Transparent Unfairness: An Approach to Investigating Machine Learning Bias

Ryan Daher

A THESIS Submitted in partial fulfillment of the requirements for the degree of BACHELOR OF SCIENCE in Data and Business Analytics.

Thesis Advisor: Manoel Fernando Gadi
IE School of Science and Technology
Madrid, Spain

May 06, 2022

Transparent Unfairness: An Approach to Investigating Machine Learning Bias

Ryan Daher

Abstract

According to research vice president of Gartner, Chris Howard, “four years ago, AI implementation was rare, only 10 percent of survey respondents reported that their enterprises had deployed AI or would do so shortly. For 2019, that number has leaped to 37 percent — a 270 percent increase in four years” (Gartner, 2019). With the immense advancements in the complexity and subsequent performance of models, Artificial Intelligence (AI), has entered almost every industry today. While the benefits seem abundant, enabling organizations to drastically increase efficiency and reduce cost, societal drawbacks are starting to emerge. With the increase in performance introduced by modern AI systems comes obscurity, creating a “black box” which is void of explainability. As such instances of discrimination and unfairness have emerged from the use of these uninterpretable systems numerous policies have followed to regulate its use and limit its scope. This paper assimilates through research conducted by numerous leading scientists, professors, organizations, and governments and presents a cohesive framework used to identify areas of unfairness, understand their sources, determine those most affected, and recommend subsequent isolated acts to rectify discrimination. In our quest to do so, we explore and employ numerous machine learning techniques, create statistical fairness tests derived from scientific research, and utilize the power of explainability techniques through their experimentation on real-life datasets.

Contents

1 Introduction

In a bid to speed up their recruitment process, in 2014, Amazon Inc tasked a team of leading data scientists to automate candidate screenings using Natural Language Processing (NLP), and Machine learning (ML), rating candidates on a 5-star scale based on their resemblance to previously accepted candidates. Within just 2 years it was apparent that the model's prediction for technical jobs was greatly hindering women's ability to obtain a job at the company, favoring men in the process (Kodiyan, 2019). Upon further investigation, the Amazon team discovered a great imbalance between the number and acceptance rate of male and female applications in the training data leading to a model that incorrectly associated women with negative prospects. To further exacerbate the bias, candidates in the "extreme" such as those part of all-girl schools and universities, or who joined a women's club were penalized even more heavily leaving them with little to no chance of acceptance (Kodiyan, 2019). While the model's life was very short in deployment, its impact was drastic and showed a spotlight on the complicated limbo of ethical artificial intelligence, and sparking debate in the data industry.

Ever since Warren McCulloch and Walter Pitts first opened the subject of Artificial Neural Networks (ANNs), the drive to create human-like computer systems has garnered unprecedented traction, with its applications seeping into almost every industry.[1] As with biological brains, ANNs are composed of neurons, connecting like the synapses of the brain. As information enters the neural system, it passes through the various network layers, and neurons within each, transforming as it traverses connective edges. By the time the information dissipates to the end of the network, the computations are aggregated and an output is computed. While the process is similar to the biological brains we use, these computerized systems can be modified in unhuman-like ways, allowing ANNs to learn at unprecedented rates. In addition, the network structure and composition can be modified to become more accurate and consistent within particular tasks. While these advantages allow networks to perform at exceptionally high levels, they come with the same obscurities of our decisions. What exactly happens when you make a quick, instinctual decision? What factors do you consider when making that decision? Was it possibly biased due to previous experiences and engagements, potentially resulting in unfavorable actions?

Due to the rapid integration of these automated systems, policies and legislation

[1] McCulloch and Pitts submitted the first mathematical model of a Neural Network in their 1943 paper "A Logical Calculus of the ideas Imminent in Nervous Activity."

have begun emerging, seeking to identify and limit unintended human consequences and protect already marginalized communities. As AI comes with a heavy reliance on data, has a high level of complexity, and is generally opaque, its uses have begun intersecting with several human rights legislation. Certain algorithms in use today are in direct interference with several articles of the EU Charter of Fundamental Rights such as (Article 7) Respect for private and family life, (Article 8) protection of personal data, (Article 11) freedom of information and expression, (Article 21) nondiscrimination, (Article 23) equality between men and women, among other clauses (European Union, 2012). In addition to the European Union, the US Department of Defense has officially adopted its Ethical Principles for Artificial Intelligence promising to only use responsible, equitable, traceable, reliable, and governable AI systems within their practice (U.S. Department of Defense, 2020). Based on proposed legislation, it is evident that many western governments and their organizations are moving towards AI regulation, with a major focus on ethicality and transparency.

This paper attempts to break down the complex web of research, technology, and policy of ethical artificial intelligence, and provide a cohesive framework used to investigate, assess, and audit existing ML/AI decision-making systems, alerting data scientists of potential biases and discriminatory effects within their data ecosystem. In this way, they may anticipate unfairness, track their root causes, and reroute at-risk candidates to a human second opinion, among other fairness efforts. In addition, this paper will go through a series of real-life cases examining their biases, applying statistical tests, and deriving a final framework that encompasses and connects today's multidisciplinary research into ethical AI.

2 State of Matter

2.1 Artificial Intelligence Foundations

Over the past decade, immense advancements have been made surrounding automated processes, with Artificial Intelligence at the forefront of these technologies. To summarize the definition of Artificial Intelligence we can use the European Commission's 2018 definition to define AI as the "systems that display intelligent behavior by analyzing their environment and taking action – with some degree of autonomy – to achieve specific goals" (Boucher, 2020). As with this simplistic definition, it is important to note the phrasing used, describing how these "intelligent" systems compute outcomes based on the environment they are placed in and the patterns they observe.

While they may be showcasing signs of "intelligence" it is clear that the information derives from the environment it is placed in, and the patterns it is taught to observe. To further understand this concept it is important to look back at the chronology of AI technology.

As published in the European Commissions 2020 paper "Artificial intelligence: How does it work, why does it matter, and what can we do about it?", the author, Philip Boucher, describes a three-phase breakdown of AI progression, starting from Symbolic AI, to data-driven Machine Learning (ML), and finally speculative future movements into general AI (Boucher, 2020). As a policy analyst at the European Parliament's Panel for the Future of Sciences and Technology (STOA), Boucher presents a reliable entry point into the field of AI. In addition, the publication is made in 2020, providing a recent analysis of historical advancements, however, it is geared towards a European audience and will be used purely for its historical contents.

As explained by the paper, in the first phase of symbolic AI, human subject-matter experts create rule-based algorithms that can be followed automatically by a machine, determining confidence levels regarding situations, and replicating closely the work of the expert, resulting in similar outcomes in just a fraction of the time. Such systems seem to thrive in stable environments but seem to break down when more variation is introduced.

To increase flexibility in algorithm performance, the second wave of AI centered around capturing and using data to train highly complex, robust, and flexible models capable of making assessments over a wide range of industries, and with dynamism in the face of variability. The automation of these processes moves away from the human experts, as seen in the first wave, and instead relies on the automation of learning, allowing algorithms to observe large amounts of data simultaneously, identify crucial patterns quickly, and come to accurate predictions. Among these advanced systems was the creation of the Artificial Neural Network (ANN) which is constructed to mimic the functions of the anatomical brain. Similarly, these ANNs are fed inputs that are converted into signals, trickling down through a structure of artificial neurons, processing tremendous amounts of computations, and resulting in an output as a response. By adding further complexity, increasing the depth of our ANNs, and using certain evolutionary concepts to automate learning, we merge into Deep Learning which continuously and gradually improves itself to optimize its performance at rates unseen before.

Speculating into the future of ANNs the shift is seemingly going from "narrow" AI which serves very specific purposes, to Artificial General Intelligence (AGI) which in turn focuses on achieving exceptional results for a wider range of tasks. Such technology is still speculative, however, allows us a glimpse into the future of the technology as well as the potential risks that may come with it. In either case, as these machine systems continue to increase in complexity, and are given greater freedom over the tasks they work on, it is crucial to implement explanatory procedures to avoid giving machines too much ability, without the proper understanding of their decisions, potentially leading to catastrophic results, with no subsequent understanding.

2.2 Barriers for Widespread Adoption

Having covered some of the base timelines of Artificial Intelligence, it is now interesting to understand the applications of such systems, the benefits that they have brought to industries, as well as the observed challenges that have come from their implementation. Looking at the paper "Opportunities of Artificial Intelligence" (Eager et al., 2020) presented by the European Commission's Policy Department for Economic, Scientific, and Quality of Life Policies, we are presented with a recent European perspective on the issue through the analysis of the common uses of AI within European Industry. While the perspective may be biased towards one region, the widespread adoption of automated technology within Europe has numerous parallels drawn to other countries and regions. As described by the authors, the main uses of AI have been placed into one of two categories (Eager et al., 2020):

1. "Applications that enhance the performance and efficiency of processes through mechanisms such as intelligent monitoring, optimisation and control."

2. "Applications that enhance human-machine collaboration."

Within both of these areas, AI has been expected to provide a high level of impact within different industries, greatly reducing costs and waste, improving standards of production, generating product customization, and overall increasing customer and staff experience. In addition, the culmination of these efforts has brought further improvements in employee safety, as well as significant improvements in labor productivity anticipated to increase by between 11% and 37% by the year 2035 (Eager et al., 2020). In a concrete example of AI effectiveness, we can look at the identification of the antibiotic "Halicin" which was reported in February 2020 using machine learning concepts. By utilizing molecular data of antibacterial properties against E. Coli, Massachusetts Institute of Technology (MIT) researchers were able to screen thousands of new molecules, isolating numerous antibacterial candidates with prospective effectiveness against E. Coli[2] (Eager et al., 2020). Through the use of Machine Learning, algorithmic learning procedures were applied to a very specific medical case, allowing for unprecedented progress toward an E. Coli vaccine. It is clear that the advantages that have been observed over the years are certainly plentiful and are seeping into almost every industry, replacing and automating certain lagging systems, and greatly augmenting human capabilities.

While the widespread adoption of AI seems rather evident it is important to note the drawbacks of such powerful technology. In today's AI landscape, there are four major transparency concerns as outlined by the European Parliamentary Research Service (EPRS). Among them is the absence of explainability of certain advanced ML algorithms, referred to as "black-box" models (Boucher, 2020). As further described, the shift from using rule-based algorithms to data-driven machine learning presented a leap in dynamism, speed, and performance, it also introduce opaqueness, as the model executes tremendous amounts of computations, training itself to identify critical patterns and optimize its performance. While the quality of automated decisions may be high, we are left at a disadvantage trying to understand the precise factors that affected these decisions. In highly sensitive areas of law, finance, healthcare, etc... it is highly critical that we understand the systems we use to prevent disastrous outcomes that we can't comprehend.

In addition, certain ethical and legal challenges have been outlined by the Policy Department for Economic, Scientific, and Quality of Life Policies such as the right to privacy and data protection, transparency and accountability of models, no discrimination, and AI security (European Commission and Directorate-General for Communications Networks, Content and Technology, 2019). It is clear that while AI is here to stay, transparency is at the forefront of ethical AI research and discussion. To bring us into its perspective and understand the need for such policy, we can look at the study conducted by the independent, nonprofit newsroom, ProPublica, of the COMPAS Recidivism Algorithm which sought to predict the propensity of a criminal to commit a re-offence, and which is published by the Washington Courts Directory (Larson et al., 2016). Comparing the predictions made on over 10,000 defendants in a specific Florida county with the actual rates of recidivism two years later, it was clear that the model was blatantly discriminating against people of color. It was evident that black defendants were being over labeled as being high-risk offenders

[2] The same model was used to screen over 100 million molecules to discover new antibacterial compounds.

even in cases when they did not go on to commit more crimes, in addition, the model underestimated the level of risk posed by the white defendants often giving them a pass in times when they did re-offend (Larson et al., 2016). Not only did black defendants have a disproportionate level of recidivism classifications, but they were also labeled as being potentially more dangerous than white defendants who had similar backgrounds of crime. While there is certainly plenty of benefits in predicting re-offense likelihood, the impact of a single mistake can have tremendous implications and long-lasting consequences. As governments and companies continue to compete for global AI dominance, tradeoffs will need to be made between the level of privacy, security, and transparency accepted within models, and the level of performance the systems bring to the table.

