2020

Implementation Considerations for Mitigating Bias in Supervised Machine Learning

Bardia Bijani Aval
College of Saint Benedict/Saint John's University, bbijaniav001@csbsju.edu

Implementation Considerations for Mitigating Bias in Supervised Machine Learning

by

Bardia Bijani Aval

The College of Saint Benedict and Saint John’s University

May 7, 2020

Implementation Considerations for Mitigating Bias in Supervised Machine Learning

By Bardia Bijani Aval

Approved By:

Dr. Noreen Herzfeld

Thesis Advisor, Professor of Computer Science and Theology

Dr. Erica Stonestreet

Faculty Reader, Associate Professor of Philosophy

Dr. Peter Ohmann

Faculty Reader, Assistant Professor of Computer Science

Dr. Imad Rahal

Chair, Department of Computer Science

Dr. Catherine Bohn-Gettler

Co-Director of Undergraduate Research – All-College Thesis

Dr. Mary Stenson

Co-Director of Undergraduate Research – All-College Thesis

Abstract

Machine Learning (ML) is an important component of computer science and a mainstream way of making sense of large amounts of data. Although the technology is establishing new possibilities in different fields, there are also problems to consider, one of which is bias. Due to the inductive reasoning of ML algorithms in creating mathematical models, the predictions and trends found by the models will never necessarily be true – just more or less probable. Knowing this, it is unreasonable for us to expect the applied deductive reasoning of these models to ever be fully unbiased. Therefore, it is important that we set expectations for ML that account for the limitations of reality.

The current conversation of ML regards how and when to implement the technology to mitigate the effect of bias on its results. This thesis suggests that the question of "whether" should be addressed first. We tackle the issue of bias from the standpoint of justice and fairness in ML, developing a framework tasked with determining whether the implementation of a specific ML model is warranted. We accomplish this by emphasizing the liberal values that drive our definitions of societal fairness and justice, such as the separateness of persons, moral evaluation, freedom and understanding of choice, and accountability for wrongdoings.[1]

[1] An immense "thank you" to Dr. Noreen Herzfeld for her limitless encouragement and mentorship throughout the development of this work.

Contents

1 Introduction 6
2 Background 7
2.1 Machine Learning 7
2.1.1 Supervised Learning 7
2.1.1.2 Methods and Black-Box Models 9
2.2 Biases 10
2.2.1 Dataset Bias 10
2.2.2 Algorithmic Bias 11
2.2.3 Hardware Bias 12
2.3 Cognitive Source of Dataset and Algorithmic Bias 12
2.4 Inductive and Deductive Reasoning 14
2.4.1 The Problem of Induction 15
2.5 Justice and Fairness in ML 15
2.6 Full Accommodation of Fairness 16
2.7 Problem Statement 17
3 Current State of the Field 19
3.1 Fairness Verification 19
3.2 Non-Empirical Feedback Loops 19
3.3 Current Fairness Criteria 20
3.3.1 Anti-Classification 20
3.3.2 Classification Parity 21
3.3.3 Calibration 22
3.4 Mutual Exclusivity of Fairness Criteria 22
3.5 Theoretical Versus Empirical Observations 23
4 Development of Framework 24
4.1 Distributive Justice 24
4.1.1 John Rawls' Justice as Fairness 24
4.2 Compensatory Justice 28
4.3 Moral Evaluation 29
4.3.1 Adam Smith's Impartial Spectator 30
4.3.2 Algorithmic Auditing 31
4.4 Freedom of Choice 31
4.4.1 Choice of Task Avoidance 32
4.4.2 Choice of Non-ML Alternatives 33

4.4.3 Explainable and Interpretable ML 35
4.5 Accountability 37
4.6 Pre-Implementation Flowchart 39
4.7 Post-Implementation Accountability 39
4.7.1 Restorative Justice 40
4.7.2 Criminal Justice 41
4.7.3 Additional Comments 42
5 Reflections on Work and Future Areas of Research 43
6 Conclusion 45
7 References 46

1 Introduction

In the past few decades, humans have made technological progress on an exponential scale. Moore's law, although unable to precisely predict future advancements, emphasizes the remarkable improvements we have made in the field of computer science. We have witnessed exponential growth in both computational power and clock speeds, which inevitably are expanding the possibilities within the field.[2]

As we are broadening the capabilities of computers and specifically artificial intelligence (AI), we are exploring new applications. Some of these aid us in addressing the shortcomings of humans, increasing both time- and task efficiency. An example is the concept of self-driving cars. In 2018, the number of traffic deaths exceeded 40,000 for the third consecutive year; about 110 deaths daily. In 92 (±2) percent of cases, the accident could be attributed to human error. This means that AI, although it would add a small computing error, has the potential to save many lives by removing all or most of the unreliable human factor in car accidents.[3] Our respective economies will have to adapt to changes posed by these new developments. Nevertheless, there is definite potential to help human beings lead better lives.

However, these advancements do not come without repercussions. In the endeavor of continuing our technological explorations and addressing more complex issues, we are facing new challenges in different aspects of development and application. One of these challenges is bias in artificial intelligence (AI) and automation, whose foundation is machine learning. These biases are not mere theoretical obstacles but have grave practical consequences for many undeserving individuals (which will become evident as we proceed). For the purpose of this work, we intend to explore the shortcomings of the machine learning models that lead to these biases, and how we could address them by being more conscientious about how, but especially when we implement the technology.

In this endeavor, we have tasked ourselves with developing a framework that will consider the current impact of bias in machine learning and its relevance to societal injustices, to give recommendations on whether implementation of a given machine learning model is warranted in the given context. In the following sections, we will explore fundamental themes that such a framework should include and use a bottom-up approach to establish its components accordingly.

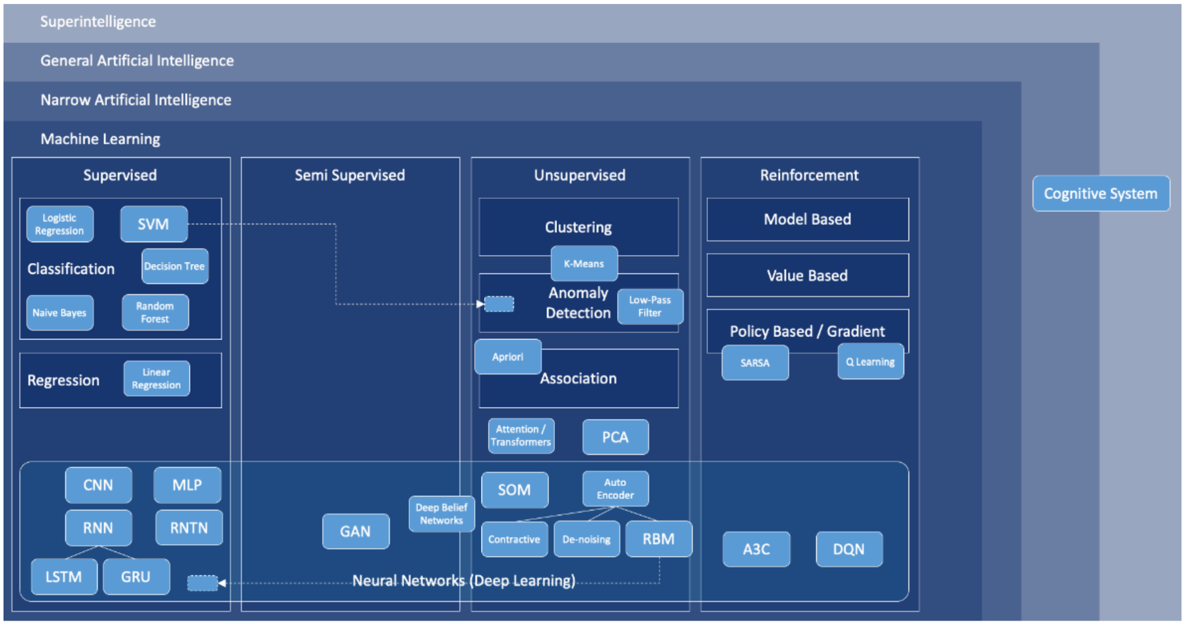
[2] Roser, Max, and Hannah Ritchie. "Technological Progress." Our World in Data, 2020. https://ourworldindata.org/technological-progress.

[3] "Traffic Safety Facts: A Brief Statistical Summary." *U.S Department of Transportation,* 2015. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115

2 Background

2.1 Machine Learning

Machine learning (ML) is the ability for computers, through analysis of extensive datasets, to find trends and make predictions using statistical models. It is the foundation of how many AI applications learn, find patterns in data and predict outcomes based on previous subjects. The figure below outlines only a subset of all available ML techniques.



Abbreviations: SVM – Support Vector Machine, CNN – Convolutional Neural Network, RNN – Recurrent Neural Network, LSTM – Long Short-Term Memory, GRU – Gated Recurrent Units, MLP – Multilayer Perceptron, RNTN – Recursive Neural Tensor Network GAN – General Adversarial Network, PCA – Principal Component Analysis, SOM – Self-Organizing Map, RBM – Restricted Boltzmann Machine, SARSA – State-Action-Reward-State-Action, DQN – Deep Q Network, A3C – Asynchronous Advantage Actor Critic

Figure 1: A subset of available ML techniques. [4]

2.1.1 Supervised Learning

There are distinctions to be made between different types of ML. The main type of learning we will consider is supervised learning, which makes use of pre-defined labels to classify and find patterns in data. In other words, unlike unsupervised learning, there are right and wrong answers for the model to consider for each subject fed to it. An example of supervised learning is feeding an algorithm a dataset filled with photos containing certain objects, along with the categories they belong to. ML would attempt to categorize future subjects using what was seen in previous examples. Not surprisingly, it is a complicated task that could go wrong. This was exactly the case when, in 2015, Google received widespread criticism for Google Photos erroneously labeling two Black individuals as gorillas.[5] It is these types of discrimination and unfairness we will continue discussing.

[4] Ackermann, Nils. “Artificial Intelligence Framework: A Visual Introduction to Machine Learning and AI.” Medium. Towards Data Science, December 15, 2018. https://towardsdatascience.com/artificial-intelligence-framework-a-visual-introduction-to-machine-learning-and-ai-d7e36b304f87.

[5] Pachal, Pete. “Google Photos Identified Two Black People as 'Gorillas'.” Mashable. Mashable, July 1, 2015. https://mashable.com/2015/07/01/google-photos-black-people-gorillas/?europe=true.

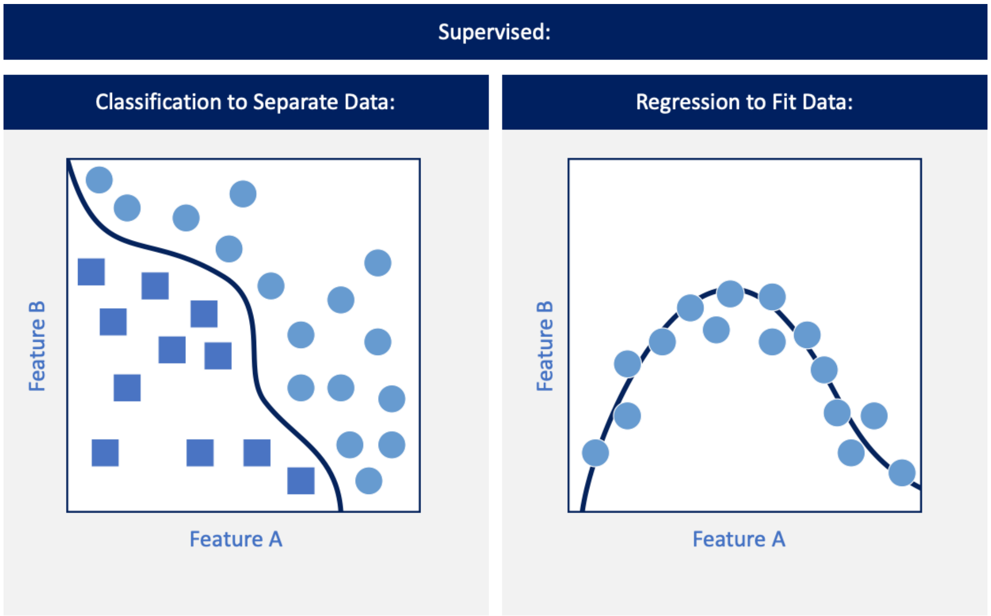


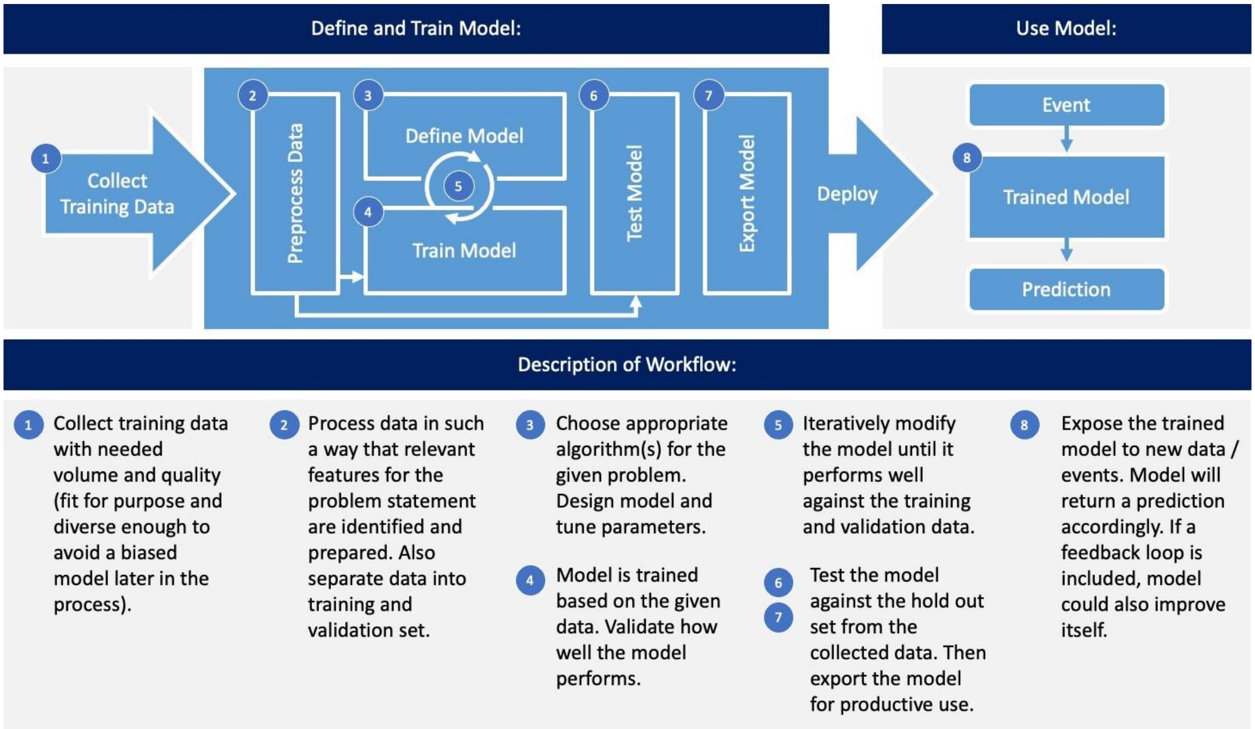
Figure 2: Classification and regression – the two types of supervised learning tasks. [6]

2.1.1.1 Tasks

The type of task performed in the above-mentioned example is called classification. It focuses on categorizing subjects into different classes based on how different they are from other classes. The other type of task is regression, which looks at similarities in the data that could help predict a certain numerical value. Predicting how likely prisoners are to recidivate by assigning risk scores to them is an example of a regression task. Worth noting is that regression tasks could be considered a type of classification, in that a numerical threshold could be assigned to decide on which category a certain numerical value belongs to (in the recidivism example these categories could regard the approval or denial of bail or parole). In other words, classification is discrete (categorical), whereas regression is continuous.[7]

[6] A Visual Introduction to Machine Learning.

