



Bias in Machine Learning Algorithms for Automotive Applications and Related Ethical Concerns

BACHELORARBEIT

zur Erlangung des akademischen Grades

Bachelor of Science

im Rahmen des Studiums

Software Information Engineering

eingereicht von

Martin Nang'ole

Matrikelnummer 11776827

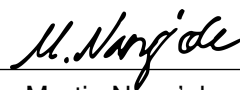
an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Mag. Dr. Margrit Gelautz

Mitwirkung: Dipl.-Ing. / BSc Dominik Schörkhuber

Wien, 12. August 2022



Martin Nang'ole

Margrit Gelautz



Bias in Machine Learning Algorithms for Automotive Applications and Related Ethical Concerns

BACHELOR'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

Software Information Engineering

by

Martin Nang'ole

Registration Number 11776827

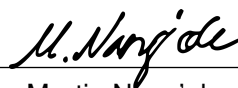
to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Mag. Dr. Margrit Gelautz

Assistance: Dipl.-Ing. / BSc Dominik Schörkhuber

Vienna, 12th August, 2022


Martin Nang'ole

Margrit Gelautz

Erklärung zur Verfassung der Arbeit

Martin Nang'ole

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 12. August 2022



Martin Nang'ole

Danksagung

Mein besonderer Dank gilt meinem Betreuer Dominik Schörkhuber und Professorin Margrit Gelautz für die Beratung und die wertvollen Anregungen, die das Schreiben meiner Arbeit zu diesem Thema überhaupt erst möglich gemacht haben.

Außerdem möchte ich mich aufrichtig bei meiner Familie für die durchgehende Unterstützung und Motivation bedanken, die während der Bachelorarbeit und während meines gesamten Studiums entscheidend waren.

Diese Arbeit ist in Anlehnung an die Forschungsprojekte „SyntheticCabin“ und „SmartProtect“ der Computer Vision Lab (CVL) Forschungseinrichtung im Institut für Visual Computing & Human-Centered Technology der TU Wien entstanden.

Acknowledgements

I would like to give special thanks to my supervisor Dominik Schörkhuber and professor Margrit Gelautz for providing me with the valuable input and guidance that made writing my thesis on this topic possible in the first place.

I would also like to sincerely thank my family for their consistent support and motivation which were crucial during the writing of my Bachelor's thesis and throughout my studies.

This work is following the research projects "SyntheticCabin" and "SmartProtect" of the Computer Vision Lab (CVL) research unit at the Institute of Visual Computing & Human-Centered Technology at the TU Wien.

Kurzfassung

Machine Learning (ML) ist ein ständig wachsendes Forschungsgebiet, das in den letzten Jahrzehnten bahnbrechende Leistungen bei komplexen Aufgabenstellungen erzielt hat, die einst als ausschließlich dem Menschen vorbehalten galten. Angesichts der steigenden Nachfrage nach automatisierten Entscheidungssystemen basierend auf ML bereitet sich die Automobilindustrie jetzt auf die Entwicklung und Einführung autonomer Fahrzeuge vor. Gleichzeitig gibt jedoch eine wachsende Zahl von Problemen im Zusammenhang mit Bias von ML Algorithmen Anlass zur Sorge über den Grad an Fairness von Algorithmen. Außerdem stellt sich die Frage inwiefern man ML Algorithmen vertrauen kann, faire Entscheidungen unter realistischen Umständen zu treffen, insbesondere, wenn die Sicherheit eines Menschen auf dem Spiel steht.

Ziel dieser Arbeit ist es, Literatur rund um das Thema algorithmischer Biases in Computer Vision (CV) zu behandeln und Informationen über mögliche Ursachen von algorithmischem Bias, die praktischen Auswirkungen von Bias in CV und konkrete Strategien zur Überwindung von Bias in Automobilanwendungen zusammenzufassen. Dabei werden relevante Arbeiten behandelt, beispielhafte Fälle von algorithmischem Bias in der Praxis aus jüngster Zeit diskutiert und einige der relevanten Theorien aus den Feldern Computer Vision und maschinellem Lernen thematisiert.

Abstract

Machine learning (ML) is an ever-expanding field of research that has garnered attention over the past decades for breakthrough performances in the completion of complex tasks that were once thought to be only achievable to humans. With an increasing demand for automated decision systems based on ML, the automotive industry is now preparing for the development and widespread adoption of autonomous vehicles. At the same time, however, a growing number of issues related to bias in the performance of ML algorithms is raising concerns about the current state of fairness in algorithms and whether or not ML algorithms can be trusted to make fair decisions under realistic circumstances, particularly when the safety of a human is at risk.

This thesis aims to review literature surrounding the topic of algorithmic bias in computer vision (CV) and provide information about possible causes of algorithmic bias, the practical effects of bias in CV and concrete strategies for overcoming bias in automotive applications specifically. In doing so, we delve into relevant works, discuss notable instances of algorithmic bias in practice within recent memory and cover some of the underlying theories in computer vision and machine learning.

Contents

Kurzfassung	xi
Abstract	xiii
Contents	xv
1 Introduction	1
1.1 Research Questions	1
1.2 Thesis Structure	2
2 Algorithmic Bias in Automotive Machine Learning	3
2.1 Introduction to Computer Vision Algorithms	3
2.2 Types of Bias	6
2.3 Examples in Automotive Applications	9
2.4 Bias Recognition	12
2.5 Bias Reduction	16
3 Current Real-Life Impacts of Algorithmic Bias in Computer Vision	23
3.1 Object Recognition	23
3.2 Facial Recognition	27
3.3 Gaze and Pose Estimation	29
3.4 Scenario Clustering	31
4 Ethics of Algorithmic Bias in Automotive Applications	33
4.1 Moral Autonomous Decision Making	33
4.2 Ethical Deployment and Consumer Trust	35
4.3 Safety in Tail- and Open-Class Scenarios	36
5 Conclusion	39
List of Figures	41
Bibliography	43

Introduction

With the beginning of a new era of AI and Machine Learning applications in the automotive sector, it is now more important than ever to address the ethical issues that come with the reinvention of personal transportation. As the development of self-driving vehicles has progressed, however, autonomous decision systems have at the same time been increasingly under scrutiny for displaying bias in practice. With this in mind, this paper attempts to provide insight into the concept of algorithmic bias, summarize its effects on contemporary computer vision (CV) models and present ethical methodologies capable of mitigating biased algorithms in autonomous driving. To begin with, this chapter will first describe the research questions this work sets out to answer, before providing an overview for the general structure of this thesis.

1.1 Research Questions

The aim of this thesis is to answer the following research questions regarding bias in automotive and computer vision applications:

1. **How does bias emerge in machine learning algorithms?**

We attempt to answer this question by establishing a list of potential sources of algorithmic bias as well as factors that encourage the introduction of new biases in algorithms. Also, we explain during which of the developmental phases these biases are integrated or become apparent. Furthermore, we establish when algorithmic bias is desirable and when it is problematic. In doing so, we consider in which cases the presence of algorithmic bias can lead to unfair losses in accuracy as well as if and how algorithmic bias can be used to combat bias in the real world. Methods for reducing unwanted algorithmic bias are also mentioned. In this context, we explore the potential usefulness of data synthesis and explainable AI.

2. What are the practical consequences of bias?

We try to assess the recent impacts of bias in machine learning research as well as current solutions and proposals for bias mitigation. While our focus lies on autonomous driving systems, we also investigate the effects of bias in other computer vision fields. There, we examine which classes and demographic groups tend to be disadvantaged by algorithmic bias in computer vision and pay particular attention to the short to mid-term effects of deploying biased autonomous decision systems for those negatively affected by the bias.

3. How can bias in autonomous driving systems be addressed?

A summary of ethical areas of concern regarding the systemic reduction or elimination of bias in autonomous driving is given. We provide an outlook over the methodologies and technological advances that could improve such efforts, taking into account the decisions manufacturers have to make when defining the priorities of safety and performance. Subsequently, we examine criticisms of algorithm design choices made by researchers and industry experts along with recommendations on how to ensure public safety during deployment and increase trust in autonomous vehicles (AV).

1.2 Thesis Structure

In Chapter 2, a basic rundown of algorithmic biases in machine learning is given. Chapter 3 then focuses on the practical effects of algorithmic bias in a variety of Computer Vision disciplines. Following that, Chapter 4 discusses the ethical issues relevant to the dismantling of algorithmic bias in automotive applications as well as the potential dangers of failing to do so. Finally, Chapter 5 summarizes the findings of this thesis and provides a conclusion, including suggestions for future research.

Algorithmic Bias in Automotive Machine Learning

In this chapter, we examine commonly used categories for defining different types of bias in machine learning algorithms and computer systems in general. As an introduction to machine learning algorithms in computer vision, we first take a look at fundamental terms and methods relevant throughout the development stages of a computer vision model, before considering taxonomies of algorithmic bias and at which stages of the development process different bias types are introduced. In addition, we discuss real-world examples of algorithmic bias in automotive applications. Finally, methods both for recognizing and reducing bias are explained.

2.1 Introduction to Computer Vision Algorithms

2.1.1 Design

According to Tsotsos et al. [1], Computer Vision algorithms can be distinguished in terms of two different approaches: Classical, theory-driven, and modern, data-driven algorithms.

Theory-driven algorithms are built on the idea that human perception can be approximated with a sufficient theoretical understanding of the geometry and physics relevant to the captured scene. These models base their perception of the real world on scientific knowledge, capturing causal relationships between physical variables, and can be understood as the more scientifically conventional approach compared to data-driven algorithms according to Karpatne et al. [2]. In machine learning, this scientific approach deepens the theoretical knowledge of image processing and analysis by developing general theories that can be applied to other systems, but typically struggles in more complex systems with many unique cases. Karpatne et al. note that models that rely on theory-based or -guided algorithms ensure a consistency with scientific knowledge

in the simulation of real world data, which allows for better generalization to other domains. Examples of field-defining theory-driven algorithms include the classic works of Rosenfeld[3] and Faugeras[4], which significantly pushed the envelope in computer vision both in theory and in practice. Another more recent example is the work of Tommasi et al.[5], which attempts to create an algorithm capable of learning general information across 12 datasets for multiple computer vision domains. For this purpose, unique attributes for object categories (i.e. "features") were manually pre-defined in order for the algorithm to capture a general understanding and retain accuracy for unseen data that significantly deviates from the training data.

