From Data to Decisions: Bias, Fairness, and Security in AI for Self-Driving Cars

Nada Madkour
Eastern Michigan University
nada.e.madkour@gmail.com

Alice Nicoleta Pavaloiu
Open Ethics Initiative
a.pavaloiu@openethics.ai

Nikita Lukianets
Open Ethics Initiative
n.lukianets@openethics.ai

Abstract

Understanding and addressing the negative effects of bias in autonomous systems is critical, particularly in use-cases with the potential to cause detrimental long-lasting, consequences. In this paper, we aim to illustrate the wide landscape of bias and offer a deeper understanding of its relationship to fairness in the context of autonomous systems, particularly self-driving cars (SDCs). This is achieved by setting the emphasis on bias that may stem from the data management phase within the Machine Learning Lifecycle (MLLC) to pinpoint the areas of convergence between bias and unfair outcomes. The paper will also cover some elements of bias in model training. We introduce the concept of the Fairness-Bias Matrix (FBM), that can support effective assessment and understanding of Artificial Intelligence (AI) bias and fairness. Adversarial attacks on fairness, ethical considerations and concerns regarding SDCs are explored, factoring in cybersecurity threats, security risks, and unfair outcomes of bias in SDCs. An assessment of fairness metrics and bias evaluation methods are provided, as they become the nexus for building and implementing ethical AI. Various mitigation and bias management methodologies are compared to reveal significant gaps that need to be addressed. In support of addressing these gaps, transparency-based approaches for the evaluation and risk assessment of AI systems are recommended.

1. Introduction

Recent advances in AI applications, especially those used in high-stake scenarios such as loans or medical diagnostics (Mehrabi et al., 2021), have led to a dialogue on the ethical and societal implications of AI. While this discussion has produced many guidelines and frameworks addressing the subject, there seem to be many gaps that need to be addressed (Hagendorff, 2020). Guidelines and ethical codes often seem to have little to no effect on the decision-making and behavior of software engineers (McNamara et al. 2018), highlighting the importance of practical methods and solutions for AI ethics. Furthermore, historical experiences show that Machine Learning (ML), the main driving technical engine of AI, can contribute to the exacerbation of racial profiling towards specific segments of the population, such as “Stop and Frisk,” which was later tried in courts and found to be illegal. In that respect, we have a lot to learn from decades of social science and humanities research that may offer guidance in dealing with similar issues in the modern AI context. This further illuminates the importance of building effective tools for implementing ethical AI and reducing biases that may be reminiscent of the “Clever Hans” and its resulting observer-expectancy effect (Crawford, 2021).

The ramifications of unaddressed AI bias can lead to significant violations of human well-being in the form of individual fairness and group fairness. These outcomes can have detrimental effects on

freedom (Dressel & Farid, 2018; Mehrabi et al., 2021), medical care (Obermeyer et al., 2019; Vyas et al., 2020), job opportunities (Kodiyan, 2019), and criminal justice (Benedict, 2022).

The high-level focus of this paper is the implications of AI bias in the automotive industry. Implications of AI bias are even more crucial to address in the context of Self-Driving Cars (SDCs), notably with the emergence of Adversarial Machine Learning (AML) attacks against the decision-making of SDCs, some of which target the ethical tenets of such decisions (Chang et al., 2020; Jagielski et al., 2021; Mehrabi et al., 2021). This paper (i) maps data-to-model bias types for SDCs, (ii) explores common fairness metrics on SDC-relevant tasks, and (iii) briefly reviews attack surfaces and mitigations.

2. Bias and AI

Bias refers to systematically skewed perspectives, preferences, or patterns of behavior that can lead to unfairness, discrimination, prejudice, or unsupported conclusions. AI bias refers to the existence of disparities and flaws within AI systems that may lead to unfair outcomes for certain sub-groups or communities. Often, AI is reliant on data that is generated and labeled by humans, which can transfer and exacerbate certain biases and lead to unfair outcomes (Aquino, 2023; Ferrara, 2023).

The controversy surrounding AI ethics can be exhibited through the unfair outcomes that are produced by such biased systems, which can perpetuate harmful stereotypes and cause social problems (Wan et al., 2023). Bias in AI can also lead to unfair treatment of certain populations or groups of people as well as deepen existing systematic inequities (Bohdal et al., 2023). These inequities often result in real-world problems that are magnified over time, creating a bias-feedback loop. The loops can result from the AI models “learning” certain biases from the human interaction, then the biased AI output influencing human judgment, resulting in a more prominent bias in subsequent decisions (Stray, 2023). The phenomenon is especially significant in recommendation systems, where user interactions can magnify initial biases in the data or model (Glickman & Sharot, 2024). These phenomena can potentially solidify bias over time if model outputs replace human annotations as a source of supervision (Taori & Hashimoto, 2022). These negative outcomes can create societal unrest and amplify bias that can intensify over time (Ferrer et al., 2021; Ferrara, 2023).

Effectively addressing and mitigating biases requires interdisciplinary collaboration, transparency of the models and data used, inclusive and representative datasets, and accountability (Ferrara, 2023).

2.1 Examples of Bias in AI

There have been many documented instances of AI systems demonstrating bias and ethical violations. These violations include value violations (or misalignment) such as human autonomy and mobility, safety, security, trust, transparency, universal usability, and human welfare (Friedman et al., 2006; Graubohm et al., 2020).

A popular example of AI discrimination and bias is the COMPAS (Mattu, 2016) recidivism tool. The tool was meant to provide “risk scores" to aid in the decision to grant parole. But COMPAS ended up showing significant bias against African-American defendants. They were twice as likely to be falsely labeled as “high risk”, and white defendants were twice as likely to be falsely labeled as “low risk” (Mehrabi et al., 2021, Dressel & Farid, 2018).

Facial recognition software is another AI tool that has many reported incidents of bias or discrimination. In 2020 a black man was wrongfully arrested after a facial recognition software falsely identified him as the perpetrator of a crime that occurred in the store Shinola (Perkowitz, 2021). This is one instance of many, consistent inaccuracies that have been found in facial recognition algorithms when applied to African Americans according to research done by MIT, Microsoft, and NIST. Those inconsistencies are even higher in African American women, with an error rate of 35% (Perkowitz, 2021). Insufficient representation of non-white individuals in the training data may result in their inconsistencies. The datasets IJB-A and Adience are comprised of mostly non-white people (79.6% and 86.2% respectively) (Buolamwini & Gebru, 2018). While in general, commercial facial recognition software is much more effective on lighter males and least effective on darker females, in the context of criminal justice the oversaturation of people of color has led to false arrests (Perkowitz, 2021).

AI hiring tools have also been found to exhibit gender bias. Lambrecht & Tucker (2019) discovered that hiring tools on average advertise STEM jobs to men 20% more than they do women. This may be possibly due to the increased cost of advertising to young women. Similar issues have been exhibited in the Amazon AI tool used for screening resumes. The tool downgraded phrases such as “women's club”, resumes with all female colleges, and even vocabulary commonly used by women (Dastin, 2018). This led to the system favoring male candidates.

SDC pedestrian detection systems have also exhibited gender bias, along with racial and age bias. A study done by Brandao (2019) found that 72% of the methods examined had a significantly higher miss rate on female pedestrians. Other studies have also found disparity in miss rates to the detriment of dark-skinned, elderly, and children pedestrians (Bae & Xu, 2023; Brandao, 2019; Li et al., 2023).

Natural language processing (NLP) tools such as word embeddings have also been shown to reflect and amplify gender and racial bias (Bolukbasi et al., 2016; Caliskan et al., 2017; Jiang & Fellbaum, 2020). Another instance of bias in NLP was showcased by the journal Science (Caliskan et al., 2017), portraying how a model that was trained with standard internet text corpora acquired cultural biases. These ranged from a morally neutral bias measured by the Implicit Association Test (Project Implicit, n.d.), in preferences: such as “insect” or “flower”, to a problematic, discriminatory, racial and gendered tapestry. Historical injustices will continue to replicate and be held as imprints in the word embeddings for as long as the text corpora contain bias (Borenstein et al., 2023).

Generative AI is not exempt from bias either. A study from Carnegie Mellon University (Zhou et al., 2024) shows the landscape of bias across three popular generative AI platforms for visual content, DALL-E 2, Midjourney, and Stable Diffusion. Their analysis showed systematic bias in gender and race and subtle biases from appearances and facial expressions, which are derived across all platforms.