[7] Brownlee, Jason. "Difference Between Classification and Regression in Machine Learning." Machine Learning Mastery, May 21, 2019. https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/.



1 Collect training data with needed volume and quality (fit for purpose and diverse enough to avoid a biased model later in the process).

2 Process data in such a way that relevant features for the problem statement are identified and prepared. Also separate data into training and validation set.

3 Choose appropriate algorithm(s) for the given problem. Design model and tune parameters.

4 Model is trained based on the given data. Validate how well the model performs.

5 Iteratively modify the model until it performs well against the training and validation data.

6 7 Test the model against the hold out set from the collected data. Then export the model for productive use.

8 Expose the trained model to new data / events. Model will return a prediction accordingly. If a feedback loop is included, model could also improve itself.

Figure 3: The process of ML.[8]

2.1.1.2 Methods and Black-Box Models

As seen in Figure 1, we use numerous approaches to complete classification and regression tasks.[9] Oftentimes, these tasks become complex to the extent that the models by which they are solved become hard to make sense of without additional tools. These models are referred to as black-box models due to their high complexity. We will specifically address these models later.[10] There are chiefly two types of complexity that contribute to black-box models – dimensionality and the type of ML technique.

2.1.1.2.1 Dimensionality

Dimensionality is the number of attributes (or factors) that are considered by a ML algorithm. For complex problems, such as predicting types of cancer in human beings based on gene expression, it is not uncommon to consider datasets with more than 1,000 different genes, with each of them being considered an independent attribute.[11]

In ML, it is commonly known that as the number of attributes considered increases, so does the difficulty of analyzing the data in question. This is often referred to as the curse of dimensionality due to the resulting complications, most of which regard the difficulty of assessing how the model in question arrives at a certain decision, but also the increased

[8] Visual Introduction to Machine Learning.
[9] See 2.1.
[10] See 4.4.3.
[11] Molla, Michael, Michael Waddell, David Page, and Jude Shavlik. “Using Machine Learning to Design and Interpret Gene-Expression Microarrays.” *AI Magazine on Bioinformatics*, 2004.

likelihood of the model reflecting the specifics of its training dataset and not performing as well on other datasets made for the task in question.[12]

2.1.1.2.2 Technique Type

Some ML techniques are obviously better for certain tasks and some techniques are more complex and ambiguous than others. Neural networks, for instance – the technique by which deep learning (a subset of ML) is deployed – are heavily used to display complex behavior in computers, as seen in AI.[13] They attempt to model the way our brains arrive at conclusions, which is a complicated and diverse process due to how little we know about the brain.[14]

These techniques are usually called explainable models, as they are not intuitive in the same way as other models and require measures (such as writing a summary program) to explain them.[15] Not all explainable models are supervised learning, although some are and will be covered in detail as we discuss the impact of black-box models on bias. [16]

2.2 Biases

Figure 3 describes the process of properly deploying ML. There is bias to consider in all mentioned steps, but they can be classified as two distinct types: dataset bias and algorithmic bias.

2.2.1 Dataset Bias

Everything a ML algorithm learns is based on input data, which means that if our data are flawed, we will never receive an acceptable output.[17] Each ML model must first be given a training dataset, which is used to help the model understand the relationships between all attributes in the data and the classifications given to each subject. If the dataset has biases, so will the model.

A commonly discussed application of ML, which has received much criticism for its uneven performance, is classifying gender in human beings. Gender Shades is a project, conducted by MIT Media Lab, which explores gender- and racial bias in well-known facial recognition algorithms. The project specifically chose to consider facial recognition software made by three companies: IBM, Microsoft and Megvii.

At first glance, the algorithms display high overall gender classification accuracies (between 94% and 87% for all three). However, when we consider the misclassified cases, biases become

[12] Tan, Pang-Ning, Anuj Karpatne, Vipin Kumar, and Michael Steinbach. *Introduction to Data Mining*. Harlow: Pearson, 2020.
[13] Ibid.
[14] "5 Unsolved Mysteries about the Brain." Allen Institute for Brain Science, March 14, 2019. https://alleninstitute.org/what-we-do/brain-science/news-press/articles/5-unsolved-mysteries-about-brain.
[15] Molnar, Christopher. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* GitHub, 2020.
[16] Mols, Bennie. "In Black Box Algorithms We Trust (or Do We?)." ACM, March 16, 2017. https://cacm.acm.org/news/214618-in-black-box-algorithms-we-trust-or-do-we/fulltext.
[17] Sharma, Dhruv. "Problems in Machine Learning Models? Check Your Data First." Medium. Towards Data Science, August 31, 2019. https://towardsdatascience.com/problems-in-machine-learning-models-check-your-data-first-f6c2c88c5ec2.

evident. With Microsoft's software, 93.6% of times faces were mistaken, the subject was Black. With Megvii's Face++ software, 95.9% of misclassifications were made when the subject was a woman. And, with IBM's software, the misclassification rate of darker women was 34.4% higher than that of lighter men.[18] These are not small margins.

A large reason for the bias seen in these models is the lack of proper subject distribution in common training datasets.[19] Minorities and women are usually less represented, which leads to ML algorithms overemphasizing White men. Joy Buolamwini, the founder of the Algorithmic Justice League and leader of the Gender Shades research team, noticed that in the IJB-A, a dataset commonly used in facial recognition by governmental bodies, 75.4% of all subjects were male, and at least 79% of the subjects had lighter skin.[20] ML algorithms learn what they see, which makes it no surprise that facial recognition currently performs best on light-skinned men.

To address the issue of dataset bias, we must consider what an optimized dataset looks like. There are general guidelines on how to develop those, which we will not fully outline. The gist is that the distribution of subjects by pre-defined group membership should be equal (or close to it) for all involved groups, to make sure that the ML model in question does not overemphasize certain portions and features of a dataset, as seen in our previous example. [21] [22]

2.2.2 Algorithmic Bias

Part of the ML process is choosing and designing a learning algorithm that takes input datasets and learns based on their contents.[23] Algorithmic bias is the bias that arises from discrimination in the learning algorithm and, eventually the resulting ML model. Algorithmic bias is a broader, more general term than dataset bias, in that other factors than the learning algorithm (including dataset bias itself) have a direct impact on algorithmic bias.[24] Regardless of how solid an algorithm is, it can only be as good as the dataset fed to it, and the choices made in defining it. These choices may entail deciding on which algorithm is best suited to complete the task in question, the weights assigned to the learning algorithm, the structure of the algorithm itself, or anything closely related. All of it has an impact on algorithmic bias.[25]

[18] Buolamwini, Joy. "Gender Shades." Gender Shades, 2018. http://gendershades.org/overview.html.

[19] Buolamwini, Joy A. "Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers." *MIT Media Lab, 2017,* 2017.

[20] Buolamwini, Joy. "Artificial Intelligence Has a Racial and Gender Bias Problem." Time. Time, February 7, 2019. https://time.com/5520558/artificial-intelligence-racial-gender-bias/.

[21] Torralba, Antonio, and Alexei A. Efros. "Unbiased Look at Dataset Bias." *Cvpr 2011*, 2011. https://doi.org/10.1109/cvpr.2011.5995347.

[22] Deeper Look at Dataset Bias (PDF)

[23] "Machine Learning Crash Course | Google Developers." Google. Google. Accessed March 29, 2020. https://developers.google.com/machine-learning/crash-course/.

[24] Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Aram Galstyan, and Kristina Lerman. "A Survey on Bias and Fairness in Machine Learning." *USC, Information Sciences Institute*, September 17, 2019. https://doi.org/10.1145/3341161.3342915.

[25] Hao, Karen. "This Is How AI Bias Really Happens-and Why It's so Hard to Fix." MIT Technology Review. MIT Technology Review, February 4, 2019. https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.

2.2.3 Hardware Bias

Not all bias is software- or data-related. For instance, in the case of Global Navigation Satellite System (GNSS) positioning (the most famous example of which is GPS), there are small deviations when reading phase changes that emerge from bias in hardware (specifically the satellite and receiver).[26] For this thesis, we will disregard hardware bias.

2.3 Cognitive Source of Dataset and Algorithmic Bias

Many of the problems regarding both dataset and algorithmic bias would be thoroughly mitigated with some theoretically basic solutions, and extensive research is being made on the details of these issues. A year after doing their first audit, the Gender Shades research team did a second audit on the same three companies and found that they all had significantly improved the performance of their models. Microsoft, for instance, which had a 20.8% difference in performance between darker females and lighter males, brought that same number down to 1.5% a year later. Face++ and IBM also made significant improvements, although the most disadvantaged in the context were still female subjects and those with darker skin. In terms of bias directly related to algorithms, improvements are also being made. Recent research has found a way to mitigate bias in the algorithms themselves, even with poor datasets. This is done by means of latent structure analysis, which allows for the importance of some attributes in training datasets to be redistributed and for bias to be reduced as a result.[27]

Nevertheless, as we witness improvements, some of the major problems remain. In the second audit, Gender Shades also investigated two newer facial recognition models, made by Kairos and Amazon respectively. Amazon clearly had the worst performance of all audited models, with a 31.4% performance difference between darker females and lighter males; almost as underwhelming as IBM in 2018.[28]

Despite the evident biases, Amazon found it convenient to pitch their software to the Immigration and Customs Enforcement (ICE) for use in their operations. This is software that would more than likely misidentify people of color frequently, leading to unjust arrests. Either Amazon was unaware of these biases (which would imply serious negligence), or the company simply does not care. Our current societies and economies incentivize the release of unethical products due to their profitability, with entities rarely being held accountable. There are

[26] Håkansson, Martin. "Hardware biases and their impact on GNSS positioning.", 2017.
See also: "What Is GNSS?" European Global Navigation Satellite Systems Agency, August 29, 2017. https://www.gsa.europa.eu/european-gnss/what-gnss.

[27] Amini, Alexander, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019. https://doi.org/10.1145/3306618.3314243.

[28] Hao, Karen. "Making Face Recognition Less Biased Doesn't Make It Less Scary." MIT Technology Review. MIT Technology Review, February 15, 2019. https://www.technologyreview.com/s/612846/making-face-recognition-less-biased-doesnt-make-it-less-scary/.

numerous examples, with Cathy O'Neil, Virginia Eubanks and others having written books containing many of them (some of which will be covered briefly in this work).[29] [30]

This brings us to the fundamental problem of bias in supervised ML: cognitive biases.[31]Although dataset and algorithmic biases are concerns, they directly derive from our implicit biases. These become apparent in how much we expect from the ML model. How much can it discriminate and still be acceptable (or fair)? How lenient can we be in our requirements on data collection? How much empirical research do we need before including a feature in our dataset? How confident are we that the chosen learning algorithm is the best suited one? Do we understand the algorithm well enough to know how it arrives at decisions? And lastly, the question that we hope to answer: For which contexts and with which solutions is ML warranted at all? These are only some of the questions that should be answered every time ML is deployed and, oftentimes, we fail to answer all of them correctly (we will find that sometimes there is no "correct" answer), which leads to severe consequences.

No matter how hard we try to erase these biases, they will remain present. Usually, we do not recognize our implicit biases and their significant impacts on outcome. A prime example came up in Amazon's attempt to make a program that would aid in recruiting. It was quickly detected that the program was biased against women, which was assumed to be due to women being directly used as an assessment variable, or feature, in the program. However, when group membership became an excluded feature, the program was still discriminating against women. After extensive investigation it was discovered that the program was, unbeknownst to the developers, seeing a pattern in resume word usage among men and women respectively, which made the program favor men's resumes. The training set used consisted of resumes from successful Amazon employees, most of whom have been male in the past, which skewed the results.[32]

Initially this seems like a farfetched bias and stereotype but it has empirical evidence behind it.[33] Out of the many factors impacting a complex model, the first thought is rarely to assume that word usage is the main contributing factor to bias. It is evident that when faced with multiple factors (attributes) to consider (in some cases thousands), humans will be unable to account for all of them. For this and other reasons, the thesis will largely be focusing on the human limitations that come with ML. The limitations of our programs start and end with us. If our

[29] Peterson, Andrea, and Jake Laperruque. "Amazon Pushes ICE to Buy Its Face Recognition Surveillance Tech." The Daily Beast. The Daily Beast Company, October 23, 2018. https://www.thedailybeast.com/amazon-pushes-ice-to-buy-its-face-recognition-surveillance-tech.

[30] Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: Picador, St. Martins Press, 2019.

[31] Turner-Lee, Nicol, Paul Resnick, and Genie Barton. "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms." Brookings. Brookings, October 25, 2019. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

[32] Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." Reuters. Thomson Reuters, October 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[33] Jackson, Abby. "12 Words That Are More Familiar to Women than Men." Business Insider. Business Insider, March 24, 2017. https://www.businessinsider.com/gender-and-vocabulary-analysis-women-2017-3?r=US&IR=T.

limitations and flaws are accounted for, doing the same for our ML models will become a simpler task.

In assessing biases, we will mainly be assessing the output, except when discussing the interpretability of an algorithm.[34]Long-term, it is the output of a certain model that defines whether the dataset and algorithm were appropriately chosen and deployed. It is by considering the output that we can properly assess whether the input data and/or ML technique must be improved. This, of course, disregards how understandable the model is to the general public (and prioritizes performance ahead of interpretability) but allows us to see how often the models in question get things right, regardless of how they arrive at decisions.

