news, the author's goal is to set realistic expectations of ML applications and to explain why it is important to be wary of ML at this stage in development.

The author defines ML as "any statistical method where the parameters of a given model are 'learnt' from a dataset through an iterative process, usually to predict an output (a given value or category)" (p. 737). Machine learning makes analyzing huge data sets very efficient because the ML models and calculations are more complex than a human can manually perform. This complexity creates a unique challenge in which the engineer must understand and appropriately fit a huge number of parameters and hyperparameters to a selected model. Thus, ML requires a human to utilize their experience and subjectivity to tell a machine what it should learn and how. Then the machine's role "is to

carry out calculation[s] that would be much too long and much too complicated for researchers to perform themselves" (p. 737).

The author reviews and debunks four common misconceptions surrounding ML: (a) they can create novelties like original artwork, (b) they can learn new concepts like a human can, (c) they are neutral and objective, and (d) they can solve sensitive ethical and cultural issues. The author stresses that these misconceptions are often a result of cherry-picking the best examples of ML. These misconceptions may lead to dangerous results, such as avoiding debate, naturalizing biases, overlooking the roles and responsibilities of ML creators and users, and disregarding some of the potential benefits of ML. The author concludes by asserting that ML will never replace the important role that humans play in critical thinking and research, but that it will supplement and revolutionize analysis and research.

This source is relevant to this study because it explains ML specifically for the general audience's understanding. It will also be useful for ML developers, system analysts, and data analysts to recognize common misconceptions, understand the realities of ML applications, and set realistic expectations for their use of ML.

**Explanations for Bias in Machine Learning**

Fortmann-Roe, S. (2012). Understanding the bias-variance tradeoff. Retrieved from http://scott.fortmann-roe.com/docs/BiasVariance.html

**Abstract.** When we discuss prediction models, prediction errors can be decomposed into two main subcomponents we care about: error due to "bias" and error due to "variance". There is a tradeoff between a model's ability to minimize bias and variance.

Understanding these two types of error can help us diagnose model results and avoid the mistake of over- or under-fitting.

**Summary.** The scope of this article covers ML prediction errors caused by bias and variance. The author defines *errors due to bias* as over-fitting a model, which is when a model's average prediction is far from the correct value. *Errors due to variance* are described as under-fitting a model, which occurs if the model results in a great deal of variability in the predictions. To illustrate these errors, the author includes a graphical visualization shown in Figure 1.

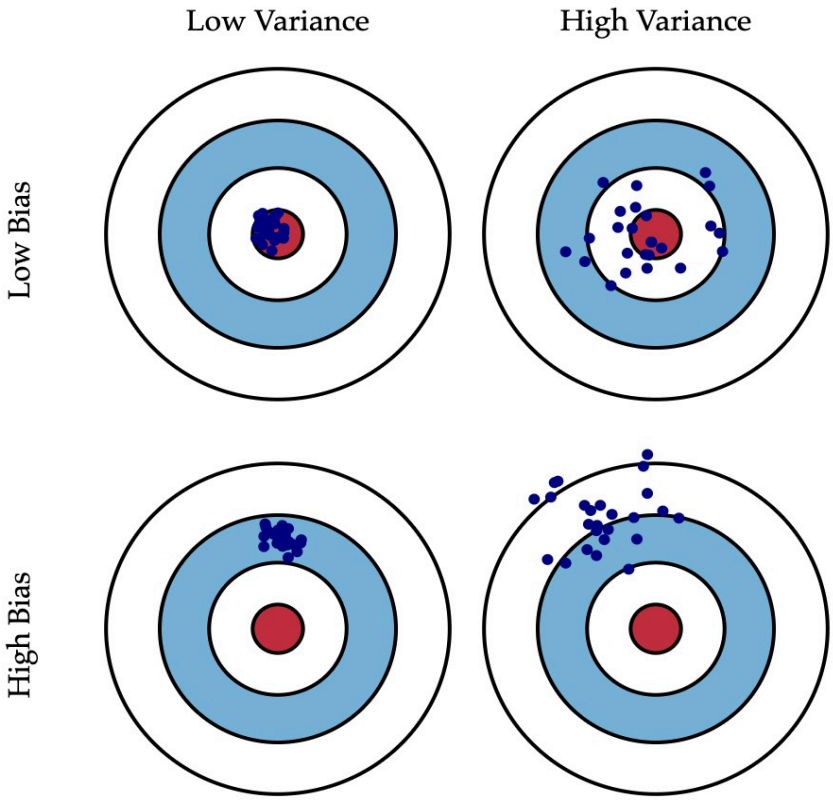



*Figure 1.* Graphical illustration of bias and variance. The bullseye represents a model that predicts correct values and the dots represent errored predictions. Retrieved from "Understanding the Bias-Variance Tradeoff" by S. Fortmann-Roe, 2012 (http://scott.fortmann-roe.com/docs/BiasVariance.html).

Using an example of a model that predicts voting intentions, the author provides a description of sources that lead to bias and variance. For example, bias was caused because the sample of people surveyed were chosen from a phone book, which means that anyone without a listed number would not be part of the sample. Additionally, the sample size was small, which led to higher variance. The author also describes the ML technique called k-*Nearest Neighbor* to represent the sweet spot between bias and variance. The k-Nearest Neighbor technique works by taking an original data point, using an arbitrary number *k*, and finding the nearest *k* number of data points in the training dataset. These *nearest neighbors* would then be predicted to be in the same category as the original data point. With this technique, the trade-off between bias and variance is described as "*increasing k* will *decrease variance* and *increase bias*. While *decreasing k* will *increase variance* and *decrease bias*" (Section 3.2, para. 3).

The article concludes with considerations for managing bias and variance. The author warns that while it may be tempting to minimize bias in any way possible, this approach should not be pursued at the expense of variance, since both errors by bias and errors by variance should be equally avoided. He states that "the sweet spot for any model is the level of complexity at which the increase in bias is equivalent to the reduction in variance" (Section 4.4, para. 2).

This article is an important resource for this study because it defines bias in a statistical context and presents the fact that overcompensating for bias may also result in inaccurate results caused by variance. This is an important consideration when attempting to mitigate bias in ML.

Hao, K. (2019, February 4). This is how AI bias really happens—and why it's so hard to fix. *MIT Technology Review.* Retrieved from https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

**Abstract.** Note: Abstract provided by the author of this annotated bibliography in the absence of a published abstract. This article examines the occurrence of bias in AI. First, the author describes three different stages in which bias can be introduced: framing the problem, collecting the data, and preparing the data. Next, the author discusses the reasons why bias is difficult to fix in AI, including the machine's lack of social context and the complexities of defining fairness. The author concludes with examples of how researchers are addressing this problem, including developing algorithms that detect bias and attempting to standardize the definition of fairness.

**Summary.** The author begins by acknowledging that bias in AI exists and addresses the fact that bias can be introduced in three key stages of the AI process: framing the problem, collecting the data, and preparing the data. In both the problem framing and data preparation stages, the ML engineer must define specific attributes to be used for the inputs and outputs of the model. The author states that these decisions are subject to both explicit and implicit biases. The second stage in which the data is collected may result in a collection of data that does not accurately represent reality or that reflects existing biases.