In contrast, **data-driven algorithms**, according to Bach et al.[6], rely on data and data analytics for their design and development. They depend on large amounts of data as input to recognize patterns and derive accurate predictions. Modern data-driven algorithms benefit from having access to large open-source image datasets, which in recent years has allowed for data-driven approaches to dominate and replace theory-centric approaches in Machine Learning and Computer Vision fields, according to observations from Tsotsos et al.. A similar sentiment is expressed by Karpapartne et al., describing a recent shift in the role of data-science models for scientific discovery from simple analysis tools to complete frameworks. Data-driven approaches are capable of addressing more complex systems, but lack universality in their theoretical understanding of a system's fundamental rules. One example of a data-driven model type are so-called **Deep Neural Networks (DNN)**, which, as pointed out by Luckow et al.[7], are especially useful for interpreting unstructured data such as text, speech and images. DNNs are modeled after the human brain and consist of multiple layers of "neurons", which each take multiple inputs and generate one output, aiming to iteratively and cooperatively reach a desired overall output. Due to the fact that most data in the automotive and computer vision fields are unstructured, DNNs have established themselves as effective methods for tackling common CV problems. Another variant of neural networks called **Convolutional Neural Networks (CNN)** is primarily used for image-based pattern recognition and profits from image-specific optimizations in its architecture that significantly improve scalability and efficiency compared to other neural networks, as described by O'Shea and Nash[8].

Computer Vision systems, as explained by Luckow et al., are commonly used in combination with deep-learning techniques in the automotive industry for a variety of purposes. Some of the general use cases that are relevant to the automotive field listed by the aforementioned authors include autonomous driving systems (ADS), which often require the learning of driving scenarios and behaviors using vast amounts of input data, conversational user interfaces, which attempt to provide more natural interactions between user and vehicle through voice commands using the Internet of Things (IoT), and maintenance of large datasets during data collection. In the domain of autonomous driving, more specific use cases include the learning of driving scenarios (i.e. learning and predicting driving behavior), object detection and object tracking (e.g. road, vehicles, pedestrians, lanes) and, in more complex systems, planning long-term decision-making based on

real-time data.

2.1.2 Data Acquisition and Labeling

The effective use of modern, data-driven algorithms for automotive computer vision systems not only requires extensive gathering, but also the labelling of data points. Data labeling is the process of assigning the correct solution or classification to a given data point. Luckow et al. note that the availability of labeled data is crucial for efficient training in deep learning systems, especially for advanced use cases such as autonomous driving. However, given that for data-driven algorithms larger datasets typically yield better training results as well as the high costs associated with labeling data, Wang et al. [9] propose incorporating unlabelled datasets into the training of deep learning models. A survey on the data management and machine learning landscape conducted by Roh et al. [10] suggests that data acquisition can be categorized into discovery (i.e. collecting data manually from unstructured data lakes, on the web or by accessing shared, collaborative datasets), augmentation (i.e. adding information to existing data) and generation. Data labeling methods on the other hand, are broadly distinguished into manual labeling, weak labeling (i.e. low-quality labeling in large quantities) and semi-supervised learning, the latter of which is explained in Section 2.1.3. Gathering sufficient data of high quality for deep learning is considered by Roh et al. to be a bottleneck in current research, in part due to manual searching becoming unsustainable at a larger scale as well as a lack of overall training data in novel applications. In addition to that, despite proposals for autolabeling processes such as that of Piewak et al. [11], labeling massive datasets for data-driven Computer Vision algorithms remains a challenge.

2.1.3 Model Training and Evaluation

Regarding the training of CV algorithms, O'Shea and Nash explain that **supervised learning** is typically applied for successful classifications in image-focused pattern-recognition tasks. For this learning method, a neural network model is provided with pre-labelled data as input and trains towards arriving at the correct output using the provided input vectors. **Unsupervised learning**, on the other hand, does not use labels, but measures success in the classification of a data point by the increase or decrease of a pre-defined cost function instead. This method, explain Sathya and Abraham [12], is used to find hidden patterns in unlabeled samples without providing an error signal during the solution finding process, which allows it to find unconsidered patterns. According to O'Shea and Nash, however, unsupervised learning is less suitable for image pattern recognition than supervised learning. **Semisupervised learning**, as explained by Wang et al. [9] is a hybrid-method that attempts to make efficient use of available data, while keeping acquisition cost for data low. First, a deep learning model is trained on a labelled dataset of manageable size using a supervised learning algorithm. Afterwards, unlabelled data is introduced to further refine detection accuracy.

After a machine learning model has been initially trained and fine-tuned on corresponding

training and validation datasets, it is tested against a test dataset, a subset of the available data that is separate from the training and validation sets. Once completed, evaluation is typically done using the confusion matrix, which provides information on True/False Positives/Negatives regarding the prediction of the classifier and the actual value in the form of a table. Using this data, one can then calculate measurements such as precision and accuracy (see Section 2.4).

Within all of the aforementioned development stages in machine learning, different types of algorithmic bias can be introduced, which we will discuss in detail in the following section.

2.2 Types of Bias

Friedman and Nissenbaum [13] developed three distinct categories of bias present in computer systems: Pre-existing, technical and emergent bias.

- **Preexisting bias** does not originate from the creation of the computer system itself but instead from either societal, institutional, or personal practices. It can enter a system through both conscious and unconscious efforts, making even algorithms developed with the best intentions susceptible to this type of bias.
- **Technical bias** on the other hand stems from reaching technical constraints or making misjudgments in the application of an algorithm. These biases, according to Friedman and Nissenbaum, can be further categorized into bias stemming from the computer tools used (ie. software, hardware and peripheral limitations), decontextualized algorithms, which fail to consider all significant conditions fairly for every group, imperfect random number generation, and bias arising due to an attempt to formalize a process that is non-formal or ambiguous to humans.
- The third and final category presented by Friedman and Nissenbaum is referred to as **emergent bias**. This type of bias generally becomes apparent after the system has been developed and actively used for a period of time. It can be caused by a significant mismatch between the population that uses the system and the one that was assumed by developers regarding expertise, moral and ethical values or otherwise. Newly emerging knowledge in society that was not accounted for in the system's initial design can also be a cause of emergent bias.

Alternatively, Danks and London [14] provide a taxonomy of algorithmic bias, according to which bias in machine learning models can be categorized as follows: Training Data Bias, Algorithmic Focus Bias, Algorithmic Processing Bias, Transfer Context Bias and Interpretation Bias.

- **Training Data Bias** is introduced to a model through bias in the training or input data it is provided with. Even when assuming a “neutral” algorithm, meaning

an algorithm that does not inherently contain any bias in its rules, training on a biased dataset can result in significant deviance from the expected outcome, thereby displaying strong bias in practice. This type of bias can go unnoticed for some time, but once the finished model is applied within a context that lies outside of the training data's scope, it becomes apparent immediately. A prime example for training data bias found within autonomous driving is the following: When a model based on a relatively neutral driving algorithm is trained exclusively with data from one part of the world and then used in a different part of the world, the local and regional driving behaviors the algorithm has been subjected to will become visible as they are outside of their original context.

- **Algorithmic Focus Bias** is a type of bias that is generated when differentiating between certain information and if or how it ought to be used in the input data. As an example, Danks and London present the case of a neutral algorithm that is capable of using legally protected information in a decision-making process. If the model chooses to apply the protected information in order to appear statistically unbiased, it presents bias towards the current legal standard, whereas if the model does not use the information, the outcome becomes statistically biased. This is an example of a “forced choice” between two different algorithmic biases, since, in this case, any decision on how to design the training data leads to a biased algorithm due to the contradiction taking place between legal and statistical norms.
- **Algorithmic Processing Bias** encompasses bias that is not introduced to, but rather generated by an algorithm. This inherent bias can happen accidentally or on purpose through the use of a statically biased estimator as part of the algorithm, for instance. Therefore, whether a bias of this type has beneficial or hurtful effects depends on the statistical results the algorithm delivers and whether or not they represent the desired outcome.
- Another category of bias Danks and London put forward in their taxonomy is called **Transfer Context Bias**. This algorithmic bias is caused as a result of utilizing an algorithm outside of the context it was intended for or merely assumed to function in.
- Lastly, **interpretation bias** is a class of algorithmic bias that is characterized by the misinterpretation of an algorithm's function and occurs when the information of an algorithm's output does not meet the information requirements posed by the user or system in question. When the information presented by output values of an algorithm is left to statistical, moral or legal interpretation before it can be applied, biases are typically present.

While Friedman and Nissenbaum's autonomy categorizes biases for computer systems in a broader, more socio-technical context, Danks and London's more recent taxonomy is grounded in the practice of machine learning applications. Despite this, however, we find sufficient overlap between the taxonomies to combine these two taxonomies, as depicted

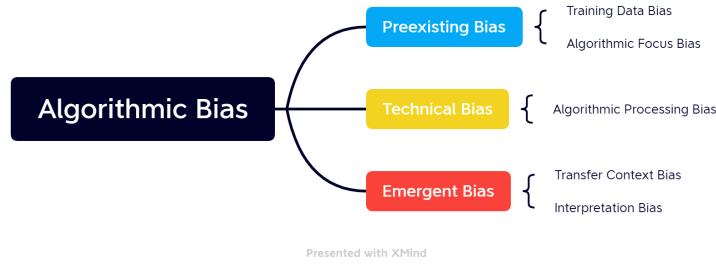


Figure 2.1: A merge of the taxonomies for algorithmic biases provided by Friedman and Nissenbaum and Danks and London.

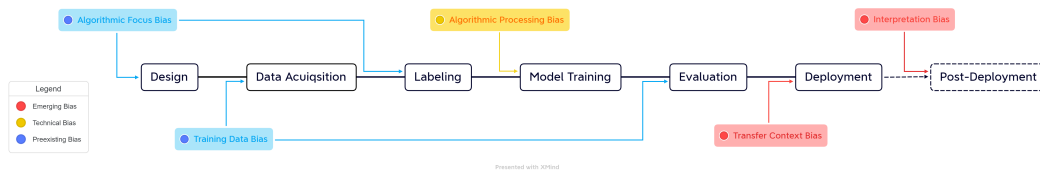


Figure 2.2: A graph showcasing at which stages of development certain types of algorithmic biases are introduced.

in Figure 2.1. Algorithmic Focus Bias and Training Data can be considered preexisting biases, Algorithmic Processing Bias can originate from preexisting bias in its design, but can otherwise be ascribed to technical bias stemming from limitations in technologies or decontextualized algorithms and, finally, Transfer Context Bias and Interpretation Bias can be categorized as emergent biases, due to their problematic nature only becoming clear during or after deployment. One notable exception to this mapping is the long-tail effect, a subset of training data bias, which is further discussed in "Long-Tail Distribution" in Section 2.2.2.