2.3 Bias, Fairness, and Discrimination

This paper will isolate and dive deeper into the topic of algorithmic biases, their negative impacts, as well as their root causes which have seeped into several critical industries. In her paper, "A Survey on Bias and Fairness in Machine Learning" (Mehrabi et al., 2019), Ninareh Mehrabi, a Ph.D. candidate at the Information Sciences Institute at the University of Southern California, describes the results of her survey on practical machine learning biases. With numerous publications regarding algorithmic fairness within Natural Language Processing (NLP) and Machine Learning (ML), Mehrabi explores how biases have affected certain real-world scenarios, digging deeper into ways that have entered, and creating a taxonomy on the findings. As defined by Mehrabi, decision-making fairness "is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired Characteristics. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people" (Mehrabi et al., 2019). Examples of such unfair algorithms can be seen in the aforementioned COMPAS case in which African American defendants were penalized more heavily than their Caucasian counterparts for the same misdemeanors (Larson et al., 2016). In addition, looking at Beauty.AI, an automated beauty pageant judge favored lighter-skinned candidates, making it significantly more difficult for darker-skinned contestants to compete[3] (Mehrabi et al., 2019). Mehrabi looks over a series of real-life bias examples, classifying them as either arising from the data or the algorithm itself. In the first case, the data is littered with direct or indirect biases; a perfect model would simply seek to reinstate the observed

[3] The Beauty.AI issue boiled down to a lack of minority representation in the dataset. Out of 44 winners, only one had dark skin.

biases as it assumes them to be the "truth", negatively impacting already marginalized communities. In the latter case, the data may be void of the most severe biases, however, nuances in the machine learning processes may inhibit them from making fair decisions.

Based on the intensive research into biases conducted by the author, a series of 23 sources of data bias have been identified including but not limited to (Mehrabi et al., 2019):

1. Historical Bias: Already existing biases are found throughout data, even when proper sampling is conducted. Examples may include a historically lower ratio of female CEOs in Fortune 500 companies.[4]

2. Representation Bias: Observed disparities between the representation of different groups within a dataset such as using only Caucasian people to train an image classifier.

3. Evaluation Bias: The improper evaluation of data through a sole data labeler, biased labels, etc...

4. Aggregation Bias: When conclusions are drawn for a subgroup based on the observations collected from a different subgroup, leading to false assumptions and faulty results.

Among either of the cases listed by Mehrabi, it is evident that a significant proportion of biases that may enter the model would come from the data itself either through natural disparities observed between different groups, the data collection technique that may be flawed, the labeling and classification of the training data, as well as the further processing of the data. In the end, the model is only capable of learning from the provided data; with disparities and biases hidden within it, we can anticipate a similar level of disparity within our model outputs. Given that the results of these biases introduce certain levels of unfairness, it is crucial to also generate an understanding of unfairness descriptions.

As outlined by the same bias and fairness survey, there are 6 key types of discrimination:

1. Direct Discrimination: When certain "sensitive" attributes result in direct unbalanced results resulting in a model that predicts based on such protected variables.

[4] As of 2021, female CEOs lead only around 8% of Fortune 500 companies.

2. Indirect Discrimination: Stemming from the idea of bias proxies, indirect discrimination results in unbalanced outcomes for certain groups, even when the model only considers non-sensitive attributes. Such non-sensitive attributes may still be correlated to their sensitive counterparts, acting as a proxy and instigating the same bias.

3. Systemic Discrimination: A form of discrimination stemming from the policies or customs acquired by a group. As the group favors people of similar characteristics, new, potentially different candidates may be excluded or hindered from advancing based on these policies.

4. Statistical Discrimination: A form of discrimination that comes from the grouping of individuals based on their specific set of characteristics, in particular those favorable in a situation. Again, these characteristics may not need to be targeting "sensitive" characteristics, but may certainly serve as a proxy for them.

5. Explainable Discrimination: A gray area of discrimination in which the unequal results between two sensitive groups, differences may be attributed to a non-sensitive factor such as average working hours, income level, etc...

6. Unexplainable Discrimination. Contrary to explainable discrimination, this form contains the same levels of disparity between sensitive groups but lacks the explanations leading to an explicitly discriminatory model.

While the discriminatory themes are plentiful, it seems evident that the major focus is not on eliminating biases, as they will certainly exist in numerous forms; rather it is to identify the root causes of disparities between groups and find ways to rectify or minimize the resulting discrimination that arises.

By conducting and combining research published by the Association for Computing Machinery (ACM) during their annual Conference on Fairness, Accountability, and Transparency (ACM FAccT), authors Ben Hutchinson and Margaret Mitchell, go through 50 years of fairness taxonomy and testing, examining their social implications, and building a series of modern terminologies for efficient tests (Hutchinson and Mitchell, 2019). As Mehrabi further scrutinizes their findings, she presents a list of ten prominent fairness definitions as seen throughout the research including but not limited to (Mehrabi et al., 2019):

1. Equalized Odds: All sensitive groups should have equal rates for true positives and false positives.

2. Equal Opportunity: All sensitive groups should have equal true positive rates.

3. Demographic Parity: The likelihood of a certain class should be the same regardless of the group.

4. Treatment Equality: All sensitive groups receive the same ratio of false negatives and false positives.

5. Counterfactual Fairness: A decision should be conducted in the same way if an individual were to be a part of either sensitive group.

6. Conditional Statistical Parity: People in sensitive groups should have an equal probability of being classified as a positive outcome, given a legitimate set of factors.

Drawing upon these findings, the paper "The Impossibility Theorem of Machine Fairness. A Causal Perspective." presented by Elemental Cognition Research Scientist, Kailash Karthik Saravanakumar under the supervision of Columbia University, goes into more depth about the fairness definitions presenting a broken down summary of three core fairness metrics which serve as the most significant and difficult of the tests (Saravanakumar, 2020). These metrics are drawn upon the research of Mehrabi, among others, and consider an additional metric, predictive parity, which takes into account the sensitivity of the model. If this condition is satisfied, we are alerted that the "probability of correctness of a prediction is the same for all values of the sensitive attribute" (Saravanakumar, 2020) indicating similar levels of distribution as we would expect in our original dataset.

Further outlined in these papers are similar mathematical formulas used to compute such metrics which can be derived through the results of confusion matrices. By converting these concepts into codable logic, we can further apply these definitions to different datasets and models in a bid to observe fairness impacts. It is important to note, however, that the research continues to show how it may be near impossible to satisfy all definitions simultaneously due to "the inherent incompatibility of two conditions: calibration and balancing the positive and negative classes." Due to this, the author argues that it is important to relate such fairness definitions to the case at hand, ensuring the right use of a definition to investigate particular instances.

To complement and compare the research conducted by Mehrabi, we can take a look at the work of Dr. Frederik Zuiderveen Borgesius, a distinct law professor at Radboud University, and a researcher at the University of Amsterdam, with research publications regularly presented as part of national and international conferences. In his study, “Discrimination, Artificial Intelligence, and Algorithmic Decision-Making” presented under the Council of Europe, Professor Borgesius goes into depth about the different ways biases may seep into datasets and machine learning outcomes, in particular, how such biases may lead to discrimination. In contrast to Mehrabi’s work, this study categorizes the bias entries into one of six categories including training data biases, target classification, data collection procedures, feature selection, proxies, and intentional discrimination (Borgesius, 2018). Comparing this classification with Mehrabi’s work, we see a cohesive image of biases entering models either through the data itself or through the selection and execution of model processes. It is also apparent that the biases may present themselves in a direct form in which sensitive characteristics are directly responsible for differences in outcome, or indirectly when such characteristics may not be directly present but rather hidden within proxy variables.

Finally, to consider an alternative view through the works of the Center for Data Ethics and Innovation (CDEI), a UK government expert body seeking to promote the use of trustworthy data and AI, we are informed of their classification of four main bias types including historical bias, data selection bias, and algorithmic design bias (UK Government Centre for Data Ethics and Innovation, 2020). In addition, this body has introduced the further idea of human oversight in which a human supervisor may influence model outcomes through the implementation, deployment, retraining, or other processes throughout the training and execution stages (UK Government Centre for Data Ethics and Innovation, 2020). While the overall themes seem consistent with the descriptions provided by the previous two works, the CDEI also stressed the importance of monitoring bias amplification; as discriminatory models may become more severe in their disparities as it incrementally retrains itself on new biased data, further amplifying disparities (UK Government Centre for Data Ethics and Innovation, 2020).

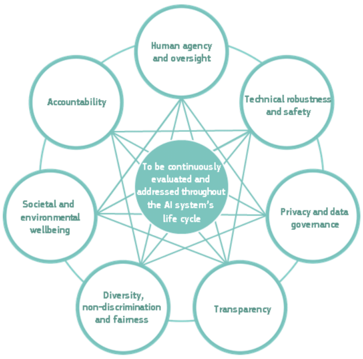
Sifting through these different perspectives, we are provided with an extensive investigation into bias, we have considered larger themes of bias, unfairness, and discrimination based on independent research conducted by authoritative figures and organizations across three large western bodies. By combining the different ideas into overarching categories, we are presented with an automated decision-making timeline that consists of data collection and preprocessing, model selection and training, and model deployment. We have observed how biases are able to enter, directly and indirectly, within each stage in the process. As these biases may enter in manners in which we may have very little influence, such as income inequality, or systemic injustices, it is then a more imperative responsibility for the data scientist to identify the existence of disparities within models. To do so, three core ideas have been common in the literature; the idea of representation, ability, and performance disparities. By categorizing the common fairness metrics as outlined by Mehrabi, Saravanakumar, and other researchers, into these categories, we can conceptualize a process of identifying different sources of unfairness. Looking at these sources in more depth, we can investigate deeper into their root causes, and recommend feasible actions to mitigate associated risks. In this way entities using automated decision making may initially assume the existence of biases within their data/model ecosystem and seek to expose and report them, ensuring full accountability, and transparency amongst stakeholders.

2.4 Legal Outlook

Not only is it an organization's moral responsibility to audit its own machines and ensure transparency, but government regulations have also already begun to emerge to regulate the use of artificial intelligence given the societal downsides that have been documented. Leading the space of AI regulation is the European Union whose research has been the most abundant, complete, and impactful thus far. With their primary set of privacy and security regulations, the General Data Protection Regulations (GDPR), establishes a set of rules applicable to countries around the world that collect data from members of the EU. With its establishment in 2018, the GDPR has promised fines against companies that disregard them, placing certain boundaries of accountability on organizations that process data (European Union, 2016). Within its documentation, Article 5 states its first primary principle of lawful data processing, stating that personal data should be "processed lawfully, fairly and in a transparent manner in relation to the data subject" (European Union, 2016). Given the black-box nature of the aforementioned models, a problem emerges in the ability to create highly accurate systems while still providing transparency to the "data subject". In addition, it places the responsibility of fairness on the controller rather than the government or subjects.

Subsequent work conducted by over 350 organizations, and under the wing of the European Union, culminated in the High-Level Expert Group on Artificial Intelligence (HLEG) "Ethics Guidelines for Trustworthy AI" which states a series of requirements for "Trustworthy AI". Among the requirements are human agency and oversight; transparency; diversity, non-discrimination, and fairness; societal and environmen-

Figure 1: Trustworthy AI Requirements (European Commission and Directorate-General for Communications Networks, Content and Technology, 2019).