Even though there are specific stages where bias can be introduced into ML, the author highlights four main reasons why mitigating bias is difficult to achieve. First, the impacts of data and choices made by the engineer are not always obvious until much later in the

process. Second, the tools and practices are not currently designed to detect and mitigate bias. For example, current processes require splitting data to use for training and validation. Using the same data to test and validate means that biases will be found in both steps and errors may not be flagged. Third, the systems and models may be valid for one social context but may be completely wrong for another, and this variety of social contexts is not currently accounted for in ML processes. Finally, fairness is a very nuanced concept that has resulted in philosophical debates for centuries. Yet, fairness must be defined in mathematical terms in ML, which requires the engineer to make difficult decisions that may not always reflect the way society thinks about these issues. This article is useful for this study because it details specific areas of ML development where bias can be introduced. Additionally, the author points out specific decisions that engineers and analysts make that can result in bias. This information will be important for all decision-makers to enable the recognition and mitigation of bias.

Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Communications of the ACM*, *60*(2), 21–23. doi:10.1145/3022181

**Abstract.** The article discusses the impact of biased data on risk assessment and predictive policing. Topics include the use of risk-based assessment tools such as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) to aid U.S. states in sentencing criminals, concerns regarding the possibility of computerized risk-assessment algorithms penalizing racial minorities by overpredicting their likelihood of recidivism, and the use of poverty, postal codes, and employment status by COMPAS.

**Summary.** The scope of this article is focused on the use of AI and ML in criminal justice and policing in the United States. The author states that policing tools like

recidivism risk-assessment algorithms that predict the likelihood of a convicted criminal to reoffend, and predictive policing, defined as using analytics to determine where and when a crime might occur, are being used on a widespread basis. Reports from police departments who use these tools have noted successes in preventing crime and rearrests with the use of these tools. In 2012, New York State reported the recidivism risk-assessment tools had an accuracy rate of 0.71 area under curve (AUC), where the optimal AUC value is 1.0. PredPol, Inc., a company that develops a predictive policing tool, reported a 32% drop in burglaries and a 20% drop in vehicle theft in a jurisdiction that used the PredPol tool.

Critics argue the algorithms used in these tools may produce biased results. For example, "factors such as poverty, postal codes, and employment status can be used as proxies for race, as some are more highly correlated with minorities" (p. 22). The author states that there are predictive tools that will create more accurate predictions over time. For example, the author cites CommandCentral, a tool that begins by using historical data for predictions, but continually receives feedback and inputs from new crime data as it is used by police, learning from the feedback and creating more accurate predictions over time. However, he emphasizes that these tools should only supplement, not replace, the knowledge and training of professional police officers.

This article is useful for this study because it references specific real-world applications of ML in which bias may appear, explains the causes of the bias, and describes the potential impact of these biases on ML applications that are actively being used.

Nasraoui, O., & Shafto, P. (2016). Human-algorithm interaction biases in the big data cycle: A Markov chain iterated learning framework. *ArXiv E-Prints*. https://arxiv.org/abs/1608.07895v1

**Abstract.** Early supervised machine learning algorithms have relied on reliable expert labels to build predictive models. However, the gates of data generation have recently been opened to a wider base of users who started participating increasingly with casual labeling, rating, annotating, etc. The increased online presence and participation of humans has led not only to a democratization of unchecked inputs to algorithms, but also to a wide democratization of the "consumption" of machine learning algorithms' outputs by general users. Hence, these algorithms, many of which are becoming essential building blocks of recommender systems and other information filters, started interacting with users at unprecedented rates. The result is machine learning algorithms that consume more and more data that is unchecked, or at the very least, not fitting conventional assumptions made by various machine learning algorithms. These include biased samples, biased labels, diverging training and testing sets, and cyclical interaction between algorithms, humans, information consumed by humans, and data consumed by algorithms. Yet, the continuous interaction between humans and algorithms is rarely taken into account in machine learning algorithm design and analysis. In this paper, we present a preliminary theoretical model and analysis of the mutual interaction between humans and algorithms, based on an iterated learning framework that is inspired from the study of human language evolution. We also define the concepts of human and algorithm blind spots and outline machine learning approaches to mend iterated bias through two novel notions: antidotes and reactive learning.

**Summary.** The authors of this article examine the links between human iterative learning and algorithm iterative learning, the method of machine learning with repeated steps in which a prediction is made, feedback or new inputs are received, classifications of the data are adjusted, and a new prediction is made based on those adjustments. They argue that biases may arise during this learning process from both the algorithms and the humans involved in the selection of the data that are training the algorithms. By analyzing the process of iterative learning and the behavioral psychology of choice, the authors discover potential blind spots that result in selection bias, or the inaccurate selections of information and data.

They first discuss iterated learning with filter-bias dependency, which is the type of learning that is commonly used in recommendation systems like websites that recommend certain products to their users. The authors argue that bias may occur because potentially important data is filtered out; for example, an algorithm used on a commercial website gathered data from a user profile, correlated the data with potentially relevant items, and recommended these items to the user. However, there may be a potentially relevant items that the algorithm already deemed irrelevant and the opportunity for learning and adjusting is lost. Due to confirmation bias, which occurs when a person selects data that they believe to be true without questioning or attempting to falsify the hypothesis, the algorithm will only learn from the potentially inaccurate data that is available. This type of learning has allowed the algorithm to create a blind spot because some data is never seen by the human.

The second type of iterated learning the authors discuss is active-bias dependency, which is the type of learning that occurs when an algorithm offers all available data and a

human is actively providing feedback to train the algorithm to become more accurate. This type of learning is often used in search engine optimization by adjusting search query results based on past performance. The authors assert that humans are likely to introduce a blind spot when they may either be biased against acting, resulting in insufficient information from which the algorithm can learn; or the human may perceive a greater need for action in some situations but not others. For example, a person may not provide feedback for a restaurant unless the person had a bad experience, thus the algorithm is primarily receiving negative results.

By recognizing these types of blind spots, the authors offer suggestions for antidotes which can fix the bias after it occurs. One antidote the authors suggest is to use a reverse-Rocchio approach, which "uses selected data to change the set of relevant and non relevant *[sic]* instances" (Section 2.4.1, para. 1). The traditional Rocchio approach uses feedback to only gather relevant data, but this antidote would also use non-relevant data to balance the results of the blind spots. The authors also offer suggestions for reactive learning, which attempts to mitigate bias by intervening with the algorithm during the process.

This article is useful to this study because it identifies how humans may unintentionally create and perpetuate algorithms that produce biased results. By recognizing these potential blind spots, the ML engineer can attempt to mitigate the issues and the end users can be aware and cautious of the predictions generated by the algorithms.