2.4 Inductive and Deductive Reasoning

We have many ways of classifying the knowledge we receive from ML applications. However, conclusions reached based on output from ML algorithms are oftentimes defined by either their predictive or descriptive nature.

Predictive tasks have, as indicated by the name, the goal of predicting outcomes or characteristics in future subjects, based on trends and data seen in previous subjects. In more technical terms, these tasks assume dependence of certain attributes in datasets on other independent attributes and try to make predictions based on the given relationships.[35] A more well-known application is using historic stock market trends to try to predict future outcomes.[36]

Descriptive tasks, on the other hand, try to highlight trends and patterns in data for various useful purposes. Different statistical measures, such as clustering or association rule mining, consider what attributes have in common and set them apart, and find appropriate ways to present relevant conclusions.[37]

Although there is much to separate these types of tasks, they have one trait in common: the construction of all models related to descriptive and predictive ML tasks are based on inductive reasoning. Inductive, or bottom-up reasoning, is the concept of using statements and instances of different sorts to provide evidence in support of a certain claim (as opposed to deductive reasoning, where we make use of informational input to reach, by way of logical discourse and while assuming the truth of the input, fully true conclusions).[38] We then apply our inductively reasoned models to deduce conclusions about future subjects, or to find patterns within data.

Both inductive and deductive reasoning have value, but the former can be misinterpreted, whereas the latter cannot. Deductive reasoning, if done correctly, leads to what is called necessarily true conclusions, which are true regardless of circumstance, so long as the logical foundation of the issue in question (as well as the assumptions that accompany them) does not

[34] See 4.4.

[35] ”Features” and ”attributes” will be used interchangeably in this work.

[36] Kompella, Subhadra, and Kalyana Chakravarthy Chilukuri. “Stock Market Prediction Using Machine Learning Methods.” *International Journal Of Computer Engineering And Technology* 10, no. 3 (2019). https://doi.org/10.34218/ijcet.10.3.2019.003.

[37] *Introduction to Data Mining*.

[38] Bradford, Alina. “Deductive Reasoning vs. Inductive Reasoning.” LiveScience. Purch, July 25, 2017. https://www.livescience.com/21569-deduction-vs-induction.html.

change. Inductive reasoning, however, only provides evidence for or against a claim, without making it necessarily true. The best inductive reasoning can do is to make certain claims more (or less) probable.

2.4.1 The Problem of Induction

It is here that the problem of induction comes into play. English philosopher C.D Broad famously stated that "induction is the glory of science and the scandal of philosophy."[39]

Posed by many prominent philosophers (including David Hume), the main criticism of inductive reasoning is that it can never provide epistemic certainty.[40] There will always be parts of the inductively reached conclusions that remain incomplete. For this reason, since the deductive reasoning applied by ML models is based on inductive reasoning, there will be unaccommodated gaps of knowledge in our conclusions.

The problem of induction targets descriptive and predictive ML tasks in different ways. For descriptive tasks, it criticizes that generalization based on patterns seen in data never will provide pure knowledge (since we can never assume full truth of the input, or data), even in cases where all subjects point to the same classification (which would be considered the best-case scenario). For predictive tasks, the criticism is directed toward the impossibility of accurately predicting all future outcomes based on previous ones.

In ML, we rarely encounter best-case scenarios. If we have information that assertively suggests a generalization or future outcome, we will likely not use ML to confirm it. We would, for instance, not need ML to tell us that the probability of each outcome on a regular, fair-weighted die is one-sixth. Although various shapes and designs will have slightly different outcomes (engraved dots, for instance, may change the weight distribution of a die), we neglect these attributes due to their lack of relevance in the larger scheme of things. Since trial and error has given us no reason to think otherwise, we have, through inductive reasoning, concluded that in the case of a six-sided die with (relatively) even weight distribution, the respective probabilities will remain constant. ML is applied to cases in which the answers are not as obvious.

Due to this problem of induction, we cannot generalize and, with full certainty, make necessarily correct assumptions about outcomes for future subjects, or the attributes by which they are assessed. In more technical terms, the answers we get from ML are of a probabilistic, non-deterministic nature. It is thus unreasonable to expect perfection from our ML models and applications. Imperfection is therefore not part of the problem for bias and should not be perceived as such. It is the way imperfection is dealt with that becomes the main issue.

2.5 Justice and Fairness in ML

Theoretically, the problem of induction suggests that due to the inductive reasoning of ML model development, we will find ourselves in situations where not everyone is treated equally by the resulting models. This theoretical discrimination will, however, also have a practical impact

[39] Broad, C. D. *The Philosophy of Francis Bacon: an Address Delivered at Cambridge on the Occasion of the Bacon Tercentenary, 5 October 1926*. New York: Octagon Books, 1976.

[40] Hume, David, and Tom L. Beauchamp. *An Enquiry Concerning Human Understanding: a Critical Edition*. Oxford: Clarendon Press, 2009.

on people’s lives, which we have seen in previous examples and seek to explore further. In light of this issue, the interpretation of what is fair and just becomes relevant. Much of the general discussion regarding justice is about how we establish equity, which warrants those same discussions taking place in the context of ML. But, before we proceed, we must highlight the distinction between justice and fairness.

These terms are often used interchangeably, although doing so has created confusion in the past, as it makes the task of distinguishing between moral obligations, versus considerations, difficult. Justice is the moral concept of what is right. Fairness, on the other hand, is the personal evaluation of justice. As Goldman and Cropanzano state, "“Justice” denotes the conduct that is morally required, whereas “fairness” denotes an evaluative judgment as to whether this conduct is morally praiseworthy.” [41] In this sense, justice is a concept independent of fairness, in that it can exist without satisfying fairness. This does not, however, necessarily mean that we have access to this definition of justice, but even if we did, it would be difficult to attain due to our biased evaluations. The question then becomes to what extent we can accommodate fairness, while having solid considerations for ML justice. We will clarify the reason(s) for this specific approach in the coming sections.[42]

2.6 Full Accommodation of Fairness in ML

In situations of complete consensus, we seem to be less concerned with investigating whether an outcome is fair. In these instances, we assume that since every individual in question has the same definition of fairness, that the outcome indeed must be fair. Unfortunately, this best-case scenario is almost never a reality – especially when we have many people with different backgrounds involved (which oftentimes applies to ML). The question then becomes: “In situations of disagreement, is there a way for ML to accommodate everyone’s fairness definitions?” One can argue that some fairness definitions are bad and that some are good, but completing such assessments would require one to make assumptions beyond what we know to be necessarily true. For that reason, taking an approach which does not intend to accommodate all definitions of fairness is selective without justification.

Nonetheless, the answer to this question has its foundation in whether one finds fairness subjective or objective. If we find that fairness is objective, we should only look to account for what is objectively fair and just. However, by the definition used in 2.5, fairness is the personal evaluation of justice, which is a subjective concept. This means, in turn, that any disagreement on what fair course of action is would imply that fairness is not fully accommodated in that instance (we believe most people encounter such situations frequently). This would consequently apply to ML, as any approach to accommodate fairness (or any disagreement with such approach) would only be a personal evaluation of what is considered just, and not necessarily what is actually just. Therefore, regardless of situation and how close to perfect an application is, there will be instances of unfairness due to us having different definitions of what is fair. In other

[41] Goldman, Barry, and Russell Cropanzano. “‘Justice’ and ‘Fairness’ Are Not the Same Thing.” *Journal of Organizational Behavior* 36, no. 2 (2014): 313–18. https://doi.org/10.1002/job.1956.

[42] Especially 4., where we address the extent to which justice and fairness could be the same thing.

words – no, we cannot accommodate everyone with one singular solution – with or without ML models. This should, however, not stop us from trying to optimize our conditions.

When we discuss fairness, we often discuss equality. It may sound like a simple measure of fairness, and sometimes, that may be true. But, what do we exactly mean by equality? Is it the equality of goods? The equality of opportunity? The equality of treatment? Ronald Dworkin explored this topic in his two essays, "The Equality of Welfare" and "The Equality of Resources". He discovered that equality oftentimes is not what we are looking for. Equality, just for the sake of it, is counterproductive and unethical.[43] Assuming limited resources, what point is there in providing insulin for someone who does not have diabetes?

Although it is much more difficult to define, as it requires us to consider the diversity of perspective and needs, the concept of equity has taken precedence over equality. For our purposes, we will define equity as "giving everyone what they need to be successful" moving forward. [44]

Sometimes, equity and some type of equality mean the same thing. Sometimes, giving everyone what they need to be successful implies treating everyone the same, or giving everyone the same opportunity. When we discuss access to human rights, this is often the case. The Universal Declaration of Human Rights, for instance, states that "the inherent dignity and […] the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world."[45] However, for situations in which equality is not equity, there are disagreements on what decisions would be considered fair, which most certainly becomes relevant in the application of ML. All these concerns will help set the foundation for justice in our framework of ML implementation considerations.

2.7 Problem Statement

In the previous sections, we have discussed the presence of bias, the nature of it, and the reasons as to why it inevitably impacts fairness and justice in our societies. In our attempts to address these problems, we often ask how and when we can ethically use ML. However, we rarely ask whether it is ethically justified to implement ML models to complete tasks in certain contexts. The questions of "how" and "when" to ethically use a certain application of ML are dependent on the existence of an affirmative answer to the question of "whether".

Some tasks cannot be ethically completed in certain contexts, which will become evident in some of the examples mentioned below.[46] Another problem is that we oftentimes focus on the performance of a task, rather than the context in which a task is being performed. We tend to

[43] Rothman, Joshua. "The Equality Conundrum." The New Yorker. The New Yorker, January 21, 2020. https://www.newyorker.com/magazine/2020/01/13/the-equality-conundrum.

[44] Sun, Amy. "Equality Is Not Enough: What the Classroom Has Taught Me About Justice." Everyday Feminism, September 25, 2014. https://everydayfeminism.com/2014/09/equality-is-not-enough/.

[45] "Universal Declaration of Human Rights." United Nations. United Nations. Accessed March 30, 2020. https://www.un.org/en/universal-declaration-human-rights/.

[46] Chamorro-Premuzic, Tomas. "Four Unethical Uses Of AI In Recruitment." Forbes. Forbes Magazine, May 30, 2018. https://www.forbes.com/sites/tomaspremuzic/2018/05/27/four-unethical-uses-of-ai-in-recruitment/#6baa3dac15f5.

assume that the implementation is wrong and that it could be remedied. This is not always the case.

To address these issues, we have tasked ourselves with developing a framework that will consider the current problems of fairness and justice in ML, along with the evident technological and cognitive bias limitations, to give recommendations on whether implementation of ML is warranted in a given context, with a given ML model. In doing so, the framework also intends to set expectation standards for ML, both pre- and post-implementation. We hope that this will encourage more self-awareness in programmers and users of ML technology about the impact of their actions, but also optimism about the potential for societal and technological improvement.

The framework in question will be assuming liberalism as deployed in most Western countries as the pursued virtue. This means that values such as freedom of choice, separateness of persons and theoretical impartiality will be commonly encountered themes. This is a reasonable approach for a general framework, as the theoretical ambition should not be set any lower than full accommodation for all individuals. Now, with a set foundation of background knowledge, we consider the current situation before commencing the development of the framework.

3 Current State of the Field

3.1 Fairness Verification

As the problem of fairness in ML has been increasingly highlighted publicly, measures have been taken to actively assess how well applications account for fairness. At the University of Wisconsin – Madison, a team of researchers have worked on a bias verification program called FairSquare, which is a mathematical approach to assessing bias presence. The gist of the practice is to assume a certain fairness criterion, defined by mathematical constraints, and check to see whether a program is successfully fulfilling that criterion.[47]Assuming that the fairness criteria are "correct", this is a truly helpful practice to ensure less biased software use. Nevertheless, there are many situations in which we are not certain about the "right" definition of fairness; likely because our definitions almost never are fully fair, and much less so when we attempt to model them mathematically (which is the case with ML). In the coming paragraphs, we will outline the problems with our current definitions of fairness from most to least critical.

3.2 Non-Empirical Feedback Loops

A common problem in ML regards the features by which a ML model attempts to complete a task or answer a question. It is common for governmental entities to occasionally include attributes in their datasets that have no empirically shown correlation with the attempted classification. The TSA has, for instance, received criticism for discriminating against Muslim subjects by using biased attributes.[48] [49]

In the worst scenarios, we take those wrongfully included features and draw erroneous conclusions about what their contribution to a certain result should tell us. We accept results without confirming their proper function. In Washington, D.C during the 2008 recession, for instance, the city administration started assessing teacher performance using a ML program called IMPACT. This program was to make decisions on which teachers to hold responsible (and release) for poor learning results among students in the district. The program was responsible for 50 percent of the assessment for each subject. Although the program's features were not revealed to the public, the results surprised many, as some of the top-rated teachers in the district (as judged by administrators and parents) received the lowest performance scores overall and were fired as a result. It became evident that one of the higher weighted features of IMPACT was changes in standardized test scores for students, from year to year. USA Today later also discovered that there was a significant level of erasures on the standardized tests assessed at many of the schools in the district (almost a fourth of them, specifically). This implied cheating

[47] Albarghouthi, Aws, Loris Dantoni, Samuel Drews, and Aditya V. Nori. "FairSquare: Probabilistic Verification of Program Fairness." *Proceedings of the ACM on Programming Languages* 1, no. OOPSLA (December 2017): 1–30. https://doi.org/10.1145/3133904.

[48] Handeyside, Hugh. "New Documents Show This TSA Program Blamed for Profiling Is Unscientific and Unreliable - But Still It Continues." American Civil Liberties Union, April 22, 2019. https://www.aclu.org/blog/national-security/discriminatory-profiling/new-documents-show-tsa-program-blamed-profiling.

[49] Gillum, Jack, and Marisol Bello. "When Standardized Test Scores Soared in D.C., Were the Gains Real?" USA Today. Gannett Satellite Information Network, March 30, 2011. http://usatoday30.usatoday.com/news/education/2011-03-28-1Aschooltesting28_CV_N.htm.

by students, and possibly teachers allowing them to check their own answers. In the investigations, teachers were not asked if they contributed to this high level of erasure (specifically wrong-to-right answers). [50] And, the teachers who were fired never received explanations as to how IMPACT arrived at their low scores.