2.2 Types of Bias in Data Management and Model Training

To better understand the potential social and societal impact of AI bias, it is important to be able to evaluate and identify the different types of bias that may arise in those systems within the data management and model training phases of the MLLC. Many types of bias exist, depending on what standard is being used. For example, moral bias can be a deviation from moral standards. The same logic can be applied to statistical bias, legal bias, and algorithmic bias (Danks & London, 2017).

One method of AI system bias categorization is to group bias based on where it originated within the MLLC. Three main areas have been identified by Mehrabi et al., (2022), data to algorithm bias, algorithm to user bias, and user to data bias. This paper will refer to "data to algorithm" as "data to model" to better reflect the effect of training data on the machine learning models rather than only on the utilized algorithms. The paper's focus is bias that affects data management and model training. Therefore, it will only describe the types of bias associated with those phases of the MLLC.

Data-to-model bias describes the bias that may be the result of the data used to train the ML model. Within this section are different variations of bias. This includes Measurement Bias, which is a result of the misuse of particular features. An example of this was found in the crime recidivism tool COMPAS where prior arrests and friend/family arrests were used to determine the "riskiness" of a crime. Omitted variable bias describes bias that is a result of variables missing from the training dataset (Wilms et al., 2021). Representation bias is found when the population sample collected lacks diversity. An example was found in the BDD100K object detection training dataset, where there were three times as many Light Skinned individuals as there were Dark Skinned. This resulted in lower performance of the model when it came to identifying darker-skinned individuals (Mehrabi et al., 2022). Similar to representation bias is Sampling Bias, which is when trends based on the majority subgroup skew the relationships within the minority subgroups, resulting in a biased regression due to unbalanced data (Mehrabi et al., 2022). This type of bias may arise in SDCs in the context of movement prediction, particularly within pedestrians and other vehicles. Aggregation bias is when false conclusions are made about individuals from observing an entire population. This example is prominent in medicine, where assumptions about individual patients are made based on the assessment of entire populations, without taking into account individual differences such as age, gender, or family history (Mehrabi et al., 2022).

Algorithmic bias (Algorithm to user type) is present when the bias is not due to the training data and the trained ML model, but due to the algorithmic choices made by the software developers. This can happen by choosing certain optimization functions, the hand-picking of sub-groups, and the use of statistically biased estimators (Baeza-Yates, 2018; Mehrabi et al., 2022).

User-to-data includes cases where users use existing systems and are influencing data in a manner similar to a self-fulfilling prophecy. Ranking bias may arise when "top-ranked" results are prioritized and shown more than others (Baeza-Yates, 2018). This is similar to Popularity bias, but popularity bias is prone to manipulation with fake reviews or bots (Nematzadeh et al., 2017). Historical bias is the bias that already exists within social structures that is reflected in derived datasets or historically labeled data (Suresh & Guttag, 2021). A great example of this is the use of historically gathered and labeled data to train the COMPAS recidivism prediction tool resulting in unfair outcomes for African-American prisoners (Flores et al., 2016).

2. 3 Training Data and Bias

Biased output of ML models could both originate from the training data, and from choices regarding systematic elements of the model training and deployment. The former include selection biases, when the dataset used to train the model is not representative of the population it aims to generalize to. Sampling and Proxy Biases, similarly, lead to skewed representations of certain groups or phenomena in the dataset. Annotation/labeling biases, when human annotators may have subjective interpretations

or preconceptions that influence how they label data points. Historical biases, which may be present in the data, can perpetuate and amplify through the training process even if no labeling is happening in the MLLC. The latter include Algorithmic biases, when certain approaches/algorithms may inherently favor or prioritize certain types of features in the data statistically, or as a result of human-introduced logic; Deployment biases, even if a model is unbiased during training, biases can emerge when the model is deployed in real-world settings due to differences between the training and operational environments.

In many instances of AI unfairness, the bias can be linked to the training data and its representation. While it's typical to attribute unfair outcomes to discriminative practices, it's not always the discrimination that leads to unfairness, but the unequal initial conditions. Technically speaking, these conditions may be omitted in the datasets, failing to provide value to meaningfully separate the data points. Sometimes, a model may indirectly learn and reinforce biases through proxy variables that are correlated with attributes such as race, gender, or socioeconomic status. Proxy variables are hard to avoid because they themselves are used as ways to economize the data collection efforts. In this case, a careful and balanced examination of features is needed to make sure that the biased outcomes won't affect those data points described by the proxy variables.

In one of the studies that analyzed more than 1000 human-graded student responses from male and female participants using fine-tuned versions of BERT and GPT-3.5, the Equalized Odds metric analysis suggested that mixed-trained models generated more fairness outcomes compared with gender-specifically trained models (Latif, 2023). The outcomes of the study generally suggest that unbalanced data does not necessarily generate bias but can enlarge disparities and reduce fairness.

The ability of the ML model to create accurate, often impressive, predictions is largely reliant on the data it is trained on. It is often the case that the data these models are trained on come with existing biases that are mirrored and amplified by the model, leading to unfavorable outcomes and unfairness (Utsab, 2022). Bias in data has two main causes: underrepresentation and incompleteness. These can be the result of data that is retrieved from sources with existing bias, missing important components, containing errors, or incomplete (Ferrara, 2023). It’s important to mention that models can be subject to various sources of bias in the deployment phase as well. An example of perpetuated bias caused by data collection is the city of Boston's pothole identification system. When the city of Boston deployed a smartphone app that would automatically detect and report potholes (Street Bump | Boston.Gov, 2017), the city failed to account for a large portion of the population, those who do not have smartphones. This resulted in potholes being identified and reported only in higher-income areas (McCarthy, 2023).

2.3.1 Unsupervised Learning and Bias

Unsupervised learning does not rely on labeled data, it rather looks for patterns and structures within unlabeled data (Wang, 2016). This makes the model more susceptible to representation bias as it can easily overlook underrepresented data or skew results towards overrepresented data. It is critical to make sure the data used to train an unsupervised model is representative of the population it is being trained for. For example, the foundation of language representation learning for the Large Language Models (LLMs) is built using self-supervised, unsupervised pretraining techniques on raw textual data (Jain et al., 2023), which leaves plenty of room for perpetuated bias and disparities in output.

Self-supervised learning is considered a subset of unsupervised learning and, therefore is susceptible to the same risk of bias, particularly when exposed to biased training data (Pan et al., 2023). If the training data is biased or largely overlooks certain demographics or perspectives, this will be reflected in the model output and even become perpetuated by the model (Utsab, 2022).

Unsupervised learning models may also exhibit clustering bias, inadvertently grouping together data points based on similarities that lead to biased clusters. This can also result in the prioritization of certain features over others, ignoring important distinctions between groups (Pan et al., 2023).

Other unsupervised model biases include performance disparity, which underpins how the model produces different outcomes without explicit programming across different demographic strata. Occurrences in performance disparity can be noticed across different domains, varying from business, in the hiring process, or finance, in loan approval (Craggs, 2023), to medicine, in medical image segmentation (Gaggion et. al., 2023); all of which can lead to having the decision-making process altered and can be conducive to unfair treatment, such as enforcing already existing inequalities, enhancing economic disparities.

Correspondingly, unsupervised learning models face interpretability bias. The output generated by the unsupervised model, especially from big volumes of unlabeled data used in clustering, generative sampling, or abstractions may increase opacity and could prove too complex for the model to be understood and interpreted or explained (Watson, 2022). Henceforth, potential harm could arise if decisions are taken without human oversight and with a heavy reliance on automated processes.

Selection bias is also a concern present in unsupervised models. The dataset used for training the model is not reflective of the real-world scenario it needs to analyze. It can also appear in semi-supervised learning, where the data that is labeled comes from a biased sample of the entire data distribution (Willets et al., 2021). Selection bias can lead to inaccurate results, decisions, and unfair outcomes.

2.3.2 Data Annotation and Bias

Bias can also be introduced during the data annotation process in several forms. Multiple studies have discussed this concept, notably, Stoev et al. (2023) present evidence of the data labelers' state influencing the annotation process.