Both taxonomies also list categories of bias that occur in various stages of development of computer vision algorithms, including design, data collection, testing and (post-)deployment. Hence, on further examination, we can also derive a timeline of potential points of entrance for different bias types during development stages, as shown in Figure 2.2: Algorithmic Focus Bias is mainly introduced either during the design phase of a CV algorithm or during labeling of data, as developers decide which data categories should become fundamentally relevant to the algorithm. Training Data Bias originates from data acquisition, but can arguably also appear in another form during evaluation, since datasets used for evaluation can also display preexisting biases. Furthermore, Transfer Context Bias becomes visible during the deployment of the computer vision system, while Interpretation Bias only becomes apparent after the system has been in use for extended periods of time. In the following subchapter, we will exemplify the mentioned categories of bias using examples from automotive applications.

2.3 Examples in Automotive Applications

2.3.1 Pre-existing Bias

Training Data Bias

Training Data bias is introduced through data that is skewed or inaccurate in its representation of the real world. When gathering data for automotive machine learning models, simply the uncommonness of important scenarios can already allow for this bias to be introduced. As mentioned by Samuel et al. [15], real-world data is predominantly unbalanced and follows a long-tail distribution pattern containing head and tail classes. While head classes represent categories that are frequently found in the dataset, tail classes include conditions or domains which are either inherently uncommon or less frequently captured during data gathering. Liu et al. [16] extend this concept in their description of Open Long-tailed Recognition: In addition to tail classes, which require a transfer of knowledge for the generation of good feature representations, and head classes, Liu et al. also consider the open-endedness of real-world data through open classes, which are samples that are completely unknown to a trained recognition system and require sensitivity to novelty to be reasonably dealt with. The long-tail problem can be considered as both a pre-existing and technical bias as it can originate from technical constraints during development (e.g. manual data acquisition, labeling) as well as from the inherently biased nature of information in the real world (see Figure 2.2). In other instances, training data bias is introduced subconsciously to an image dataset while collecting real-world data. A comparison study by Torroalba and Efros [17] indicates that object recognition datasets are likely to contain inherent bias, describing the evolution of image datasets as a vicious cycle in which one dataset is often created as a reaction to or rejection of current standards set by contemporaries, thereby creating new biases or perpetuating old ones in a different way. To test this hypothesis, cross-data generalization, a process in machine learning in which a machine learning model is trained on one dataset and tested with data from an entirely different dataset, was conducted between five object detection datasets, with the authors concluding that only little generalization is taking place partly due to a clear presence of bias in the selection of images within each dataset. This type of bias is further exemplified in a study from De Vries et al. [18], which suggests that geographical and language bias during data collection are some of the leading causes for significant performance loss in object recognition for images from undersampled geographical locations (see Section 3.3 for more details). As one concrete automotive example of training data bias, Marathe et al. [19] note the need for research on structural pre-existing biases such as gender and age bias in current pedestrian detection systems.

Algorithmic Focus Bias

As previously mentioned, algorithmic focus bias occurs in its simplest form when a parameter that is irrelevant to the model's performance is taken into consideration during its decision making (e.g. considering the market price of a vehicle in traffic

while establishing the risk distribution over traffic participants). However, more nuanced variants of this form of bias become noticeable in scenarios similar to the "Trolley Problem", in which a decision between life and death is to be made and where a decision on which factors are relevant will inevitably unveil moral bias in an algorithm's design. In such instances, Danks and London suggest, any design choice made by the developers of a machine learning algorithm based on moral standards could realistically be considered biased and inappropriate by other parties. As an example, if AVs were to treat the security of everyone involved in traffic equally, its decisions would likely be perceived as morally biased by those arguing that harm for the vehicle's own passengers must be minimized at all times and that the algorithm is therefore failing to adequately disregard irrelevant factors such as the safety of other traffic participants. The Trolley Problem is a heavily debated topic in the field of self-driving cars due to its ethical nature and requires centering in further research on ethical AV algorithms according to Geisslinger et al. [20]. Another notable example of this type of bias comes from a paper by Yudkowsky [21], mentioning the parable of a computer vision model trained for the purpose of distinguishing between images of camouflaged tanks and those of forests. According to this version of the story, the model ended up discriminating between the two categories perfectly on the given training and test dataset. However, once the model was tested against new images from other datasets, researchers found that the neural network had learned to differentiate between cloudy days and sunny days, as all images of tanks in the original dataset were shot in on cloudy days, while the pictures of plain forest were only taken on sunny days.

2.3.2 Technical Bias

Long-Tail Distribution

Limitations in the technologies and methods used in data collection can be, next to pre-existing bias, another cause of the long-tail effect. In image datasets for object detection purposes, technical capabilities can create obstacles for the gathering of tail class data as well as negatively affect the quality of tail class samples, which can include events such as extreme or atypical weather, certain times of day or objects appearing out of the ordinary. For instance, if the camera system used to capture image samples is not tuned to accommodate for low-light scenarios, an artificial scarcity of high-quality low-light data is created, despite the abundance of low-light conditions in the real world. While this scarcity in part stems from a decision made by the researchers creating the training dataset, technical constraints such as the inability of a given camera system to produce high quality samples for all lighting conditions under the same settings, can also be at fault. Examples of long-tail scenarios for autonomous driving datasets in the case of Wang et al. [22], include sudden accidents, rain or erratic human behavior (i.e. jaywalking). Other potential examples include the presence of less common traffic participants such as bicycles, buses, strollers, walkers, wheelchairs or emergency vehicles. When a machine learning model is trained on an image dataset which does not account for data imbalance, the head classes tend to benefit from disproportionately high accuracy

when compared to the average performance of the model, while tail classes suffer from disproportionately low accuracy. This is due to the larger amount of head class encounters during representation learning (also known as feature learning), the process in which a machine learning model learns to extract features that are representative of certain classes. This leads to lower-quality representations for tail classes, which causes higher uncertainty and failure rates when attempting to recognize them, which in turn creates overconfidence towards head classes. Even worse, in the case of computer vision models, it is common for popular CV datasets to fail at including rare, but crucial classes such as wheelchairs, leading to very frequent misclassification of wheelchairs, two-wheeled vehicles and bicycles, as illustrated by Zhang et al. [23] in their attempt at enabling accessibility-driven vision-based autonomous system. Referring to popular diverse benchmarks COCO and Cityscape, the authors found long-tail taxonomies for vehicle and person categories to lack a category for samples regarding mobility aid or person type. Yu et al. [24] illustrate the long-tail distribution problem in object detection for Unmanned Aerial Vehicle (UAV) images and further find that typical re-balancing methods, which would mitigate class bias in natural image datasets, do not yield the same results in the case of UAV datasets. Recent research by Wilson et al. [25] also addresses the lack of diversity in current image datasets for autonomous vehicles (AV) by creating a sizeable dataset with a focus on diverse and complex scenarios in an attempt to push current boundaries of autonomous training datasets - both in size and quality - while sufficiently representing commonly raised examples of under-represented long-tail classes. In spite of that, since collecting data on scarce or unknown tail classes to balance a dataset can come with astounding, exponentially increasing costs, as mentioned by Yu et al., efforts of improving accuracy and robustness for tail classes often focus on ways to reduce bias by manipulating existing data, which will be further explained in Section 2.5.

Algorithmic Processing Bias

As explained in the taxonomy provided by Danks and London, algorithmic processing biases can occur on accident in the form of estimations that are unintentionally statistically biased, in which case their origin is a preexisting bias, but is otherwise created deliberately as a means of correcting other instances of bias. Methods for mitigating harmful biases caused by the long-tail distribution problem (e.g. re-weighting, two-stage fine-tuning), which are discussed in chapter 2.5, all fall into the technical category of algorithmic processing biases.

2.3.3 Emergent Bias

Due to the fact that emergent biases per-definition require social and user context as well as some time following the release of the product, the listed examples of emergent biases in automotive applications are merely hypothetical.

Transfer Context Bias

Transfer context bias stems from the application of an algorithm outside of its intended environment or use case. This bias could appear in an ADS after its release, when the assumptions made about a region's traffic laws do not align with the true context of its utilization. This would be the case, according to Danks and London, if an autonomous driving model were trained with input data purely from right-driving countries and then deployed in a left-driving country, for instance. This inappropriate deployment of an autonomous driving system would presumably be noticeable in its lacking ability to accurately follow a law-abiding route on the road. Furthermore, an AV might exhibit transfer context bias when assumptions on local driving practices or common weather conditions in its training data are inaccurate compared to those found in the real world. Transfer learning, in which according to Thuang et al. [26] knowledge from other related domains is transferred to a new domain, can also be considered as a potential source of transfer context bias since the derived knowledge might not be relevant in the new context.

Interpretation Bias

Since interpretation bias is caused by misinterpreting the output of an algorithm in practice, Danks and London find one example of this bias in autonomous vehicles in the case of an AV's output regarding the risk of fatal injury in a given traffic scenario. Depending on how the information is presented, the algorithm might influence developers differently in how the risk is perceived and whom to prioritize. If the algorithm were to output the risk of every traffic party involved in an order that prefers, for example, passengers of its own vehicle over other individuals, even if the risk between them was equal, the user or other sections of the system that rely on its output as input might falsely interpret that to mean that the risk of others is lower or to be less prioritized. Similarly, if the results are not displayed in a consistent way for every scenario, the system or the user might misinterpret the information. While this example also includes the aforementioned Trolley Problem, interpretation bias specifically stems from a misunderstanding of the algorithmic output and not the output itself.

2.4 Bias Recognition

2.4.1 Metrics for Bias Recognition

Reducing bias in computer vision algorithms with imbalanced datasets requires metrics that allow for the quantification and estimation of discrepancies in prediction accuracy between classes. Therefore, research by Fawcett [27] suggests that instead of purely relying on accuracy, a measurement which is inadequate for imbalanced class distribution as it hides class-specific losses, metrics such as precision, recall and F1 score ought to be considered as well.