Human agency and oversight
Accountability
Technical robustness and safety
To be continuously evaluated and addressed throughout the AI system's life cycle
Societal and environmental wellbeing
Privacy and data governance
Diversity, non-discrimination and fairness
Transparency

tal wellbeing; and accountability (European Commission and Directorate-General for Communications Networks, Content and Technology, 2019). Such requirements again place responsibility on the data controller to ensure their systems are both fair, and understandable. These requirements were later merged into the European Commission's Artificial Intelligence Act of 2021 as a baseline for ensuring proper due diligence of organizations operating "high-risk" models which tend to have more severe implications on the subject such as creditworthiness, recruitment ability, law enforcement, and more (European Commission, 2021). The most severe societal themes within European policy on Artificial Intelligence seem to center around the transparency, fairness, and trustworthiness of the systems at hand, with severe consequences in cases deemed as "high risk".

In contrast to the EU, the United States has a series of commercial regulations to ensure fairness within commerce such as the Fair Trade Commission Act (FTCA), the Fair Credit Reporting Act (FCRA), and the Equal Credit Opportunity Act (ECOA). As machines continue to automate human processes, they can be directly implicated with such policies. As stated in section 5 of the FTCA, "unfair or deceptive acts or practices in or affecting commerce" (United States Federal Trade Commission, 2006) are prohibited, which places the deployment of unfair biases within that category as they have the potential to cause significant damages to US citizens. In addition, the FCRA aims to regulate the way credit reporting agencies collect and utilize consumer data, ensuring proper privacy, fairness, and accuracy (United States Federal Trade Commission, 2018); it is of equal relevance when considering algorithms that serve the purpose of determining credit, employment, insurance, among other fundamental decisions. Finally, the ECOA strictly outlaws discrimination on the basis of "race, color, religion, national origin, sex, marital status, age, receipt of public assistance, or good faith exercise of any rights under the Consumer Credit Protection Act" (United States Federal Trade Commission, 1974). While more indirect in comparison to EU regulation on AI, the concepts are clear; under no circumstances will discrimination be permitted within the US or EU, whether on the part of humans, or machines.

To culminate the research thus far, we have introduced the concepts of machine learning, the reasons for its widespread application throughout various industries, the rise of accuracy and subsequent opaqueness of certain advanced techniques, the downsides which have risen from this opaqueness, the damages that they have caused in particular with relation to bias, fairness, and discrimination, and finally the subsequent legal guidelines that have followed suit in a bid to regulate what is now being observed as the downsides of using black-box models. It is clear that companies, more specifically their data team, will need to implement procedures to identify, report, and rectify their automated techniques before widespread implementation, with the most severe implications surrounding "high-risk" applications which have larger implications on its subjects. It is now important to shift focus to the actionable frameworks and principles enacted by companies and organizations in order to comply with the legal policies and promote trustworthy algorithms.

2.5 Corporate Outlook

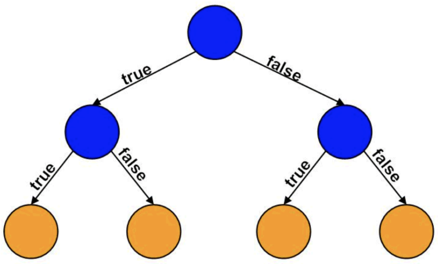
CEO of Alphabet, Sundar Pichai, published a series of principles established by Google to display its efforts in being prudent with its AI development. Among their core beliefs is that AI should be socially beneficial, steer away from enforcing or creating unfair bias, be accountable and transparent, and be built with privacy at the forefront (Pichai, 2018). To complement these efforts, Microsoft has published their approach on "Responsible AI" in which they outline the core principles that they will enact in their practice. Highlighted in these principles are the concepts of "fairness", "transparency", and "accountability" (Microsoft, 2022). A common theme between these companies is the need for explainable systems rather than pure "black-box" models which have until now been the main player in the AI scene. To further add to this analysis, looking back at the HLEG "Ethics Guidelines for Trustworthy AI", the authors propose a series of technical methods that can be practically implemented to ensure trustworthiness within their systems. One of the most crucial is "explanation methods" in which the HLEG states that in order for a system to be deemed trustworthy: "we must be able to understand why it behaved a certain way and why it provided a given interpretation" (European Commission and Directorate-General for

Communications Networks, Content and Technology, 2019). However they go on to state that "today, this is still an open challenge for AI systems based on neural networks. Training processes with neural nets can result in network parameters set to numerical values that are difficult to correlate with results" (European Commission and Directorate-General for Communications Networks, Content and Technology, 2019), bringing back the complicated limbo between model performance, explainability, and neural network opaqueness.

2.6 Introduction to Explainable Artificial Intelligence

To remedy this contradiction, immense advancements have been made in the field of Explainable Artificial Intelligence (XAI) which has the sole purpose of explaining artificial intelligence systems. Through the pursuit of this objective, we make a fundamental shift in the way we interact with our automated systems, placing a major emphasis not only on the outcomes of these high-performing models but also on the factors that contributed to them. The Natural Language Processing and Chinese Computing (NLPCC) presents its perspective on XAI through its publication "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges" (Xu et al., 2019). In the paper, the authors describe how AI has traditionally been explainable with the first automated systems following a distinct set of rules in which outcomes can be explained by applying back the rules (Xu et al., 2019). An example of such an explainable structure is the decision tree which checks candidates against particular rules, and moves across the tree accordingly, resulting in a final classification.

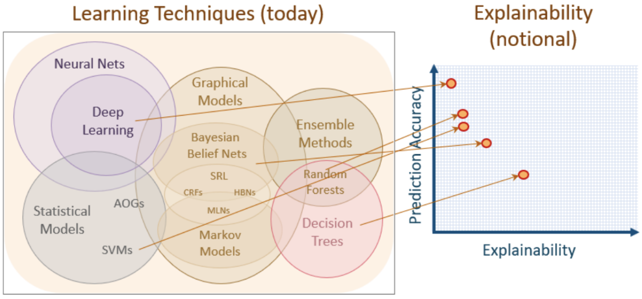
Figure 2: Decision Tree (Systems Applications and Products in data processing, 2019).



However, as we've progressed towards modern Deep Neural Networks (DNNs), we have introduced opaqueness making them unexplainable by "the neural network itself, nor by an external explanatory component, and not even by the developer of the system" (Xu et al., 2019). As illustrated by the Defense Advanced Research

Projects Agency (DARPA), a research and development organization under the US Department of Defense, the explainability of a model is in inverse correlation to the performance of the model which can be observed below.

Figure 3: Performance Explainability Tradeoff (Turek, 2020).

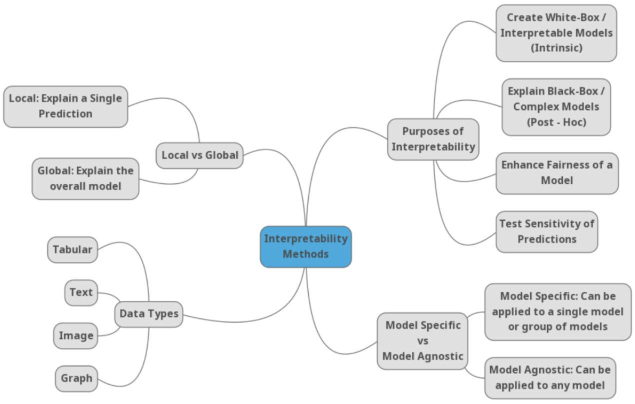


While certain simplistic models such as decision trees have high levels of explainability, their performance lags behind more opaque models. However, in the commercial and governmental use of automated decision-making, it is crucial to also have high-performing models.

To bridge this gap two main areas of XAI have risen, the first concerned with enabling transparent design within models, and the second focused on post hoc explanations, inferring the explanations that lead to a result, allowing for a higher level understanding into the decision making process of the opaque system. Presented as part of the Entropy international and interdisciplinary peer-reviewed open access journal, machine learning engineer Pantelis Linardatos proposes a taxonomy of four main purposes of explainability: understanding black-box models, creating white-box models, promoting fairness, and analyzing model prediction sensitivity (Linardatos et al., 2021), seemingly following the same outlined framework of the NLPCC (Xu et al., 2019). In both cases, efforts into XAI are focused on the translation of opaque systems, and the creation of effective transparent counterparts. The Entropy publication provides an additional branch of fairness and sensitivity which is in line with legal frameworks and societal studies mentioned previously.

Digging deeper into the category of black-box interpretability, we narrow it down to model agnostic interpretability methods, which can be used to infer explanations of any model. A primary component of prominent explainability research has been additive feature attribution methods which seek to approximate the weight of each feature based on the outputs of the models, creating in turn, a post hoc explainer of the black-box model. Examples of such features are explained in the framework "A Unified Approach to Interpreting Model Predictions", conducted by Microsoft Senior

Figure 4: Interpretabilitability Methods (Linardatos et al., 2021).


Interpretability Methods
Local vs Global
Local: Explain a Single Prediction
Global: Explain the overall model
Data Types
Tabular
Text
Image
Graph
Purposes of Interpretability
Create White-Box / Interpretable Models (Intrinsic)
Explain Black-Box / Complex Models (Post - Hoc)
Enhance Fairness of a Model
Test Sensitivity of Predictions
Model Specific vs Model Agnostic
Model Specific: Can be applied to a single model or group of models
Model Agnostic: Can be applied to any model

Researcher Scott Lundberg, in which he outlines the "LIME", "DeepLIFT", "Layer-Wise Relevance Propagation", and "Classic Shapley Value Estimation" (Lundberg and Lee, 2017). He then describes how the Shapley value estimation method is the only one that has maintained properties of: "local accuracy, missingness, and consistency" (Lundberg and Lee, 2017) making it a powerful estimator for feature importance within black-box models. In addition, the research conducted was not only more accurate in "differentiating among the different output classes, but also in terms of better aligning with human intuition when compared to many other existing methods" (Lundberg and Lee, 2017). Based on this analysis, it seems evident that using the Shapley value estimation in the context of machine learning can provide tremendous benefit in black-box explainability.

Putting together the complicated web of research into the societal consequences of deep learning, human and algorithmic biases, the legal policies seeking to protect the individual, and the apparent limbo between model interpretability and performance, we can now understand the deep interconnectivity between transparency and discrimination. It is evident that organizations may no longer collect and process data without proper self-auditing of their own machines.

2.7 Literature Remarks

Congregating the various theoretical concepts into a practical framework, we can follow the "theoretical lens of a 'sense-plan-act' cycle", as described by the HLEG framework (European Commission and Directorate-General for Communications Networks, Content and Technology, 2019). Applying this concept to the problem of ML fairness, we can break down three core steps in providing robust, and responsible artificial intelligence: Identify, Understand, and Act (IUA).

1. Identify: The process of exposing direct or indirect biases within a dataset and/or model.
2. Understand: The process of isolating impactful scenarios and obtaining transparent explanations for outcomes.
3. Act: The process of reporting and rectifying identified disparities within the automated system.

By understanding the philosophical forms of unfairness as defined by our review of the literature and categorizing our prominent fairness metrics into the overarching categories of representation, ability, and performance, we can establish a series of tests to "identify" levels of disparities between sensitive groups at different levels. Merging these findings with the explainability of our models through the use of white-box models, or Shapley value estimation for black-box models, we can dig deeper into the model's predictions, "understanding" how classifications were made, and how they varied from the natural dataset exposing both natural biases as well as added model differences. Finally, by probing further into levels of misclassification, in particular looking at negative outcomes, we can isolate groups most at risk and set up a series of "actions" that can be taken to mitigate the effects. Given this three-step framework which combines societal, legal, and technical considerations, the paper will then go through a series of cases, and examine the proposed framework.