For example, the mental fatigue of the data labeler can have a significant detrimental impact on the quality of data labeling work. Fatigue reduces concentration spans (Csathó et al., 2012), making it challenging to maintain focus and accuracy when reviewing complex examples. Subtleties are more likely to be missed by a fatigued labeler, resulting in incorrect or missing labels. Tired minds struggle with pattern recognition, losing consistency in identifying features that determine a label, possibly leading to mislabeled or contradictory examples.

Other types of bias associated with data annotation are, Entity Bias, which is a result of certain entities appearing more frequently than others (Qian et al., 2022), and Labeler Bias, which refers to the influence of the existing biases harbored by the individuals who labeled the data (Haliburton et al., 2023). Moreover, studies have shown that stereotypes held by labelers as well as labeler demographics,

can influence the process and impact chosen labels (Shen et al., 2019). Order dependent biases have also been exhibited in sequential labeling tasks where the order of labels affects correlation. Lack of labeler experience may also cause Noisy labeling (Zhang et al., 2015).

To address these types of biases, two mitigation strategies stand out. Demartini et al., (2021) bring forth the narrative of bias management rather than bias removal, proposing a transparency-based strategy that will allow users to make informed choices (Havens et al., 2022). Another notable methodology suggests including data labeler characteristics and background information with the dataset (Demartini et al., 2021; Lukianets & Pavaloiu, 2021). These two methods suggest that addressing labeler bias may be achieved by implementing transparency mechanisms throughout the ML training process. This can help identify and possibly mitigate biases at early stages of the lifecycle, helping avoid cascading harms stemming from bias. For example, if an AI tool is trying to identify a picture of a mule, but the only labels "horse" or "donkey" exist, then the model would have to mislabeled the mule as a horse "horse" or a "donkey," potentially further distorting model training and accuracy (Christian, 2020). Labeling is such a critical part of the MLLC as it creates a pre-set ontology that has to be followed, further emphaizing the need to address harmful biases early in the MLLC.

2.4 The Fairness-Bias Matrix (FBM)

Although freedom from bias and fairness are often used as interchangeable terms (Reagan, 2021), for the purpose of this paper, they hold different meanings. The simplest way to make a distinction is by referring to the ML lifecycle. The ML lifecycle consists of 5 basic phases, data management, model training, model testing, model deployment, and using and monitoring (Yang et al., 2021). Bias is something that may be present in the data, training, or model. While fairness is determined by the outcomes which depend on the deployment & use context. An ML model may be biased, and it may have fair or unfair outcomes, based on how fairness is defined. Figure 1. represents the sequence of components of the ML lifecycle and demonstrates the areas where bias may exist and the areas where unfair outcomes may arise. When evaluating a system for bias, you would look within the data management, model training, and model testing. When evaluating the system for fairness of the outcome, you would look within the deployment and use.

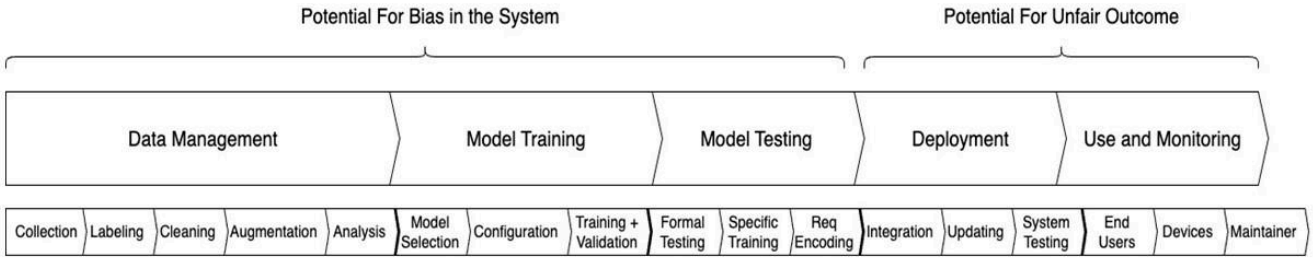

Figure 1: Bias and Fairness in the Machine Learning Lifecycle

It is important to mention that not all bias is unfair, and some fair outcomes rely on a certain level of bias in the model.

Here we introduce the Fairness-Bias Matrix (FBM) as a proposed conceptual method for evaluating and understanding bias within AI. Figure 2 represents the four quadrants of the Fairness-Bias Matrix.

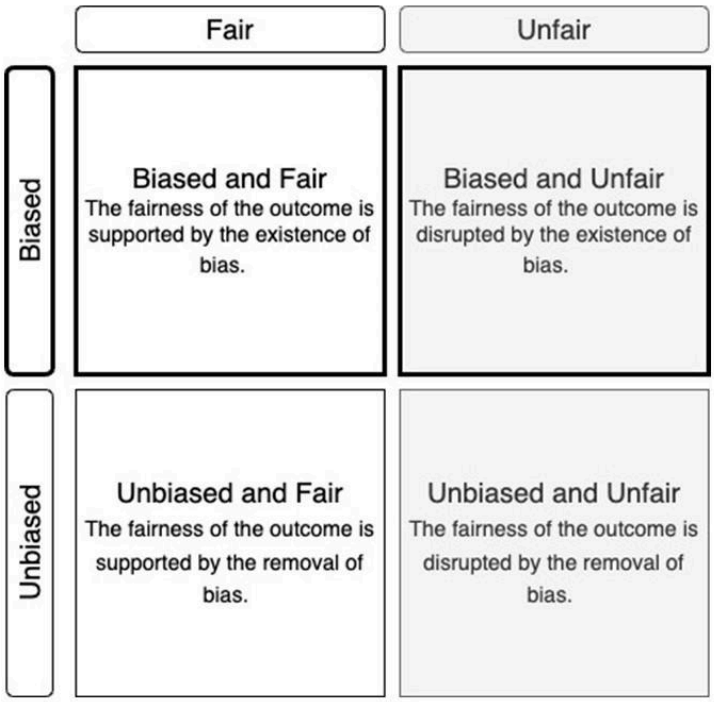

Figure 2: Fairness-Bias Matrix (FBM)

Quadrant one, Biased and Fair:

Encompasses situations where bias in the model serves the purpose of the outcome and aids in system fairness. An example of this can be found in gender-specific medical applications. A model that is used to track and provide insights into women's health should be biased in favor of women.

Another hypothetical example can be underpinned in the retail industry, where in order to get a store credit card, a credit-scoring criterion needs to be met beforehand. The system is biased against people who do not meet the requirement of being able to pay, which aligns with the store policy. (IEEE, 2023)

Quadrant two, Biased and Unfair:

Represents the bias in models that contributes to unfair outcomes. An example that would fit well into this quadrant is the COMPAS recidivism tool, where bias in the historically labeled data led to unfair predictions and racial disparity (Dressel & Farid, 2018).

The study of scoring results demonstrates that gender-unbalanced training has insignificant traces of gender bias. The findings are that models trained on mixed-gender data, as opposed to gender-specific data, produced lower mean score gaps by gender (MSG) when compared with human-generated results, implying that training data lacking gender balance might lead to algorithms exacerbating gender disparities (Latif, 2023).

Quadrant three, Unbiased and Fair:

Represents models that are free of bias which contributes to its fair outcomes. It can be viewed as a Northstar of responsible AI.

Quadrant four, Unbiased and Unfair:

Represents models that are free of bias which contributes to unfair outcomes. An illustration of this concept can be found within the context of the example in quadrant one. A model that is meant to track and provide insight into women's health may be negatively affected by the incorporation of men's health data, therefore contributing to a less biased model but less fair outcomes. Focusing on a single audience, however, could be also interpreted as avoidance of the deployment bias. Another example is in cases where correctly functioning algorithms, acting upon good data, set prices without human intervention, leading to charging poor people and minorities more for essential goods (Lukianoff, 2019). Sometimes, unbiased algorithms are a mirror of a biased world, which leads to even greater unfair outcomes, by having inequalities of the past be sustained and perpetuated (Ngomo, 2019).

2.5 Adversarial Attacks on Fairness

Adversarial Machine Learning (AML) is composed of a set of techniques that allow an adversary to exploit an ML system by manipulating its data, model, or both. The impact of AML attacks may result in the model learning the wrong things or making wrong decisions (Qayyum et al., 2020). Multiple studies have experimented with running adversarial attacks that target model fairness. One of these methods found that existing representation bias increases the effectiveness of the attack (Jagielski et al., 2021). This section highlights some of the most notable examples.