Recall is defined by Fawcett as the ratio of true positives, meaning the amount of positives that were correctly classified, divided by the total amount of actual positives present.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Precision on the other hand is equated by Fawcett to the definition of the positive predictive value (PPV), which is the ratio of true positives relative to the overall amount of positive predictions made by the model.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Both recall and precision are capable of providing insight into the failure rates of classes as they differentiate between false positive and false negative errors in the predictions of a model, allowing for an analysis of which of the classes in the confusion matrix - positive or negative - tends to be misclassified more frequently.

The **F1 score**, defined by Jeni et al. [28] as

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

combines both recall and precision values in its scoring, allowing it to indicate the overall accuracy of an algorithm, while also signalling when precision, recall or both are low. This metric enables a better understanding of inconsistencies in accuracy between classes than precision, recall or accuracy on their own. Hence, the F1 score is recommended by Jeni et al. for measuring performance and observing algorithmic bias towards certain classes when training with very imbalanced datasets.

2.4.2 Adversarial Learning

One method which allows for recognition of bias as early as possible is adversarial learning. The general idea behind adversarial learning, as explained by Zhang et al. [29], is to test whether a deep neural network (DNN) can reach a certain prediction with high probability, despite attempts of an adversary, experimenting with different inputs, to get it to produce a misclassification. Using an adversary allows for exploiting and uncovering biases that could potentially provoke a DNN that was trained to deliver fair predictions to instead predict according to stereotypes of protected characteristics. For instance, an adversarial attack against a network created to predict an individual's income bracket would attempt to create inputs that lead to results based on characteristics such as gender or zip code, according to Zhang et al.. In computer vision specifically, so-called generative

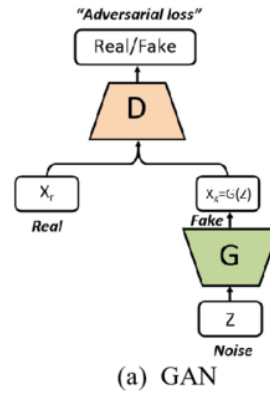


Figure 2.3: An illustration used by Sampath et al. in [30] to describe the basic architecture of vanilla generative adversarial networks. G: Generator, D: Detective

adversarial networks (GAN) apply the methodology of adversarial learning to detect and learn the underlying distributions of classes, according to Sampath et al. [30]. The basic architecture of a classic GAN consists of two neural networks: the generator, who fulfills the role of generating samples that are indistinguishable from the original samples, and the detector, whose job it is to tell synthetic samples from real-world samples. Over time, as both networks learn and improve at their respective task, the output of the generator, given high-quality samples as input, becomes capable of producing realistic samples, even for minority classes.

Hence, as pointed out by Sampath et al., GANs are commonly used in computer vision applications for creating synthetic samples for minority classes based on their recognition of the distribution of head and tail class data. One example of this usage can be found in the work of Song et al. [31], which relied on adversarial learning to recognize minority classes and generate diverse samples for the purpose of increasing the robustness of an object tracking-by-detection framework.

In the case of autonomous driving, adversarial training is also relevant for discovering and pre-emptively countering potential attacks against DNNs that would appear invisible or insignificant to humans, but can significantly affect the ability of autonomous vehicles to recognize traffic signs and other elements of traffic in its environment, as demonstrated by Cao et al. [32]. As systemic misclassifications caused by even simple adversarial attacks on autonomous driving models can have fatal consequences for passengers and the public at large, as demonstrated by Boloor et al. [33], adversarial training is a key tool in recognizing and in turn mitigating unwanted, risk-inducing biases. Zhang et al. also note that protected characteristics, which are historically vulnerable to discrimination (e.g. ethnicity, gender) could be incorporated during adversarial training to verify that certain attacks do not increase the loss functions of any demographic above a specified threshold, thereby reducing the extent to which negative pre-existing biases could be manifested.

2.4.3 Explainability

Another concept that becomes relevant in the field of bias recognition in algorithms is that of **Explainable AI**. With one of the main weaknesses of machine learning applications being the lack of insight of most of its users into its general workings, explainable artificial intelligence, as defined by Goebel et al. [34], proposes models that fundamentally implement tools capable of explaining the way an algorithm arrives at a solution to a human comprehensibly. Such models have proven to be an important tool in combating algorithmic bias in applications, as they not only improve acceptance among the population but also allow for better insight into an algorithm’s decision making process, in case of a suspicion of unwanted bias. As an example, Merrick and Taly [35] postulate a conceptual framework called “formulate, approximate, explain” (FAE), which unifies and generalizes the methods used in previous algorithms that attempt to explain a machine learning model’s prediction to humans. These methods rely on the Shapley-values approach, which originated in cooperative game theory and is used to distribute the payoff of a cooperative game fairly to its players, based on their individual contribution. In Machine Learning, Shapley values are used in explanation algorithms to attribute a score to each feature learned by a model based on the feature’s contribution to the model’s prediction. The FAE framework functions by (1) formulating features that are relevant and clearly contrast an event against one or more assumed norms, (2) approximating the Shapley values for randomly selected samples according to those features and (3) presenting a summary of the calculated Shapley values (e.g. sample mean). Merrick and Taly claim to have demonstrated through multiple case studies that their general game formulation allows for a straightforward quantification of how certain methods affect a machine learning model’s results, while also enabling the user to understand those scorings in comparison to reference inputs. For instance, after training machine learning models on real Adult Income datasets, the FAE framework was successfully applied to derive information about the impact of relationship, capital gain, education, marital status and age features on the models’ predictions. In the case of autonomous driving, explainable frameworks could help provide insight into the decision making process of autonomous vehicles, which in turn can be used to locate sources of and mitigate algorithmic biases, allowing for an increases in fairness and safety. One example of a computer vision model already showcasing the potential for explaining visual output to humans based on natural language, mentioned by Joshi et al. [36], is the Dall-E image generation model, which takes text captions as inputs and attempts to interpret the concept conveyed by them in order to generate corresponding images. In addition to that, explainable AI, according to Merrick and Taly, can also help increase an algorithm’s situational awareness by allowing for more direct and comprehensible feedback in unexpected situations. This could not only lower the risk of injury in the case of a sudden accident, but also increase the public’s trust in autonomous vehicles by allowing for traffic participants to have a deeper understanding of AV behavior.

2.5 Bias Reduction

Despite standardization still being in its early stages, a variety of methods and tools to reduce the presence of harmful bias in machine learning research have been presented.

2.5.1 Addressing Pre-existing Biases

Training Data Bias

One of the publicly available tools for quantifying and mitigating current biases in visual datasets is REVISE, provided by Wang et al. [37]. This tool assists in the reveal of potential biases along three axes: object-based, meaning biases regarding size, context or diversity of an object's representation, gender-based, which includes stereotypical representations of gender in datasets or geography-based biases, describing the lacking variety of geographical locations. Once the evaluation has taken place, REVISE can then offer specific recommendations for researchers on how to reduce detected biases.

Another approach to reduce the presence of pre-existing bias in machine learning algorithms altogether would incorporate establishing broadly accepted dataset standards and increasing the availability of data to the public, according to Norori et al. [38]. This could be achieved by relying on responsible data sharing frameworks that focus on openness while also protecting the privacy of individuals, particularly in cases of sensitive data. Through international cooperation of automotive research teams, AVs could ideally rely on a standardized, global data pool for sufficient data in the most relevant categories in order to improve fairness without sacrificing too much robustness. Norori et al. also note that, since public datasets as of now are mostly failing to provide diverse and interoperable datasets, clear and inclusive data standards must be set by open data sharing frameworks to mitigate training data bias in models.

Data synthesis could be another potential method for closing data gaps and reaching inclusive dataset design, argue Norori et al.. In this process, data is artificially generated to enhance datasets and accommodate for a lack of real data. The SYNTHIA dataset authored by Ros et al. [39], for instance, consists of more than 200.000 synthetic images taken from a virtual city under different seasonal, weather and lighting conditions and thus can be used to alleviate the scarcity of such tail-class samples in the real world. While this could find use in diversifying existing autonomous driving datasets, Norori et al. also mention that a highly sensitive, participant-centered approach would have to be facilitated to ensure a realistic and context aware usage.

Algorithmic Focus Bias

Efforts of inclusion in development teams are promising in the reduction of algorithmic focus bias. Regarding gender bias, research by Stathoulopoulos and Mateos-Garcia [40] indicates that differences in AI research papers can be found depending on female participation in authorship. Papers that are co-authored by women tend to have a higher semantic relation to social as well as political issues, suggesting, alongside other related

evidence, that women in AI research tend to be more engaged in societal research than men. Another study by Garcia-González et al. [41] assessing the perception of gender equality in Spain and the UK further suggests that understanding of gender bias in research differs between men and women, with surveys showing that male researchers generally perceive less gender inequality in their own departments than women in both countries, regardless of age or position. Leavy [42] notes that a majority of the people leading the field in bias mitigation in artificial intelligence are female, concluding that those affected by bias are more likely to recognize and attempt to resolve it. In the case of visual computing algorithms used in AVs, such gender biases, including bias towards the gender binary, can be reflected in the language used in the algorithm’s training data as well as the algorithm’s perception of gender in recognition. In order to effectively combat the presence of gender bias in machine learning, Levy therefore concludes, it is essential for women to be at the center of the discussion surrounding the definition of fairness in AI as it is a necessity for gaining sufficient insight for the development of gender-inclusive machine learning applications. Similarly, Livingston [43] suggests that agencies at a federal level must prevent the manifestation of racial bias in artificial intelligence systems by increasing racial diversity in AI designers. For context, Livingston notes that, in 2018, Black and Hispanic workers made up about 8.1% and 5.8% of the U.S. computer and math workforce respectively, while 2.8% and 3.3.% of technologists at Google and Microsoft respectively were Black. Increasing racial diversity in development teams could assist in preventing the perpetuation of racially biased driving behavior from the real world in autonomous driving systems since findings by Goddard et al. [44] suggest that driver behavior causes longer waiting times at crosswalks for Black pedestrians than for White pedestrians. Centering the needs and perspectives of users with physical or mental disabilities by co-creating accessible design approaches with disabled people can also be an important debiasing technique, according to Whittaker [45]. For this approach, Whittaker emphasizes the need for developers to re-think who is meant by the term “user”, since interactions with technologies can come out of compulsion or reluctance, while non-use can stem from rejection or exclusion of disabled individuals.