Sweeney, L. (2013). Discrimination in online ad delivery. *ACM Queue*, *11*(3), 1-19. https://arxiv.org/abs/1301.6822

**Abstract.** A Google search for a person's name, such as "Trevon Jones", may yield a personalized ad for public records about Trevon that may be neutral, such as "Looking for Trevon Jones?", or may be suggestive of an arrest record, such as "Trevon Jones, Arrested?". This writing investigates the delivery of these kinds of ads by Google AdSense using a sample of racially associated names and finds statistically significant discrimination in ad delivery based on searches of 2184 racially associated personal names across two websites. First names, assigned at birth to more black or white babies, are found predictive of race (88% black, 96% white), and those assigned primarily to black babies, such as DeShawn, Darnell and Jermaine, generated ads suggestive of an arrest in 81 to 86 percent of name searches on one website and 92 to 95 percent on the other, while those assigned at birth primarily to whites, such as Geoffrey, Jill and Emma, generated more neutral copy: the word "arrest" appeared in 23 to 29 percent of name searches on one site and 0 to 60 percent on the other. On the more ad trafficked website, a black-identifying name was 25% more likely to get an ad suggestive of an arrest record. A few names did not follow these patterns. All ads return results for actual individuals and ads appear regardless of whether the name has an arrest record in the company's database. The company maintains Google received the same ad text for groups of last names (not first names), raising questions as to whether Google's technology exposes racial bias.

**Summary.** The scope of this article covers the topic of apparent racial discrimination found in dynamic online ad delivery through Google searches. The author performs an experiment in which she performs a search on Google.com and Reuters.com, which uses a Google platform for searches, using *black-identified* and *white-identified* names and

analyzes the patterns of ads that result. The author gathered *black-* and *white-identified* names using from the results of a job discrimination study performed by Marianne Bertrand and Sendhil Mullainathan in 2003.

The author searched 2,184 names over the span of a month. The author performed the searches at different times of the day, different days of the week, and using different IP addresses and machine addresses across the United States. She describes 19 observations as a result of her searches, including the finding that ads containing the word *arrest* appear in 60% of black-identifying name searches as opposed to 48% of white-identifying name searches. These results were determined to be statistically significant because the author determined there is less than a 0.1% probability that the data can be explained by chance.

Explaining why this discrimination is occurring is beyond the scope of this article, but the author posits that it may be due to the *Google Algorithm*, which learns over time which ads get more clicks. She explains that these types of search results can have negative effects in instances such as an employer querying a name of an interviewee and the interviewee being associated with the word *arrest* in the search results, even if the interviewee never had a criminal history.

This article is useful for this study as it proves the occurrence of bias in commonly used technologies. These findings are important in establishing that bias does exist in machine learning algorithms while considering the consequences of bias in ML.

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 5365-5372). Maui,

Hawaii: Hawaii International Conference on System Sciences.
doi:10.24251/HICSS.2018.668

**Abstract.** Biases in AI and machine learning algorithms are presented and analyzed through two issues management frameworks with the aim of showing how ethical problems and dilemmas can evolve. While "the singularity" concept in AI is presently more predictive than actual, both benefits and damage that can result by failure to consider biases in the design and development of AI. Inclusivity and stakeholder awareness regarding potential ethical risks and issues need to be identified during the design of AI algorithms to ensure that the most vulnerable in societies are protected from harm.

**Summary.** The aim of the authors is to consider the ethical implications of bias in ML in the context of two frameworks: Marx's 7 Stages of Issues Development and Fink's Four Stages of Crisis Management. The authors set up their arguments by introducing the history of AI and ML and examining many examples of bias found in current applications of ML. For example, it was reported in 2015 that searching *CEO* in Google primarily resulted in pictures of white men. The authors state that a significant issue that contributes to this problem is that ML algorithms are often in a *black-box* of secrecy. This issue may be due to private companies wanting to protect their intellectual property, but it can also occur because of the sheer complexity of these algorithms. "The algorithms are also often so complex that even the engineers and designers that have access to the formulas may struggle or fail to predict the outcome and effects of the algorithms *[sic]* results" (p. 5367).

Using the models by Marx and Fink, the authors review the current status of ethics in ML. As of the time of publication of the article in 2018, the authors believe that the United States is at stage three of Marx's seven stages: interest group development, in which industry and academic stakeholders partner to develop best practices and ethical standards. Through the first two stages, felt-need and media coverage, enough evidence has been presented to convince major players to create initiatives to address the ethical issues at hand. This evidence includes Facebook's potential role in influencing the 2016 United States Presidential election, and *ProPublica's* investigative report on recidivism risk-assessment tools perpetuating significant racial biases and discrimination. The authors also believe that the United States is only in stage one of Fink's model, the pre-crisis stage, in which warning signs and symptoms of a major issue are arising. The authors believe that the steps being taken through interest groups, including representatives of Amazon, Facebook, Google, IBM, and Microsoft developing *The Partnership on Artificial Intelligence to Benefit People and Society,* will help prevent stage two: the acute crisis stage in which a crisis event has occurred and caused major damage (Sellnow & Seeger, 2013).

The article concludes with a caution to AI experts and stakeholders. "As has been illustrated by the *ProPublica* investigation and other examples offered here, inclusivity and stakeholder awareness of impending ethical risks and issues are crucial in the design of AI to ensure that the most vulnerable in our society are protected from harm" (p. 5370). They urge value-based self-regulation to mitigate bias.

This article is useful to my study as it highlights the potential ethical dangers of ignoring bias in ML. It presents a call for action to all involved in ML.

**Solutions for Mitigating Bias in Machine Learning**

Amini, A., Soleimany, A., Schwarting, W., Bhatia, S., & Rus, D. (2019). Uncovering and mitigating algorithmic bias through learned latent structure. *Proceedings of the 2019 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* (pp. 1-7). Honolulu, HI: Association for the Advancement of Artificial Intelligence/Association for Computing Machinery. Retrieved from http://hdl.handle.net/1721.1/121101

**Abstract.** Recent research has highlighted the vulnerabilities of modern machine learning based systems to bias, especially for segments of society that are under-represented in training data. In this work, we develop a novel, tunable algorithm for mitigating the hidden, and potentially unknown, biases within training data. Our algorithm fuses the original learning task with a variational autoencoder to learn the latent structure within the dataset and then adaptively uses the learned latent distributions to re-weight the importance of certain data points while training. While our method is generalizable across various data modalities and learning tasks, in this work we use our algorithm to address the issue of racial and gender bias in facial detection systems. We evaluate our algorithm on the Pilot Parliaments Benchmark (PPB), a dataset specifically designed to evaluate biases in computer vision systems, and demonstrate increased overall performance as well as decreased categorical bias with our debiasing approach.

**Summary.** The scope of this article is focused on the development of an algorithm that integrates debiasing capabilities directly into an ML model training process. This algorithm learns the desired task; learns the latent, or hidden structure of the training data; and adapts to changes without supervision. To test their technique, the authors use a

commonly problematic application of ML: facial recognition systems, which have been found to provide significantly less accurate results with dark-skinned people.

Using their statistical algorithm, the authors define the latent variables, like skin tone and age, in a dataset and use these latent variables to resample the training data. This step will then allow the facial recognition model to recognize a higher frequency of previously latent variables and increase the probability that rarer data will be selected for training; a higher frequency of rarer data will then increase the sampling of under-represented variables. Experiments with facial recognition systems indicated that adaptive resampling of rare instances using the authors' algorithm resulted in increased classification accuracy on *dark male* subjects, confirming their hypothesis and validating their algorithm.