3 Methodology

3.1 Data Selection

It is clear that based on the current research, it is imperative to account for many philosophical, legal, and technical considerations when moving towards transparent and fair machine learning. While the literature review allowed us to explore many of the most pressing discussions and research conducted in the field of ethical machine learning, it is now important to take a look at these concepts in practice through the use of several datasets. In this way, we can apply the concepts and observe, first-hand, the existence of numerous biases, the magnitude of resulting unfairness, and the root causes of such discrimination.

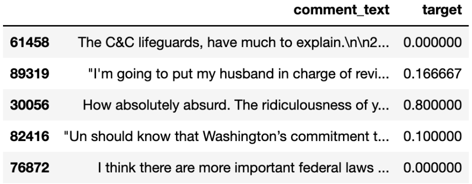
The first of these datasets is presented by Civil Comments, a commenting plugin that was used by 50 independent English-language news sites from around the world between 2015 and 2017 in a bid to reduce online commenting "toxicity", or generally offensive language. Before allowing users to comment on news articles, the plugin would invite them to label the "civility" of three other randomly selected comments. When the platform shut down in 2017, the comments and their labels were made publicly available, with all usernames removed (Jigsaw/Conversation AI, 2019).

The dataset includes roughly two million comments generated within two years and includes metadata identifying each unique comment, its creation date, the article which it falls under, and user reactions to the comments [funny, wow, sad, like]. In addition, it includes a civility "rating" label based on the ratings given to the comment by other users (Jigsaw/Conversation AI, 2019). For the purposes of faster processing, a subset of the data will be used.

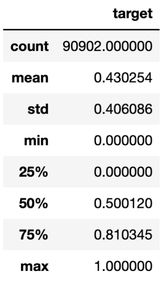
To obtain this civility attribute, each comment was randomly shown to numerous annotators who were asked to rate the comment on a scale:

- Very Toxic: "A very hateful, aggressive, or disrespectful comment that is very likely to make you leave a discussion or give up on sharing your perspective."
- Toxic: "A rude, disrespectful, or unreasonable comment that is somewhat likely to make you leave a discussion or give up on sharing your perspective."
- Not Toxic
- Hard to Say

In addition, toxicity sub-attributes could be identified such as "severe_toxicity", "obscene", "thread", "insult", "identity_attack", and "sexual_explicit". The ratings were then compiled together to represent the overall toxicity of a comment (Jigsaw/Conversation AI, 2019). To add further dimensionality to the dataset, independent annotators were asked to identify specific identity attributes about the comments such as the mentioned genders, sexual orientation, race, religion, and ability. This paper will use only the comments and their target. It is important to note that while the target is portrayed as a probability, we will use a threshold of 0.7 to classify a comment as being toxic.



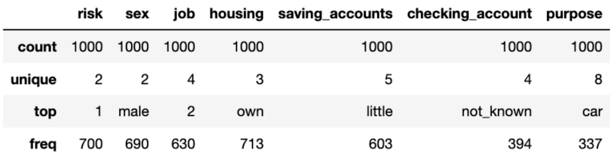
	comment_text	target
61458	The C&C lifeguards, have much to explain.\n\n2...	0.000000
89319	"I'm going to put my husband in charge of revi...	0.166667
30056	How absolutely absurd. The ridiculousness of y...	0.800000
82416	"Un should know that Washington's commitment t...	0.100000
76872	I think there are more important federal laws ...	0.000000



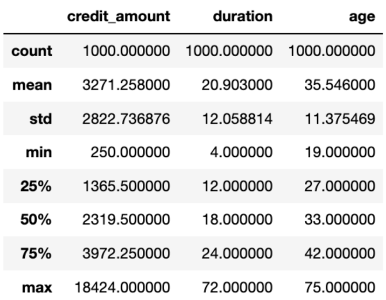
	target
count	90902.000000
mean	0.430254
std	0.406086
min	0.000000
25%	0.000000
50%	0.500120
75%	0.810345
max	1.000000

Given the public nature of this dataset, and the free ability of anonymous users to write comments, we will have tremendous flexibility in breaking down this dataset to understand how biases have managed to enter. Utilizing machine learning algorithms we will then create a classifier to predict toxicity labels based on comment text. Looking at the results we will evaluate both the positive and negative impacts of using such a system. Diving deeper, we will then examine the decision-making process itself, uncovering potential, unintended sources of bias that our model could have picked up, and reexamining the implications of such effects on a larger scale application. While this data includes mainly human-created content, another dataset will be used to showcase the same bias potential with other, more empirical data.

The second dataset is presented by Dr. Hans Hofmann, a professor at the University of Hamburg's Institute of Statistics and Econometrics. Presented in 1994, the dataset is definitely old in its capture and was filled with inconsistencies and symbolic attributes that are hard to decipher, however, is one of the most-used datasets and is present as part of numerous python and R packages (Hofmann, 1994). In its original form, the dataset consisted of 20 mainly categorical attributes and over 1000 entries, each representing a bank client who has taken credit, each is then classified as bearing either good or bad credit risk. To make the data more accessible, the University of California, Irvine has modified the dataset to make it readable and usable, including the removal of several obscure variables. In the end, the adapted dataset used in this paper is composed of 9 variables (Hofmann, 1994).



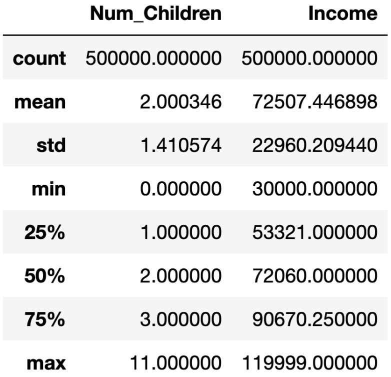
	risk	sex	job	housing	saving_accounts	checking_account	purpose
count	1000	1000	1000	1000	1000	1000	1000
unique	2	2	4	3	5	4	8
top	1	male	2	own	little	not_known	car
freq	700	690	630	713	603	394	337



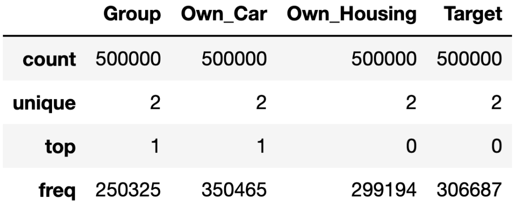
	credit_amount	duration	age
count	1000.000000	1000.000000	1000.000000
mean	3271.258000	20.903000	35.546000
std	2822.736876	12.058814	11.375469
min	250.000000	4.000000	19.000000
25%	1365.500000	12.000000	27.000000
50%	2319.500000	18.000000	33.000000
75%	3972.250000	24.000000	42.000000
max	18424.000000	72.000000	75.000000

Looking at the summary of our variables, we get an overview of several demographic attributes, financial status, bank status, and classification of risk. Using this dataset, we will explore how the decisions to classify a candidate as being high or low risk are being made by firstly observing the dataset and its natural disparities, running a model to simulate the decision making, and comparing the outcomes for both males and females, testing for any disparities in the process. Finally, we will use the power of SHAP to isolate the most impacted group and understand the reasons for their impact.

To further cement the proposed framework, we move to the third case study consisting of a synthetically constructed credit card approval database mimicking real-life credit card applicants. With 500,000 records and 6 dimensions of data, each applicant has personal information about themselves including the number of children in their household, their income, and whether or not they own a car or a house (Cook et al., 2021). In addition, the members have been placed into one of two groups: "A", and "B". We can gather a series of descriptive statistics and observe that around 40% of the applicants were approved for a card.



	Num_Children	Income
count	500000.000000	500000.000000
mean	2.000346	72507.446898
std	1.410574	22960.209440
min	0.000000	30000.000000
25%	1.000000	53321.000000
50%	2.000000	72060.000000
75%	3.000000	90670.250000
max	11.000000	119999.000000



	Group	Own_Car	Own_Housing	Target
count	500000	500000	500000	500000
unique	2	2	2	2
top	1	1	0	0
freq	250325	350465	299194	306687

With such a synthetic dataset, we have the freedom to experiment with "raw" data without the potential of exposing identities. With this data, we will once again observe natural disparities present in the dataset based on the attribute "Group", putting an added emphasis on the concepts of implicit and explicit bias. In this way, we will have explored the ideas of bias, unfairness, and discrimination as illustrated in the literature review in a practical manner, and observe its impacts on user-generated content.

3.2 Fairness Considerations

Having looked over the primary datasets that will be examined as part of this paper, it is important to isolate and consider the factors of bias and unfairness that will be used to identify such disparities. Before going into them, it is important to briefly cover the confusion matrix and the different metrics that can be extracted from it. Simply put, the confusion matrix consists of a table that is used to denote model classifications compared to their true labels. From this matrix, we can extract four key metrics: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Merging these concepts with the works of Mehrabi, and Saravanakumar we may come to a series of prominent fairness metrics present throughout today's literature that can be described in terms of confusion matrix metrics. By looking at the research intersections we had identified three key areas of interest in bias detection: representation, ability, and performance. Within each of the categories we can introduce the particular metrics brought up by the research, in particular:

- Demographic Parity: The level at which the outcome is dependent on the sensitive group. Satisfying this parity would require the Positive Rate to be the same regardless of group. It can be denoted mathematically as: $P(\hat{Y}|A=0) = P(\hat{Y}|A=1)$ (Mehrabi et al., 2019)

- Equalized Odds: The level at which the positive outcome is dependent on the sensitive group, given the accuracy of the outcome. It would thus require the TPR and FPR to be equal between the groups and can be denoted mathematically as: $P(\hat{Y}=1|A=0,Y=y) = P(\hat{Y}=1|A=1,Y=y), y\epsilon\{0,1\}$ (Mehrabi et al., 2019)

- Equal Opportunity: A variant of Equalized Odds, it indicates the level in which the positive outcome, while correct, is dependent on the sensitive group, satisfying the condition if the True Positive Rate (TPR) is the same for either group. It can be denoted mathematically as: $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$ (Mehrabi et al., 2019)

- Predictive Parity: The level at which a candidate's predicted value and true value are dependent on the sensitive group. This would mean analyzing the disparity between group model precision and can be denoted as follows: $P(Y = v|\hat{Y} = v, A = a) > P(Y = v|\hat{Y} = v, A = d)$ (Saravanakumar, 2020)

In addition, a few more common metrics will be included to further enrichen the analysis such as group representation disparity which looks at differences between the number of members of each sensitive group present in the dataset as inequalities between their representation indicates differences in opportunity for the model to learn.

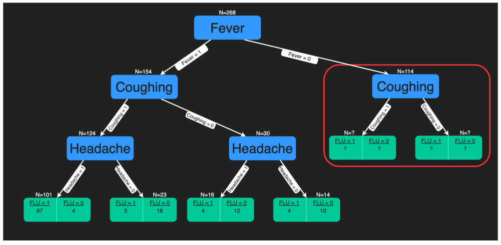
While this list and structure of these fairness metrics seem the most concrete and commonly used metrics within research, their traditional implementations have thus far been in the form of percentage outputs in which case the data scientist will need to interpret their results and determine acceptable levels of differences between group metrics. In addition, as presented by Saravanakumar, it is near impossible to satisfy all of our conditions, making the subjectivity of the data scientist at the core of the outcome, again introducing human bias to the equation (Saravanakumar, 2020). To avoid this, this paper will use these calculations and compare them in the form of a chi-squared test, assessing the relation between group scores and predictive outputs. Based on the extracted p, we can assess the level of disparity significance present within the metric disparities and in turn determine a proper course of action.