2.5.2 Addressing Technical Biases

One approach at lessening the impact of technical constraints on bias against tail- and open class scenarios that is especially relevant in the field of autonomous driving is the incorporation of redundancy in sensor systems for real-time image-based object recognition. Chetan et al. [46] thoroughly describe how the use and low-level integration of multiple viable sensor technologies alleviate the biases that each individual technology might exhibit under various conditions. Infrared cameras, for instance, are considered to be the primary technology needed for night vision on self-driving vehicles, but are reliant on other sensory inputs for both accuracy and robustness due to pre-existing visual assumptions about lanes. Light detection and ranging technology (LIDAR) enables real-time 3D-mapping of a vehicle’s environment by measuring the distance between object and sensor, works at night and detects pitches and slopes in roads, but is often paired together with radio detection and ranging (RADAR) on grounds of cost as well as

RADAR’s ability to obtain information on the speed of the vehicle using the doppler effect. Similarly, technologies such as stereo imaging and global positioning systems (GPS) can excel in mapping a vehicle’s environment and providing additional information under the right conditions, but struggle to perform in situations where either accurate feature identification or localization remains problematic. By integrating multiple technologies, however, autonomous vehicles could rely on the strengths of each sensor under varying circumstances, thereby improving overall accuracy and robustness for tail and open class scenarios. In order to do so, however, Chetan et al. note that an adequate implementation of said integration is required.

Regarding attempts to resolve the long-tail distribution problem altogether, Samuel and Chechik[15] explain that such state-of-the-art efforts in computer vision can be categorized into four main approaches: data-manipulation, loss-manipulation, two-stage fine-tuning and ensemble-models.

Data-Manipulation (Re-Sampling)

Data manipulation aims to accommodate for the long-tail distribution within a dataset and increase the likelihood of tail-classes encountered during training by over-sampling minority classes (i.e. copying samples), under-sampling head classes (i.e. removing samples) or generating augmented data based on examples of tail classes.

For oversampling Chawla et al.[47] present the synthetic minority over-sampling technique (SMOTE), which relies on repeatedly performing linear interpolation on all features of a real minority class data point (i.e. feature vector) and k of its closest neighbor(s). This is done using a random number between 0 and 1 as a multiplier to select a random value along a line segment between two features and repeating this for every feature in the feature vector until a new sample is created. By combining SMOTE with under-sampling, Chawla et al. were able to achieve higher accuracy for classifying tail classes than by solely under-sampling head classes. Random Oversampling (ROS) also attempts to generate realistic samples by randomly selecting existing samples from a minority class in the training data in order to duplicate them, as described in a paper by Mohammed et al.[48]. As more recent research by Shi and Zhang[49] explains, simple oversampling methods such as SMOTE and ROS are insensitive in their sampling towards information that is crucial to the classifier for recognizing different instances of tail classes and can therefore fail to generate any new data points that are relevant to the real world. This is called over-fitting, since, in this case, a model will perform exceptionally well on its own training data, while significantly underperforming on test or real-world datasets.

Li and Vasconcelos[50] demonstrate the use of **under-sampling** in their representation bias removal (REPAIR) method. By undersampling majority classes (alongside the oversampling of minority classes), Li and Vasconcelos explain that fairer training for visual recognition models are enabled as a result of representation bias in the dataset being addressed during re-sampling. Drummond and Holte[51] describe the under-sampling process as non-deterministic, since the removed samples are chosen randomly, resulting

in a random subsample of the original dataset as training data. As a consequence of this introduction of non-determinism into what is otherwise considered a deterministic learning process, additional variance in performance is created, according to Drummond and Holte. While undersampling is generally considered quite effective in reducing bias towards head classes - even more so than oversampling, as comparisons by Drummond and Holte between different naive re-sampling approaches show - the reduction in size of a training dataset also degrades a model's generalization ability, as mentioned by Samuel and Chechik.

An example of an **augmented data generation** approach is the adaptive synthetic sampling technique (ADASYN) [52], which can automatically decide and dynamically readjust the amount of synthetic samples necessary to compensate for the current skewed distribution in the data. The idea behind it is to generate more data samples based on instances of a tail class that are more difficult for the machine learning algorithm to learn compared to other minority class samples. Another method to achieve re-sampling is demonstrated by Sarkar and Czamecki [53] as they rely on a behavioral-driven rare event (RE) sampling approach in order to generate realistic traffic scenarios for testing. RE sampling allows for estimating the probabilities of rare events to occur under given system conditions as well as generating conditions that would lead to rare events. An adaptive sampling technique is then used to search for a sampling distribution that maximizes the potential for rare events like crash or near-miss scenarios. However, uncertainty and tail class events in traffic often stem from sub-optimal human behavior and driving styles due to limited capacities for real-time decision making, contrary to the behavioral designs of autonomous driving systems which are optimized according to pre-defined objectives (e.g. safety, progress towards a destination, abiding traffic rules, etc.). Therefore, Sarkar and Czamecki apply the theory of bounded rationality to model the inherent sub-optimality of human behavior in their model. This is done by implementing a so-called *quantal response function*, which outputs the probability of a discrete action being taken, given an environment state and a utility, and relies on a rationality parameter λ such that $\lambda \rightarrow \infty$ results in the event of the optimal action and $\lambda \rightarrow 0$ of a random action being taken. Extending the rationality parameter to multiple objectives, as is typically required for AVs, leads to a set of tuples otherwise known as rationality vector Λ . Sarkar and Czamecki conclude that categorizing driving scenarios according to behaviors that lie within the constraints of Λ before optimizing parameters to find the driving policies and Λ with the maximum potential of causing rare events, is an effective method for detecting and sampling rare events. While data augmentation methods avoid over-fitting tail classes and reducing generalization abilities, they also come with high development expenses.

Loss-Manipulation (Re-Weighting)

Loss manipulation approaches attempt to improve the learning of tail classes on already existing data by adjusting the scale at which tail classes are considered in the loss function (also known as the cost function). This can be done for example by scaling a model's

loss by the inverse of a class’s frequency, as shown by He and Garcia[54]. Among other methods, He and Garcia apply misclassification costs to their imbalanced dataset with less frequent classes receiving higher costs than more frequent classes. The resulting distribution function for the training dataset could then be iterated over by an adaptive boosting algorithm, which aggressively in- or decreases the cost of high-cost samples based on whether or not they are still being misclassified.

Philion[55] presents another approach for how loss can be dynamically rescaled according to difficulty in classification. The lane detection model demonstrated by Philion is capable of adapting to tail-class environments and displays high accuracy without any additional human annotation or training data by basing predictions on the following assumptions: 1) lanes are curve segment functions that use the height axis of a pixel as parameter and 2) lanes can be accurately drawn by iterating on the previous pixels that were determined to be part of a lane. During training, loss is minimized by dynamically weighing loss according to different sets of uncertainties and scales based on the objectives of binary segmentation and (pairwise) prediction. By considering the loss functions of each objective based on their level of uncertainty, accurate classifications can be learned for head- and tail classes simultaneously during training. This can outperform models that were trained separately for each task as demonstrated by Kendall[56].

Two-Stage Fine-Tuning

This type of bias reduction method, as described by Samuel and Chechik, divides the training process into representation learning, during which a deep neural network attempts to learn representations from the unmodified dataset and without re-sampling or reweighting, and classifier learning, which fine-tunes the last fully connected layers using re-sampling or unbiasing of the classifier layer. This approach is based on the idea that corrections in the classifier layer adequately remove underlying biases stemming from an imbalanced dataset. One example of two-stage fine-tuning stems from Ouyang et al.[57]. In order to improve performance in object detection for tail classes, a hierarchical learning scheme is applied in which the training dataset is first divided into groups and subgroups of classes that appear visually similar. For each group of object classes, a deep model is then finetuned with its parent group’s model acting as its initial point. During this process, each model is finetuned using only positive samples of its pre-assigned group of classes and negative samples that were previously accepted by the parent’s model. This way, the knowledge of larger groups is transferred into their subgroups. Furthermore, feature representation and classifier learning based on the feature representations occurs using different multi-class loss functions. Ouyang et al. observe consistent improvement when using clustering methods, which consider visual similarity between object classes or previous error rates on classes. Additionally, with more increasing levels inside the hierarchy, models consistently learn more specific and higher-quality feature representations, thereby increasing detection accuracy. Also, Ouyang et al. find that finetuning has benefited from gradual specialization. A different variation of the two-stage finetuning approach is exemplified in a paper by Gong et al.[58]. In this

instance, bias in a face recognition model was mitigated by automatically determining whether or not a certain network layer will be added, relying on a novel loss function to reduce bias between demographic groups within classes. While this type of approach comes at the expense of accuracy, results by Gong et al. demonstrate that performance can - at least in some cases - remain at competitive levels with state-of-the-art solutions.

Ensemble-Models

Samuel and Chechik describe ensemble-models as approaches that balance a model by grouping samples and specializing them on subsets of the dataset before assembling them again, creating a multi-expert-framework in the process. In a previously mentioned paper that applies this type of method in the field of UAV imagery, Yu et al. [24] propose the Dual Sampler and Head detection network (DSHNet), which mainly consists of Class-Biased Samplers (CBS) as well as Bilateral Box Heads (BBH). Using biased sampling, the two CBS sample with priority for head- and tail classes respectively. These sets of biased samples are then fed into the BBH, where predictions for the head and tail classes are generated on their corresponding biased sets. Loss rates over all classes are also computed during training. Afterwards, the predictions are fused to retrieve the final, best-performing results. This approach yields large precision improvements for both tail- and head-classes when compared to the base model as well as existing state-of-the-art solutions for long-tail visual recognition.

2.5.3 Addressing Emergent Biases

Transfer Context Bias

Adversarial attacks can be introduced to document and mitigate the existence of transfer context bias in autonomous driving applications. As was already mentioned, overgeneralized assumptions surrounding the context of an AV in practice can be uncovered using often simple adversarial manipulations of the environment. Boloor et al. list the placing of stickers on speed limit signs and lane markings as examples of low-effort, but highly successful physical attacks on autonomous driving systems. The importance of diverse samples in training as well as attack generation and defense mechanisms is also noted. Hence, in order to reduce this type of emergent bias in automotive algorithms, adversarial learning methods capable of producing realistic attacks (e.g. GANs) ought to be applied during and post development.