This article is useful for this study because it provides details of an algorithm that is designed to recognize potential bias in training models and correct the training models without supervision. This algorithm offers a potential use by ML engineers to debias training datasets without manual intervention.

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Paper published at the *Neural Information Processing Systems Conference* (pp. 4356-4364). Barcelona, Spain: Curran Associates Inc. https://arxiv.org/abs/1607.06520

**Abstract.** The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises

concerns because their widespread use, as we describe, often tends to amplify these biases…We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding.

**Summary.** The authors analyze word embedding, which is a framework that is used in many machine learning processing tasks. Word embeddings serve as a dictionary for computer programs by training on word co-occurrence in text corpora, or large, structured sets of text. The authors state that many papers have been written about word embeddings but "none of these papers have recognized how blatantly sexist the embeddings are and hence risk introducing biases of various types into real-world systems" (p. 1). To prove this claim, their stated goal for the paper is to "quantitatively demonstrate that word-embeddings contain biases in their geometry that reflect gender stereotypes present in broader society" (p. 3). They define a further goal of reducing gender bias in word embedding while, at the same time, maintaining the utility of word embedding.

To achieve these goals, the authors analyze word2vec, the embedding tool that is "trained on a corpus of Google News texts consisting of 3 million English words and terms into 300 dimensions" (p. 3). They use the pre-trained embedding tool on a set of texts from Google News to collect 26,377 words used for the experiments. They used a crowdsourcing platform to perform two types of experiments: "one where we solicited words from the crowd (to see if the embedding biases contain those of the crowd) and one where we solicited ratings on words or analogies generated from our embedding (to see if the crowd's biases contain those from the embedding)" (p. 6). Comparing the results from the crowdsourcing experiments with the biases presented in the word-

embedding, they evaluated the different types of gender bias that may be found in the word embeddings. The results of their analysis suggest that the biases of word embeddings are aligned with the crowd's judgement of gender stereotypes.

Based on their results, they attempted to develop a tool for debiasing the algorithm while maintaining utility in the system. They reported success in doing so; before applying their algorithm, 19% of gender-related analogies were judged as showing gender stereotype and after applying their algorithm, only 6% were deemed stereotypical.

This article directly relates to this study because the authors define specific types of bias found in a machine learning algorithm and provide a method for *debiasing* the process.

Koene, A., Dowthwaite, L., & Seth, S. (2018). IEEE P7003™ standard for algorithmic bias considerations. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, (pp. 38–41). doi:10.23919/FAIRWARE.2018.8452919

**Abstract.** The IEEE P7003 Standard for Algorithmic Bias Considerations is one of eleven IEEE ethics related standards currently under development as part of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. The purpose of the IEEE P7003 standard is to provide individuals or organizations creating algorithmic systems with development framework to avoid unintended, unjustified and inappropriately differential outcomes for users. In this paper, we present the scope and structure of the IEEE P7003 draft standard, and the methodology of the development process.

**Summary.** The authors explain that in response to growing public concerns about the use of automated decision-making tools and the limited transparency of the algorithmic processes, the IEEE Global Initiative on Ethics for Autonomous and Intelligence Systems

was launched in 2016. The goal of the founders was to create a set of ethical standards and codes of conduct for the implementation of intelligent technologies. As of early 2018, they have created a set of eleven ethical standards called the *IEEE P70XX series*. This article details the specifics of IEEE P7003, the Algorithmic Bias Considerations. The authors explain that IEEE P7003 is meant to be used when implementing systems that perform automated decision-making using personalization or individual assessment, like marketing automation applications that adjust prices or content based on an individual's behavior or preferences. The IEEE P7003 standard assists businesses in assuring users that steps have been taken to ensure fairness by providing a framework to identify and mitigate biases in algorithmic results. The framework currently includes elements like a set of guidelines for designing and implementing the ML applications, engaging various stakeholders involved in the ML project, guidelines for developers to be able to assess and address bias issues in the algorithms, and a taxonomy of algorithmic bias. The content of the P7003 standard is still in development, and the participants involved in development consist of 78 participants with expertise in a variety of fields such as computer science, law, and humanities. Once completed, the content will be submitted for approval to the IEEE-Standards Association.

This article is useful for this study as it provides an actionable framework standard for ML engineers, systems analysts, and data analysts to use to identify and mitigate bias in algorithms. This article is indicated as a *work in progress paper* as the IEEE P7003 standard is still in development.

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Paper published at the *Neural Information Processing Systems Conference* (pp. 4069-4079). Long Beach, CA: Curran Associates Inc. https://arxiv.org/1703.06856

**Abstract**. Machine learning can impact people with legal or ethical consequences when it is used to automate decisions in areas such as insurance, lending, hiring, and predictive policing. In many of these scenarios, previous decisions have been made that are unfairly biased against certain subpopulations, for example those of a particular race, gender, or sexual orientation. Since this past data may be biased, machine learning predictors must account for this to avoid perpetuating or creating discriminatory practices. In this paper, we develop a framework for modeling fairness using tools from causal inference. Our definition of counterfactual fairness captures the intuition that a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group. We demonstrate our framework on a real-world problem of fair prediction of success in law school.

**Summary.** The authors of this article examine current definitions of statistical fairness, provide examples of why fairness models might still introduce bias into predictions and offer a framework for a new type of fairness modeling. Because ML training data can often contain historical and societal prejudices, the authors state that an important role of the ML engineer is to thoughtfully select the right model of fairness to use for predictions. They suggest that some definitions of fairness may actually increase discrimination.

To explain this statement, they give a hypothetical example of a Law School Admissions Council that wants to predict successful students by using Law School Admissions Test

(LSAT) scores, grade point average (GPA) prior to entering law school, and first year average grade (FYA) as the predictors in an ML algorithm. One attempt at fairness modeling could be the use of the fairness through unawareness (FTU) model, which would remove protected attributes like race or sex from consideration. However, the authors argue that the LSAT, GPA, and FYA themselves are biased by race and sex due to social factors. Therefore, race and sex must be considered in the model as they are variables that, if changed, may produce different LSAT, GPA, and FYA results and thus, different ML predictions of student success.

The authors test their model, which they call *counterfactual fairness;* the model uses the protected attributes as a variable to predict, and they determine that their model, while providing slightly less accurate predictions than other models, provides much fairer and less biased results. Based on these results, they state that it is best to consider different social biases and actively compensate for them in predictive modeling.

This article is useful for my research because it provides a model that can be used by ML engineers to mitigate error from bias in ML predictions.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now Institute* (pp. 1–22). Retrieved from https://ainowinstitute.org/aiareport2018.pdf

**Abstract.** Our Algorithmic Impact Assessment Report helps affected communities and stakeholders assess the use of AI and algorithmic decision-making in public agencies and determine where – or if – their use is acceptable. Algorithms in government are already a part of decisions that affect people's lives, but there are no agreed-upon methods to ensure fairness or safety, or protect the fundamental rights of citizens. Our AIA report

provides a practical framework, similar to an environmental impact assessment, for agencies to bring oversight to automated decision systems.