Several other statistical tests were considered for this task including the Student's T-Test, The Kruskal-Wallis Test, as well as the chi-squared test. However, as the metric results come in the form of frequencies and are comparing two groups, rather than a center measure, both the T-test and the Kruskal-Wallis test could be unreliable. Thus, the chi-squared test is the most robust in determining the statistical differences between our compared group frequencies and the expected scenario.

3.3 Algorithmic Considerations

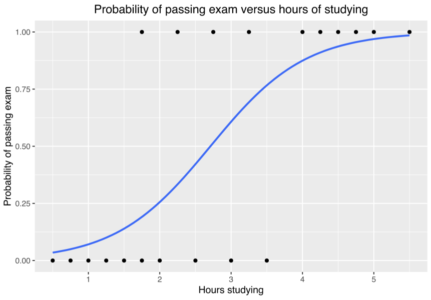
Having gone over the base datasets that will be used to test out the bias as well as the testing approach used to assess the dataset and the models, we will now outline the models that will be used throughout the cases. In this way, we will compare lessons across various models, and draw conclusions based on the similarities and differences of our findings. Among these models will include both “black-box” and “white-box” models to again investigate fairness regardless of model structure. Such algorithms include:

Figure 5: Sample Decision Tree (Hansen, 2020).



Decision Trees: A supervised learning algorithm used for both regression and classification cases that don’t involve any distribution parameters for the parameters. To obtain a prediction this algorithm infers simplistic decisions based on the provided features. In the example case in Figure 5, the algorithm predicts the flu by understanding and ranking symptoms “coughing”, and “headache”. Such trees are simplistic but yield transparent results.

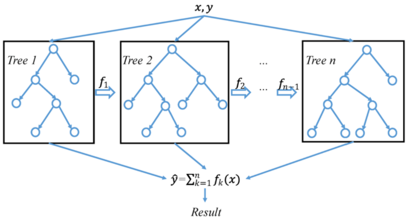
Figure 6: Sample Logistic Regression (Walia, 2021).



Logistic Regression: A popular, yet simplistic, supervised ML algorithm used to classify instances on binary conditions. Using a logistic function, we are able to model our binary outcome based on our dataset by computing the relationship between our variables, ultimately leading to an output bound between 0 and 1.

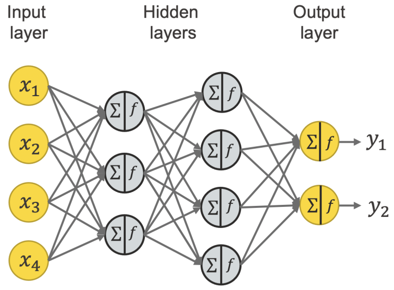
XGBoost: An implementation of a gradient boosted decision tree algorithm. Utilizing the power of ensemble techniques in which multiple models are strapped together to generate predictions, XGBoost sequentially ads models until there is no further improvement in performance. In particular, XGBoost uses the concept of gradient boosting, in which these new models predict the errors of primary models, making adjustments to minimize loss, and compiling them to generate predictions.

Figure 7: Sample XGBoost Structure. (Wang et al., 2019)



This algorithm is extremely powerful and utilizes core concepts of the "white-box" decision-tree algorithm, however, its boosting procedure introduces obscurity.

Figure 8: Sample Neural Network. (Melcher, 2021)



Neural Networks: An algorithmic system modeled based on the biological brain, allowing for advanced decision making and prediction through the use of numerous feature layers and complex non-linear computations. In the same way as the brain, these systems allow for very powerful and efficient results, however, lack the fundamental transparency needed to understand each decision, creating a "black-box".

Interpretation Techniques: While the decision tree algorithm will be interpretable by nature, allowing us to derive conclusions as to the model's decision through basic statistical means, certain models like the XGBoost or Neural Networks will be more difficult to decipher. As mentioned previously in the paper, the use of SHAP values will be important in breaking down the latter models.

SHapley Additive exPlanations (SHAP): The Shapley value is a solution concept in cooperative game theory. It was named in honor of Lloyd Shapley, who introduced it in 1951 and won the Nobel Memorial Prize in Economic Sciences for it in 2012. The objective of this value is to provide the marginal contribution an individual has to a coalition's output. By observing all the different ways we can compose this coalition, through the inclusion and exclusion of members, and finding the differences in outputs based on individual presence, the shapely value gives us the contribution each member has provided (Roth, 1988). SHapley Additive exPlanations (SHAP) apply this ideology to machine learning models, relating the individuals to features in a dataset, and the coalition output as the predicted value of the model (Lundberg and Lee, 2017).

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

g(z) represents the explanation model, approximating the model function, z'j is the coalition's vector indicating which features are included in the coalition, M is the largest coalition size, permeating the process for each feature, and oj is the attribution of a particular feature (j) (Lundberg and Lee, 2017). Through the use of the KernelSHAP, we can use approximate similar concepts within our models. By obtaining the vector of feature inclusion and exclusion, we map the included features to their true value in a particular instance and sample a random data point to use for the "excluded" features as a way to exclude their contribution. Using this technique, and some of its variations we can obtain a breakdown of feature contributions to particular explanations, allowing us to question a model's output and dissect its decision-making process.

Finally, to apply all outlined concepts within Python, we will leverage a series of popular packages to facilitate data analysis, model building, and metric exploration. Such packages include:

- Pandas: An effective, dynamic, and powerful data manipulation and analysis tool.

- Numpy: A high-level mathematical library used to operate on multi-dimensional arrays and matrices.

- Matplotlib: A visualization package allowing for the creation of dynamic, and interactive visualizations.

- Scikit-learn: A robust library for predictive data analysis.

- TensorFlow: An open-source python library used for ML and AI implementations, particularly focused on the creation and inference of DNNs.

- Keras: An open-source library built on TensorFlow that serves as a user interface for ANNs creation.

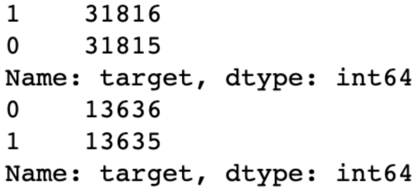
- SHAP: Aforementioned game theory approach to machine learning output explainability.

Mapping out our methodology, we culminate the literature review, identifying core fairness metrics that can be used to assess different forms of discrimination and applying statistical tests to objectify their results. Through the use of Python and its various libraries, we will be able to apply these tests directly to three datasets of different nature, utilizing a series of models in the process. Finally, we will deploy the use of SHAP to bring about explainability within obscure models, seeking to understand our insights and bring forth rectifying action.

4 Results

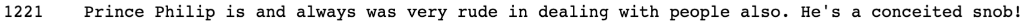
4.1 Case Study 1: Civil Comments

Starting with the Civil Comments datasets, we begin by extracting user comments and toxicity labels from the dataset, splitting them into a training and testing set for our model to train and predict. We stratify the division based on the target to ensure a proper representation of toxic and non-toxic comments in both the training and testing leaving us with around 60,000 comments to train our model on.



```
1    31816
0    31815
Name: target, dtype: int64
0    13636
1    13635
Name: target, dtype: int64
```

Before building the model, we can observe some sample comments labeled as being toxic and non-toxic, getting an understanding of how the labelers determined the classification. As humans have labeled the comments, we can already begin to observe a degree of bias from the labeler's freedom to decide on the label. Given that the labels in the Civil Comments dataset come from a combination of scores derived by professional annotators and labelers we can see how they have already tried to limit the exposure of individual biases on the target label. Extracting a random "Toxic" labeled comment prints the following:



```
1221    Prince Philip is and always was very rude in dealing with people also. He's a conceited snob!
```

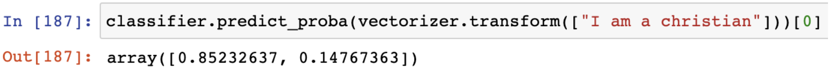
Extracting a random "Non-Toxic" labeled comment prints:

```
87341    If you want the price of homes to drop all you need to do is increase supply. Supply to demand is the only r
atio that really impacts the price of anything.
```

As anticipated we can see how toxic comments are rude, insulting, and directly attacking while non-toxic comments seem to be more general, informative, and non-inflammatory. To make predictions on the toxicity classification of a comment based on the Civil Comments dataset, we begin by utilizing a CountVectorizer to count the number of times a particular word is present within a string. From there, we can run a simple Logistic Regression on the vectorized comments and determine their toxicity based on the severity of the word as observed by the model in the training set.

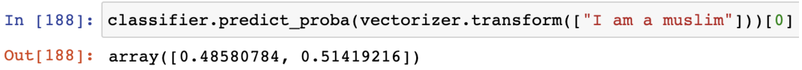
Having trained the Logistic Regressor, we create predictions on the testing set and compare the results to their true labels, accurately classifying around 93% of our comments. With this very high accuracy, it can be easy to assume this model as being near perfect, or at least better than the average human at predicting comment toxicity. However, when put to the test, we can observe a different result.

To see how the model interprets particular phrases containing sensitive groups we input a few phrases such as: “I am a christian” in which case the model certainly predicts as being non-toxic, however, when given the phrase “I am a muslim”, the model is more uncertain in its decision, slightly leaning towards the toxic classification, even though no toxic words are present.



```
In [187]: classifier.predict_proba(vectorizer.transform(["I am a christian"]))[0]
Out[187]: array([0.85232637, 0.14767363])
```

To further investigate, we look at the classifications of “I am black” and “I am white” giving similar, racially-biased outcomes.



```
In [188]: classifier.predict_proba(vectorizer.transform(["I am a muslim"]))[0]
Out[188]: array([0.48580784, 0.51419216])
```

With such a highly accurate model it seems rather odd that it can misclassify such simple and obvious examples. To begin our investigation, we will first isolate a social group, for simplicity, we will choose religion, in particular, the words “christian” and “muslim”. From there we will begin our analysis with “representation” to understand the way the two religions are represented within our dataset. To begin we look at the number of phrases containing either of the words and exclude those containing both:

```
There are 190 comments containing the word ‘christian’
There are 289 comments containing the word ‘christian’
```

Table 1: Religion Toxicity Contingency Table

	0	1
Muslim	0.267	0.733
Christian	0.505	0.495

We can immediately note a significant difference of around 100 comments, mentioning the words “muslim” indicating that our model will have had more data regarding such comments compared to “christian” phrases. It is worth noting, however, that due to the data being composed of natural language, numerous other words could have been used to symbolize either of the religions, for simplicity, we will only consider the exact words “christian” and “muslim”, regardless of capitalization. To further assess the dependency on the classification we create a normalized contingency table outlining the relation between religious classification and model output, running a chi-squared test of contingency on this table, our Null Hypothesis is the independence of the prediction and the sensitive group. The resulting p ¡ .001, rejecting our null hypothesis at most any level of alpha, indicating a significant dependence between our two variables, and failing our test of demographic parity. Running the same procedure on the true comment labels presents us with a p of .001, once again rejecting our null, in this case, we are alerted of dependencies within our base data, and not only within our predictions. This is a clear sign of biases that are present within our data, regardless of our technical procedures; even with a perfect model, we would simply come closer to replicating these witnessed disparities.