2.5.1 Examples Of Adversarial Attacks on Fairness

The first example is an experiment conducted by the National University of Singapore and Google using the COMPAS dataset with race as the sensitive attribute (Chang et al., 2020). This example focuses on the group fairness metric of equalized odds (Beigang, 2023). The paper demonstrates the effect of an adversarial data poisoning attack that changes a small fraction of the training data on equalized odds. This experiment assumes the attacker has access to the training and labeling process of the data. The attack exploited the existing conflict between fairness and robustness. The attack proved to have a much larger impact on minority groups when the model did not have fairness constraints. The data poisoning created by the most effective attack was placed into the smallest groups with the least frequent labels (Chang et al., 2020). This experiment stresses the importance of proper representation and labeling for all groups and subgroups.

The second experiment by the University of South Carolina runs attacks on other fairness metrics: Statistical Parity and Equal Opportunity (Bellamy et al., 2018) with gender as the sensitive attribute (Mehrabi et al., 2021). The results of the experiments found that the proposed attacks can be used to

harm the system in terms of fairness as well as create undetectable bias without harming model accuracy (Mehrabi et al., 2021).

The third experiment run by Solans et al. (2021) targeted Average Odds Difference (Bellamy et al., 2018) and Demographic Parity (Jacobs & Wallach, 2021) using the COMPAS dataset by injecting poisoned samples into the data. The experiment includes Black-Box and White-Box poisoning attacks, which both proved to increase unfairness with a modest effect on accuracy. The experiment findings demonstrate that these attacks can be used to introduce algorithmic bias into models as well as magnify biases that already exist. These attacks can be used without access to the specific model by using a surrogate model (Solans et al., 2021).

This next experiment uses the COMPAS dataset with race as the sensitive attribute, and the German Adult Dataset with gender as the sensitive attribute. The experiments showed effective loss of fairness by using the data poisoning attacks, adversarial sampling, adversarial labeling, and adversarial feature modification on the training data while using the test data unchanged (Van et al., 2022).

In this next experiment by Jagielski et al. (2021), the results show that sub-population fairness data poisoning attacks were more successful on some of the sub-populations more than others. When running the attacks on facial recognition datasets, the experiment found that the sub-populations easiest to attack were Latinos and Middle Easterns (an accuracy drop from 100% to 60%), which were the sub-populations with the least representation in the data. While images containing white people had an accuracy drop of 15.2% (Jagielski et al., 2021).

4. Self-Driving Car Bias, Fairness, and Security

The following section will cover concepts and considerations specific to SDCs in the context of bias, fairness, risk, and the intersectionality of those three elements.

4.1 SDC Ethical Considerations and Concerns

SDCs bring about a unique set of ethical considerations and concerns. The subject of ethical considerations regarding SDCs has been extensively discussed and debated with one of the biggest points of contention being the decision-making process involved in unavoidable crashes involving human life (Pagallo, 2022). Another concept that often emerges is accountability and determining the party responsible for unethical decision-making and unfair outcomes of SDCs (Servin et al., 2023). A growing concern in SDC ethics is the lack of attention given to cybersecurity concerns, especially considering the amount of personal information and sensitive data involved in the process (Nobles et al., 2023). Other critical points arise in these discussions such as social and political ethical implications, privacy and safety tradeoffs, and potential criminal exploitation (Arfini et al., 2022). It is overall important to address these ethical issues for the successful deployment and governance of SDCs.

4.2 Examples of Bias and Unfair Outcomes in SDCs

Several studies have shed light on the potential for unfair, potentially lethal, outcomes of biased SDCs. In 2018, a self-driving Uber vehicle crashed into and killed a woman in Tempe, Arizona. It was reported that the system design did not account for jaywalking and failed to classify Elaine Herzberg as a pedestrian, resulting in her death (NTSB, 2020). Another recent study that evaluated eight pedestrian detection algorithms with four common SDC training datasets found that on average these models have a 7-10% higher dark-skinned pedestrian miss rate and 19-16% higher children miss rate compared to lighter-skinned and adult pedestrians (Li et al., 2023). A study conducted by the University of Oxford also found that children and women had higher miss rates than adults and men (Brandao, 2019). A study by Stony Brooke University also found evidence of worse performance when predicting the pedestrian trajectory of children and the elderly compared to adults (Bae & Xu, 2023). The potential for bias is real, and the potential for real violation of human well-being is real.

5. Evaluating Bias

To properly evaluate the potential of bias in an ML model or dataset, it is necessary to understand the different types of methodologies available that target fairness and bias in AI systems. This section of the paper will highlight various approaches that address managing and evaluating bias in AI systems. Appropriately representing these methods requires the exploration of common fairness metrics and the different types of biases that may stem from different phases in the MLLC. The evaluation methods mentioned in this section include frameworks, tools, and concepts.

5.1 Bias in the ML Lifecycle

Different types of bias can manifest within the MLLC phases. The model training phase may be affected by Algorithmic bias, Ranking Bias, and popularity Bias. Model testing may be affected by Evaluation bias. In the deployment phase population and presentation bias may be exhibited. In the using and monitoring phase user interaction bias, temporal bias, social bias, and emergent bias may occur. And between the using and data management phase, historical bias may affect the model (Mehrabi et al., 2022; Yang et al., 2021). Figure 3 below shows the different phases of the MLLC and the types of bias that may originate in each of these phases.

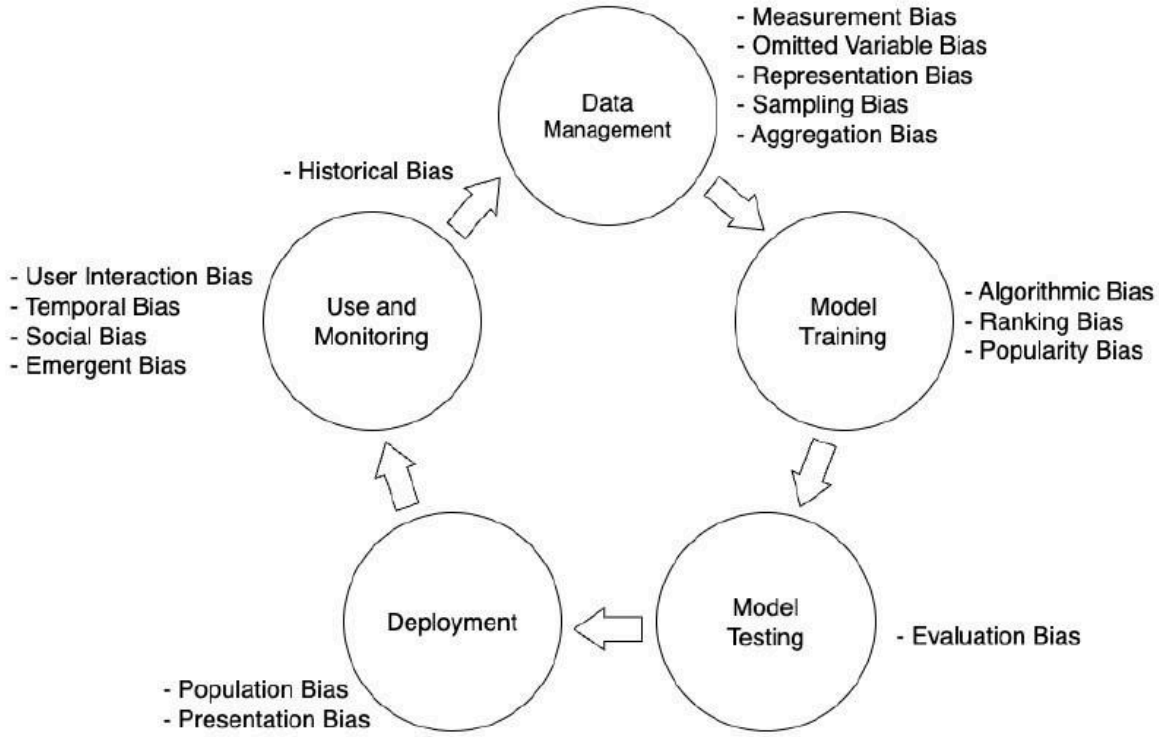

Figure 3: Types of Bias in MLLC Phases

In the context of the data management phase of the MLLC, a very specific set of biases may be exhibited that can lead to unfair consequences. Figure 4 represents the different biases in the data management phase and where they may be exhibited within that phase. Identifying where these biases are introduced will be crucial to mitigating them. These types of biases have been labeled as Data to Model (or Data to Algorithm) biases and represent bias in the data that may result in unfair outcomes when used to train the ML model. Measurement Bias can be caused by the way data is chosen and used to measure certain variables. Omitted Variable bias is caused by one or more important variables being left out of the training data. Representation bias is exhibited in the sampling and population in the data collection process. Aggregation bias is caused by false conclusions made about sub-groups or individuals when analyzing the population. Sampling bias is caused by non-random sampling of the population (Mehrabi et al., 2022).