**Summary.** The authors detail a proposed framework for public agencies to assess automated decision systems that they plan to implement. The authors explain that the use of decision-making algorithms makes it difficult for public agencies to identify or respond to bias because they are often performed in an unaudited *black box*. "The turn to automated decision-making and predictive systems must not prevent agencies from fulfilling their responsibility to protect basic democratic values, such as fairness, justice, and due process, and to guard against threats like illegal discrimination or deprivation of rights" (p. 5). They argue that implementing an Algorithmic Impact Assessment (AIA) will inform the public of the use of these systems and demand that public agencies consider all possible consequences of automated decision-making.

The AIA requires that an agency establish the scope of and clearly defines *automated decision system* within its own context. The authors argue that the definition should be published to allow the public to evaluate and comment on its justifiability. The authors stress that the definition should be addressed in terms of "human and social factors, the histories of bias and discrimination in the context of use, and any input and training data" (p. 12).

The authors stress that the agency must then perform a self-assessment, which would identify the potential impacts on the community and demonstrate how the system will have a net positive impact. The benefits of the assessment include creating opportunities for the agency to engage with the impacted communities early in the process and

requiring the agency to develop expertise in their own automated systems, both of which will ensure public trust.

The authors propose that AIA should be released for comment and scrutiny by both the public and technical experts. Before implementation, the agency must make necessary adjustments in response to these comments. Additionally, researchers and auditors should be allowed to continuously review systems once they are in use, to ensure the AIA is still in compliance. Their findings should also be published so the public continues to be involved and informed.

This article is useful for this study because it provides an assessment framework that could be used by systems and data analysts when considering implementing ML and automated decision-making systems. This framework ensures due diligence and consideration of all potentially impacted communities.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). San Francisco, California: Association for Computing Machinery. doi:10.1145/2939672.2939778

**Abstract.** Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model…In this work, we propose LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual

predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem.

**Summary.** This article presents reasons why transparency in ML models is needed and provides techniques that can be used to create transparency. The authors argue that trust is an important aspect of ML applications and that humans must be able to trust predictions and the models that create the predictions. In order to create trust, they believe that ML users should have access to textual or visual artifacts that provide explanations for how predictions were formulated. They offer an algorithm called LIME and a method called SP-LIME that they developed to provide explanations for predictions. This article includes detailed explanations of these methods and techniques and reviews potential drawbacks to using them that should be considered. The authors outline the desired characteristics of LIME and SP-LIME: interpretable; local fidelity, or being able to "correspond to how the model behaves in the vicinity of the instance being predicted" (p. 1137); model-agnostic; and global perspective, or providing an explanation for both the predictions and the models.

The authors first evaluate LIME and SP-LIME through a simulated user experiment by taking two sentiment analysis (natural language processing) datasets of books and DVDs and set the ML model to classify the product reviews as positive or negative. After the models ran, they compared the results of LIME with the results of other explanation procedures like parzen and greedy. LIME provided more trustworthy explanations of the predictions and the models than other explanation methods.

Next, they evaluated LIME with human subjects who were recruited through Amazon Mechanical Turk, a crowdsourcing marketplace. The subjects had no ML expertise and

were asked to perform three tasks with aid from LIME: (1) choose which of two classifiers generalize better; (2) perform feature engineering to improve the model based on the LIME explanations; and (3) identify and describe classifier errors using the explanations. Their results indicate that even without ML expertise, the human subjects were able to improve the model with the aid of the explanations provided by LIME. This article does not directly address bias, but it is useful for this study because it provides a successful method for creating transparency within ML, which will help in the effort to identify and remove bias from ML algorithms. Experts argue that the *black-box* of ML is a significant challenge and these methods directly address that problem.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 1-17. doi:10.1177/2053951717743530

**Abstract.** Decisions based on algorithmic, machine learning models can be unfair, reproducing biases in historical data used to train them. While computational techniques are emerging to address aspects of these concerns through communities such as discrimination-aware data mining (DADM) and fairness, accountability and transparency machine learning (FATML), their practical implementation faces real-world challenges. For legal, institutional or commercial reasons, organisations might not hold the data on sensitive attributes such as gender, ethnicity, sexuality or disability needed to diagnose and mitigate emergent indirect discrimination-by-proxy, such as redlining. Such organisations might also lack the knowledge and capacity to identify and manage fairness issues that are emergent properties of complex sociotechnical systems. This paper presents and discusses three potential approaches to deal with such knowledge and

information deficits in the context of fairer machine learning. Trusted third parties could selectively store data necessary for performing discrimination discovery and incorporating fairness constraints into model-building in a privacy-preserving manner. Collaborative online platforms would allow diverse organisations to record, share and access contextual and experiential knowledge to promote fairness in machine learning systems. Finally, unsupervised learning and pedagogically interpretable algorithms might allow fairness hypotheses to be built for further selective testing and exploration. Real-world fairness challenges in machine learning are not abstract, constrained optimisation problems, but are institutionally and contextually grounded. Computational fairness tools are useful, but must be researched and developed in and with the messy contexts that will shape their deployment, rather than just for imagined situations. Not doing so risks real, near-term algorithmic harm.

**Summary.** The authors of this article discuss the mitigation of discrimination and bias in ML caused by unfair patterns in datasets. They discuss current techniques in place, such as discrimination-aware data mining (DADM) and fairness, accountability and transparency in machine learning (FATML), which are standardized techniques that involve altering processes and introducing models to algorithms that would correct bias despite the existence of bias in the training data. They also share that there are multiple ways to define fairness, such as accuracy equity, which considers the accuracy of a predictive model for each group, and equality of opportunity, which considers the likelihood that each group would receive the same prediction regardless of their foundational differences. Although there are various ways to measure fairness and attempt de-biasing, the authors state that studies have found that it is not possible for one

model to satisfy several of these techniques at the same time. Therefore, the techniques must be intentionally selected based on the context of the ML project.

In order to successfully select a model, they argue that the ML engineer should have access to all available data to fully examine and consider all possible consequences of each model. However, due to privacy laws, such as the European Union's General Data Protection Regulation (GDPR), and legally protected characteristics including disability, race, sexuality, and gender, the access to available data may be limited. The authors propose three approaches to improving fairness despite the potential of limited data.

The third-party approach would allow another organization to hold and govern sensitive data and work with the first party to determine if and when the sensitive data can be used. This approach "is primarily useful where trust in the organization interested in model building is low, or potential reputational risk is high" (p. 12), such as in insurance or hiring.

The collaborative knowledge base approach suggests the creation of a database or wiki that features fairness issues, research, and techniques that are shared with the data science and technical community. This could be beneficial "where general uncertainty is acute, risk assessment must be undertaken preemptively, or risks are complex, changing and sociotechnical" (p. 12). The authors state that this approach would require a cultural mindset change, as many organizations are reluctant to openly discuss their models.

The exploratory approach would be used to explore unusual patterns in data in which sensitive characteristics are difficult to obtain or do not exist at all. The authors suggest that more work needs to take place to formalize methods to be used in this approach.

This article is useful for this research as it provides practical approaches to identifying and mitigating bias. These approaches can be implemented by data scientists and engineers, as well as lay people who are not as familiar with theoretic frameworks of ML.