Interpretation Bias

Explainability in AI is an important resource for mitigating interpretation bias. As a machine learning model becomes capable of providing meaningful explanations about its decision process, it not only creates more transparency for pre-existing biases, but also provides an additional layer of interpretation for users to base their understanding of the output on. With humans generally struggling to understand deep learning systems that do not focus on justifications, as described in the paper by Joshi et al. mentioned prior,

explainability could therefore clear up misconceptions about the reasoning behind the algorithm's output as well as the output's meaning. This allows for less misinterpretations in human-computer interactions, since any statistical, moral or legal interpretations had by the user are constrained by the initial interpretation provided by the system, reducing interpretation bias in the process.

In summary, as an answer to the question of how bias emerges in machine learning algorithms, we find that algorithmic bias can be introduced throughout various stages of machine learning development, including the design of the algorithm, data collection, the training, evaluation and deployment of the model and its use within a social context post deployment. We also find algorithmic bias to express itself in a number of ways, including as preexisting, technical, and emerging biases, with imbalanced datasets and the long-tail distribution problem being two unresolved, frequently mentioned causes of bias in computer vision.

Current Real-Life Impacts of Algorithmic Bias in Computer Vision

Around 1.35 million road traffic fatalities occur globally each year according to estimates by the U.S. Centers for Disease Control and Prevention[59]. With autonomous driving and related areas of research in computer vision gradually approaching (super-)human levels of reliability, this chapter inspects the effects and challenges algorithmic biases pose in the current state of automotive research as well as the social repercussions recent deployments of computer vision systems have brought with them. More specifically, we discuss the impact of algorithmic bias in the domains of object recognition, facial recognition, gaze and pose estimation and scenario clustering.

3.1 Object Recognition

3.1.1 (Geo-)Diversity in Datasets

In the field of object recognition, discrimination caused by algorithmic bias takes place in the representation of geographical coverage and income groups, according to research by De Vries et al.[18]. As part of an analysis into how well modern object recognition systems function for user demographics across countries, cultural backgrounds and income levels, the aforementioned authors set out to analyze the fairness and accuracy of five publicly available state-of-the-art object recognition systems (i.e. Microsoft Azure, Clarifai, Google Cloud Vision, Amazon Rekognition and IBM Watson). Using a subset of the Dollar Street image dataset for training, the researchers found a significantly lower accuracy for non-western than for western households in all of the tested models. In addition to that, the systems also showed relatively poor performance on household items in

low-income households compared to high-income households. Despite the training dataset containing roughly the same amount of images for each income class, the average accuracy in the lowest-income bracket is about 10 pp. lower than for the highest income bracket. Furthermore, the average accuracy of all the tested systems appears to vary significantly across countries, with Amazon’s Rekognition system being about 15 pp. more accurate on household items in the U.S. than on household items in Somalia or Burkina Faso. Potential reasons for these discrepancies are the difference in appearance of common items between regions of the world as well as the context in which these items appear. For instance, the object recognition systems performed worse on toothbrushes and soap products in households without a bathroom. De Vries et al. isolate two different sources of bias in image datasets that partly account for the measured discrepancies in error rates: the geographical distribution of data and the reliance on the English language for data collection. It is also recognized that income-based disparities in accuracy can originate from a variety of other algorithmic biases, including representation and measurement bias.

This sentiment of lacking diversity in datasets as origin of bias is echoed by Model and Shamir[60], who come to the conclusion that the homogeneity within classes of standardized object recognition benchmarks allows for the correct classification of objects even by algorithms that merely recognize patterns, given a small sub-image of the original sample as input. In their experiments, datasets that were generated in controlled environments delivered the most biased results while datasets that were created through image collection on the web generally displayed weaker bias. Therefore, the performances measured by object recognition benchmark datasets are not applicable to real-world settings, further pushing away the idea that current object recognition models are capable of delivering fair and accurate results across varying environments. In order to allow for accurate performance evaluations under real-life circumstances, Model and Shamir suggest the use of as many human data collectors as possible to increase the diversity of images during natural image collection. A similar sentiment is reflected by Wang et al., who point out that equal or equitable geo-representation should exist in datasets and that dataset representation must not follow technology’s general trend of leaving digitally underrepresented groups behind. De Vries et al. note, however, that while methods such as geography-based resampling of image datasets and multi-lingual training would reduce dataset bias, they may be insufficient to combat the imbalance in currently available data in object recognition. According to them, fair object-recognition models would most certainly require training algorithms capable of accurately learning from a small number of samples without being susceptible to statistical variations in training data. Recently published work by Prabhu et al.[61], which attempts to address this exact problem using 4 standard domain adaption (DA) methods, also observes significant performance drops after geographical shifts with the tested DA methods, resulting in only limited improvements. The authors conclude that current off-the-shelf techniques are ineffective at allowing object recognition models to retain performance in unsampled geographical contexts, emphasizing the necessity for specialized geographical solutions. Thus, until either geographical domain adaption becomes feasible or fair geo-representation in visual

datasets become a reality, western influences in datasets will likely continue to negatively affect object recognition applications.

3.1.2 Pedestrian Detection

Events of recent years have shown the importance of high accuracy on tail classes for safety in traffic involving self-driving cars. One event that exemplifies this need is a fatal car crash incident in 2018 involving a self-driving car manufactured by Uber. Kohli and Chadja's case study [62] on the Uber car crash, provides more insight on why pedestrian recognition specifically failed in this case. In their analysis of the low-light scenario surrounding the crash, using crash video footage made available by Uber following the incident, both the raw and an enhanced, image-processed version of the footage was given to their state-of-the-art Object Detection model with the intent of finding out whether or not the incident could have been avoided. Regarding image enhancement, Kohli and Chadja tested a variety of methods for deriving information out of the original, low-quality video footage, including gamma correction, motion and edge detection. The experimental results surprisingly show that none of the image processing techniques had a positive impact on pedestrian recognition, with most methods in fact reducing accuracy. However, the YOLO Object Detection framework used during testing was capable of detecting the pedestrian in the raw footage 0.86 seconds prior to the crash. Hence, Kohli and Chadja conclude that, whether or not the accident could have been prevented in this time depends of the decision making time of Ubers autonomous car as well as mechanical factors such as its emergency braking system performance. These findings however reiterate the need for fast reaction times in autonomous driving models for tail class scenarios, particularly in emergency traffic situations.

A different study by Zhang et al. [63] attempts to draw a comparison between current standards in pedestrian detection and human performance. While noting that recent progress in pedestrian detection has shown no indication of slowing down, when testing on current benchmarks, state-of-the-art detectors - including Checkerboards and RotatedFilters - were all significantly outperformed in accuracy by a human baseline set using the judgements of two domain experts. After inspecting potential sources of error for Checkerboards, the top-performer out of all detectors, Zhang et al. point towards Background Errors, which they define as the detection of pedestrians in the background without any overlap with a pedestrian as declared by ground truth annotations, as the most common type of error found in all the tested recognition systems. Out of the analysed background errors, false positives on tree leaves and traffic lights were most common, which leads the authors to suggest that detectors are lacking vertical context for a better understanding of large structures and height estimates. Localization errors, which are false positives that overlap with a bounding box given by ground truth annotations, comprised along with background errors the totality of false positives. When incorporating CNNs, a sizeable reduction in background errors was accomplished. This however came at the cost of a slight increase in localization errors surrounding the areas of an object. This is ascribed to an intrinsic limitation of convnets and the current architectures used along with them regarding the accurate detection and localization of

small objects, since performance gains eventually stall with higher proposal quality from convnets. Thus, Zhang et al. conclude that further research will be required to fully address the current technical limitations of convnets regarding localization in order to reach human performance in pedestrian detection.

3.1.3 Lane Detection

Accuracy and fairness in lane detection is crucial for the development of safe and robust autonomous driving systems. In the study by Phillion[55], previously mentioned in Section 2.4, we learn of the lack of data on weather and environmental corner cases in public annotated image datasets for lane detection as well as how current approaches of detecting lanes by fitting them into polynomials fail to accurately capture lanes and avoid localization errors in tight curves. Chetan et al. discuss this issue in their overview of recent advancements made in lane detection. In order to increase robustness and avoid bias towards certain types of traffic properties and constraints, Chetan et al. suggest most importantly relying on multiple perception sensors and sensor level integration, in which each sensor’s properties are effectively used to increase the performance of the sensor system. The reasoning behind this is that a lane detection system that is capable of integrating a combination of the most commonly used sensor technologies will obtain an advantage in performance, especially on tail classes.

By examining contemporary research, Chetan et al. identify combinations of multiple cameras, LIDAR and GPS as the best-performing examples of sensor level integration, since mapping, object detection and positional information all help increase lane detection support. Despite that, Chetan et al. identify the use of multiple cameras, RADAR and ultrasonic sensors as the most cost-effective and commercially viable option. In addition to that, algorithm level integration, meaning the integration of multiple lane detection algorithms, as well as system level integration, the combination of multiple detection systems (i.e. road detection and lane detection), are also approaches that are currently researched and, according to claims by Chetan et al., deliver promising results so far. Research by Jaipuria et al. [64] also argues that, while the goal of carefully obtaining a balanced, task-agnostic dataset is likely in vain due to the insurmountable quantity of data that would be required, breaking down dataset biases and reducing them on a task-by-task basis can effectively improve performance. The approach presented by Jaipuria et al. attempts to augment realistic data to reduce bias only in task-specific environmental factors, which is referred to as noise factor distribution bias, instead of diversifying the entire dataset. The results achieved in their experiments show that models that were partly trained using task-specifically augmented data can significantly outperform models that were trained purely on real data, therefore indicating a reduction in bias.