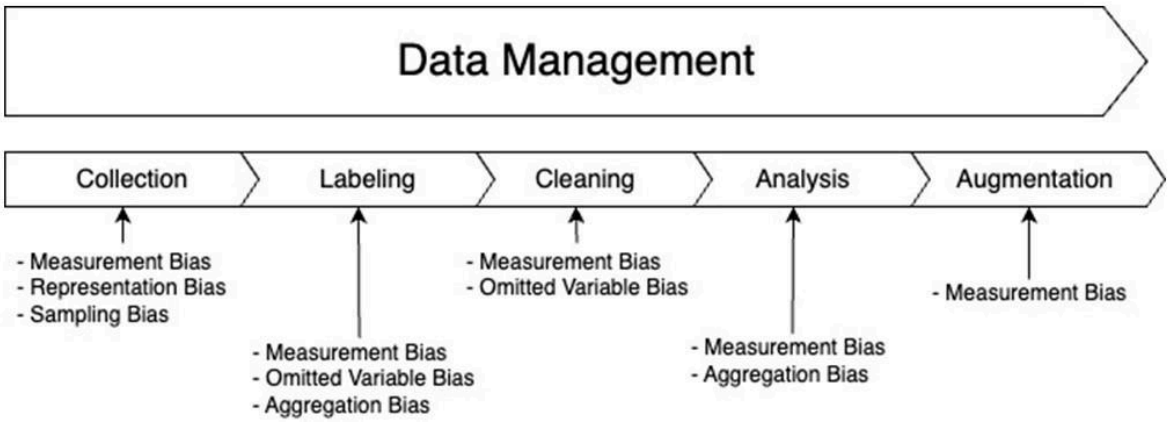

Figure 4: Types of Bias in the Data Management Phase Steps

5.2 Common Fairness Metrics

Many of the tools and methods discussed in section 5.3.4 rely on the use of fairness ML metrics. Fairness metrics can be split into two main categories. Group Fairness and Individual Fairness metrics. This paper will focus on group fairness metrics Table 1 lists the common fairness metrics used in AI fairness evaluation along with the definitions. These metrics are a powerful tool used to evaluate fairness and bias in ML systems. For example, a system that does not meet Demographic Parity constraints and demonstrates unbalanced favorable outcomes towards one gender but not the other, and exhibits gender bias. Each of these metrics measures fairness and bias differently, and each provides valuable insight into how the model may be showing bias.

Table 1: Group Fairness Metrics

Metric	**Definition**	**Source**
Demographic Parity	The difference in the probability of favorable outcomes between the underprivileged and privileged groups.	(Jacobs & Wallach, 2021)
Disparate Impact	The ratio in the probability of favorable outcomes between the unprivileged and privileged groups.	(*AI Fairness 360*, n.d.)
Equalized Odds	Requires false positive and negative rates to be similar among protected and non-protected groups.	(Jacobs & Wallach, 2021)
Average Odds Difference	This is the average odds difference in false positive rates and true positive rates between unprivileged and privileged groups.	(*AI Fairness 360*, n.d.)
Equal Opportunity	It captures differences in the true positive rate between different (advantaged and disadvantaged) demographic groups.	(Jacobs & Wallach, 2021)

Predictive Parity	looks at the raw accuracy score between groups: the percentage of the time the outcome is predicted correctly.	(*AI Fairness 360*, n.d.)

5.3 Bias Evaluation Methods

A variety of research has been done with the goal of developing bias evaluation and mitigation methods for AI (Ferrara, 2023). Assessing and understanding bias plays a critical role in mitigating those biases. The potential sources of bias may be algorithmic, data-based, or related to human decision-making (van Stein et al., 2023). The developed strategies and tools include frameworks and tools, each targeting specific subsets of bias causes in AI.

5.3.1 Evaluating Bias in Models

Multiple approaches exist to evaluate bias in AI models. One method is N-Sigma, commonly used in social sciences, and bias evaluation in facial recognition technologies (DeAlcala et al., 2023). Another method involves generating canonical tests that reveal model logic, potentially highlighting biases. This method, known as LUCID, prioritizes understanding the decision-making process and can be used with traditional fairness metrics (Mazijn et al., 2023). Metrics, such as Systematic Distance Error (SDE) and Combined Error Variance (CEV) can be used to evaluate class-wide bias (Park & Hu, 2023). Post-development bias assessment can also be implemented using a general model proposed by Fonseca & Diego (2022).

5.3.2 Evaluating Bias in Datasets

Bias in the training data can be evaluated by examining the data-gathering process. It is important to identify any disparities or biases in the sample that may lead to unfair model performance. This includes evaluating for proper representation of all data subgroups, particularly those with protected attributes, such as race or gender (Park & Hu, 2023). Managing bias in AI training data may include evaluating representation bias, testing for implicit biases, and the enhancement of data collection diversity (Lee, 2022). Other bias evaluation methods include ones that target the identification of unfair casual relationships in a dataset, and acting on the unfair casual edge cases (Ghai & Mueller, 2022).

The following is a recommended high-level step-by-step guide towards evaluating bias in an AI system derived from the concepts highlighted in this paper and the authors' expertise:

Data Examination:

- Assess Dataset Balance
- Explore data subset distributions

Model Examination:

- Examine feature relevance
- Compare performance across groups
- Analyze error distributions

- Evaluate against fairness metrics acceptable thresholds
- Examine variables and variable correlations between sensitive outcomes
- Examine the model parameters, proxies, and assigned weights.

5.3.3 Bias Evaluation Frameworks

While many proposed frameworks for Ethical AI exist, they mainly provide general guidelines for responsible technology. A number of these frameworks provide valuable insight into fairness, explainability, accountability, and privacy, but they often lack comprehensive coverage and guidance toward practical implementation (Barletta et al., 2023). For example, the U.S. Department of Commerce National Institute of Standards and Technology (NIST) has issued its own framework for addressing AI ethics, the NIST AI Risk Management Framework (Tabassi, 2023). This framework addresses fairness, privacy, safety, and transparency in the context of requirements and design, yet it fails to cover guidance for development, testing, and deployment. Many of these guidelines and frameworks addressing the subject have been proposed, but it is apparent that some gaps need to be addressed (Hagendorff, 2020). A study by McNamara et al. (2018) found that guidelines and ethical codes seem to have little to no effect on the decision-making and behavior of software engineers. Framework and guidelines alone cannot combat the issue of bias in AI.

5.3.4 Bias Evaluation Tools

Technical tools for bias evaluation and mitigation aim to reduce bias and ensure fairness in the AI development and application phases (Srivastava & Sinha, 2023). Mitigation strategies may target the algorithm, data pre-processing, and model validation (Ferrara, 2023). Examples of some tools that address bias and fairness in AI using fairness metrics include IBM AIF360 (AI Fairness 360, n.d.), Microsoft Fairlearn (Fairlearn, n.d.), Google's ML Fairness Gym (Google/Ml-Fairness-Gym, 2019/2024), TensorFlow's Fairness Indicators (Tensorflow/Fairness-Indicators, 2019/2024), and Fariness-Comparison (Algofairness/Fairness-Comparison, 2017/2024). These toolkits primarily perform metric-based evaluations and datasets and some even offer mitigation for biases found in the datasets. While many proposed frameworks for Ethical AI.