## Conclusion

Machine learning is commonly used across a multitude of industries and applications (Brennan et al., 2008; Brundage et al., 2018; Daud et al., 2017; Indrajith & Vijayakumar, 2016; Jordan & Mitchell, 2015). Experts believe that ML makes data-related tasks more efficient and safer from human error (Bakshi & Bakshi, 2018; Borana, 2016; de Saint Laurent, 2018). However, many warn that bias can be integrated into and perpetuated by ML algorithms and that humans must carefully monitor outputs and correct bias results (Amini et al., 2019; Bolukbasi et al., 2016; Brennan et al., 2009; Brundage et al., 2018; de Saint Laurent, 2018; Fortmann-Roe, 2012; Hao, 2019; Li, 2019; Kirkpatrick, 2017; Koene, Dowthwaite, & Seth, 2018; Nasraoui & Shafto, 2016; Reisman et al., 2018; Sweeney, 2013; Veale & Binns, 2017;West et al., 2019; Yapo & Weiss, 2018).

This annotated bibliography presents findings that will provide ML engineers, system analysts, and data analysts with tools and techniques to identify and mitigate bias in ML. Information is presented in the following categories: (a) understanding machine learning, (b) explanations for bias in machine learning, and (c) solutions for mitigating bias in machine learning.

### Understanding Machine Learning

Machine learning algorithms accomplish learning through methods that include supervised or unsupervised learning (Bakshi & Bakshi, 2018). Supervised learning requires inputs and outputs to consist of labeled datasets; the two major types of supervised learning algorithms are classification, in which the goal is to predict the assignment of categories to data, and regression, defined as the prediction of a continuous number, such as income (Bakshi & Bakshi, 2018). Unsupervised learning involves datasets that are not labeled and require the

algorithm itself to recognize patterns and categorizations; unsupervised learning algorithms include transformation, defined as using the existing formations of data to create new formations, and clustering, defined as the partitioning of data into separate groups of similar items (Bakshi & Bakshi, 2018).

A key step in developing an accurate and successful ML application is assessing the data and the project goals before selecting the appropriate model and continuing to reassess as the model is being used (Bakshi & Bakshi, 2018; de Saint Laurent, 2018; Hao, 2019; Veale & Binns, 2017) For example, the *k-Nearest Neighbor* algorithm is best used for binary data, or data that may be only one of two things, like the sex designation of male or female (Bakshi & Bakshi, 2018; Fortmann-Roe, S., 2012). However, this algorithm is not a good choice for datasets with many different attributes, as it is not complex enough to handle all of the variables (Bakshi & Backshi, 2018).

News headlines have popularized ML, and de Saint Laurent (2018) states that researchers and journalists often cherry-pick the best examples of ML. This practice has led to some common myths that de Saint Laurent (2018) attempts to debunk. For example, ML algorithms cannot learn new concepts in the same way that humans can, "…for a human learning to recognise *[sic]* a cat or a dog means learning the concepts of cats and dogs, for a machine it simply means be able to recognise *[sic]* patterns of pixels and matching them to a certain category" (de Saint Laurent, 2018, p. 741). Although machines may learn in a more transactional way than humans, they are lauded by experts as able to gather insights and discover patterns from large and disparate data sets in a much more efficient and accurate way than can be accomplished manually by humans (Bakshi & Bakshi, 2018; de Saint Laurent, 2018).

De Saint Laurent (2018) believes that understanding both the powers and the limitations of ML is crucial for avoiding the potential dangers and risks of ML. ML engineers, business leaders, and ML adopters should learn as much as possible about ML technologies to be able to safely employ ML to its fullest potential (de Saint Laurent, 2018).

**Explanations for Bias in Machine Learning**

Fortmann-Roe (2012) defines ML errors due to bias as over-fitting a model, or when a model's average prediction is far from the correct value. He also describes another type of common error, variance, in which a model results in a great deal of variability. Fortmann-Roe (2012) explains that there is a trade-off between these two types of errors; when bias decreases, variance increases, and vice versa. He warns that this trade-off needs to be considered when trying to mitigate bias and that it is more important to find the middle ground between bias and variance than removing bias altogether.

Finding that middle ground and mitigating bias is not an easy feat because bias can be integrated into ML applications in multiple stages of development and deployment (Hao, 2019). Hao (2019) states that bias can appear during the problem framing stage when the goals of the project are being defined, the data collection stage when relevant data is being selected to train the ML algorithm, and the data preparation stage when the input and output attributes are being defined and transformed. The problem framing stage involves many roles beyond the ML engineer, as this is when the business problem must be understood and the goals must be developed (Hao, 2019). For example, a credit card company wanting to use ML to predict *creditworthy* customers must first define what *creditworthy* means, and their definition may be influenced by profit rather than fairness (Hao, 2019).

Kirkpatrick (2017) specifically examines the data collection and preparation stages in his discussion about predictive policing. Kirkpatrick (2017) explains that factors like postal codes, employment statuses, and race are attributes used in predictive policing tools that are employed to predict high-crime geographic areas. He states that critics argue that many of these attributes can contribute to bias, as they can often be correlated with minorities (Kirkpatrick, 2017). He points out that these tools, like most ML applications, continue to collect data to learn new insights and adjust their predictions based on these new insights (Kirkpatrick, 2017). The new insights are also subject to biases, though, and Nasraoui and Shafto (2016) attribute some of these biases to human and machine blind-spots.

Nasraoui and Shafto (2016) examine ML algorithms that make recommendations to users, like a movie streaming service suggesting a movie that it predicts the user will enjoy. These predictions may be created by gathering historical data from other similar users and filtering out any movie the algorithm deems irrelevant, or by offering every movie from the start and learning the user's preferences as the user submits ratings for movies over time (Nasraoui & Shafto, 2016). When filtering out *irrelevant* data, the machine is allowing a possible blind-spot by preventing the user from seeing a potentially relevant movie (Nasroui & Shafto, 2016). On the other hand, depending on the user to provide learning data to the ML algorithm by rating different movies allows for the possibility of a user never rating a movie or only rating movies the user did not like, which results in another blind-spot and inaccurate results (Nasroui & Shafto, 2016).

Sweeney (2013) provides another example of bias found in predictive recommendations in her experiment with Google ads. She discovered that using Google to search for black-identified names results in ads that contain the word *arrest* 60% of the time compared to 48% of

the time when searching for white-identified names (Sweeney, 2013). Although her study does not identify the cause of these bias results, she suggests that the Google Algorithm receives inputs by way of clicks, learns which search results tend to have a higher click rate, and then recommends those ads more often than others when similar searches occur (Sweeney, 2013). This example of potential racial bias (Sweeney, 2013) leads to ethical implications.