3.2 Facial Recognition

The presence of algorithmic bias in facial recognition has been thoroughly documented in recent years. Leslie[65], for example, comments on the history of algorithmic bias in facial recognition technologies by explaining that patterns of discrimination and recklessness have been present since the very beginnings of computer vision. As source of this bias, Leslie refers to centuries-old patterns of racism in photography, which are reproduced by facial recognition technologies at an increasing rate. Leslie also points out that the attitude shown in research and development in the field of facial recognition has often been one of disregard towards the issue of bias, which in turn could stem from research teams mainly consisting of social demographics that do not face the worst consequences of apathy bias. As an early example for implicit racial bias, Leslie names the Coolpix S360 smart camera system released by Nikon in 2009, which was notably reported by Asian American users to incorrectly register their faces as blinking. It can be assumed that an algorithmic processing bias was responsible for this outcome, since the algorithm's design at the time was presumably not considering Asian eye types adequately. Another example from the same year stems from a bug in Hewlett-Packard webcams, which caused them to be unable to follow the movements of darker-skinned employee's faces, while those of lighter-skinned employees were tracked smoothly under the same, normal conditions. These showings of omissions, argues Leslie, are the reason for a subconscious assumption of light skin-tones as default being prevalent in modern computer vision models, similar to how it has historically been in photography. Nowadays though, the transition from rule-based to data-centric approaches in machine learning over the previous decade, Leslie explains, has made it easier for computer vision models to be introduced to bias in the form of training data. With this in mind, it appears as if current facial recognition technologies are not necessarily working towards dismantling systemic discrimination and instead manifesting preexisting biases through a variety of sources, which becomes clearer when examining them from an intersectional viewpoint.

3.2.1 Intersectionality in Datasets

In 2018, Buolamwini and Gebru[66] demonstrated the extent to which racial and gender biases coincide in their negative effects on state-of-the-art performances at the time. During their testing of Microsoft's Face Recognition Software FaceDetect, an overall error rate of 6.3% was measured. When distinguishing between women and men, a significant difference in accuracy was found with an error rate of 10.7% for female and 2.6% for male subjects. Analyzing performance according to skin-tone found a margin in error rates of 12.2 percentage points between darker skinned (12.9%) and lighter-skinned (0.7%) individuals. Once the intersection of gender and race was factored in, however, even more remarkable discrepancies in error margins were exposed: dark-skinned women had an error rate of about 20.8%, compared to 6.0% for dark-skinned men, 1.7% for light-skinned women and 0.0% for light-skinned men. With the highest error margins appearing to lie in the intersections of protected groups such as gender and race, the need for urgent focus on intersectional demographics in Facial Recognition becomes apparent. More

recent research by Majumdar et al. [67] demonstrates the presence of gender and racial bias even in face recognition models that appear to be unbiased based on training data scenarios, by introducing real-world image distortion. Once realistic image distortion operations (e.g. Gaussian Blur, Brightness, Resolution, etc.) were individually applied onto the datasets, Majumdar et al. observed an unequal drop in performance across gender and racial sub-groups. This difference in performance degradation between race and gender sub-groups is caused, according to the authors, by a shift in the models' focus away from the most discriminative facial regions (i.e. eyes, nose and face mask region) towards regions that are less discriminative for certain sub-groups. Furthermore, while each model introduces varying degrees of bias, the biases mostly appear to consistently discriminate against white and female demographics across all image distortion operations. Intersectional analysis on the effects of Gaussian blur confirms this assessment, revealing huge gaps in degradation of performance between the different intersections, with the female-and-white intersection receiving the largest relative performance loss of 30.74 pp. compared to 22.93 pp. for the female-and-black, 14.7 pp. for the male-and-white and 9.57 pp. for the male-and-black sub-set on the LCNN-29 recognition model ($\sigma = 4.0$). Hence, the authors conclude that the understanding of bias in face recognition models as consistent and intersectional is important going forward in the development of deep learning models that are unbiased under real-world conditions.

3.2.2 Double-Barrelled Discrimination

Double-barrelled discrimination, as explained by Leslie, occurs in practice when a demographic affected by algorithmic bias is denied both access to the technology's benefits as well as representation, since a process that is made convenient for favoured demographics is at the same time more complicated for marginalized groups. One recent major example of this form of discrimination mentioned by Leslie stems from Amazon's use of a biased AI tool for estimating competence in job recruitment. Amazon decided to halt the usage of its own job recruiting tool after findings [68] indicated that it was biased against women, penalizing graduates of all-women's colleges as well as resumes that contained the word "women's". While details about Amazon's abandoned job recruitment algorithm remain scarce, the hiring platform HireVue, whose software analyzes speech and facial expressions of candidates during automated video interviews, similarly decided to halt [69] the use of facial expression analysis after complaints of failure to provide transparency and guarantees of fairness in the process. In these instances, discrimination takes place both in the lowered representation of affected demographics (i.e. women within Amazon) and in the systemic penalization of those affected in the job market.

In the context of police surveillance, double-barrelled discrimination can have large effects on the public safety of minorities. According to Najibi [70], computer vision tools as of now have played an assisting role in the overpolicing of minority communities (i.e. stop-and-frisk). Garvie [71] arrives at a similar notion, finding that, due to disparities in face recognition accuracy between races as well as the disproportionately high number of African Americans in U.S. mug shot databases, racial bias will disproportionately affect

African Americans. As an example of such discrimination, Najibi mentions Project Green Light, a surveillance program started by the Detroit Police Department in partnership with local businesses, the City of Detroit and community groups. By comparing the areas in which high-definition cameras were most commonly installed to district census data from 2010, Najibi found a significant overlap between predominantly Black and Hispanic districts and surveillance spots. As a reaction to this use of surveillance technology, the U.S. district of New Jersey introduced the No Biometric Barrier to Housing Act [72] in 2019, which prevents facial recognition technologies to be used in public housing units and lists the potential impact on vulnerable communities as one of its concerns. In a similar fashion, after global protests following a series of police killings in the U.S. [73], including those of Breonna Taylor, Ahmaud Arbery and George Floyd, Microsoft and Amazon publicly announced in 2020 that they would halt development on their facial analysis tools along with sales to police departments in the United States to prevent human rights violations and allow for the introduction of stronger legal regulation [74]. On a similar note, Facebook has decided in 2021 to discontinue its facial recognition system, which contained scans of facial features from more than one billion users, citing growing concerns about the place of facial recognition technology in society [75]. Aside from discontinuing such use, Najibi recommends reducing dataset bias, optimizing the default camera settings for training images to better account for darker skin tones and educating producers of facial recognition models on racial literacy. Another more recent example of facial recognition algorithms being mis-used to over-police minorities comes from the Greek government’s new Biometrics Policing Program. Human Rights Watch (HRW) and Greek Digital rights organization Homo Digitalis report [76] that the EU-funded program plans to use face scan technology on immigrants to gather biometric information and cross-validate it against police, immigration and private sector databases. This, according to HRW, would allow Greek police forces to continue and increase racial profiling through discriminatory stop-and-frisk searches. HRW therefore suggests that the European Commission should not fund Greece’s policing program as it violates human rights standards and presents a risk to nondiscrimination protections for asylum seekers and minority groups. This, along with the aforementioned examples of double-barrelled discrimination, showcases the drastic effects of deploying computer vision algorithms in public sectors without proper legislation set in place to protect vulnerable people. In the deployment of automotive applications, these concerns over bias in facial recognition could apply in real-time driver monitoring that rely on monitoring the driver’s face specifically. For instance, the smart driver monitoring system proposed by Shaily et al. [77] incorporates analysis of facial landmarks for the detection of drowsiness, making it susceptible to bias in facial recognition datasets and algorithms.

3.3 Gaze and Pose Estimation

Toyoda, Lucas and Gratch [78] illustrate the harm gaze estimation models can have when applied without pre-emptively accounting for discriminatory bias. The aim of the referred paper is to illustrate whether or not it is possible to accurately estimate worker

efficiency by analysing gaze and facial movements on subjects. As reasoning behind this research, Toyoda, Lucas and Gratch mention the significant interest garnered for automated methods of managing workers as well as recent progress made in approaches for predicting student engagement and driver fatigue through facial expressions and eye tracking. Compared to current methods, the researchers argue that their method of estimating work performance can provide objectivity to productivity evaluation and wage promotion distribution.

First, the potential of predicting work accuracy from moment-to-moment was examined by providing 63 participants with a real-world task that is widely used to study worker engagement and accuracy. The equipment used for this experiment consists of OpenFace, used for acquiring head pose and gaze information, OpenPose for estimating body posture, the Broad Bioimage Benchmark Collection medical image dataset and a standard webcam (Logicoool C920n). The results show that, similar to filtering out workers' solutions using a self-reported confidence level in results, filtering solutions by nonverbal behavior displayed at the time the task was solved can also substantially improve the estimate. This was achieved by discarding estimates from images in which a worker was predicted to be less engaged or fatigued and therefore more likely to be inaccurate in their work. Despite further reducing the number of used estimates, the authors note that this filtering method better approximated the results derived from experts and that the model reliably predicts the accuracy of workers automatically through nonverbal behavior during repetitive tasks.

However, Toyoda, Lucas and Gratch also acknowledge that, upon investigating differences in accuracy between skin tones, their model displayed systemic racial bias that disadvantaged dark-skinned workers, while also wrongfully awarding workers with lighter skin. Examining two groups of workers based on skin-tone unveiled an average task accuracy of 86% for the light-skinned and 76% for the dark-skinned group. Once the false negative and false positive rates were analysed, a significantly higher false negative rate for those with lighter skin was revealed, while the false positive rate for those with darker skin was slightly worse. After analysing the frames in which the tracker misaligned a worker's face by skin tone, the researchers found a positive correlation between darker skin tones and tracking errors. Without identifying the source of bias, the paper refers in its conclusion to further research and data being needed to identify a source, adding that debiasing or at least auditing is required before deployment in the real world. This conclusion arguably exemplifies the need for urgency and prioritization of fairness in computer vision algorithms. Instead of relying on future debiasing solutions to justify the current progress made on biased computer vision models, the results of Toyoda, Lucas and Gratch could be interpreted as an indicator for the dangers of biased automated management of human workers as, once deployed without large-scale systemic changes and significant progress on debiasing, such an application could lead to double-barrelled discrimination against minorities.