5.3.5 Apparent Gaps

A review done by Ayling & Chapman (2022) on 46 AI ethics tools revealed a few gaps that need to be addressed. One of those gaps is risk assessment methodologies, particularly for the design phase of the MLLC. Of the tools that address risk assessment, only two had a technical design for implementation, and of the two, only one was applicable during the design phase. Another important gap that was reported was the lack of tools and frameworks that addressed specific MLLC assessment. There are also gaps in stakeholder inclusion, current methods do not go beyond the inclusion of direct and core stakeholders. The adverse effects of AI without ethical considerations affect both those who are direct and indirect stakeholders (Friedman et al., 2017). Understanding the potential ethical conflicts between all stakeholders, direct and indirect, is a crucial part of the ethical design process, and must be understood early in the design phase (Graubohm et al., 2020). The consideration of ethics is

recommended to be done early on in the design process (Friedman et al., 2017) and be integrated into AI at the design phase rather than being an afterthought.

6. Conclusion

There are many risk considerations affecting AI and SDCs. More specifically, many risks associated with data labeling and training. Deployment of SDC systems on a large scale is partially dependent on the safety and security of such systems. A substantial portion is reliant on minimizing bias and unfair outcomes. We introduced the FBM as a simple conceptual tool for evaluating and understanding bias within AI, based on contextual fairness requirements.This review has shown that bias is not only a risk that exists in datasets with biased labels and underrepresentation, but that bias can also be injected or magnified with adversarial attacks. These risks must be addressed and minimized as they do not only pose harm to ethical values but also to the security and safety of the system.

As the research sheds light on, there are still many shortcomings regarding the guidelines and frameworks proposed: some only provide a partial solution, leaving gaps that need to be addressed; others, even if more encompassing in the design phase struggle with practical implementation in the machine-learning lifecycle, at the level of development, testing, or deployment. Subsequently, the impact of some guidelines and frameworks is minimal or not significant at all for software engineers, which also emphasizes the relevance of the audience in the methodology-building process.

6.1 Future Direction

There is no single, nor simple solution to address bias in all AI sectors. Nonetheless, understanding the risks of bias as shown in the study is quintessential as a starting point for addressing it. The direction of future research will focus on building an adaptive risk assessment tool that aims to identify and prioritize ethical risks in the machine-learning lifecycle, made accessible to a wide range of stakeholders, including system developers, designers, and policymakers with varying levels of technical expertise. This tool will be modeled after TARA, a mandate of ISO 21434 encompassing vehicle cybersecurity (Bastian Kruck, 2020; Luo et al., 2021). The tool will be retailored to fit the context of an ethical risk assessment, addressing elements such as value alignment, human rights, impact, likelihood, and the phases of the MLLC. This tool can be used in alliance with the Open Ethics Data Passport.

Open Ethics Initiative proposes a cohesive approach towards prioritizing built-in transparency through the Open Ethics Data Passport (Lukianets & Pavaloiu, 2021). The iterative tool which can be used in all phases of the machine-learning lifecycle aims to address systemic bias in trained AI models, by using self-disclosure to describe the origins of models, including the dataset acquisition and data annotation practices. The development of the risk assessment tool, akin to the Open Ethics Data Passport, integrates value-based considerations, which allows the methodology to address bias in a unique way, through a human-centered approach, henceforth enabling responsible decision-making. This in return allows for safeguarding against unfair outcomes and promotes the alignment of technology with ethical principles and fundamental societal values and human rights.

The safeguarding of humanity from unaligned AI may very well determine the future of our well-being and flourishing. In order for humanity to reach its full potential, certain safeguards and

considerations need to be considered and properly researched (The Precipice, 2020). While the topic of AI ethics risk may be in its infancy, it needs not be taken for granted as our flourishing future may be in jeopardy.

Acknowledgements

This paper represents the collaborative research which originated from Nada Madkour's independent study, as part of her Doctorate in Technology from Eastern Michigan University. We would like to extend our thanks to the university, the College of Engineering and Technology, her dissertation committee (Suleiman Ashur, John Coolage, Robert Carpenter, and Konnie Kustron), and her dissertation chair Samir Tout. We also thank Samir Tout for his contributions to this paper, particularly his expertise in adversarial machine learning and autonomous vehicles.

References

AIFairness 360. (n.d.). Retrieved March 11, 2024, from https://research.ibm.com/blog/ai-fairness-360

Algofairness/fairness-comparison. (2024). [HTML]. Algorithmic Fairness. https://github.com/algofairness/fairness-comparison (Original work published 2017)

Aquino, Y. S. J. (2023). Making decisions: Bias in artificial intelligence and data-driven diagnostic tools. Australian Journal of General Practice, 52(7), 439–442. https://doi.org/10.31128/AJGP-12-22-6630

Arfini, S., Spinelli, D., & Chiffi, D. (2022). Ethics of Self-driving Cars: A Naturalistic Approach. Minds and Machines, 32(4), 717–734. https://doi.org/10.1007/s11023-022-09604-y

Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: Are the tools fit for purpose? AI and Ethics, 2(3), 405–429. https://doi.org/10.1007/s43681-021-00084-x

Bae, A., & Xu, S. (2023). Discovering and Understanding Algorithmic Biases in Autonomous Pedestrian Trajectory Predictions. Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems, 1155–1161. https://doi.org/10.1145/3560905.3568433

Baeza-Yates, R. (2018). Bias on the web. Communications of the ACM, 61(6), 54–61. https://doi.org/10.1145/3209581

Beigang, F. (2023). Reconciling Algorithmic Fairness Criteria. Philosophy & Public Affairs, 51(2), 166–190. https://doi.org/10.1111/papa.12233

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias (arXiv:1810.01943). arXiv. https://doi.org/10.48550/arXiv.1810.01943

Barletta, V. S., Caivano, D., Gigante, D., & Ragone, A. (2023). A Rapid Review of Responsible AI frameworks: How to guide the development of ethical AI. Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, 358–367.

Bastian Kruck (Director). (2020, May 28). Doing a ISO21434 TARA in YAKINDU Security Analyst. https://www.youtube.com/watch?v=qZkGGPVjsJk

Benedict, T. J. (2022). The Computer Got It Wrong: Facial Recognition Technology and Establishing Probable Cause to Arrest. Washington and Lee Law Review, 79, 849. https://heinonline.org/HOL/Page?handle=hein.journals/waslee79&id=839&div=&collection=

Bohdal, O., Hospedales, T., Torr, P. H. S., & Barez, F. (2023, April 16). Fairness in AI and Its Long-Term Implications on Society. arXiv.Org. https://arxiv.org/abs/2304.09826v2

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Advances in Neural Information Processing Systems, 29. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Borenstein, N., Stańczak, K., Rolskov, T., Perez, N. da S., Käfer, N. K., & Augenstein, I. (2023). Measuring Intersectional Biases in Historical Documents (arXiv:2305.12376). arXiv. https://doi.org/10.48550/arXiv.2305.12376

Brandao, M. (2019). Age and gender bias in pedestrian detection algorithms (arXiv:1906.10490). arXiv. https://doi.org/10.48550/arXiv.1906.10490

Buolamwini, J., Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81:1–15, 2018. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Chang, H., Nguyen, T. D., Murakonda, S. K., Kazemi, E., & Shokri, R. (2020). On Adversarial Bias and the Robustness of Fair Machine Learning (arXiv:2006.08669). arXiv. https://doi.org/10.48550/arXiv.2006.08669

Christian, B. (2020). The alignment problem: Machine learning and human values. W.W. Norton & Company. https://catalog.carnegiestout.org/Record/287404

Craggs, J. (2023, August 11). Uncovering the Veiled Challenge: unraveling bias in unsupervised machine learning models. https://www.linkedin.com/pulse/uncovering-veiled-challenge-unraveling-bias-machine-learning-craggs/

Crawford, K. (2021). Atlas of AI, Power, Politics, and the Planetary Costs of Artificial Intelligence, 2021

Csathó, Á., van der Linden, D., Hernádi, I., Buzás, P., & Kalmár, G. (2012). Effects of mental fatigue on the capacity limits of visual attention. Journal of Cognitive Psychology, 24(5), 511–524. https://doi.org/10.1080/20445911.2012.658039

Dastin, J. (2018). Insight—Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automationinsight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-womenidUSKCN1MK08G/

DeAlcala, D., Serna, I., Morales, A., Fierrez, J., & Ortega-Garcia, J. (2023). Measuring Bias in AI Models: An Statistical Approach Introducing N-Sigma (arXiv:2304.13680). arXiv. https://doi.org/10.48550/arXiv.2304.13680