Moving over to our second scope of analysis, we look at measures of ability, which can help us understand if our religiously separated comments had the same chances of being considered not toxic. We can see the TPR for both subsets:

```
True Positive Rate for comments containing ‘christian’ 87.88%
True Positive Rate for comments containing ‘muslim’ 88.43%
```

Running a chi-squared test on our TPR disparities we obtain a p of .967, with a null hypothesis indicating insignificant disparities among TPR rates for both categories, we accept the hypothesis. It seems as though, when comments are truly toxic, our model is classifying them as toxic around 88% of the time, and with minor differences based on religious use. This additionally satisfies our condition of equal opportunity as described in the analyzed works.

```
False Positive Rate for comments containing ‘christian’ 7.69%
False Positive Rate for comments containing ‘muslim’ 28.77%
```

In contrast, our false-positive rates seem rather different. Running our chi-squared test for the false-positive disparity, we obtain a p ¡ .001 rejecting our hypothesis and indicating a significant disparity between false-positive rates depending on religious classification. While we may be accurately categorizing toxic comments when they are in fact toxic, regardless of religious context, we were not able to satisfy equal odds as we have significant differences among FPRs based on sensitive groups. In this case, we are made aware that comments containing the word “muslim” are highly more likely to be classified as toxic when in reality they are not, making it harder to post safe content containing this religious word, and a clear indication of unfairness.

```
True Negative Rate for comments containing ‘christian’ 92.31%
True Negative Rate for comments containing ‘muslim’ 64.41
```

Deriving to the TNR we observe a difference between groups, with a p of .026 we reject our null at the alpha level of 1% indicating a slightly significant difference between the rates of toxic classifications, given that they were toxic. In this case, we complement our previous analysis that “muslim” containing comments are being labeled as toxic when in reality not toxic compared to “christian” comments.

```
False Negative Rate for comments containing ‘christian’ 12.12%
False Negative Rate for comments containing ‘muslim’ 11.57%
```

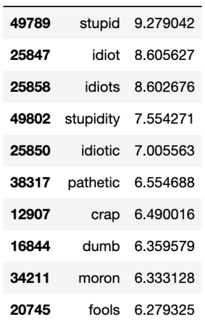
Deriving then to the FNR we obtain a p of .911 accepting out null, and indicating no significant differences between “christian” and “muslim” comments misclassification as being non-toxic.

```
Precision of comments containing ‘christian’ 92.55%
Precision of comments containing ‘muslim’ 90.09%
```

Finally, we move into “performance” by observing the precision score of comments within each group. We can see a slight difference but nothing too significant. We run our chi-squared test on precision disparity and obtain a p of .856, accepting our null of insignificance and indicating a similar representation of toxic comments within our predictions as in our base dataset. In addition, we are not alerted of major exacerbations of the disparities compared to what we would normally expect.

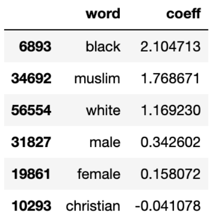
Reviewing our analysis of the Civil Comments dataset, we are notified of immediate disparities between comments containing either of two religious words. If this model were to go into production on a wider scale, it would be clear that certain discriminatory biases and unfairness claims would emerge. To understand where and how these biases would translate into unfairness, we go through the "representation, ability, and representation" framework and scrutinize both our data and our model for these sources. To begin with, we were notified of a difference between the overall representation of comments within each religious group, indicating potential differences in model learning. In addition, looking at demographic parity through the lens of a chi-squared test of independence, we are notified of a clear and significant relation between the religious classification of a text and its predicted output. We are notified of this same discrepancy within the dataset itself, making it clear that our dataset itself is biased towards one group. To further instigate, we ran tests of equal opportunity and equalized odds, and while it seems as though both types of text have the same chance of being categorized as toxic when they are in fact toxic, they do not have the same chances of being accidentally misclassified as toxic. Looking at our false-positive rates, it seems evident that "muslim" containing comments have a higher likelihood of being incorrectly labeled as toxic, especially when compared to comments containing "christian". This is further validated by our chi-squared test. Finally, to ensure no further exacerbation of the bias, we look at the predictive parity of our subgroups and notice no significant differences indicating equal model calibration and representation of the true distribution of classes. From this analysis, we know that our data is biased inherently and that it is translated into unfairness through a disproportionate rate of toxic misclassifications for "muslim" comments as opposed to "christian". To now isolate the subsample of "muslim" comments that are most affected we investigate further into the model and its factors of decision-making.

To understand the model, we must understand what it is allowed to use in order to make its predictions, in the Civil Comments case, our Logistic Regressor only sees the counts of words, extracted by comments made by random individuals as well as their toxicity classification. While the model itself will always aim to maximize its power in predicting the labeled classifications, the decisions that create the predictions stem more granularly from the comments themselves.



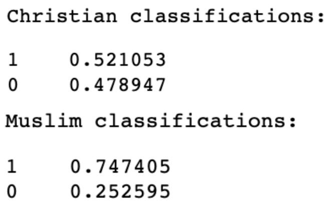
	word	coeff
49789	stupid	9.279042
25847	idiot	8.605627
25858	idiots	8.602676
49802	stupidity	7.554271
25850	idiotic	7.005563
38317	pathetic	6.554688
12907	crap	6.490016
16844	dumb	6.359579
34211	moron	6.333128
20745	fools	6.279325

In the scenario of the Logistic Regression, these "micro-decisions" are easily observable through the visible coefficients of our model. By combining the coefficients of words found in the given sentence, the model obtains a final output that will converge to a "toxic" or "non-toxic" verdict. By printing out these coefficients we can see the worst words as assumed by the model, each with a very high coefficient, pushing almost any phrase with its inclusion into the "toxicity" category. At first glance, each included word seems rather justified in its classification.



	word	coeff
6893	black	2.104713
34692	muslim	1.768671
56554	white	1.169230
31827	male	0.342602
19861	female	0.158072
10293	christian	-0.041078

However, when we capture the coefficients of sensitive groups, we can see that the words "black" and "muslim" have the harshest coefficient within the list, indicating that their inclusion in a sentence carries a negative weight while the word "christian" seems to carry a positive weight to it. In all cases, without context, neither of the words in the list have any positive or negative value, yet are assumed as having such.

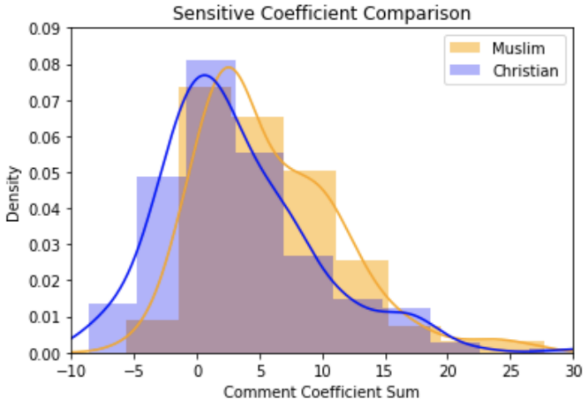


```
Christian classifications:

1    0.521053
0    0.478947
Muslim classifications:

1    0.747405
0    0.252595
```

To further understand the reasoning for this disparity, we can break down the inputs, observing how they are represented within the dataset before being consumed by the model. Looking at all the sentences which include the word "muslim", we can see that the dataset already includes (75%) of "muslim" involved comments to be toxic, compared to (52%) for the "christian" included sentences. This implies that the model will assume a generally more negative weighting for texts with the word "muslim" due to the sheer percentage of them indicated as such within the dataset. This can be due to most comments being generally more negative towards this religious group or potentially due to labeling bias. To investigate this, we can look at the distribution of coefficients among both groups.

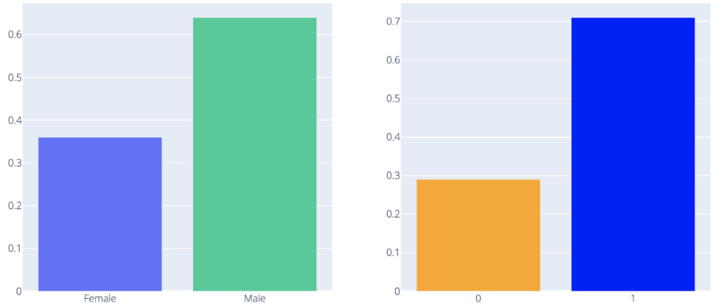


Looking at the averages, we can observe a slight difference between the average severity of the phrases in either group, however, the major differences between the two seem to stem from the stronger right tail skew visible within the "muslim" category and having a longer, more extremely negative tail seemingly shifting the Logistic Regressor into associating the word with a more extreme negative connotation. In this context we can denote that the most affected category of comments are those containing the word "muslim" that have a neutral standpoint as the sum of the coefficients will rely heavily on the more extreme weight presented by the "muslim" coefficient, weighing down our final verdict into a toxic misclassification when in reality it is a neutral comment.

Through this brief analysis, the complexity of the underlying biases seems to stem from numerous locations, from the biases of the human comments, to the labelers, to the uneven magnitude of negatively labeled comments containing sensitive groups, to the more in-depth disparities in the severity of sensitive comments. In either case, it seems as though the dataset is littered with biases at its current state; even with a perfect model, we could expect it to strive towards converging into the true classification, re-creating and potentially exacerbating the extent of the observed biases. Looking back at our entire analysis of the Civil Comments dataset, we have traversed through the principle of "identify" by gouging through our series of fairness metrics covering "representation, ability, and performance" tests, and identifying areas of bias entry as well as unfairness outcomes that have resulted from it. Moving into the second principle of "understanding" we broke down our white-box model, extracting coefficients of different words. Investigating further we understood some of the potential reasons for our previously identified discrimination and isolated the most affected subgroup. To finally act on these insights, we would need to make the findings public and transparent for stakeholders to see, and implement techniques to provide generally neutral comments containing the word "muslim" a second chance at non-toxicity.

4.2 Case Study 2: German Credit Risk

To move into our second case study of the German Credit Risk dataset, we are provided with nine, fact-based features to determine whether a candidate is a high credit risk. Among this dataset are two main sensitive groups, sex, and age. For this particular analysis, we will focus on sex as being our sensitive group. To generate predictions, we will use an XGBoost classifier which received an overall test accuracy of (76%) indicating moderate predictive performance.



In a similar fashion to the Civil Comments case, we will begin by observing representation factors by firstly looking at the representation of both the sex labels and outcome labels. We can see that there are a lot more males than females in the dataset, and in general, more people are labeled as being of high credit risk. This may indicate that our model will have more opportunities to train on such subgroups.

Table 2: Sex Risk Contingency

	0	1
Female	0.347	0.653
Male	0.258	0.742

Finally, testing for demographic parity, we look at the normalized contingency table between our sex categories and the associated predicted risk, running our chi-squared assessment we obtain a p of .823, accepting our null hypothesis of independence and indicating no significant relation between the sex and the risk classification. In addition, running the same analysis on our base dataset provides us with a p of .239, once again indicating no major dependencies of the sensitive variable on the target label. Moving then into the principle of ability, we compute our true-positive and false-positive rates.

```
True Positive Rate for Males 85.26%
True Positive Rate for Females 95.74%
```

Based on our true positive rates, it seems as though we have a small difference between sensitive groups, however, running our chi-squared test, we obtain a p of .436, accepting our null, and indicating no significant difference between our true positive rates. This would thus satisfy equal opportunity for both males and females in which both genders have similar high-risk probabilities given that they in fact present a high risk.

```
False Positive Rate for Males 57.58%
False Positive Rate for Females 52.00%
```

In addition, the false-positive rates for both groups seem rather consistent with a p of .594, allowing us to accept our null hypothesis of no significant differences between the false-positive rates and noting similar levels of high-risk misclassifications for either group.

```
True Negative Rate for Males 42.42%
True Negative Rate for Females 48.00%
```

Looking at true negative rates in which our model correctly predicted proper low-risk classifications we see similar scores for both groups supported by a p of .558, indicating similar levels of proper low-risk classifications for each gender.