Yapo and Weiss (2018) state that there has been enough evidence of ethical risks from ML applications to encourage interest groups to be created to address the ethical implications and develop best practices and ethical standards to avoid these risks. For example, representatives from major companies such as Amazon, Facebook, Google, IBM, and Microsoft have developed *The Partnership on Artificial Intelligence to Benefit People and Society* to develop evidence-based practices and guidelines related to ethics, fairness, privacy, and trustworthiness between people and ML systems (Yapo & Weiss, 2018). A significant problem with ML is that the predictions are created in a *black-box* of secrecy (Brundage et al., 2018; Deng, 2018; Koene et al., 2018; Kusner et al., 2017; Reisman et al., 2016; Ribeiro et al., 2016; Veale & Binns, 2017; Yapo & Weiss, 2018). Yapo and Weiss (2018) state that the black-box may occur because companies want to protect their intellectual property or because the complexity of the algorithms makes them difficult or impossible to understand, even by ML and data science experts. The black-box of secrecy is a significant problem, and by understanding and recognizing where bias may come from and how and when it may be integrated into ML applications, ML developers and users will be a step closer to mitigating and combatting bias in ML (Yapo & Weiss, 2018).

**Solutions for Mitigating Bias in Machine Learning**

The Global Initiative on Ethics of Autonomous and Intelligent Systems led by the Institute of Electrical and Electronics Engineers (IEEE) is an interest group that is working to develop a set of ethical standards and codes of conduct for AI and ML (Koene et al., 2018). Koene et al. (2018) examine the IEEE P7003 standard that has been developed by the group: Algorithmic Bias Consideration. Although the framework is still in development, some elements include procedures and criteria for identifying and selecting data sets for bias quality control, a taxonomy of algorithmic bias, and guidelines for developers to identify when to evaluate bias issues and suggested methods for mitigating these issues (Koene et al., 2018).

Methods that have shown promise for mitigating ML bias include a ML model Bolukbasi et al. (2016) developed to address bias found in word embeddings. Through analyzing correlative word embeddings, Bolukbasi et al. (2016) discovered that some predictions may include gender biases and stereotypes. For example, automatically generated analogies interpret *she* is to *sewing* as *he* is to *carpentry* (Bolukbasi et al., 2016). To combat these stereotypes, they detail an ML model that they developed for debiasing the algorithm which resulted in a 13% decrease in gender stereotypical results (Bolukbasi et al., 2016).

Kusner et al. (2017) also detail an ML model that they created and called *counterfactual fairness*. They examined different types of statistical fairness algorithms like the fairness through unawareness (FTU) model that removes protected attributes like race and sex from the datasets (Kusner et al., 2017). However, they argue that because of the relationship between protected attributes and the data, not accounting for these attributes may actually increase discrimination (Kusner et al., 2017). The *counterfactual fairness* model that they describe explicitly uses

protected attributes to make predictions; they report that the model produces slightly less accurate, but more fair and unbiased predictions than other fairness models (Kusner et al., 2017).

Veale & Binns (2017) also discuss fairness in ML, but from a process-oriented context. Their focus is specifically on ML bias caused by data inputs; for example, when the religious sects *Catholic* and *Protestant* are categorized under one label of *Christianity*, the relevant distinctions of each of those religions may then be lost in the ML model (Veale & Binns, 2017). They state that in certain contexts, protected attributes like race, gender, and religion should not be used in ML applications, but that it is important for the ML engineer to have the full context of all available attributes in order to accurately select the right model (Veale & Binns, 2017). To address the sensitivity of data and the necessity for knowledge of this sensitive data, they suggest three approaches, including an approach in which a third-party would be authorized to access the sensitive data and work closely with the engineers to determine which attributes they need and which model to use to create a successful ML application (Veale & Binns, 2017). They also suggest the development of a database or wiki in which ML experts share their experiences, models, research, and other helpful tools with others in the technical community as a way of spreading the knowledge while being able to maintain the protection of sensitive data (Veale & Binns, 2017). Finally, they describe an exploratory approach that could be used to explore unusual patterns in data in which sensitive characteristics are difficult to obtain or do not exist at all (Veale & Binns, 2017). They suggest that more work needs to take place to formalize methods to be used in this approach (Veale & Binns, 2017).

Amini et al. (2019) introduce an algorithm that is meant to integrate into an existing ML model, learn the latent structure of training data, recognize bias in the data, and adjust the training dataset to mitigate the bias, all without human supervision. Although many experts

stress the necessity of constant human supervision (Brundage et al., 2018; de Saint Laurent, 2018; Hao, 2019; Ribeiro, 2016), this algorithm addresses the issue of the black-box and the complex algorithm activities that are difficult for engineers and developers to fully understand (Yapo & Weiss, 2018) by allowing an algorithm to identify and mitigate the ML bias. Amini et al. (2019) tested this algorithm on facial recognition software, which resulted in increased classification accuracy when attempting to identify *dark male* subjects; facial recognition systems typically demonstrate significantly less accuracy in identifications of this population.

Given the issue of the black-box, Ribeiro et al. (2016) developed a model called LIME that attempts to open the box and explain what is occurring in ML algorithms and why they made their specific predictions. They tested the model by asking humans subjects with limited ML expertise to use the explanations provided by LIME to perform tasks like identifying and describing classifier errors in an algorithm and improving the algorithm to remove these errors (Ribeiro et al., 2016). Their analysis of the experiment results found that the subjects were usually successful in performing their tasks even though they were not ML engineers (Ribeiro et al., 2016). They concluded that their model is crucial for human and machine interactions, as the insights it provides into the ML activities are useful for accurate selection of models for an ML project, assessing trust of the predictions, and improving untrustworthy models (Ribeiro et al., 2016).

Even with tools and techniques to safeguard against bias in ML, Reisman et al. (2018) propose an Algorithmic Impact Assessment Report (AIA) that they recommend for use by public agencies to assess the agencies' use of automated decision systems and provide transparency to the community and other stakeholders. Reisman et al. (2018) argue that tools that can directly affect the public, like systems that may approve or deny mortgages to people based on their

home zip code, should be subject to scrutiny by that same affected community. The AIA would require a public agency to clearly define the scope of the technology the agency plans to implement, release the details to the public, answer questions and comments, and be subject to continued evaluation throughout the deployment of that technology (Reisman et al., 2018). Reisman et al. (2018) believe that this level of scrutiny and transparency will address potential issues of fairness, discrimination, and justice in ML.

The literature in this section offers complex techniques and tools that can be implemented by ML engineers (Amini et al., 2019; Bolukbasi et al., 2016; Kusner et al., 2017, Ribeiro et al., 2016), as well as guidelines and resources for systems analysts, data analysts, and ML engineers to mitigate and protect against bias in ML (Koene et al., 2018; Reisman et al., 2018; Veale & Binns, 2017). As many of the experts convey, the context of the ML project is key in determining how to select the right models, standards, and guidelines to improve the accuracy of ML algorithms (Kusner et al., 2017; Ribiero et al., 2016; Veale & Binns, 2017).

**Final Thoughts**

Machine learning is a complex tool that is being used in many applications in our world at this time (Brennan et al., 2008; Brundage et al., 2018; Daud et al., 2017; Indrajith & Vijayakumar, 2016; Jordan & Mitchell, 2015). Although ML engineers are responsible for the creation of ML algorithms, most humans influence the technology, whether through providing the algorithm with learning inputs by rating a restaurant online, defining a business problem and goal that will be addressed by an ML application, or utilizing the predictions by allowing a smart phone to finish a half-typed word in a text message (Bolukbasi et al., 2016; Hao, 2019; Indrajith & Vijayakumar, 2016; Nasraoui & Shafto, 2016). Through all of these types of activities, there is potential for bias to be integrated into the ML algorithms and predictions (Amini et al., 2019;

Bolukbasi et al. 2016; de Saint Laurent, 2018; Fortmann-Roe, 2012; Hao, 2019; Kirkpatrick, 2017; Koene et al., 2018; Nasraoui & Shafto, 2016; Reisman et al., 2018; Sweeney, 2013; Veale & Binns, 2017; Yapo & Weiss, 2018).