Demartini, G., Roitero, K., & Mizzaro, S. (2021). Managing Bias in Human-Annotated Data: Moving Beyond Bias Removal (arXiv:2110.13504). arXiv. http://arxiv.org/abs/2110.13504

Deng, Y., Zhang, T., Lou, G., Zheng, X., Jin, J., & Han, Q.-L. (2021). Deep Learning-Based Autonomous Driving Systems: A Survey of Attacks and Defenses (arXiv:2104.01789). arXiv. http://arxiv.org/abs/2104.01789

Ding, S., Tian, Y., Xu, F., Li, Q. A., & Zhong, S. (2019). Poisoning Attack on Deep Generative Models in Autonomous Driving. https://www.semanticscholar.org/paper/Poisoning-Attack-on-Deep-Generative-Models-in-Ding-Tian/e729a48bb462a307d17a22757adb4919c1ed30b9

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Fairlearn. (n.d.). Retrieved March 11, 2024, from https://fairlearn.org

Ferrara, E. (2023). Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies (arXiv:2304.07683). arXiv. https://doi.org/10.48550/arXiv.2304.07683

Ferrer, X., Nuenen, T. van, Such, J. M., Coté, M., & Criado, N. (2021). Bias and Discrimination in AI: A Cross-Disciplinary Perspective. IEEE Technology and Society Magazine, 40(2), 72–80. https://doi.org/10.1109/MTS.2021.3056293

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. Federal Probation, 80, 38. https://heinonline.org/HOL/Page?handle=hein.journals/fedpro80&id=116&div=&collection=

Friedman, B., Hendry, D. G., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. Foundations and Trends® in Human–Computer Interaction, 11(2), 63–125. https://doi.org/10.1561/1100000015

Friedman, B., Kahn, P., Borning, A., Zhang, P., & Galletta, D. (2006). Value Sensitive Design and Information Systems. In The Handbook of Information and Computer Ethics. https://doi.org/10.1007/978-94-007-7844-3_4

Gaggion, N., Echeveste, R., Mansilla, L., Milone, D. H., & Ferrante, E. (2023). Unsupervised bias discovery in medical image segmentation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2309.00451

Ghai, B., & Mueller, K. (2022). D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias (arXiv:2208.05126). arXiv. https://doi.org/10.48550/arXiv.2208.05126

Girdhar, M., Hong, J., & Moore, J. (2023). Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models. IEEE Open Journal of Vehicular Technology, 4, 417–437. https://doi.org/10.1109/OJVT.2023.3265363

Glickman, M., & Sharot, T. (2024). How human-AI feedback loops alter human perceptual, emotional and social judgements. https://doi.org/10.31219/osf.io/c4e7r

Google/ml-fairness-gym. (2024). [Python]. Google. https://github.com/google/ml-fairness-gym (Original work published 2019)

Graubohm, R., Schräder, T., & Maurer, M. (2020). VALUE SENSITIVE DESIGN IN THE DEVELOPMENT OF DRIVERLESS VEHICLES: A CASE STUDY ON AN AUTONOMOUS FAMILY VEHICLE. Proceedings of the Design Society: DESIGN Conference, 1, 907–916. https://doi.org/10.1017/dsd.2020.140

Greenberg, A. (2015). Hackers Remotely Kill a Jeep on the Highway—With Me in It. Wired. https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/

Hagendorff, T. (2020). The Ethics of AI Ethics—An Evaluation of Guidelines. Minds and Machines, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Havens, L., Terras, M., Bach, B., & Alex, B. (2022). Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text. In C. Hardmeier, C. Basta, M. R. Costa-jussà, G. Stanovsky, & H. Gonen (Eds.), Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP) (pp. 30–57). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.gebnlp-1.4

ISO - International Organization for Standardization. (n.d.). ISO. Retrieved March 25, 2024, from https://www.iso.org/home.html

ISO/SAE 21434. Road Vehicles - Cybersecurity Engineering (2021). ISO. https://www.iso.org/standard/70918.html

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 375–385. https://doi.org/10.1145/3442188.3445901

Jagielski, M., Severi, G., Pousette Harger, N., & Oprea, A. (2021). Subpopulation Data Poisoning Attacks. Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 3104–3122. https://doi.org/10.1145/3460120.3485368

Jain, N., Saifullah, K., Wen, Y., Kirchenbauer, J., Shu, M., Saha, A., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Bring Your Own Data! Self-Supervised Evaluation for Large Language Models (arXiv:2306.13651). arXiv. https://doi.org/10.48550/arXiv.2306.13651

Jiang, M., & Fellbaum, C. (2020). Interdependencies of Gender and Race in Contextualized Word Embeddings. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), Proceedings of the Second Workshop on Gender Bias in Natural Language Processing (pp. 17–25). Association for Computational Linguistics. https://aclanthology.org/2020.gebnlp-1.2

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. Researchgate Preprint. https://www.researchgate.net/profile/Akhil-Alfons-Kodiyan/publication/337331539_An_overview_of_ethical_issues_in_using_AI_systems_in_hiring_with_a_case_study_of_Amazon's_AI_based_hiring_tool/links/5dd2aa8d4585156b351d330a/An-overview-of-ethical-issues-in-using-AI-systems-in-hiring-with-a-case-study-of-Amazons-AI-based-hiring-tool.pdf

Koh, P. W., Steinhardt, J., & Liang, P. (2022). Stronger data poisoning attacks break data sanitization defenses. Machine Learning, 111(1), 1–47. https://doi.org/10.1007/s10994-021-06119-y

Latif, E., Zhai, X., Liu, L. (2023) AI Gender Bias, Disparities, and Fairness: Does Training Data Matter? (arXiv:2312.10833). arXiv. https://arxiv.org/abs/2312.10833

Lee, W. (2022). Tools adapted to Ethical Analysis of Data Bias. 29, 200–209. https://doi.org/10.33430/V29N3THIE-2022-0037

Li, X., Chen, Z., Zhang, J. M., Sarro, F., Zhang, Y., & Liu, X. (2023). Dark-Skin Individuals Are at More Risk on the Street: Unmasking Fairness Issues of Autonomous Driving Systems (arXiv:2308.02935). arXiv. https://doi.org/10.48550/arXiv.2308.02935

Lukianets, N., & Pavaloiu, A. (2021). GitHub—OpenEthicsAI/OEDP: Open Ethics Data Passport. https://github.com/OpenEthicsAI/OEDP

Lukianoff, M. (2019). Unbiased Isn't Necessarily Fair. Towards Data Science. https://towardsdatascience.com/confusion-over-bias-7fa46200bd08

Luo, F., Jiang, Y., Zhang, Z., Ren, Y., & Hou, S. (2021). Threat Analysis and Risk Assessment for Connected Vehicles: A Survey. Security and Communication Networks, 2021, e1263820. https://doi.org/10.1155/2021/1263820

Mattu, J. L., Julia Angwin,Lauren Kirchner,Surya. (2016). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Mazijn, C., Prunkl, C., Algaba, A., Danckaert, J., & Ginis, V. (2023). LUCID: Exposing Algorithmic Bias through Inverse Design. Proceedings of the AAAI Conference on Artificial Intelligence, 37(12), Article 12. https://doi.org/10.1609/aaai.v37i12.26683

McCarthy, M. (2023, June 13). What Pothole Detection Teaches Us About AI Bias. AutoVision News. https://www.autovision-news.com/industry/ai-bias-pothole-detection/

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 729–733. https://doi.org/10.1145/3236024.3264833

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 115:1-115:35. https://doi.org/10.1145/3457607

Mehrabi, N., Naveed, M., Morstatter, F., & Galstyan, A. (2021). Exacerbating Algorithmic Bias through Fairness Attacks. Proceedings of the AAAI Conference on Artificial Intelligence, 35(10), Article 10. https://doi.org/10.1609/aaai.v35i10.17080

Nematzadeh, A., Ciampaglia, G. L., Menczer, F., & Flammini, A. (2017, July 3). How algorithmic popularity bias hinders or promotes quality. arXiv.Org. https://doi.org/10.1038/s41598-018-34203-2