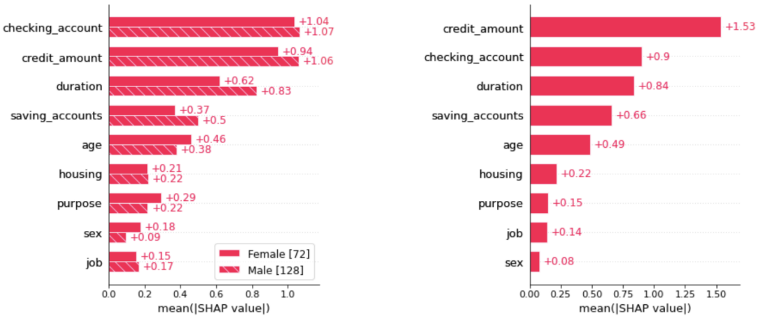
```
False Negative Rate for Males: 14.73%
False Negative Rate for Females: 4.25%
```

Finally looking at false-negative rates we get a different picture, it seems as though males are being incorrectly labeled as being of low risk at a much higher rate than females, supporting this is a p of .016, while this may be insignificant for alpha levels of 10% and 5% it does not satisfy at 1% indicating a potential advantage for men receiving more credit in cases when they should not be, especially when compared to their female counterparts who are not given this privilege.

```
Precision for Men: 81.00%
Precision for Women: 77.59%
```

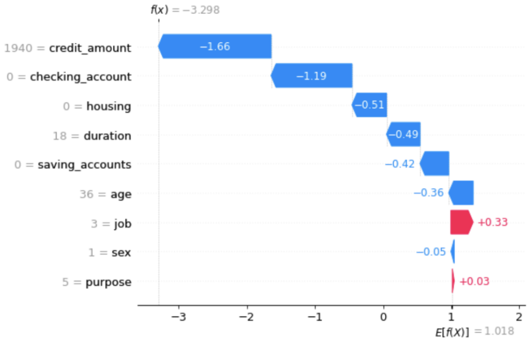
Finally, to observe model exacerbation of biases through the lens of predictive parity, we notice slight differences in the precision scores for both groups. Looking at the chi-squared test, however, we obtain a p of .786 signaling low significance of disparity, we are informed that the model is not greatly hindering either group, but rather enforcing existing disparities. This is complimentary of our previous analysis in which we only hinted at a slight significant disparity among the false-negative rates.

Having understood the basis of the slight disparity detected, we can isolate the most affected group as being males that are incorrectly classified as having low credit risk. To further investigate, we must dive into the workings of the black-box model. To do so, we will reintroduce the idea of SHAP through the SHAP library. Our first result is to observe key feature importance based on our sex cohorts.

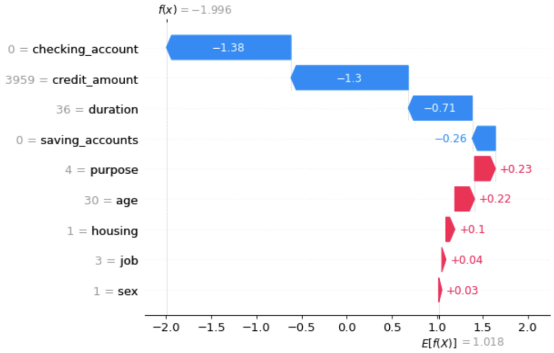


Observing the graph on the left, we can see the overall feature importance, and on the right, we see the importance of the isolated male, false-negative cases. In general, it seems as though the same factors are considered at roughly the same weight between genders. However, we do notice a larger gap within duration seeing that it is of higher importance to the model with males. Comparing the graph to the isolated FN cases, we immediately notice a big difference in importance between the credit amount observed in the overall outputs as opposed to the subgroup. It is a clear indication that a certain sub characteristic of the male FN group is contributing to the incorrect labels.

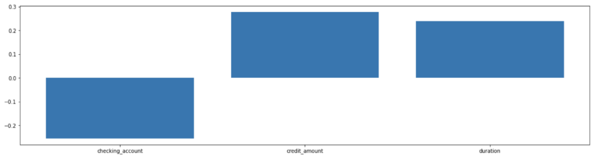
Thanks to SHAP's local interpretability we can further isolate individual cases to understand how their decisions were made.



Selecting a random case, we can immediately see our hypothesis in play, as the credit amount of the individual had a significant influence on the model's decision to predict the candidate as having low risk. However, noticing their checking account label of 0, and cross-checking it with our dataset original labels, we are aware of the candidate being labeled as having a "little" checking account.



Looking at another random case we again see a similar picture. Although the candidate has a very small checking account, the credit amount and duration pushed the model towards a low-credit risk classification.

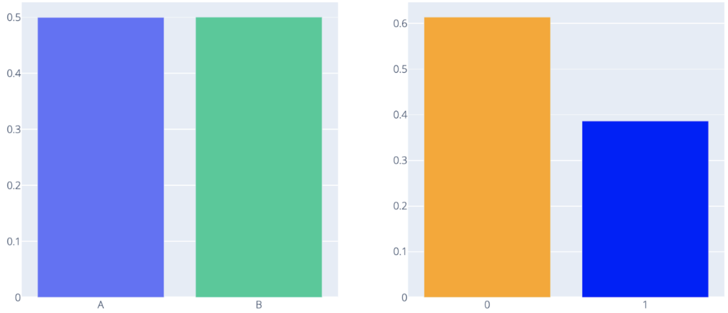


Finally, by looking at the differences in the average values of the three primary model factors, we can see that the average male, incorrectly labeled as being of low risk tends to have lower checking account labels, a higher credit amount, as well as duration as opposed to people who were correctly labeled as having no risk. Moving into the final "act" stage of our framework, we can hypothesize that certain men, with small checking accounts, but with higher than average credit amounts and durations could be subject to further inspection to minimize the privilege and minimize risk for the bank.

Finally, to culminate the research results, we go into our final case of the synthetic credit card approval dataset in which we will use a Neural Network to cover another model and provide a different view of opaqueness.

4.3 Case Study 3: Synthetic Credit Card Approval

Within the Credit Approval dataset, we are presented with five variables regarding the characteristics of candidates seeking credit card approval, among them is the sensitive variable "Group" which is used as a representation of a sensitive group. In the same process, we begin with representation, starting with the natural representation of the sensitive classes as well as the natural labels.



We can see an equal representation of both groups within the dataset, and more people being rejected for credit cards represented by the class “0”. While the model will have an equal chance to learn from members of both groups, it will have more information about candidates who were rejected. Moving into demographic parity, we again create our contingency matrix. Looking at our matrix, we can suspect a relation between the group label and the predicted class. Confirming this is a p ¡ .001, rejecting our hypothesis at all levels, and indicating significant dependence of the predicted output on the group label, disqualifying demographic parity. In addition, comparing it with the base relation found in the dataset we find the same hypothesis rejected with p ¡ .001 again indicating dependencies between the group and true label. Even a perfect model would seek to replicate such disparities and mimic the relation.

Table 3: Group Approval Contingency

	0	1
Group A	0.796	0.204
Group B	0.435	0.565

Table 4: Group Label Contingency

	0	1
Group A	0.805	0.195
Group B	0.423	0.577

Moving then into ability, we compute the TPR, FPR, TNR, and FNR.

```
True Positive Rate for Group A: 94.75%
True Positive Rate for Group B: 96.54%
```

With a p of .897, we are fairly confident with our hypothesis of no significant disparity within the true-positive rate, satisfying equal opportunity and providing similar chances of acceptance to qualified participants, regardless of group.

```
False Positive Rate for Group A: 2.42%
False Positive Rate for Group B: 1.89%
```

With very low false-positive rates overall, we don't notice major differences between groups. This is confirmed by our test *p* of .795 symbolizing minor differences in incorrect approvals, and indicating no majorly privileged group.

```
True Negative Rate for Group A: 97.58%
True Negative Rate for Group B: 98.11%
```

Reversing the FPR we obtain our TNR and a subsequent *p* of .969 again accepting our hypothesis that similar levels of candidates are being properly rejected regardless of group.

```
False Negative Rate for Group A: 5.25%
False Negative Rate for Group B: 3.46%
```

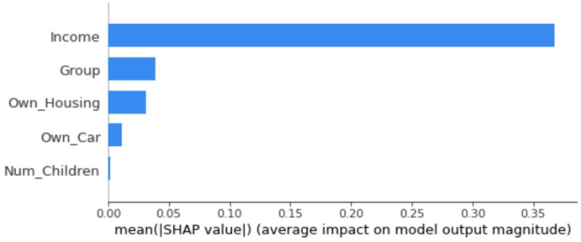
Finally, derived from the TPR, we obtain our FNR, running our test we obtain a *p* of .544 indicating no major differences in improper candidate rejection between our groups. Based on these tests, we can assess our model as having passed equalized odds, thus providing reasonable ability for all candidates regardless of group classification.

```
Precision for Group A: 90.45%
Precision for Group B: 98.59%
```

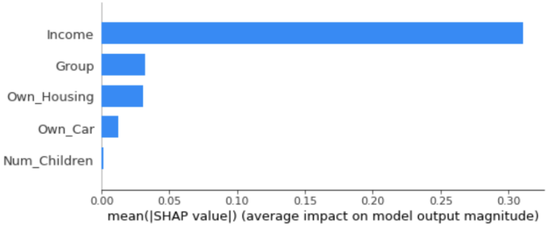
Lastly, looking at the predictive parity for both groups, we compute our precision and compare them observing no major differences. Computing our *p*, we obtain a value of .554, once again confirming the lack of model introduced bias and passing predictive parity.

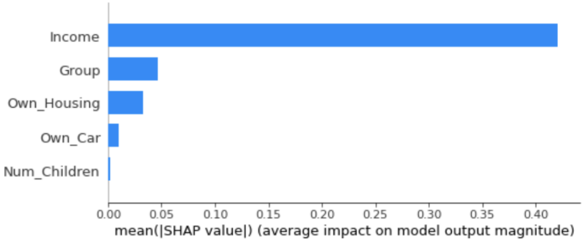
In this scenario, we have identified a strong relationship between the group labels and their corresponding predictive outcome both in our dataset and in our model predictions signaling a clear sign of bias from within the dataset itself, in this particular scenario it seems as though members of Group A are being rejected more often, most likely due to the proportion of Group A rejections from within our base dataset labels, however, in terms of the opportunity to obtain credit, qualified candidates are receiving credit in similar rates between groups. If we were to act upon this and allow for more Group A members to obtain acceptance, we would introduce our own biases as we create misclassifications and allow more unqualified Group A members to obtain credit, recreating another form of unfairness. In this case, to “act”, it is more crucial to report these disparities and determine the underlying reasons for them.

To do so, we initialize our SHAP kernel explainer, positioning it on a subsample of our data to increase speed.

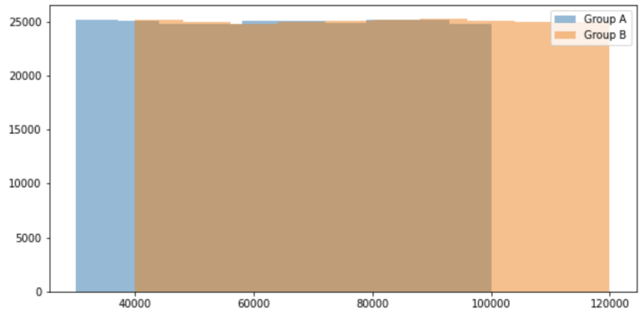


We can see that the largest contributor to credit card approval is the income of the individual, from a higher level this seems to make sense as a higher income would indicate a lower chance of defaulting on the credit.