IMPACT assumed that better grades implies more learning, which there is no general empirical evidence for (it may apply in certain schools, but not all). This is what O'Neil refers to as a "WMD [Weapon of Math Destruction] feedback loop", which accepts a ML model arriving at the right answer without having empirical backing. O'Neil discusses a more general example:

> Employers, for example, are increasingly using credit scores to evaluate potential hires. Those who pay their bills promptly, the thinking goes, are more likely to show up to work on time and follow the rules. In fact, there are plenty of responsible people and good workers who suffer misfortune and see their credit scores fall. But the belief that bad credit correlates with bad job performance leaves those with low scores less likely to find work. Joblessness pushes them toward poverty, which further worsens their scores, making it even harder for them to land a job. It's a downward spiral. And employers never learn how many good employees they've missed out on by focusing on credit scores.[51]

Usually, we assess the success of a classification task by considering misclassification rates. These answer the question: "A posteriori, how often did the model get things right?" In the examples seen above, this question cannot be answered as there is no misclassification rate to consider. We do not know how many times we got things wrong, which acts as positive reinforcement for the developers and users; all they know is that it gets some job done.

3.3 Current Fairness Criteria

There are three approaches currently used with intention to maximize fairness in our ML models. These are anti-classification, classification parity and calibration. Corbett-Davies and Goel (2018) did a comprehensive analysis of the potential flaws that each approach may have.[52] We will analyze the effectiveness of these using the example of a program called COMPAS, developed in 2016, which assigns inmates recidivism risk scores for sentencing by considering more than 100 different factors that supposedly play a role in an inmate's probability of reoffending. Its outcomes were passionately discussed after its implementation, which has made for increased statistical availability.

3.3.1 Anti-Classification

Anti-classification disregards which pre-defined group (such as ethnicity or sex) one belongs to, with the intention of not having that impact the result of the model. In other words, pre-defined group membership is not considered a feature of the developed ML model. However, although

[50] Turque, Bill. "'Creative ... Motivating' and Fired." The Washington Post. WP Company, March 6, 2012. https://www.washingtonpost.com/local/education/creative--motivating-and-fired/2012/02/04/gIQAwzZpvR_story.html.

[51] O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin Books, 2018.

[52] Corbett-Davies, Sam, and Sharad Goel. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *Stanford University, 2018*, 2018.

this approach has good intentions, anti-classification could have an adverse effect if classifications are disregarded.

For instance, results show that it is inevitable that COMPAS will discriminate against women if anti-classification is enforced. Without considering the classification of the subjects in question, women are assigned significantly higher risk scores for the same subject recidivism rates. In other words, this implies that women who do not recommit crimes are given harder treatment by the ML model (and – in turn – the courts) if the assessment does not directly consider the subject's gender. This is due to implicit biases of our data inevitably leading to our learning algorithms catching patterns indirectly related to group membership, changing the treatment of certain subjects accordingly.

Usually, we do not recognize our implicit biases and their significant impacts on outcome. This is the danger with anti-classification. The aforementioned Amazon recruitment system is enough evidence that anti-classification does not mean anti-discrimination.[53]

3.3.2 Classification Parity

Classification parity is the approach that intentionally regards group membership and ensures that the evaluation measure for each subject's performance is equalized across the relevant groups. This uses confusion matrices – a concept commonly used in ML to determine how often a model classifies correctly – for each pre-defined group classification to assess how well the ML model in question works (usually by considering metrics such as precision and false positive rates) and to equalize in accordance.

The problem with classification parity is that our measures of risk distribution will be different for every relevant group classification. In that regard, if we make decisions on relevance based on a single threshold for all groups, we are not exercising classification parity (as there will be disparity in how each group is treated, based on their respective risk distribution). The solution is to assign a different threshold for each group that accounts for the relative risk distributions of each classification. When this approach is taken, however, we see large differences in optimal thresholds for different demographics, meaning that classification parity also could result in some demographics being significantly more roughly treated than others. For COMPAS, regardless of whether one wanted to minimize recidivism or achieve most equal treatment across all relevant groups, this resulted in a 16-17% optimal threshold for Black subjects (meaning that less than a fifth of Black subjects would have high enough risk scores to be detained), compared to 31% for White subjects – a significant difference (which will be discussed briefly later).[54]

With classification parity, it is also difficult to know whether the problem is actual discrimination or underlying problems that lead to the impression of discrimination. In some situations, the possibility exists that the risk distribution that results for a certain group is close to true, but that vast disparities give us reason to believe that the distribution is biased.

[53] See 2.3.

[54] "Critical Review of Fair Machine Learning."

Classification parity disregards this possibility by reasoning that bias must always be accounted for by equating subjects by group membership, when this is not always the case.

3.3.3 Calibration

Calibration is a tool used to ensure that the predictions of a certain model align well with actual outcomes (and that if they do not, that the model gets "calibrated" accordingly). In other words, it sets a standard for what the risk score in question is supposed to mean across all subjects. Thus, calibration implies that the same predicted classifications for two different groups should mean the same (or very similar) thing in real life and modifies the model until that is the case. Calibration for COMPAS, for instance, would mean that the same risk score for two subjects of different groups would imply the same recidivism rates for those respective groups.

One problem with calibration is its susceptibility to direct discrimination.[55] Redlining (that is, the systematic approach taken by governmental entities to limit resources for minority communities), for instance, has been heavily associated with calibration, in that it allows the modification of each subject's risk score to gather around a distribution that is less beneficial for a certain group.[56] The flexibility provided by calibration also oftentimes implies that standards are a non-necessity in the context, which defeats the point of one of our current objectives – to set a standard for what fairness looks like in ML.

3.4 Mutual Exclusivity of Fairness Criteria

As pointed out, all currently used mathematical definitions of fairness have their respective biases. One may then wonder about the possibilities of combining the strengths of all three measures to account for all their flaws. It turns out that this is oftentimes impossible.

There are instances in which two or more different criteria, which all cover legitimate concerns and most likely should contribute to the scenario's holistic definition of fairness, contradict one another mathematically. The reason COMPAS became a topic of discussion was because of the disagreements in what was considered fair.

Race was not one of the features, meaning that the program developer, Northpointe, was exercising anti-classification with COMPAS. On the one hand, Northpointe correctly argued that COMPAS was accurate in its assessment ("defendants assigned the highest risk score reoffended at almost four times the rate as those assigned the lowest score"), and that the program achieved this without considering race. On the other hand, news organization ProPublica accurately claimed that the program treated Black people who did not recommit crimes rougher than White people in the same category (on average, Black people in this category were given a risk score twice as high).[57]

[55] Ibid.

[56] Yu, Joe. "Assessing Fairness in COMPAS: Impossibility Theory, Calibration, and Redlining." Medium. Medium, March 19, 2019. https://medium.com/@qy002/assessing-fairness-in-compas-impossibility-theory-calibration-and-redlining-2d32a8d6f9af.

[57] Corbett-Davies, Sam, Emma Pierson, Avi Feller, and Sharad Goel. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear." The Washington Post. WP Company, October 17, 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

We could argue that both Northpointe and ProPublica’s perspectives should be part of the holistic fairness definition and be included in the program’s assessment. Race should, by our moral measures, likely not be an attribute of the program, and Black people who do not reoffend should not be considered higher risk than White people in the same category. However, based on a mathematical analysis of the program, COMPAS would be unable to achieve the constraints set out by both companies. Fulfilling one fairness criterion would require us to exclude the other fairness criterion, making it mathematically impossible to achieve mutual inclusivity of all fairness criteria in the case of COMPAS. It becomes evident that if this is a reality for one ML model, that it is a possibility for others as well. This example points out the shortcomings of mathematical models in achieving what we would consider fully fair outcomes.

3.5 Theoretical Versus Empirical Observations

Theoretically, the problem of induction already implied that ML would not be able to achieve full accommodation of fairness.[58] Now, we have seen empirical evidence in support of that claim. No matter how hard we try, perfection will never exist in ML, and for that reason, it should never be the expectation. Our task is to develop a framework that accounts for both theoretical and practical limitations and sets a standard for the expectations that should be placed on ML models and tasks we hope to solve.

[58] See 2.6.

4 Development of Framework

We have now established the difficulties of fully accommodating fairness, and thus establishing equity. Thus, it is of importance to emphasize solutions that aim to maximize the parts of fairness we define (or should define) similarly (we will elaborate on this in 4.1).

For proper application of the framework, three input objects must be defined:

1. The task to be solved
2. The ML model by which the task is to be completed
3. The implementation context of the task

The considerations of the framework will not be discussed in the order that they are addressed in the framework itself; that order is established for our convenience in our condensed flowchart representation.[59] As seen in above-mentioned examples, the definition of optimal distribution (for establishing equity) is the main question that remains unanswered in supervised ML tasks. We will begin by addressing that question, and see if our answer resembles any of the previously outlined criteria of fairness currently used in ML.

4.1 Distributive Justice

The question of optimal distribution raises the concept of distributive justice, which is the branch of justice that concerns itself with the just distribution of goods – both tangible (products/materials) and non-tangible (services).

The problem in ML is often the distribution of mistakes. For reasons that require no explanation, the discrimination against women and dark skin as seen in Gender Shades is a prime example of distributive injustice in ML.[60] The question then becomes: "What would constitute distributive justice in this and any other supervised learning context?" John Rawls can help us set a foundation.

4.1.1 John Rawls' Justice as Fairness

Since fairness looks different to every individual, based on their respective nature and nurture, we want to consider philosophical theories that account for the separateness and biases of individuals, and allow us to set an objective standard for distributive justice. We also want to explore the extent to which we can find consensus despite the biases we have regarding fairness. Rawls recognized the presence of human bias and thought of ways to theoretically mitigate its effects on our definitions of justice.

In *A Theory of Justice* (1972), John Rawls introduces theories on what distributive justice should look like to a reasonable human being. In other words, Rawls describes the foundation of distributive justice, by outlining what everyone, regardless of personal preferences and definitions of fairness, can (or should) agree on. The conclusions he reaches are independent of

[59] See 4.6.

[60] See 2.2.1

virtue, or the common good – they are focused on justice. He calls the theory "Justice as Fairness."[61]

In his work, Rawls famously describes the concept of the original position; a thought experiment that assumes what is called a veil of ignorance. A veil of ignorance is the concept of removing the partial human characteristics of the individuals/parties in question, thus rendering decision making as impartial as possible. Rawls makes the case that this theoretical impartiality will aid in establishing the foundation of justice, since one would make decisions knowing nothing about oneself or other individuals. In summary, the original position is the one all people would (theoretically) take, had they not known anything about themselves or others, and should therefore be used as the foundation of distributive justice.

In exploring this position of supposed complete impartiality, Rawls comes up with two basic principles of justice that he considers applicable, regardless of one's biased stance on fairness. In his 1985 book *Political Liberalism*, where he refines some of the arguments made in *A Theory of Justice*, he describes these as follows:

a. Each person has an equal claim to a fully adequate scheme of equal basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value **[liberty principle]**.

b. Social and economic inequalities are to satisfy two conditions: first, they are to be attached to positions and offices open to all under conditions of fair equality of opportunity **[fair equality of opportunity principle]**; and second, they are to be to the greatest benefit of the least advantaged members of society **[difference principle]**.

Rawls makes a strong case for all, claiming that from the original position, these are conditions all reasonable individuals would want for themselves and others. He also notes that a) takes precedence over b), should these two interfere.[62]

Rawls claims that all reasonable individuals, regardless of their backgrounds and outlooks on the common good, would be inclined to agree with the abovementioned principles, due to what he calls overlapping consensus. He defines a reasonable individual as someone who is willing to work with other individuals in their respective societies to reach mutual agreements that fairly accommodate all.[63]

His ideas of overlapping consensus are derived from Immanuel Kant's concept of the categorical imperative, which claims that reason within each individual allows us, by logical discourse, to arrive at some principles of morality that are applicable to all, regardless of one's views in regard to everything else.[64] Rawls makes the case that his outlined argument in "Justice as Fairness" is as far as we can go in establishing the categorical imperative from a societal standpoint.[65]

[61] Rawls, John. *A Theory of Justice*. Cambridge, MA: Harvard Univ. Pr., 1980.
[62] Rawls, John. *Political Liberalism*. New York: Columbia University Press, 1985.
[63] Ibid.
[64] Kant, Immanuel. *Kant: Groundwork of the Metaphysics of Morals*. Provo, UT: Renaissance Classics, 2012.
[65] Rawls, John. "Justice as Fairness." Cambridge, MA: Harvard Univ. Pr., 1971.

In establishing the difference principle, Rawls recognizes the distinction between equality and equity, and that not all individuals from the beginning of time recognize unconditional equality as the best definition of fairness. This is also part of Rawls' Pareto argument for inequality, which makes the claim that inequality is virtue if it benefits everyone in the context.[66] He points out that sometimes, regardless of how conscientious we are in accommodating fairness, we will have inequalities. In those cases, he claims that such inequalities should favor "the least advantaged."[67]

Rawls' theories were mainly intended for optimizing societal structures to accommodate endeavors of all individuals. Nevertheless, the goods produced by our societies that directly affect individuals should also reflect the values by which these societal structures are established. In other words, if we have goods produced by the market that limit enforcements of the "basic structures of society", that is a problem. For that reason, the same societal structures developed by Rawls are also applicable to ML and its respective operations.

4.1.1.1 Choice of Fairness Criteria

From this, one could assume that Rawls likely would prefer the approach of classification parity to the other two, as the inequality in question will often favor what we consider to be the least advantaged group (with COMPAS for instance, we could strongly argue that Black people are less advantaged in Western societies than White people).[68] The important question to answer, regardless of situation, is if the solution in question favors what we consider to be the least advantaged in the context of the issue.

One major weakness of Rawls' discourse: he never confidently defines the least advantaged. This is especially a weakness since we would need to properly identify the least advantaged for classification parity to fulfill the requirements of Rawls' justice principles. He does describe the potential traits of a relatively disadvantaged person (most of which focused on social and economic class differences) but does not go onto introducing a formal definition. In a later paper, he changes his description of the least advantaged individuals, and characterizes them by how undeserving people are of the position they are in. His description begs the question: "Is this individual better or worse off, based on the initial conditions that they were given?"[69] There are weaknesses in both descriptions, but moving forward, we will assume that the definition of "the least advantaged" is sufficiently accounted for and that the hypothetical context of the ML model implementation makes it possible to accurately define the least advantaged group of individuals.

One can also argue the extent to which inequality between groups can be considered acceptable, even in the case of favoring the least advantaged (and thus abiding by Rawls' difference principle). With classification parity, both commonly used measures of parity (demographic parity and utility-maximizing equalization of false positives) resulted in a 16-17% threshold for

[66] Fisher, A.R.J, and E.F McClennen. "The Pareto Argument for Inequality Revisited.", 2011.
[67] We will address some concerns with this wording in 4.1.1.1.
[68] See 3.3.2.
[69] Weatherford, Roy C. "Discussions Defining the Least Advantaged." *Equality and Liberty*, 1991, 37–45. https://doi.org/10.1007/978-1-349-21763-2_4.