Nobles, C., Burrell, D. N., Burton, S. L., Waller, T., Nobles, C., Burrell, D. N., Burton, S. L., & Waller, T. (2023). Driving Into Cybersecurity Trouble With Autonomous Vehicles (driving-into-cybersecurity-trouble-with-autonomous-vehicles) [Chapter]. Https://Services.Igi-Global.Com/Resolvedoi/Resolve.Aspx?Doi=10.4018/978-1-6684-7207-1.Ch013;IGIGlobal.https://www.igi-global.com/gateway/chapter/www.igi-global.com/gateway/chapter/321022

Ngomo, A. (2019, December 11). Can biased people create unbiased algorithms? https://www.linkedin.com/pulse/can-biased-people-create-unbiased-algorithms-alicia-ngomo/

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1356. https://doi.org/10.1002/widm.1356

NTSB. (2020). Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018 (ACCIDENT REPORT PB2019-101402). NATIONAL TRANSPORTATION SAFETY BOARD. https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science (New York, N.Y.), 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Open Ethics, Lukianets, N., Pavaloiu, A. (2021). OpenEthicsAI/OEDP: Public release for the Open Ethics Data Passport v0.2-alpha (v0.2-alpha). Zenodo. https://doi.org/10.5281/zenodo.5137849

Pagallo, U. (2022). The Politics of Self-Driving Cars: Soft Ethics, Hard Law, Big Business, Social Norms. In R. Jenkins, D. Cerný, & T. Hríbek (Eds.), Autonomous Vehicle Ethics: The Trolley Problem and Beyond (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780197639191.003.0010

Pan, Z., Niu, L., & Zhang, L. (2023). Geometric Inductive Biases for Identifiable Unsupervised Learning of Disentangled Representations. Proceedings of the AAAI Conference on Artificial Intelligence, 37(8), Article 8. https://doi.org/10.1609/aaai.v37i8.26123

Park, Y., & Hu, J. (2023). Bias in Artificial Intelligence: Basic Primer. Clinical Journal of the American Society of Nephrology, 18(3), 394. https://doi.org/10.2215/CJN.0000000000000078

Perkowitz, S. (2021). The bias in the machine: Facial recognition technology and racial disparities. MIT Case Studies in Social and Ethical Responsibilities of Computing, Winter 2021, 1-16. https://doi.org/10.21428/2c646de5.62272586

Project Implicit. (n.d.). Retrieved April 7, 2024, from https://implicit.harvard.edu/implicit/index.jsp

Qayyum, A., Usama, M., Qadir, J., & Al-Fuqaha, A. (2020). Securing Connected & Autonomous Vehicles: Challenges Posed by Adversarial Machine Learning and The Way

Forward. IEEE Communications Surveys & Tutorials, 22(2), 998–1026. https://doi.org/10.1109/COMST.2020.2975048

Qian, K., Beirami, A., Lin, Z., De, A., Geramifard, A., Yu, Z., & Sankar, C. (2022). Annotation Inconsistency and Entity Bias in MultiWOZ (arXiv:2105.14150). arXiv. http://arxiv.org/abs/2105.14150

Reagan, M. (2021, April 2). Understanding Bias and Fairness in AI Systems. Medium. https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

Rouzrokh, P., Khosravi, B., Faghani, S., Moassefi, M., Vera Garcia, D. V., Singh, Y., Zhang, K., Conte, G. M., & Erickson, B. J. (2022). Mitigating Bias in Radiology Machine Learning: Data Handling. Radiology: Artificial Intelligence, 4(5), e210290. https://doi.org/10.1148/ryai.210290

Ruth, C. (2023, August 24). The bias against bias: using bias intentionally in artificial intelligence. IEEE Standards Association. https://standards.ieee.org/beyond-standards/the-bias-against-bias/

SAE International. (n.d.). Retrieved March 25, 2024, from https://www.sae.org/site/

Servin, C., Kreinovich, V., & Shahbazova, S. N. (2023). Ethical Dilemma of Self-driving Cars: Conservative Solution. In S. N. Shahbazova, A. M. Abbasov, V. Kreinovich, J. Kacprzyk, & I. Z. Batyrshin (Eds.), Recent Developments and the New Directions of Research, Foundations, and Applications: Selected Papers of the 8th World Conference on Soft Computing, February 03–05, 2022, Baku, Azerbaijan, Vol. II (pp. 93–98). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-23476-7_9

Solans, D., Biggio, B., & Castillo, C. (2021). Poisoning Attacks on Algorithmic Fairness. In F. Hutter, K. Kersting, J. Lijffijt, & I. Valera (Eds.), Machine Learning and Knowledge Discovery in Databases (pp. 162–177). Springer International Publishing. https://doi.org/10.1007/978-3-030-67658-2_10

Srivastava, S., & Sinha, K. (2023). From Bias to Fairness: A Review of Ethical Considerations and Mitigation Strategies in Artificial Intelligence. International Journal for Research in Applied Science and Engineering Technology, 11(3), 2247–2251. https://doi.org/10.22214/ijraset.2023.49990

Stoev, T., Yordanova, K., & Tonkin, E. L. (2023). Experiencing Annotation: Emotion, Motivation and Bias in Annotation Tasks. 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops), 534–539. https://doi.org/10.1109/PerComWorkshops56833.2023.10150364

Stray, J. (2023). The AI Learns to Lie to Please You: Preventing Biased Feedback Loops in Machine-Assisted Intelligence Analysis. Analytics, 2(2), Article 2. https://doi.org/10.3390/analytics2020020

Street Bump | Boston.gov. (2017, April 21). https://www.boston.gov/transportation/street-bump

Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Equity and Access in Algorithms, Mechanisms, and Optimization, 1–9. https://doi.org/10.1145/3465416.3483305

Tabassi, E. (2023). AI Risk Management Framework: AI RMF (1.0) National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Taori, R., & Hashimoto, T. B. (2022). Data Feedback Loops: Model-driven Amplification of Dataset Biases (arXiv:2209.03942). arXiv. https://doi.org/10.48550/arXiv.2209.03942

Tensorflow/fairness-indicators. (2024). [Jupyter Notebook]. tensorflow. https://github.com/tensorflow/fairness-indicators (Original work published 2019)

The precipice: Existential risk and the future of humanity. (2020). Hachette Books.

UN Regulation No. 155—Cyber security and cyber security management system | UNECE. (2021). https://unece.org/transport/documents/2021/03/standards/un-regulation-no-155-cyber-security-and-cyber-security

Utsab, Khakurel., Ghada, Abdelmoumin., Aakriti, Bajracharya., Danda, B., Rawat. (2022). Exploring bias and fairness in artificial intelligence and machine learning algorithms. 12113:1211324-1211324. doi: 10.1117/12.2621282

Van, M.-H., Du, W., Wu, X., & Lu, A. (2022). Poisoning Attacks on Fair Machine Learning. In A. Bhattacharya, J. Lee Mong Li, D. Agrawal, P. K. Reddy, M. Mohania, A. Mondal, V. Goyal, & R. Uday Kiran (Eds.), Database Systems for Advanced Applications (pp. 370–386). Springer International Publishing. https://doi.org/10.1007/978-3-031-00123-9_30

van Stein, B., Vermetten, D., Caraffini, F., & Kononova, A. V. (2023). Deep-BIAS: Detecting Structural Bias using Explainable AI (arXiv:2304.01869). arXiv. https://doi.org/10.48550/arXiv.2304.01869

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms. The New England Journal of Medicine, 383(9), 874–882. https://doi.org/10.1056/NEJMms2004740

Wan, Y., Wang, W., He, P., Gu, J., Bai, H., & Lyu, M. (2023, May 21). BiasAsker: Measuring the Bias in Conversational AI System. arXiv.Org. https://arxiv.org/abs/2305.12434v1

Wang, L. (2016). Discovering phase transitions with unsupervised learning. Physical Review B, 94(19), 195105. https://doi.org/10.1103/PhysRevB.94.195105

Watson, D. (2023). On the Philosophy of Unsupervised Learning. Philosophy & Technology, 36(2). https://doi.org/10.1007/s13347-023-00635-6

Willetts, M., Roberts, S. J., & Holmes, C. C. (2019). Semi-Unsupervised Learning: Clustering andClassifying using Ultra-Sparse Labels. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1901.08560

Wilms, R., Mäthner, E., Winnen, L., & Lanwehr, R. (2021). Omitted variable bias: A threat to estimating causal relationships. Methods in Psychology, 5, 100075. https://doi.org/10.1016/j.metip.2021.100075

Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). Bias in generative AI. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2403.02726