Experts have referenced ML algorithms as a *black-box* (Brundage et al., 2018; Deng, 2018; Koene et al., 2018; Kusner et al., 2017; Reisman et al., 2016; Ribeiro et al., 2016; Veale & Binns, 2017), but they also stress the necessity of human intervention to mitigate bias (Amini et al., 2019; Bolukbasi et al. 2016; de Saint Laurent, 2018; Fortmann-Roe, 2012; Hao, 2019; Kirkpatrick, 2017; Koene et al., 2018; Nasraoui & Shafto, 2016; Reisman et al., 2018; Sweeney, 2013; Veale & Binns, 2017; Yapo & Weiss, 2018). The literature presented in this annotated bibliography provides tools, techniques, and guidelines to be able to open that black-box and enable people to understand what is occurring in the machines. ML engineers, system analysts, and data analysts are in a position to be explicitly involved in ML applications and the literature presented will help develop the capabilities to identify and mitigate bias in machine learning.

**References**

Amini, A., Soleimany, A., Schwarting, W., Bhatia, S., & Rus, D. (2019). Uncovering and mitigating algorithmic bias through learned latent structure. *Proceedings of the 2019 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* (pp. 1-7). Honolulu, HI: Association for the Advancement of Artificial Intelligence/Association for Computing Machinery. Retrieved from http://hdl.handle.net/1721.1/121101

Bakshi, K., & Bakshi, K. (2018). Considerations for artificial intelligence and machine learning: Approaches and use cases. Paper published at the *2018 IEEE Aerospace Conference* (pp. 1-9). Big Sky, MT: Institute of Electrical and Electronics Engineers. doi:10.1109/AERO.2018.8396488

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Paper published at the *Neural Information Processing Systems Conference* (pp. 4356-4364). Barcelona, Spain: Curran Associates Inc. https://arxiv.org/abs/1607.06520

Borana, J. (2016). Applications of artificial intelligence & associated technologies. In *Proceedings of International Conference on Emerging Technologies in Engineering, Biomedical, Management and Science,* 64-67. Retrieved from https://www.cs.buap.mx/~aolvera/IA/2016_Applications%20of%20IA.pdf

Brachman, R. J. (2006). AI more than the sum of its parts. *AI Magazine*, *27*(4), 19–19. doi:10.1609/aimag.v27i4.1907

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior, 36*(1), 21–40. doi:10.1177/0093854808326545

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., … Amodei, D. (2018). The malicious use of artificial intelligence: forecasting, prevention, and mitigation. https://arxiv.org/abs/1802.07228v1

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *AI Magazine, 26*(4), 53–53. doi:10.1609/aimag.v26i4.1848

Center for Public Issues Education. (2014). Evaluating information sources. *University of Florida*. Retrieved from http://www.piecenter.com/wp-content/uploads/2014/08/evaluateinfo.pdf

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: origins of bias and theory of debiasing. *BMJ Quality & Safety, 22*(Suppl 2), ii58-ii64. doi:10.1136/bmjqs-2012-001712

Data analyst job description. (n.d.). *Workable.com*. Retrieved May 18, 2019, from: https://resources.workable.com/data-analyst-job-description

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW 17 Companion,* (pp. 415-421). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. doi:10.1145/3041021.3054164

de Saint Laurent, C. (2018). In defence of machine learning: Debunking the myths of artificial intelligence. *Europe's Journal of Psychology, 14*(4), 734–747. doi:10.5964/ejop.v14i4.1823

Deng, L. (2018). Artificial intelligence in the rising wave of deep learning: The historical path and future outlook [perspectives]. *IEEE Signal Processing Magazine*, *35*(1), 180–177. doi:10.1109/MSP.2017.2762725

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. doi:10.1145/2347736.2347755

Fortmann-Roe, S. (2012). Understanding the bias-variance tradeoff. Retrieved from http://scott.fortmann-roe.com/docs/BiasVariance.html

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism*, 69, S36–S40. doi:10.1016/j.metabol.2017.01.011

Hao, K. (2019, February 4). This is how AI bias really happens—and why it's so hard to fix. Retrieved from https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Indrajith, G., & Vijayakumar, K. (2016). Automatic mathematical and chronological prediction in smartphone keyboard. *International Journal Of Engineering And Computer Science, 5*(5), 16714-16718. doi:10.18535/ijecs/v5i5.64

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 253–255. doi:10.1126/science.aaa8415

Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Communications of the ACM*, *60*(2), 21–23. doi:10.1145/3022181

Koene, A., Dowthwaite, L., & Seth, S. (2018). IEEE P7003™ standard for algorithmic bias considerations. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), (pp. 38–41). doi:10.23919/FAIRWARE.2018.8452919

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Paper published at the *Neural Information Processing Systems Conference* (pp. 4069-4079). Long Beach, CA: Curran Associates Inc. https://arxiv.org/1703.06856

Li, M. (2019, May 13). Addressing the biases plaguing algorithms. Retrieved May 24, 2019, from https://hbr.org/2019/05/addressing-the-biases-plaguing-algorithms

Lighthill, J. (1973). Artificial intelligence: A general survey. *Artificial Intelligence: A Paper Symposium*. London: Science Research Council. Retrieved from http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm

Machine learning engineer job description. (n.d.). *Workable.com*. Retrieved May 18, 2019, from: https://resources.workable.com/machine-learning-engineer-job-description

McCarthy, J., & Feigenbaum, E. A. (1990). In memoriam: Arthur samuel: Pioneer in machine learning. *AI Magazine, 11*(3), 10–10. doi:10.1609/aimag.v11i3.840

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now Institute* (pp. 1–22). Retrieved from https://ainowinstitute.org/aiareport2018.pdf

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). San Francisco, California: Association for Computing Machinery. doi:10.1145/2939672.2939778

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 71–105. doi:10.1147/rd.33.0210

Sellnow, T. L., & Seeger, M. W. (2013). *Theorizing crisis communication*. Retrieved from eBook Collection (Ebook Central Academic Complete) database.

Sweeney, L. (2013). Discrimination in online ad delivery. *ACM Queue*, *11*(3): 1-19. https://arxiv.org/1301.6822

System analyst job description. (n.d.). *Workable.com*. Retrieved May 18, 2019, from: https://resources.workable.com/system-analyst-job-description

Turing, A. M. (1995). Lecture to the London Mathematical Society on 20 February 1947. *M.D. Computing: Computers in Medical Practice*, *12*(5), 390–397. PMID: 7564963

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 1-17. doi:10.1177/2053951717743530

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. *AI Now Institute* (pp. 1–33). Retrieved from https://ainowinstitute.org/discriminatingsystems.html

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 5365-5372). Maui, Hawaii: Hawaii International Conference on System Sciences. doi:10.24251/HICSS.2018.668