Black people and 31% for White people, meaning that the treatment of White people would be almost twice as rough.[70] Whether this is acceptable should be seriously discussed.

Finally, the problem previously mentioned in the outlining of classification parity persists.[71] Sometimes, in a particular context, people may not actually be discriminated against. It may be simply the reality of the situation. Rawls would likely argue that we may have been led to this reality due to discrimination in other contexts of life. For instance, if we were to assume that significantly more crimes are being committed in poorer neighborhoods, the fact that some individuals are poor sometimes warrants no other action. Rawls would claim that we counteract the "negative" discriminations in our societies by enforcing "positive" discriminations where it is not possible to have equality of opportunity.

4.1.1.2 Other Concerns

Some will make the case that the difference principle is utilitarian. In *A Theory of Justice,* Rawls specifies in detail how his philosophy differs from utilitarianism. An important part of the difference principle is that it was not suggested with the intention to maximize utility; it was suggested because it was logically derived from the original position, and still takes into account the separateness of persons (something utilitarianism does not). G.A Cohen, an analytical Marxist, provided an extensive critique of the same principle, although there are other philosophers who claim that the difference between the two is not their stances on egalitarianism, but more the way they seek to achieve it.[72]

It becomes evident that other people may arrive at different conclusions than Rawls – even when considering the original position – which confirms that bias is not fully removed even with a veil of ignorance. Some of Rawls' conclusions, for instance, were reached while implying that humans will always put their own needs and desires first (which, if we consider the original position, he could not have known). Nevertheless, he is one of few to have fully expressed his thoughts and attempted to apply deontology, to its possible extent, to a consequentialist reality. For this reason, for a first attempt, Rawls' principles seemed to be the best by which to develop the distributive justice considerations of this framework. This does not ever mean that his (or our) framework is free from flaws.

Knowing that this definition of distributive justice is not carved in stone, it is essential that John Rawls' ideas are treated as merely a starting point for distributive justice. His approach is deontological, which makes for a difficult practical application, but gives an idea of what we theoretically should consider fairness. His claims are basic, since that is only as far as one can go without making further assumptions about the specific task and model at hand. Nevertheless, those hypotheticals will always exist, which requires us to deal with them. Therefore, this is a framework of considerations and not criteria (as it is immensely difficult to deal with the hypotheticals of each specific scenario in a general framework). The models in question must be

[70] Critical Review of Fair Machine Learning.
[71] See 3.3.2.
[72] Smith, Paul. "Incentives and Justice." *Social Theory and Practice* 24, no. 2 (1998): 205–35. https://doi.org/10.5840/soctheorpract19982422.

evaluated with the context of their respective applications in mind (the situation specificity of each solution will be partly addressed in 4.3).

4.1.2.1 Considerations

So, the above analysis suggests this consideration for the framework:

CONSIDERATION: Does the model in question fulfill the defined requirements of distributive justice?

For this consideration, three questions must be answered about the solution in question, to address the three principles of distributive justice put forth by Rawls:

SUBCONSIDERATION 1: Does the model allow each individual equal rights and liberties?

If no, then we do not have distributive justice as defined by Rawls. If yes, proceed to 2).

SUBCONSIDERATION 2: Does the model allow each individual equal opportunity in the context of its application?

If no, proceed to 3). If yes, 3) need not be considered and we have distributive justice as defined by Rawls.

SUBCONSIDERATION 3: Does potential inequality favor the least advantaged of the involved groups?

If no, then we do not have distributive justice as defined by Rawls. Otherwise, we do.

In a sense, subconsideration 3 is a type of compensatory justice. However, there are more aspects of compensation to consider in the context of ML.

4.2 Compensatory Justice

Compensatory justice concerns itself with ensuring that subjects are properly compensated for unfair disadvantages. Let us assume that we achieve distributive justice as described above. Now, further assume that one is part of the subset that the ML model still works against. As previously mentioned, due to the problem of induction, ML will never yield perfect results. Regardless of how well a model is implemented, there will always exist individuals who get unfairly excluded from the applicability of the model in question. In other words, although a ML model may avoid discriminating against groups, it may still discriminate against certain individuals who are considered anomalies, or outliers. It is of utmost importance that these outliers are not punished for the shortcomings of the model, and that the entities who choose to deploy these models compensate accordingly.

In ML, this is a prevalent problem. In her book *Automating Inequality* (2018), Virginia Eubanks explores scenarios in which the application of automation had undesirable effects. One of the cases in question had a program make decisions on welfare distribution to individuals in Indiana, which resulted in terrible outcomes for many individuals. One million citizens had benefits denied during the first three years of implementation, 54% more denials than the three years prior to implementation. One specific case that gained attention was that of Omega Young, who in late 2008 had her benefits revoked for not having booked an appointment, due to her being in the hospital getting treated for cancer. The automated system interpreted this as a “failure to cooperate.” Young died less than a year later due to her disease complications.[73] In this case, the system generalized its decision-making, based on the assumption that most people who fail to book an appointment do not have a good enough reason to do so. Omega Young was an unperceived outlier who had done nothing wrong.

The concept of compensatory justice ensures that there are ways to accommodate situations like that of Omega Young. It ensures that in the case of classification anomalies, the individuals in question do not get unjustly punished. So, in ML, compensatory justice would be established when, even though we may have theoretical discrimination against individuals by the model, there are failsafe ways of ensuring that this is redressed, and the practical outcome is equitable. In Young’s case, compensatory justice could be ensuring that high-risk decisions are double-checked by a human worker, or that outlier detection is used to have certain decisions never be made by a ML model to begin with.

4.2.1.1 Considerations

Thus, the next consideration:

CONSIDERATION: Does the model in question fulfill the defined requirements of compensatory justice?

For this consideration, only one question must be answered:

SUBCONSIDERATION: In the case of theoretical discrimination against individuals by the model, does the developer have failsafe ways of ensuring that the practical outcome is equitable?

If the answer is yes, we have compensatory justice. Otherwise, we do not.

4.3 Moral Evaluation

Theoretically, we have now established the definitions and considerations of distributive and compensatory justice respectively. It is obvious, however, that determining whether these are carried out in practice is a more difficult task due to our differing moral views and limitations in distancing ourselves from our biases. As 18th century philosopher and economist Adam Smith wrote in his *Theory of Moral Sentiments:*

> There are some situations which bear so hard upon human nature that the greatest degree of self-government, which can belong to so imperfect a creature as man, is not able to stifle, altogether, the voice of human weakness, or reduce the violence of the passions to that pitch of moderation, in which the impartial spectator can entirely enter into them.[74]

[73] Edes, Alyssa, and Emma Bowman. “'Automating Inequality': Algorithms In Public Services Often Fail The Most Vulnerable.” NPR. NPR, February 19, 2018. https://www.npr.org/sections/alltechconsidered/2018/02/19/586387119/automating-inequality-algorithms-in-public-services-often-fail-the-most-vulnerab.

[74] Smith, Adam. *The Theory of Moral Sentiments*. Oxford: Clarendon, 1759.

With certainty, we can say that ML falls into this category of situation. So, how do we make these decisions? How do we deal with biases in ML, knowing that they will be present in some shape or form regardless of how hard we try to eliminate them? How do we optimize the mitigation of these biases, considering the diversity of views that must be considered? These are the questions moral evaluation theories try to answer, both for individuals and collectives. Smith's impartial spectator theory is one worth considering.

4.3.1 Adam Smith's Impartial Spectator

Smith strongly believed that our implicit biases cloud our judgment in specific situations. He believed that these biases ignite feelings in us, which inevitably lead us to act with emotion and not logic. He did, however, have some thoughts on how certain mindsets could partially remedy the influence of bias. In the previously outlined quote, Smith mentioned an "impartial spectator", which he claims is the mindset by which individuals can distance themselves from their biases most efficiently. The impartial spectator is a mindset in the form of a theoretical person, who can assess the situation by considering the perspectives of everyone involved. He believed that by distancing oneself from one's individual feelings about a certain situation, one could reach more sound conclusions on reasonable courses of action.

As mentioned, the impartial spectator is theoretical and has abilities beyond what humans are capable of. This was highlighted in the previously outlined quotation. Smith continues, however, by stating that one may make a serious attempt to emulate the abilities of the impartial spectator by optimizing in the presence of human limitations:

> Though in those cases, therefore, the behaviour of the sufferer fall short of the most perfect propriety, it may still deserve some applause, and even in a certain sense, may be denominated virtuous. It may still manifest an effort of generosity and magnanimity of which the greater part of men are incapable; and though it fails of absolute perfection, it may be a much nearer approximation towards perfection, than what, upon such trying occasions, is commonly either to be found or to be expected.[75]

Adam Smith discussed, in short, how to at least attempt to optimize conditions by trying to take the perspectives of all into consideration. Smith claims that regardless of whether we are considering real-world or hypothetical conditions, the "presence of the impartial spectator, the authority of the man within the breast, is always at hand to overawe them into the proper tone and temper of moderation."[76]

The strength of Smith's impartial spectator model is that it is not merely theoretical. Research has been done on how to enforce it empirically, especially in professional settings.[77] There is, however, only an extent to which his theories can be carried out in practice. Nevertheless, Smith's views emphasize how we can consciously admit our biases and do everything in our

[75] Ibid.

[76] Ibid.

[77] Konow, James. "Is Fairness in the Eye of the Beholder? An Impartial Spectator Analysis of Justice." *SSRN Electronic Journal*, 2006. https://doi.org/10.2139/ssrn.980356.
See also: Szmigin, Isabelle, and Robert Rutherford. "Shared Value and the Impartial Spectator Test." *Journal of Business Ethics* 114, no. 1 (October 2012): 171–82. https://doi.org/10.1007/s10551-012-1335-1.

power to mitigate them by means of moral evaluation beyond ourselves. For this reason, we must try to emulate the impartial spectator in our process of assessing the sufficiency of ML models.

Based on his writings on the limitations of human beings, it is also reasonable to believe that Smith would agree that an actual impartial spectator of a specific scenario would be more proficient in assessing its fairness, as opposed to subjects who themselves are directly involved. This warrants the discussion of algorithmic auditing.

4.3.2 Algorithmic Auditing

Algorithmic auditing is the process by which an algorithm and its respective impact are assessed. There is no unequivocal template to consider in completing this audit, although Smith's impartial spectator implies that analyzing the direct impact on the people involved should be the main priority in assessing our models. After all, his entire moral evaluation argument regards our lack of empathy, and that actions should be taken to improve our abilities in that respect.

Cathy O'Neil (the author of *Weapons of Math Destruction*) has started her own company, ORCAA (O'Neil Risk Consulting & Algorithmic Auditing), which primarily focuses on "the people who will be impacted by the algorithm's success or failure", and makes an assessment of the risks associated with the impact on those groups.[78] One of the tools ORCAA uses in auditing algorithms and assessing their respective impacts is an ethical matrix.[79] This assessment apparatus is used, oftentimes in collaboration with affected groups, to render reasonable conclusions on what is considered fair in the specific context. The "respect for justice" principles of the ethical matrix are also developed specifically with John Rawls' fairness definitions in mind, which makes this approach a suitable fit for our purposes.[80]

It is necessary that the auditors be some of the very best computer scientists and programmers in the field. The difficulties of assessing the weaknesses of a previously unencountered complex algorithm will become evident otherwise, as these are developed by some of the most proficient programmers in the world. It is important that when full transparency is offered by the audited entity, that the auditor itself fully understands the process and implications.

We do not directly endorse ORCAA's approach to algorithmic auditing, as there may be other approaches that better emulate the impartial spectator theory (however, we have not found one). We are also not including any directly defined considerations in the framework for moral evaluation, as this mentality should be encouraged throughout the process and not in a specific portion of the framework. We realize the necessity of algorithmic auditing in future practices of ML implementation, and hope that it becomes common practice for all entities soon.

4.4 Freedom of Choice

There are situations in which distributive and compensatory justice are not necessarily accounted for by ML, yet one may still consider it to be the best solution to the problem available. If a cancer patient is told that there is an ML assessment tool that discriminates in performance

[78] "ORCAA." ORCAA. Accessed March 29, 2020. https://orcaarisk.com/.

[79] Mepham, Ben, Matthias Kaiser, Erik Thorstensen, Sandy Tomkins and Kate Millar. "Ethical Matrix Manual." *LEI,* 2006.

[80] Ibid.

against minorities, but that still performs much better than any physician's diagnosis, one may still want to consider it. It becomes a question of whether one would want to forego the importance of distributive and compensatory justice for the sake of contextual performance. This would not be an easy decision to make, as one may trust one's own physician more than any statistical advantage a ML model can provide, but that is the point – the decision should not be made for the subject, but by the subject. Thus, we must consider the impact freedom of choice has in different scenarios of ML implementation.

There are mainly two questions of freedom of choice that must be answered for us to get the full scope:

(1) "Can the task in question be avoided altogether?" and
(2) "Are there other non-ML alternatives to completing the task in question?"

4.4.1 and 4.4.2 respectively will address these questions, as well as the reasons for which they must be answered.

4.4.1 Choice of Task Avoidance

Throughout this work, the emphasis has been people and the severity of impact on their lives as a result of how some tasks are pursued. Eventually, this framework is developed for the betterment of human wellbeing. For this reason, the framework needs to ensure that the context of a situation is accounted for and that increased caution is advised for certain situations.

Then, one may ask which circumstances warrant additional caution and restriction. One way to approach this question is to define tasks that must be addressed for the survival of the individual in question and can therefore not be avoided. Here, Abraham Maslow's Hierarchy of Needs can help us define when we should be increasingly conscientious about our ML applications. Maslow, a prominent 20th century psychologist attempted in his *Theory of Human Motivation* (1943) to define our needs by their necessity. He did so to emphasize the source of motivation in human beings, but his hierarchy is applicable in many contexts.[81] In the context of ML, we must ensure that more good is done than harm, and that in the event of failure we ensure that negative consequences are not beyond what we would consider reasonable. Below is a visual of Maslow's Hierarchy of Needs:

[81] Maslow, Abraham H. *Theory of Human Motivation*. S.l.: Wilder Publications, 2018.