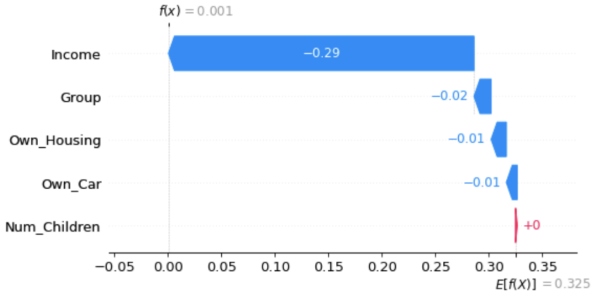


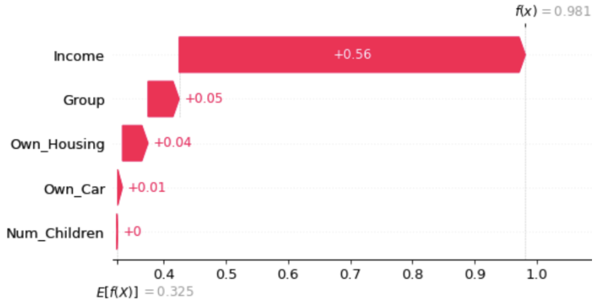


Breaking it down further, we isolate explanations for group A, on the left, and group B, on the right. While it seems as though the order of importance is the same, and it is clear that income is the primary and most significant factor in determining acceptance, we notice that the magnitude for group B seems to be higher than that of group A, especially in relation to the income. Thus, we can anticipate differences in the income distributions of both groups, potentially leading to the indirect biases we had identified previously.



To test this out, we look at the distribution of income among the different groups and can see that while the distribution seems unified, Group B certainly has a longer right tail, symbolizing higher levels of extreme wealth, whereas Group A has a longer left tail symbolizing more extreme levels of low income.





Looking at individual cases, the graph on the top represents a Group A candidate while the bottom graph represents a Group B candidate. We are immediately alerted of a red flag, as the model has opposing views when comparing groups, slightly penalizing Group A members, while slightly boosting Group B members purely based on the group label. This is certainly an issue worth noting, however, shies in comparison to the indirect effects introduced from the income level which immediately placed candidate A into the rejection category. In contrast, the income of the random group B candidate was by far the largest contributor to their acceptance. Looking back at the data, we can see that our first candidate had a below-average income, in comparison, our second candidate had the same magnitude of difference, however, above the average income. In such cases, it seems evident why the model would place each candidate in their respective placements.

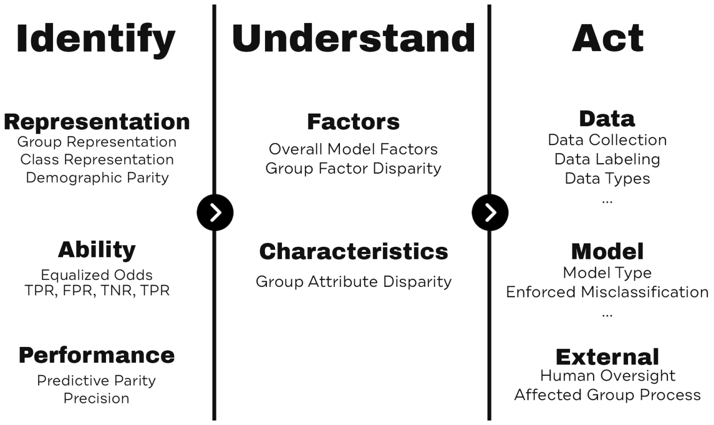
In this scenario, it seems as though no group is being majorly disadvantaged in their ability to obtain a credit card, given that they are qualified. In addition, looking at predictive parity, neither of our groups is being more severely impacted than what was present in our dataset. However, due to the natural disparities between the income levels of each group, we have noticed a subsequent difference in the ratio of each group's members that manage to get an acceptance, which was flagged by our demographic parity test. In such a case, an organization would need to decide whether it is more important to obtain equality of outcome, or maintain their equality of opportunity, assessing in the process the different factors that may come with it.

5 Conclusions & Discussions

5.1 Research Synthesis

Having gone through an immense web of research surrounding the topic of Artificial Intelligence, as it is in use today, we have observed, first-hand, the immense benefits that have arisen from its use. Over time, the performance of such models has increased drastically, even allowing for the discovery of particular antibiotics through the use of complex permutations, and extremely rapid computations; enabling a machine to learn on its own, identify patterns and replicate true label classifications with very high ability. While these advancements are certainly here to stay, we were also introduced to the downsides that have risen from it, in particular the idea of data/model bias which can seep in from various steps in the data process, with over 20 listed by Mehrabi (Mehrabi et al., 2019). In such scenarios, the biases can enter from the data collection, the labeling of data, and the natural disparities and biases that exist in the real world, among other forms. While the research suggests the ability to counteract some biases, this paper focuses on the critical goal of bias identification, and unfairness materialization, rather than mitigation itself. We believe that the first step in taking proper action to address ML bias is to know where it is, who it affects, and how it presents itself in deployment. To further reiterate the pressingness of this issue, we outline a series of policies that have been enacted around the world, particularly concerning discrimination as they truly have the potential to make detrimental life-changing decisions. As we reach the crossroad between ML performance, and explainability, as presented by DARPA's diagram (?), it is clear that we will no longer be able to use black-box models without further precautions. As a means of investigating the bias, we have extracted the core, overarching, fairness metrics as discussed by numerous authors, converting the metrics into statistical tests using a chi-squared test. In this way, we can obtain unique, objective, insights into group fairness. Diving deeper, we implement the SHAP value into Python and use it to explain our model factors as well as individual decisions. From there, we were equipped with enough insights to draw preliminary conclusions and instill a need for further action.

Applying this methodology to a series of three unique cases, we analyzed natural language, credit card approvals, as well as credit risk, taking into account a particularly sensitive group within each. Following the same steps, we reached extremely different conclusions for each case, however, similar to all was our ability to extract the key insights of unfairness, run objective statistical tests, and isolate groups and characteristics most affected and in need of further support. All in all, from the conducted study, a framework has emerged, covering the fundamental principles of bias, unfairness, and discrimination, as assessed through our literature, and directly complimentary to the various policies enacted across the world.



To follow the framework, we must go through three primary stages, "identify, understand, and act". Within the first stage, we will need to compute several insights relating to the group representation, their ability to obtain classification, and their differences in model performance. Through this three-step process, we can gather the vast majority of information regarding the potential discrimination present in the dataset and can link it to a subgroup, helping us isolate the source. To get to this stage, we must go through a series of statistical tests of demographic parity, equalized odds, and predicted parity as described by our literature review, utilizing a chi-squared test to ensure an objective interpretation of disparities. Having identified the area and group of concern, we move into our second phase of "understanding" in which we scrutinize our model's structure and feature importance, comparing baseline levels to the previously identified subgroup. Through this analysis, we can note differences in model factors, as well as differences in individual characteristics, allowing us to further generalize affected people based on these characteristics. Finally, moving into our "act" phase, we create actionable plans to mitigate identified biases in a way that supports the negatively affected subgroup and does not introduce further detriments, such as changing the way we label and collect our data, the selection of models we use, or by simply using human oversight to give "high-risk" misclassified cases a second chance.

5.2 Limitations and Future Works

While the scope of this paper has been to scour through the complicated picture of automated unfairness and to derive some form of concrete, framework-based analysis, certain improvements can be made to further enhance the accessibility and power of such a framework. For instance, the implementation of these tests and frameworks into a public python library allows users around the world to utilize the framework tests on their own data problems and enables them to make modifications and changes to improve the development of fair ML. In addition, while this study is mainly focused on decision-making algorithms that work binarily, these concepts must be applied to a wider scope of projects, allowing for thorough analyses into regression disparities, as well as multi-group classifications.

As seen in the paper "The Impossibility Theorem of Machine Fairness. A Causal Perspective" the ability to obtain a perfectly fair model across the three key tests of demographic parity, equalized odds and predictive parity seems to be near impossible, as they appear to be in a causal relationship among themselves (Saravanakumar, 2020). While still serving the scope of this project as it allows us to identify areas of unfairness, using it to determine the total fairness or unfairness of a model and dataset is not clear and can not be concluded from such an analysis. In addition, while the statistical tests provide us with concrete details regarding the key disparities, the subsequent analyses conducted on the reasons for these biases, as well as the identified most affected subgroups, are based on subjective hypotheses on the data and subsequent insights, it could be more objectively beneficial if further tests were implemented to determine the relation of the factors with the areas of disparities, helping us draw more granular statistical conclusions.

Finally, due to the sheer size of neural networks, and the enormously large datasets in use today, using a permutative function such as the SHAP on a large dataset, or with a very complicated network will prove extremely slow as it attempts to permeate through every combination of factors to determine the importance, and making it difficult to obtain quick results for complicated models. While the advancements in this area will come from the further development of interpretation techniques, at its current state, the SHAP explainer seems effective in smaller datasets or subsamples, but still lacks for larger ones that use more complicated network structures.

All in all, while this framework is certainly not complete, and may be modified and appended to maximize its value, currently it serves as a starting point for any individual, organization, or company with high concerns for discrimination that would like to continue using ML algorithms. Utilizing such a framework will allow them to break down the convoluted web of bias, and turn them into simple actional steps that extract tremendous amounts of insights, isolating risk, and making the action a great deal easier. While it may be impossible to satisfy all our proposed test conditions, as further exemplified by Saravanakumar, it is more crucial to understand the meaning behind a failed test and to then determine the proper course of action to minimize further disturbances through human manipulation (Saravanakumar, 2020).

References

Borgesius, F. J. Z. (2018). Discrimination, artificial intelligence, and algorithmic decision-making.

Boucher, P. (2020). Artificial intelligence: How does it work, why does it matter, and what can we do about it?

Cook, A. et al. (2021). Synthetic credit card approval.

Eager, J., Whittle, M., Smit, J., Cacciaguerra, G., Lale-Demoz, E., et al. (2020). Opportunities of artificial intelligence.

European Commission (2021). Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

European Commission and Directorate-General for Communications Networks, Content and Technology (2019). *Ethics guidelines for trustworthy AI*. Publications Office.

European Union (2012). Charter of fundamental rights of the european union.

European Union (2016). General data protection regulation.

Gartner (2019). Gartner survey shows 37 percent of organizations have implemented ai in some form.

Hansen, C. (2020). Decision tree explained (classification).

Hofmann, H. (1994). Statlog (german credit data) data set.

Hutchinson, B. and Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.

Jigsaw/Conversation AI (2019). Jigsaw unintended bias in toxicity classification.

Kodiyan, A. A. (2019). An overview of ethical issues in using ai systems in hiring with a case study of amazon's ai based hiring tool.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compass recidivism algorithm.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635.

Melcher, K. (2021). A friendly introduction to [deep] neural networks.

Microsoft (2022). Responsible ai.

Pichai, S. (2018). Ai at google: our principles.

Roth, A. E. (1988). *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, Cambridge.

Saravanakumar, K. K. (2020). The impossibility theorem of machine fairness - A causal perspective. *CoRR*, abs/2007.06024.

Systems Applications and Products in data processing (2019). Decision tree expression.

Turek, M. (2020). Explainable artificial intelligence (xai).

UK Government Centre for Data Ethics and Innovation (2020). Review into bias in algorithmic decision-making.

United States Federal Trade Commission (1974). Equal credit opportunity act 15 u.s.c. §§ 1691-1691f.

United States Federal Trade Commission (2006). Federal trade commission act incorporating u.s. safe web act amendments of 2006.

United States Federal Trade Commission (2018). Fair credit reporting act 15 u.s.c § 1681.

U.S. Department of Defense (2020). Dod adopts ethical principles for artificial intelligence. *U.S. Department of Defense*.

Walia, J. (2021). Logistic regression.

Wang, Y., Pan, Z., Zheng, J., Qian, L., and Mingtao, L. (2019). A hybrid ensemble method for pulsar candidate classification. *Astrophysics and Space Science*, 364.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, Jun", e. J., Kan, M.-Y., Zhao, D., Li, S., and Zan, H. (2019). Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing*, pages 563–574.

6 Appendix

Link to Jupyter Notebook codes, datasets, and technical analyses.
https://github.com/ryandaher/ethicalAI_Thesis