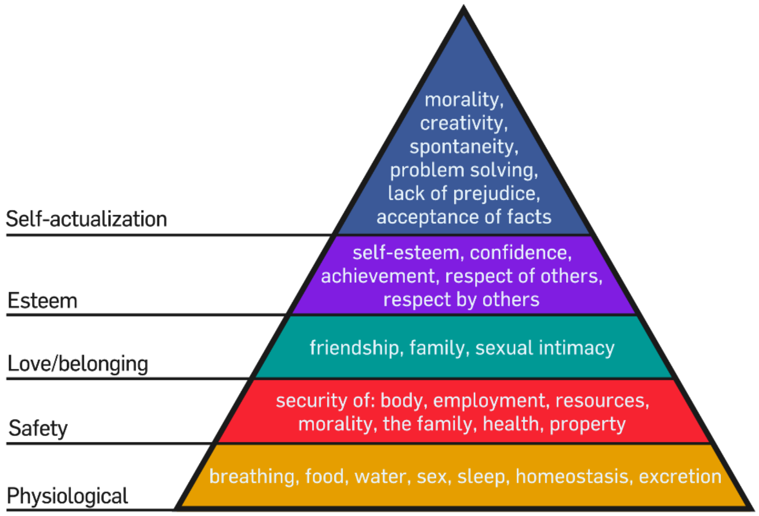

morality,
creativity,
spontaneity,
problem solving,
lack of prejudice,
acceptance of facts
Self-actualization
self-esteem, confidence,
achievement, respect of others,
respect by others
Esteem
friendship, family, sexual intimacy
Love/belonging
security of: body, employment, resources,
morality, the family, health, property
Safety
breathing, food, water, sex, sleep, homeostasis, excretion
Physiological

Figure 4: Maslow's Hierarchy of Needs. [82]

The case we make in using the hierarchy for our ML framework, is that tasks included in the first two steps of Maslow's hierarchy are tasks one does not have the choice to escape. One's survival is dependent on physiological needs, which in turn are dependent on the mentioned safety needs. Long-term, without attending to our physiological and safety needs, we die. Thus, we must ensure that decisions made by ML models about these needs are taken very seriously. For the remainder of this work, we will refer to those two first steps of the hierarchy as "life-defining needs".

We do have our respective objections to Maslow's hierarchy. For one, sex cannot reasonably be considered a physiological need in the way food and sleep are. Simply put, one will not die without sex. In addition, sexual intimacy is already accounted for in the love/belonging stage. Second, for the purpose of ML application, mental health should be included in the safety stage. Maslow would reasonably justify the exclusion of mental health as the entire pyramid intends to address that aspect of human beings. Nevertheless, for the purpose of our framework, it is of essence that mental health is considered and should likely be placed in the safety need step of the hierarchy.

4.4.1.1 Considerations

The first consideration of our framework flowchart (as this will be considered before the considerations on distributive and compensatory justice) then becomes the following:

CONSIDERATION: Is this model making decisions on life-defining needs?

4.4.2 Choice of Non-ML Alternatives

Now that we have properly defined what constitutes a task for which the choice exists to avoid it altogether, we must consider scenarios in which the task itself is not as vital. This question is

[82] McLeod, Saul. "Maslow's Hierarchy of Needs in pyramid form with explanations and examples." *Simply Psychology,* 2007. https://commons.wikimedia.org/wiki/File:Maslow%27s_Hierarchy_of_Needs_Pyramid.png

important to consider, as it helps us define appropriate expectations on distributive and compensatory justice for different scenarios, but also the understanding the public should have of an algorithm. Now, the central question becomes whether there is a choice in completing the task by other means than ML. The reason we explicitly ask for the existence of non-ML alternatives, is due to the assumption that unless a solution is of ML nature it will, at the very least, be interpretable.[83] That is, the rationale by which the decision was made will be understandable and that if it is not, there will be full accountability for the mistakes made. For instance, in the case of a physician's recommendation, we always expect a reason as to why they have arrived at a specific recommendation. In ML, we will not always have a fully outlined rationale. So, these are the three scenarios to consider:

4.4.2.1 Life-Defining Need Without Non-ML Alternatives

This is the scenario in which we must be most cautious about interpretability and justice requirements. If the task regards a life-defining need, there is no escaping its completion. Furthermore, if one does not have non-ML alternatives to consult, then the task must be completed exactly as outlined by the model considered by the framework, or some other ML model that the framework has deemed inferior. For this reason, it is required that our model fulfills the recommendations of compensatory and distributive justice, but also that there is interpretability.

4.4.2.2 Life-Defining Need With Non-ML Alternatives

Here, the recommendations on distributive and compensatory justice are not required, nor is the recommendation of interpretability. All remain preferences, but if there are other ways of completing the life-defining task, then the ML implementation does have flexibility in abiding by distributive and compensatory justice, as the potential user has the freedom to avoid the implementation or compare the implementation's results with those arrived at through non-ML methods.

4.4.2.3 Not a Life-Defining Need

This scenario has the same consequences as 2. First, a distinction need not be made between a non-life-defining task with or without alternatives, as the freedom for the individual exists to avoid the task altogether. Second, it is possible that a difference should be established between the restrictions on life-defining tasks with alternatives and non-life-defining tasks with or without alternatives. The rationale here is that in all three cases, the potential user is making a choice to prioritize performance over justice. There are instances in life-defining needs where we are aware that our solution is far from perfect, but that it is the best approach currently available. The point is that the choice being made, although of varying importance, is of the same nature.

[83] We will discuss interpretability in detail in 4.4.3.

4.4.2.4 Considerations

So, to address all three scenarios, the next consideration is as follows:

CONSIDERATION: Are there non-ML alternatives for accomplishing the task in question?

4.4.3 Explainable and Interpretable ML

The requirement of freedom of choice is insufficient, if we do not understand the choices at our disposal. This is based on the idea that freedom of choice is not fully free unless the choice has the opportunity to be fully informed and reasoned (that is, that all decision-making information is at one's disposal).[84] Thus, we must discuss black-box models, and contrast them with interpretable ones. Black-box models are oftentimes criticized for their inability to provide a rationale as to how they arrive at a certain conclusion. This is why the general framework will advise against implementation of black-box models when the user intends to complete a life-defining task and has no other non-ML alternatives.

There are, however, exceptions to the restricted use of black-box models in safety and physiological needs – if the subject in question has a choice in whether they are exposed to it. Once again, assume that a subject has cancer and is looking for the optimal treatment. They will get recommendations from a certified physician (who can explain the rationale behind their recommendation) and a black-box ML model. Based on previous data, the ML model gives recommendations that lead to remission 30% more often than those of the physician. It is likely that some would be willing to place more trust in a model that seems to do better than the average physician (even in the absence of rationale), while some may find the physician's rationale to be very reasonable and place trust in the explanation instead. This warrants a comparison between interpretability and explainability.

In ML, both interpretability and explainability are about making sense of how and why a certain model arrives at certain conclusions. Interpretability refers to when an algorithm is intuitive and understandable without full expertise on the topic (this usually refers to a certain subset of methods, such as Naïve Bayes classifiers and rule-based ML). Explainability is, as the word implies, when there are ways to make sense of and simplify complex and non-intuitive algorithms (among these are neural networks).[85] Worth noting is that the type of algorithm is not always what defines the interpretability or explainability of a model – sometimes it is the number of features in the considered dataset. With explainable ML[86], computer scientists oftentimes must write programs to help explain the model (due to the complexity of the data and algorithm), whereas with interpretable ML the structure and rationale of the model makes the classification process more transparent and directly understandable.

[84] Wells, Thomas. "Free Choice Is Informed Choice: Why We Need Ethical Warning Labels on Animal Products." ABC Religion & Ethics. Australian Broadcasting Corporation, June 2, 2016. https://www.abc.net.au/religion/free-choice-is-informed-choice-why-we-need-ethical-warning-label/10096936.

[85] Lawton, George. "UX Defines Chasm between Explainable vs. Interpretable AI." SearchEnterpriseAI. TechTarget, December 24, 2019. https://searchenterpriseai.techtarget.com/feature/UX-defines-chasm-between-explainable-vs-interpretable-AI.

[86] We will use "Explainable ML" and "black-box" interchangeably.

ML scientist Cynthia Rudin brings up an array of issues with explainable ML.[87] Commonly, the case is made that explainable ML is deployed because of its superior performance. This may be true occasionally, which is why black-box algorithms are still warranted from time to time – but not always. Rudin claims from experience that in most contexts, there are interpretable ML implementations that would have equal or better accuracy compared to their explainable counterparts.

Furthermore, the fact that programs are needed to make sense of black-box algorithms implies that even the explanations provided for the model in question are not doing the full model justice. If the original model were fully explainable, we would not need a program to explain it for us. As Rudin herself puts it:

> An explainable model that has a 90% agreement with the original model indeed explains the original model most of the time. However, an explanation model that is correct 90% of the time is wrong 10% of the time. If a tenth of the explanations are incorrect, one cannot trust the explanations, and thus one cannot trust the original black box. If we cannot know for certain whether our explanation is correct, we cannot know whether to trust either the explanation or the original model.[88]

Explainable ML also has weaknesses in not being able to properly accommodate the specific circumstances of each subject, as well as increasing the probability of human error due to the extensive preprocessing needed to ensure that all data is correct. Preprocessing data for interpretable models with fewer features is doable – doing the same for a program like COMPAS with more than 130 features is difficult in comparison.

Rudin advocates for more interpretable ML models instead of explainable ones, as they make us able to more efficiently analyze its flaws. They are sometimes more difficult to construct algorithmically but allow subjects who are not necessarily experts on the topic of ML to intuitively understand how they are being assessed.[89]

Occam’s Razor, a principle used extensively and that has empirically been proven very useful in analysis, becomes relevant here. The principle has multiple different formulations, but the main one attributed to theologian William of Ockham is: “Plurality must never be posited without necessity.”[90] The gist of the statement is that if one has competing solutions to a problem, then the simplest (most understandable) solution is preferred. The case is made that excessive complexity leads to more potentially unfounded assumptions, which in turn reduce the credibility of one’s solution. For this reason, it would never make sense to choose a black-box model when other interpretable models have similar performance. Furthermore, it would also be unwise to consult explainable ML if interpretable ML has not been attempted. This is why the framework

[87] Rudin, Cynthia. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” *Nature Machine Intelligence* 1, no. 5 (2019): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

[88] Ibid.

[89] Ibid.

[90] Kneale, William, and Martha Kneale. *The Development of Logic*. Oxford: Clarendon Press, 2008.

advises that if the task in question never has been completed using ML, that the first attempt should be with an interpretable model.

4.4.3.1 Considerations

Thus, the following considerations:

CONSIDERATION: Is this a black-box model?

CONSIDERATION: Is this the first time the task is being completed with ML?

CONSIDERATION: Do interpretable models with similar performance exist?

Due to the importance of having interpretability, this framework recommends that if a certain task never has been pursued previously, that the first attempt is made with an interpretable model before consulting other, potentially black-box solutions.

Assuming that explainable models only are implemented when performing better than interpretable ones, when we are faced with a choice between an interpretable model (ML or non-ML) and an explainable model, it is essentially a choice between interpretability and performance. But, in the case of black-box models, you really do not know for certain what the choice is. Since black-box models cannot be fully rationalized and interpreted, there is no way to justify them being the sole option under any circumstance in which avoiding the task altogether is not a choice (that is, if it addresses a life-defining need). Thus, the framework does not allow black-box models for use in life-defining tasks with no other alternatives. In these cases, there is not a choice to be made between interpretability and performance, and in those situations (as elaborated on previously), we prioritize interpretability. Although our societies and judicial systems oftentimes are consequentialist, we still place importance on how consequences come about. Otherwise, our courts would consider manslaughter and murder to be the same thing since the outcomes are nearly identical – but they do not. Clearly, the rationale by which a decision is made is significant, which the prioritization of interpretability emphasizes in this context.

Do note that this framework does not ensure that should one opt for a non-ML solution, one will avoid the injustices that come with the model in question. It just ensures that ML is not responsible for the injustices, and that if it is, that one has chosen to subject oneself to them. This does not, however, mean that the responsibility lies entirely on the user. There should also exist accountability for entities to provide the right information and having the right intentions.

4.5 Accountability

In our current societies, there are many incentives for entities to deploy ML for various tasks. It saves money because, unlike humans, it does not require financial compensation. It saves time because it can make inferences from immense amounts of data in a fraction of the time required for a human being.

Most importantly for entities, however, ML saves accountability (although this applies to automation and AI holistically). Due to ML being applied to new and unexplored contexts constantly, there are few to no ways of deciding legally who is responsible for failures in distributive and compensatory justice. This has become a large discussion topic in the context of self-driving car accidents and accompanying liabilities.[91] But, as the Voltaire-inspired Peter Parker principle famously states: “With great power comes great responsibility.”[92] This rationale has been used by ethicists to justify the accountability being placed on developers of the decision-making algorithms. Kirsten Martin, business ethics professor at George Washington University, says:

> As such, firms should be responsible not only for the value-laden-ness of an algorithm but also for designing who-does-what within the algorithmic decision. As such, firms developing algorithms are accountable for designing how large a role individual will be permitted to take in the subsequent algorithmic decision. Counter to current arguments, I find that if an algorithm is designed to preclude individuals from taking responsibility within a decision, then the designer of the algorithm should be held accountable for the ethical implications of the algorithm in use.[93]

Responsibility for failure in these models should be put on a defined set of individuals prior to launch. There should never be ambiguity in whom to consult (or blame) in situations of injustice, and the incentives for entities and individuals should not exist to avoid accountability through ML (especially if they are already benefitting from savings of money and time).

4.5.1.1 Considerations

Thus, the following consideration:

CONSIDERATION: Are the individuals behind the model’s development claiming full responsibility for failure in fulfilling the requirements of distributive and compensatory justice?

If the answer is yes, the model could be allowed implementation in some form. If there are flaws in distributive and compensatory justice prior to launch, these should be publicly disclosed for freedom of choice to be a reality. If the answer is no, the model should not be allowed implementation without exception.

Worth noting is that accountability only applies to the impactors and impacted of the model in question. Thus, if a supposedly unjust ML model is to be implemented on a smaller scale, if consensus can be found among all exposed individuals, the model should still be allowed implementation due to it being within the freedom of choice of aforementioned individuals to subject themselves to the model. This is helpful, as there may exist businesses and individuals who can benefit from the use of ML on a small scale, but do not have the resources to create robust and generally equitable models.

[91] Nyholm, Sven R., and J. Smids. ”Automated cars meet human drivers: responsible human-robot coordination and the ethics of mixed traffic.” *Ethics and Information Technology,* 2020. https://doi.org/10.1007/s10676-018-9445-9

[92] “With Great Power Comes Great Responsibility.” Quote Investigator, May 28, 2019. https://quoteinvestigator.com/2015/07/23/great-power/.

[93] Martin, Kirsten E. “Ethical Implications and Accountability of Algorithms.” *SSRN Electronic Journal,* 2018. https://doi.org/10.2139/ssrn.3056716.

4.6 Pre-Implementation Flowchart

And thus, we arrive at the below flowchart representation for considerations prior to implementation:

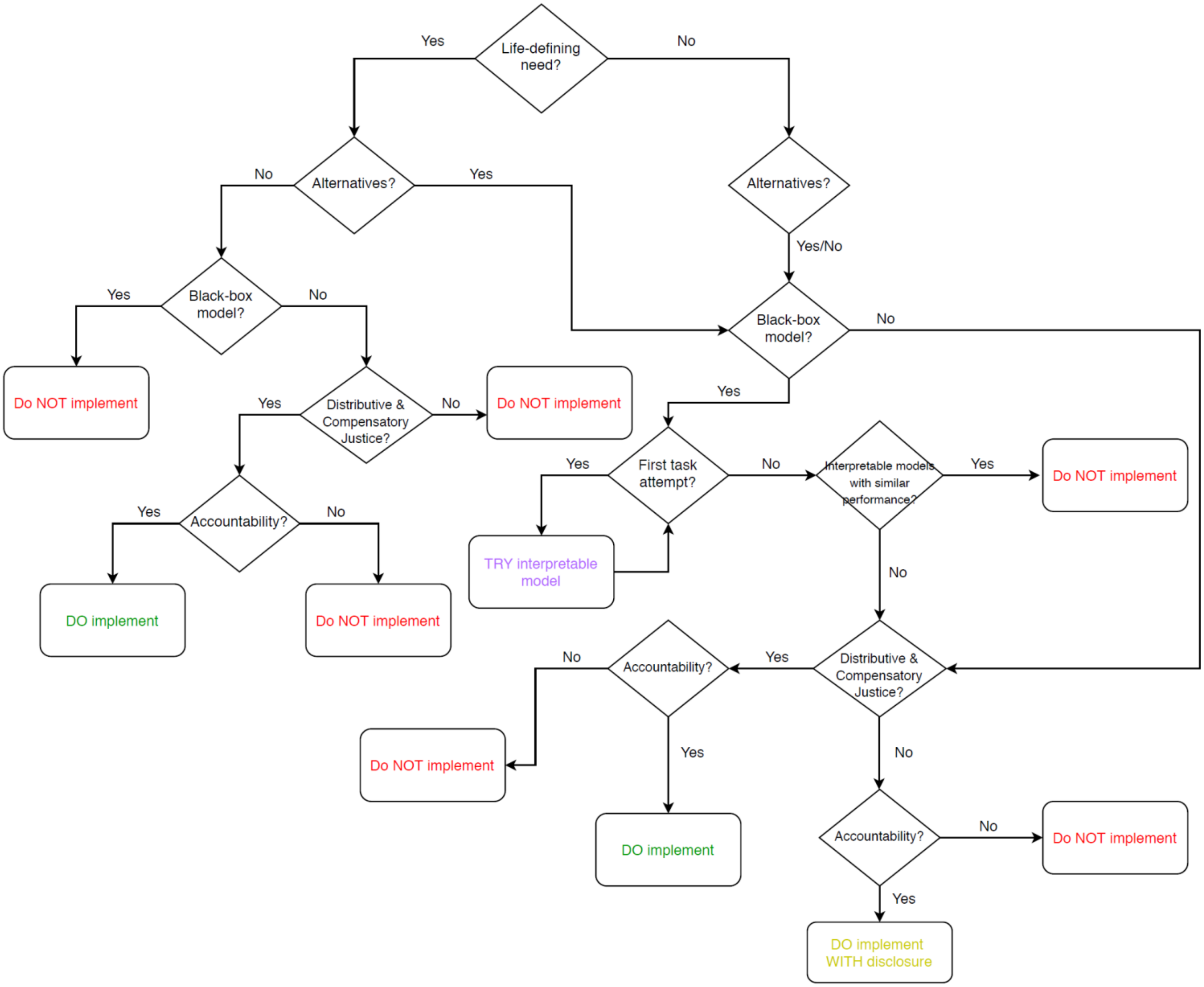


Figure 5: Flowchart representation of the implementation framework.

4.7 Post-Implementation Accountability

The above flowchart only accounts for considerations before implementing a certain ML model. Nevertheless, there are considerations for situations that may occur after implementation as well. After all, many of the issues that arise in ML result from use of the respective models in applied contexts. Thus far, the framework has encouraged increased general conscientiousness and accountability in developing and deploying ML, with additional caution being advised in situations of life-defining needs (to avoid catastrophic situations like that of Omega Young).[94] Finally, before concluding our analysis of the issue at hand, we must ensure the same exactitude and accountability as the models are applied in the real world.

The requirements of transparency, distributive and compensatory justice are truly difficult to satisfy for most ML models. They must be. There is a reason only a select few models warrant implementation for life-defining needs when no other non-ML alternatives are available. Therefore, this cannot be understated: good ML models are works in progress and must be treated as such. This is where restorative justice plays a role.

4.7.1 Restorative Justice

To echo Adam Smith's thoughts regarding virtue, accountability does not imply achieved perfection.[95] This framework openly allows implementation even in cases where distributive and/or compensatory justice requirements are not met. Accountability implies the ambition of perfection. It implies that if justice requirements are not met, that the producer of the model intends to improve the algorithm, with intentions to make it more just with time. Restorative justice can aid us in doing that efficiently.

Restorative justice concerns itself with holding responsible subjects accountable for wrongdoings by having them fix their problems through cooperation with and understanding of victims. It is the idea that offenders should be allowed the opportunity to right their wrongs without additional criminal punishment. Howard Zehr, a pioneer in the field of restorative justice, outlines the concept and its rationale using these three principles:

- Crime is a violation of people and of interpersonal relationships.
- Violations create obligations.
- The central obligation is to put right the wrongs.[96]

Admittedly, restorative justice was developed as a new way to deal with crime. Many of the wrongs made in good faith and by mistake in ML will not be considered crimes by our legal systems (and our framework), and so "crime" may not be the right word to use in this context. Replacing "crime" with "injustice" in the context of ML will make these principles strongly applicable to what we want to achieve. In the case of failure in fulfilling the outlined requirements of distributive and/or compensatory justice, it is important that there exists an expectation for the entity responsible to improve the model. We do not want to punish entities for honest mistakes – instead, we want to encourage a deeper understanding of the users' problems, so that they can efficiently be dealt with.

It is here that we argue that competition provided by free markets in liberal societies will, to a certain extent, automatically encourage companies to improve their models if the flaws are publicly disclosed – even if that were not the initial intention.[97] If people affected know the flaws of a certain application, they will move toward other solutions that do not discriminate, should

[94] See 4.2.
[95] See 4.3.
[96] Zehr, Howard. *The Little Book of Restorative Justice*. Vancouver, B.C.: Langara College, 2016.
[97] See 4.5.

they appear. Having the flaws publicly exposed will make the lack of discrimination a truly marketable trait of a model. Free markets – to the extent that they are reasonable – are also especially important, to increase the likelihood of full accommodation of fairness.[98] If there are more options to consider, the likelihood that an individual will align with the fairness definition of at least one of them is larger than if options are fewer. In that regard, when there are no "right" answers, markets can aid us in providing more solutions (of different nature) to the same problem, making more people content in the process.[99]

4.7.2 Criminal Justice

We mentioned previously that there are incentives for entities to pursue ML solutions other than proficiency in solving a problem. For that reason, if a model saves an entity time and money the way it already is, there would not be much incentive to fully disclose the issues with it, as well as dealing with the concerns of users. If there is no intention to improve the deficient model, it shows that the company did not make the model with the improvement of the human condition in mind, but for the other incentives provided by the market. This is where we get to the issues of entities and individuals failing to hold themselves accountable and required criminal justice as a result.

Due to the inductive nature of ML and the inherent biases of human beings, there should be room for trial and error in developing what we consider a just application. Nevertheless, these trials should be done in good faith. There should be a distinction between the treatment of those who mistakenly implemented inadequate applications, and those who did so knowingly. With accountability comes the intention to work towards better – not cheaper and more time efficient – solutions.

4.7.2.1 Considerations

Thus, the following consideration:

CONSIDERATION: In the case of failure in fulfilling the requirements of distributive and compensatory justice, was there negligence or reluctance to deal with obvious problems?

If no, then restorative justice measures should be taken. If yes, then criminal justice measures should be taken.

One may have differing views on what "obvious" problems are. We argue that the definition should regard the problems brought to light during algorithmic auditing. For instance, assuming one uses ORCAA's approach, obvious problems can be defined as those brought up in the development of the ethical matrix.[100] Nevertheless, it would be dependent on how algorithmic auditing is done.

[98] Some of these forecasts may seem too optimistic due to the influence we believe future human-centered economies will have on our current incentive structures. This is further addressed in 5.

[99] The consideration in regard to restorative justice will be addressed in 4.7.2.

[100] See 4.3.

4.7.3 Additional Comments

To reiterate, restorative justice measures will have to continue until there are no evident biases in the respective programs, which likely will be – never. Criminal justice measures are only warranted if restorative justice has been encouraged and denied by the entity in question.

Although we have developed a framework based on deontological arguments, our current legal frameworks are based on consequentialism. The interplay between restorative and criminal justice accounts for the consequentialist nature of our legal statutes, while still emphasizing the importance of a deontological approach to societal issues. It also allows for a distinction between good and bad intentions.

We will not elaborate on the nature of and approach to criminal justice, as we believe measures can be taken to make criminal justice a non-necessity in the future (as we will discuss briefly later). Nevertheless, assuming the current structure of our economy, it is important that the criminal punishment is severe enough, so that it disincentivizes the act of wrongdoing. As 2020 U.S Presidential Candidate Andrew Yang expresses on his website when discussing the lack of accountability for pharmaceutical companies in the opioid crisis: "Purdue Pharma has made more than $35 billion in revenue since releasing OxyContin in 1995. The fine of $635 million for false advertising around claims of non-addictiveness and tamper-proofing is barely a slap on the wrist."[101] If we are to take the route of criminal justice, we need to ensure that the repercussions are not merely a "slap on the wrist" to incentivize just measures.

[101] Yang, Andrew. "Hold Pharmaceutical Companies Accountable - Yang2020 - Andrew Yang for President." Yang2020. Accessed March 29, 2020. https://www.yang2020.com/policies/holding-pharmaceutical-companies-accountable/.

5 Reflections on Work and Future Areas of Research

Conceiving a new framework from scratch has been challenging, and multiple different approaches were taken before arriving at the one seen in this thesis. Considering that this was a first attempt at an ethical framework for the implementation of ML, it is more than possible that there are other thoughts and philosophical theories that could contribute to the framework, either by replacing certain parts or adding substance to the current structure. Hopefully this serves as a reason for other models to be developed, which take new stances on the topic in question.

The main importance was to develop considerations that had a common thread of logic, without making the framework overwhelmingly technical. This work is intended for anyone affected by ML, either as impactor or impacted. Today, that is essentially everyone. For this reason, it was important to develop guidelines that are, to a large extent, understandable to most or all human beings. We believe that this work has succeeded in this regard, and that with additional constructive feedback, can help spur discussion about a field that needs more serious consideration in the near future.

One weakness of the current framework is that it does not consider the hierarchy of importance in Rawls' three principles of distributive justice. By this framework, you either fulfill the requirements (along with compensatory justice) or you do not. It may be the case that if the model does not fulfill the first subconsideration, that implementation should not be advised in any context.[102] A future ambition would be to take this first attempt and after extensive feedback elaborate on some of the gaps in collaboration with other peers. In 2017, IEEE started developing a "Standard for Algorithmic Bias Considerations" to provide a guideline for involved program architects and project leaders.[103] We hope to see more analysis of the IEEE Standard, along with other frameworks (including ours), to assess their strengths and weaknesses in addressing bias in ML.

We recognize that the current framework needs legal backing of some nature to serve a substantive purpose. Entities will rarely abide by an ethical framework if the economy incentivizes other action. This has only recently become recognized in law, with the European Union establishing action against algorithmic bias in its 2018 General Data Production Regulation (GDPR).[104] The U.S has now also understood the necessity for legal action, with a bill – the Algorithmic Accountability Act – being introduced to the House of Representatives in 2019. [105] [106] These measures are indeed important, but it will be of essence to ensure that the

[102] See 4.1.2.1.

[103] Koene, Ansgar, Liz Dowthwaite, and Suchana Seth. "IEEE P7003™ Standard for Algorithmic Bias Considerations." *Proceedings of the International Workshop on Software Fairness - FairWare 18,* 2018. https://doi.org/10.1145/3194770.3194773.

[104] "What Does the GDPR Mean for Business and Consumer Technology Users." GDPR.eu, February 13, 2019. https://gdpr.eu/what-the-regulation-means-for-everyday-internet-user/.

[105] "Booker, Wyden, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms." Cory Booker | U.S. Senator for New Jersey, April 10, 2019. https://www.booker.senate.gov/?p=press_release&id=903.

[106] Robertson, Adi. "A New Bill Would Force Companies to Check Their Algorithms for Bias." The Verge. The Verge, April 10, 2019. https://www.theverge.com/2019/4/10/18304960/congress-algorithmic-accountability-act-wyden-clarke-booker-bill-introduced-house-senate.

regulations in question are carried out efficiently and with realistic expectations, which will take trial and error.

In this regard, one area that needs serious research and consideration is incentive structures within our respective economies that would make the implementation of the above outlined framework practically possible. First, it may be the case that even if there are compensatory justice requirements to consider, that the criminal justice measures are too mild for companies to want to avoid them. In other words, there exist scenarios in which companies would skip the considerations of compensatory and distributive justice for the purpose of profitability. The reason we chose not to elaborate on the specifics of criminal justice is because we believe that in the right economy, we barely need criminal justice. In a human-centered economy, where there are incentives to justly and ethically implement technology for the benefit of all, we believe negligence and reluctance would be non-factors in the greater scheme of things, and that restorative justice would take precedence. Currently, our economy has financial incentives. A prime example in the context of ML is data collection. Gathering datasets specifically made for the context in question is much more expensive than to train the model with already existing, more general datasets. Due to the financial incentives of companies, and the fact that they are the ones investing in such a service, they will see little reason to invest additionally to make up for what they consider small deficiencies in their programs. This would be different in an economy that prioritizes the wellbeing of its people. The idea of human-centered capitalism has been popularized by individuals like 2006 Nobel Peace Prize winner Muhammad Yunus and 2020 U.S Presidential Candidate Andrew Yang.[107] We realize that this suggested research may arrive at non-capitalist recommendations, which would require changes in our framework as it currently relies on our contemporary incentive structures. Nevertheless, more research on how to practically implement a human-centered economy will alleviate some of the issues not only in ML, but in other contexts of our societal reality. That could only be a good thing.

[107] Bornstein, David. “Giving Capitalism a Social Conscience.” The New York Times. The New York Times, October 10, 2017. https://www.nytimes.com/2017/10/10/opinion/giving-capitalism-a-social-conscience.html.
See also: “Humanity Forward.” Humanity Forward. Accessed March 30, 2020. https://movehumanityforward.com/.

6 Conclusion

In a world of booming technology, it is imperative that actions are taken to ensure that our advancements are not used for purposes other than the improvement of the human condition. In that regard, computer scientists have taken measures to try to accommodate the importance of justice and fairness in our ML applications. Our analysis confirms that both concepts are not always simple to integrate into our mathematical models, which should encourage us to be increasingly conscientious about how and when we implement ML and discourage us from using ML in every technically applicable context. In this endeavor, we completed a first effort to develop a framework that accommodates for present limitations in both human beings and technology and sets what we believe are realistic expectations for developers and entities to follow.

The practical measures to be taken for this theoretical framework to properly work are many, and the framework is limited in its ability. Nevertheless, our developed implementation considerations provide a foundation to build on and start the conversation on a topic worthy of further discussion and elaboration. The potential gaps of the thesis do not detract from the fact that we need a framework with which to assess the implementation of machine learning, and that just as in the field it intends to target, it will require trial and error to optimize. This should only serve as the beginning of a field in desperate need of growth. We hope to witness this progression in the coming months and years.

7 References

"5 Unsolved Mysteries about the Brain." Allen Institute for Brain Science, March 14, 2019. https://alleninstitute.org/what-we-do/brain-science/news-press/articles/5-unsolved-mysteries-about-brain.

Ackermann, Nils. "Artificial Intelligence Framework: A Visual Introduction to Machine Learning and AI." Medium. Towards Data Science, December 15, 2018. https://towardsdatascience.com/artificial-intelligence-framework-a-visual-introduction-to-machine-learning-and-ai-d7e36b304f87.

Albarghouthi, Aws, Loris Dantoni, Samuel Drews, and Aditya V. Nori. "FairSquare: Probabilistic Verification of Program Fairness." *Proceedings of the ACM on Programming Languages* 1, no. OOPSLA (December 2017): 1–30. https://doi.org/10.1145/3133904.

Amini, Alexander, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019. https://doi.org/10.1145/3306618.3314243.

"Booker, Wyden, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms." Cory Booker | U.S. Senator for New Jersey, April 10, 2019. https://www.booker.senate.gov/?p=press_release&id=903.

Bornstein, David. "Giving Capitalism a Social Conscience." The New York Times. The New York Times, October 10, 2017. https://www.nytimes.com/2017/10/10/opinion/giving-capitalism-a-social-conscience.html.

Bradford, Alina. "Deductive Reasoning vs. Inductive Reasoning." LiveScience. Purch, July 25, 2017. https://www.livescience.com/21569-deduction-vs-induction.html.

Broad, C. D. *The Philosophy of Francis Bacon: an Address Delivered at Cambridge on the Occasion of the Bacon Tercentenary, 5 October 1926*. New York: Octagon Books, 1976.

Brownlee, Jason. "Difference Between Classification and Regression in Machine Learning." Machine Learning Mastery, May 21, 2019. https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/.

Buolamwini, Joy. "Artificial Intelligence Has a Racial and Gender Bias Problem." Time. Time, February 7, 2019. https://time.com/5520558/artificial-intelligence-racial-gender-bias/.

Buolamwini, Joy. "GenderShades." Gender Shades, 2018. http://gendershades.org/overview.html.

Buolamwini, Joy A. "Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers." *MIT Media Lab, 2017,* 2017.

Chamorro-Premuzic, Tomas. "Four Unethical Uses Of AI In Recruitment." Forbes. Forbes Magazine, May 30, 2018. https://www.forbes.com/sites/tomaspremuzic/2018/05/27/four-unethical-uses-of-ai-in-recruitment/#6baa3dac15f5.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, and Sharad Goel. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear." The

Washington Post. WP Company, October 17, 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

Corbett-Davies, Sam, and Sharad Goel. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *Stanford University, 2018,* 2018.

Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." Reuters. Thomson Reuters, October 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Edes, Alyssa, and Emma Bowman. "'Automating Inequality': Algorithms In Public Services Often Fail The Most Vulnerable." NPR. NPR, February 19, 2018. https://www.npr.org/sections/alltechconsidered/2018/02/19/586387119/automating-inequality-algorithms-in-public-services-often-fail-the-most-vulnerab.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: Picador, St. Martins Press, 2019.

Fisher, A.R.J, and E.F McClennen. "The Pareto Argument for Inequality Revisited.", 2011.

Gillum, Jack, and Marisol Bello. "When Standardized Test Scores Soared in D.C., Were the Gains Real?" USA Today. Gannett Satellite Information Network, March 30, 2011. http://usatoday30.usatoday.com/news/education/2011-03-28-1Aschooltesting28_CV_N.htm.

Goldman, Barry, and Russell Cropanzano. "'Justice' and 'Fairness' Are Not the Same Thing." *Journal of Organizational Behavior* 36, no. 2 (2014): 313–18. https://doi.org/10.1002/job.1956.

Handeyside, Hugh. "New Documents Show This TSA Program Blamed for Profiling Is Unscientific and Unreliable - But Still It Continues." American Civil Liberties Union, April 22, 2019. https://www.aclu.org/blog/national-security/discriminatory-profiling/new-documents-show-tsa-program-blamed-profiling.

Hao, Karen. "Making Face Recognition Less Biased Doesn't Make It Less Scary." MIT Technology Review. MIT Technology Review, February 15, 2019. https://www.technologyreview.com/s/612846/making-face-recognition-less-biased-doesnt-make-it-less-scary/.

Hao, Karen. "This Is How AI Bias Really Happens-and Why It's so Hard to Fix." MIT Technology Review. MIT Technology Review, February 4, 2019. https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard to-fix/.

"Humanity Forward." Humanity Forward. Accessed March 30, 2020. https://movehumanityforward.com/.

Hume, David, and Tom L. Beauchamp. *An Enquiry Concerning Human Understanding: a Critical Edition*. Oxford: Clarendon Press, 2009.

Håkansson, Martin. "Hardware biases and their impact on GNSS positioning.", 2017.

Jackson, Abby. "12 Words That Are More Familiar to Women than Men." Business Insider. Business Insider, March 24, 2017. https://www.businessinsider.com/gender-and-vocabulary-analysis-women-2017-3?r=US&IR=T.

Kant, Immanuel. *Kant: Groundwork of the Metaphysics of Morals*. Provo, UT: Renaissance Classics, 2012.

Kompella, Subhadra, and Kalyana Chakravarthy Chilukuri. "Stock Market Prediction Using Machine Learning Methods." *International Journal Of Computer Engineering And Technology* 10, no. 3 (2019). https://doi.org/10.34218/ijcet.10.3.2019.003.

Kneale, William, and Martha Kneale. *The Development of Logic*. Oxford: Clarendon Press, 2008.

Koene, Ansgar, Liz Dowthwaite, and Suchana Seth. "IEEE P7003™ Standard for Algorithmic Bias Considerations." *Proceedings of the International Workshop on Software Fairness - FairWare 18*, 2018. https://doi.org/10.1145/3194770.3194773.

Konow, James. "Is Fairness in the Eye of the Beholder? An Impartial Spectator Analysis of Justice." *SSRN Electronic Journal*, 2006. https://doi.org/10.2139/ssrn.980356.

Lawton, George. "UX Defines Chasm between Explainable vs. Interpretable AI." SearchEnterpriseAI. TechTarget, December 24, 2019. https://searchenterpriseai.techtarget.com/feature/UX-defines-chasm-between-explainable-vs-interpretable-AI.

"Machine Learning Crash Course | Google Developers." Google. Google. Accessed March 29, 2020. https://developers.google.com/machine-learning/crash-course/.

Martin, Kirsten E. "Ethical Implications and Accountability of Algorithms." *SSRN Electronic Journal*, 2018. https://doi.org/10.2139/ssrn.3056716.

Maslow, Abraham H. *Theory of Human Motivation*. S.l.: Wilder Publications, 2018.

McLeod, Saul. "Maslow's Hierarchy of Needs in pyramid form with explanations and examples." *Simply Psychology*, 2007. https://commons.wikimedia.org/wiki/File:Maslow%27s_Hierarchy_of_Needs_Pyramid.png

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Aram Galstyan, and Kristina Lerman. "A Survey on Bias and Fairness in Machine Learning." *USC, Information Sciences Institute*, September 17, 2019.

Mepham, Ben, Matthias Kaiser, Erik Thorstensen, Sandy Tomkins and Kate Millar. "Ethical Matrix Manual." *LEI*, 2006.

Molla, Michael, Michael Waddell, David Page, and Jude Shavlik. "Using Machine Learning to Design and Interpret Gene-Expression Microarrays." *AI Magazine on Bioinformatics*, 2004.

Molnar, Christopher. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* GitHub, 2020.

Mols, Bennie. "In Black Box Algorithms We Trust (or Do We?)." ACM, March 16, 2017. https://cacm.acm.org/news/214618-in-black-box-algorithms-we-trust-or-do-we/fulltext.

Nyholm, Sven .R, and J. Smids. "Automated cars meet human drivers: responsible human-robot coordination and the ethics of mixed traffic." *Ethics and Information Technology,* 2020. https://doi.org/10.1007/s10676-018-9445-9

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin Books, 2018.

Pachal, Pete. “Google Photos Identified Two Black People as 'Gorillas'.” Mashable. Mashable, July 1, 2015. https://mashable.com/2015/07/01/google-photos-black-people-gorillas/?europe=true.

Peterson, Andrea, and Jake Laperruque. “Amazon Pushes ICE to Buy Its Face Recognition Surveillance Tech.” The Daily Beast. The Daily Beast Company, October 23, 2018. https://www.thedailybeast.com/amazon-pushes-ice-to-buy-its-face-recognition-surveillance-tech.

Rawls, John. *A Theory of Justice*. Cambridge, MA: Harvard Univ. Pr., 1980.

Rawls, John. “Justice as Fairness.” Cambridge, MA: Harvard Univ. Pr., 1971.

Rawls, John. *Political Liberalism*. New York: Columbia University Press, 1985.

Robertson, Adi. “A New Bill Would Force Companies to Check Their Algorithms for Bias.” The Verge. The Verge, April 10, 2019. https://www.theverge.com/2019/4/10/18304960/congress-algorithmic-accountability-act-wyden-clarke-booker-bill-introduced-house-senate.

Rothman, Joshua. “The Equality Conundrum.” The New Yorker. The New Yorker, January 21, 2020. https://www.newyorker.com/magazine/2020/01/13/the-equality-conundrum.

Roser, Max, and Hannah Ritchie. “Technological Progress.” Our World in Data, 2020. https://ourworldindata.org/technological-progress.

Rudin, Cynthia. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” *Nature Machine Intelligence* 1, no. 5 (2019): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Sharma, Dhruv. “Problems in Machine Learning Models? Check Your Data First.” Medium. Towards Data Science, August 31, 2019. https://towardsdatascience.com/problems-in-machine-learning-models-check-your-data-first-f6c2c88c5ec2.

Smith, Adam. *The Theory of Moral Sentiments*. Oxford: Clarendon, 1759.

Smith, Paul. “Incentives and Justice.” *Social Theory and Practice* 24, no. 2 (1998): 205–35. https://doi.org/10.5840/soctheorpract19982422.

Sun, Amy. “Equality Is Not Enough: What the Classroom Has Taught Me About Justice.” Everyday Feminism, September 25, 2014. https://everydayfeminism.com/2014/09/equality-is-not-enough/.

Szmigin, Isabelle, and Robert Rutherford. “Shared Value and the Impartial Spectator Test.” *Journal of Business Ethics* 114, no. 1 (October 2012): 171–82. https://doi.org/10.1007/s10551-012-1335-1.

Tan, Pang-Ning, Anuj Karpatne, Vipin Kumar, and Michael Steinbach. *Introduction to Data Mining*. Harlow: Pearson, 2020.

Tommasi, Tatiana, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. “A Deeper Look at Dataset Bias.” *Domain Adaptation in Computer Vision Applications Advances in Computer Vision and Pattern Recognition*, 2017, 37–55. https://doi.org/10.1007/978-3-319-58347-1_2.

Torralba, Antonio, and Alexei A. Efros. “Unbiased Look at Dataset Bias.” *Cvpr 2011*, 2011. https://doi.org/10.1109/cvpr.2011.5995347.

"Traffic Safety Facts: A Brief Statistical Summary." *U.S Department of Transportation*, 2015. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115

Turner-Lee, Nicol, Paul Resnick, and Genie Barton. "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms." Brookings. Brookings, October 25, 2019. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

Turque, Bill. "'Creative ... Motivating' and Fired." The Washington Post. WP Company, March 6, 2012. https://www.washingtonpost.com/local/education/creative--motivating-and-fired/2012/02/04/gIQAwzZpvR_story.html.

"Universal Declaration of Human Rights." United Nations. United Nations. Accessed March 30, 2020. https://www.un.org/en/universal-declaration-human-rights/.

Weatherford, Roy C. "Discussions Defining the Least Advantaged." *Equality and Liberty*, 1991, 37–45. https://doi.org/10.1007/978-1-349-21763-2_4.

Wells, Thomas. "Free Choice Is Informed Choice: Why We Need Ethical Warning Labels on Animal Products." ABC Religion & Ethics. Australian Broadcasting Corporation, June 2, 2016. https://www.abc.net.au/religion/free-choice-is-informed-choice-why-we-need-ethical-warning-label/10096936.

"What Does the GDPR Mean for Business and Consumer Technology Users." GDPR.eu, February 13, 2019. https://gdpr.eu/what-the-regulation-means-for-everyday-internet-user/.

"What Is GNSS?" European Global Navigation Satellite Systems Agency, August 29, 2017. https://www.gsa.europa.eu/european-gnss/what-gnss.

"With Great Power Comes Great Responsibility." Quote Investigator, May 28, 2019. https://quoteinvestigator.com/2015/07/23/great-power/.

Yang, Andrew. "Hold Pharmaceutical Companies Accountable - Yang2020 - Andrew Yang for President." Yang2020. Accessed March 29, 2020. https://www.yang2020.com/policies/holding-pharmaceutical-companies-accountable/.

Yu, Joe. "Assessing Fairness in COMPAS: Impossibility Theory, Calibration, and Redlining." Medium. Medium, March 19, 2019. https://medium.com/@qy002/assessing-fairness-in-compas-impossibility-theory-calibration-and-redlining-2d32a8d6f9af.

Zehr, Howard. *The Little Book of Restorative Justice*. Vancouver, B.C.: Langara College, 2016.