In an automotive context, racial bias in gaze and pose estimation could, much like with facial recognition tools, influence the accuracy of driver monitoring and pedestrian detection models, thereby causing disparities in safety improvements for different demo-

graphics. For example, while the model presented by Shirpour et al. [79] for approximating driver gaze using head pose estimation was capable of accurately detecting the gaze area of drivers with 95% confidence 82.5% of the time when tested on 15 drivers, analysis of potential performance differences between racial groups is missing. Similarly, the system proposed by Murphy-Chutorian and Trivedi [80] for monitoring driver awareness reports overall performance improvements over state-of-the-art competitors when tested on 14 drivers of different ages, race and sex, but is not supported by an age-, sex- or race-specific error analysis, leaving questions regarding bias related to these categories open. Furthermore, research by Ngxande [81] evaluating the accuracy of three models (ResNet, VGG-19 and VGG-Face) used for detecting driver drowsiness against publicly available datasets finds a general presence of bias against a wide range of races and cultural contexts representative of South African society, originating from training data bias. Ngxande concludes that while standard bias correction techniques using GANs manage to improve fairness, no fairness guarantees can be given as biases may still be present in the models. Regarding pedestrian detection systems, Aledhari et al. note that state-of-the-art pedestrian detection systems for AVs struggle to detect people with darker skin tones due to algorithmic bias, proposing an architecture combining a CNN with clustering for skin detection with promising experimental results on pedestrians, both across skin tones and clothing colors. It is worth mentioning that in the case of Aledhari et al., clothing colors were found to be a more determining factor for pedestrian recognition than skin color. In their conclusion, it is pointed out that contemporary research on the topic of pedestrian detection frequently does not account for diversity in their datasets, despite attempting to tackle the issue of bias, which potentially worsens training data bias.

3.4 Scenario Clustering

The creation and simulation of driving scenarios is of utmost relevance during testing and benchmarking of autonomous driving systems. Hauer et al. [82] explain that during scenario-based testing of ADS, one source of algorithmic bias can be found in the generation of test scenario types. These scenario types are required in order to generate new scenario instances, which contain variety in their properties, for stress tests such as near crashes or abrupt changes in speed. Establishing a complete list of possible scenario types therefore allows for an increase in confidence in the testing procedure and as a result also in the driving system. The process of deriving such scenario types, however, usually occurs through researchers manually listing categories that are considered relevant, which allows for the introduction of human biases, since certain types of driving scenarios might be overlooked or the used mental model might prove to be insufficient for practical use. Though the alternative of deriving scenario types automatically through collected real-world data, based on a given mental model, almost eradicates this manual derivation bias, it is still affected by the open-endedness of the real world and can consequently also leave out relevant types. Hence, the proposal presented by Hauer et al. aims to complement and extend a pre-specified mental model by filling in

all missing scenario types on varying levels of granularity without any additional human input on the structure. Hauer et al. note that this automated approach for clustering scenarios into types is enabled by the growing amount of publicly available real-world driving data, with the results most likely resembling highly accurate groups that are not representative of any to humans semantically meaningful scenario types. To address this, an automated cluster interpretation model is afterwards used to assign semantic meaning to the clusters. While their approach also involves human bias, Hauer et al. argue that their methodology minimizes human input while assisting in the completion of manual test scenario generation by providing redundancy for industry experts. The authors, however, point towards further research to answer the question of which level of granularity is required for optimal test scenario generation. Zhao et al. [83] expand on the approach presented by Hauer et al., sharing the goal of removing handcrafted feature extraction from the clustering process, but introducing a deep neural network to train on the unlabelled data, which is also referred to as unsupervised deep learning. After testing their clustering implementation on an augmented dataset and comparing performance with the benchmark provided by Hauer et al., the researchers report a True Positive Rate of 99.88% and a False Positive Rate of 0.10%, compared to the respective rates of 77.27% and 3.0% of the benchmark method. Furthermore, the technical limitations of the benchmark methods with regard to its maximum dataset size of 100 were overcome by Zhao et al. as their newly proposed method is scalable on much larger datasets. In order to further reduce the potential for bias during the clustering of driving scenarios, the authors recommend for further research to include more information provided by the scenarios, like movement information on bikes, pedestrians, and crosswalks, in the automated clustering process.

To answer the research question regarding the practical consequences of unintended bias in current automotive research: algorithmic bias appears to be largely contributing negatively towards the accuracy of models in practice, being not only noticeable, but of significant concern to automotive research. Considering the recent social impacts of harmfully biased CV algorithms, including the discrimination of marginalized demographics through facial and object recognition systems, unexpected impacts on public safety stemming from the deployment of biased automotive systems are plausible.

Ethics of Algorithmic Bias in Automotive Applications

This chapter discusses ethical areas of concerns for bias in autonomous driving at large as well as corresponding methodologies that can be applied to reduce the negative impact of algorithmic biases. We categorize these concerns into three main categories: 1) avoiding bias in the creation of morality and ethics guidelines 2) preventing unethical use and deployment and 3) increasing safety in tail and open class scenarios.

4.1 Moral Autonomous Decision Making

By containing negative moral biases in their decision systems, autonomous vehicles could potentially jeopardize their passengers and other traffic participants on the road. One way for this to occur would be through consumer preferences incentivizing an unfair distribution of risk between road users. While conducting six online surveys with a total of $n = 1928$ participants, Bonnefon et al. [84] found out that while consumers would appreciate the presence of self-sacrificing AVs in traffic, they also prefer the prioritization of occupants' safety when considering the purchase of a self-driving car. This gives AV producers an incentive for breaking fairness in risk distribution for the sake of maximizing profits. Thus, a conflict of interests for AV manufacturers could occur, when presented with the opportunity of increasing profit margins in their target group at the cost of the safety of vulnerable, lower-income demographics. In addition to that, research by Lim and Taeihagh [85] postulates how AV manufacturers might choose to minimize or keep liability claims constant by changing an autonomous vehicle's driving style depending on the average income of the district it is driving in. As a consequence, AVs would shift safety risks from areas with a higher average income to areas with a lower average income as part of another form of income-based discrimination through morality in autonomous driving systems.

In order to avoid potential moral bias and match societies' expectations of ethics in machine learning algorithms, Awad et al. [86] developed a method for quantifying moral decisions. The so-called "moral machine" is a now prominent online platform designed to collect global data and collective cultural preferences in moral dilemmas. In this "serious game", participants are presented with a fatal AV scenario and a choice between two different outcomes. These scenarios were created mainly by exploring the following nine factors in the decision on what life should be spared: humans vs pets, staying on course vs swerving, passengers vs pedestrians, more lives vs fewer lives, men vs women, young vs elderly, pedestrians who cross legally vs jaywalkers, fit vs less fit and people with higher social status vs people with lower social status. After gathering 39.61 million decisions in ten different languages and 233 countries and territories, Awad et al. found global preferences to be the strongest for saving human life over that of animals, saving more lives over few and sparing young life over elderly. Due to this, Awad et al. argue that these preferences ought to be fundamental for creating autonomous driving and machine learning ethics in general, suggesting that even if legal requirements do not clearly mandate these settings, the public's view will likely be strongly in favour of them.

Furthermore, Awad et al. detected no significant differences between demographics according to the attributes of age, education, gender, income, politics and religion, with the most notable correlations being that male respondents are marginally less likely (0.06 pp.) to spare women than female respondents and that religious people are slightly more inclined (0.09 pp.) to spare humans over animals. Culturally speaking, the findings were grouped into three main clusters: Western (Christian cultural groups in North America and Europe), Eastern (Confucianist and Islamic cultural groups in Asia, Oceania and the Middle East) and Southern (Latin America and French-influenced countries). While the results indicate regional overlap in ethics preferences, analysing the differences between clusters reveals large variance, with the only remaining commonality across all examined cultures being the weak tendency to spare pedestrians over passengers and lawful over unlawful traffic participants. Awad et al. see this as potentially problematic for the deployment of AVs, noting that manufacturers and policymakers should at least be aware of the vastly differing views on AV ethics in the countries they are providing them in as consumer behavior and the public's tolerance towards AVs will likely be affected by perceived moral bias. Finally, the data gathered through the moral machine displayed high correlation between a country's economic and cultural status and its ethics preferences. One important finding in this aspect is the split between individualistic cultures, who show stronger preference towards sparing the lives of as many people as possible, and collectivistic cultures, which display weaker preference towards sparing younger lives. Furthermore, a correlation between a country's prosperity and stronger preference against jaywalkers was found. Finally, countries with higher inequality between rich and poor classes also tend to treat rich and poor people differently. This carries over to differential treatment of women and men and its relation to the size of a country's gender gap in health and survival. In conclusion, while the results of the moral machine present significant differences in preferences regarding moral dilemmas between different societies and cultures, they also note that broad agreements across regions of the globe

would suggest that challenges in establishing global ethics guidelines for autonomous driving systems are surmountable.

The adoption of ethics in autonomous systems in practice, however, poses another challenge, as is addressed by Vakkuria et al. [87]. In their case study on five different companies and their methods on how to implement AI ethics into their solutions, the researchers found that in all of the companies', developers lacked a systemic way of integrating ethics into practice, with developers generally considering the issue of ethics as important but far removed from their level of work. This is in part due to a lack of agreement on what the term ethics means in the context of autonomous driving systems in the industry, as well as strict outside regulations on ethical issues, which leave little room for input from developers. Bonnefon et al. [88], critical of the status quo, emphasize the importance of developers and engineers having a voice in the ethical discussion of autonomous machine learning systems and explain the practical, statistical trolley dilemma engineers face when designing a moral system. While the trolley dilemma appears to many as irrelevant on a technical level, reframing it in a statistical format makes understanding the impact of moral bias in AVs on the amount of yearly passenger and pedestrian fatalities easier, thus revealing the true importance of ethics in autonomous driving. Additionally, Bonnefon note the potential impact of AV consumer preferences on the reality of risk distribution in traffic without adequate standardization across the industry. If, for instance, two self-driving car models, which are equally effective in reducing the amount of fatalities while driving, released on the market, but one of them distributes risk equally between passengers and pedestrians while the other heavily favours the safety of passengers, the average consumer might prefer the second product, which would gradually and disproportionately increase the risk for all pedestrians. Thus, there are realistic ways for decisions made on a technical level to statistically influence public safety on the road. Hence, bridging the gap between theoretical research and industry practice is crucial, according to Vakkuria et al., in solving problems caused by prominent algorithmic biases. Consequently, the authors suggest both developing tools and practical methodologies for industry experts to help implement AI ethics as well as furthering the discourse on AI ethics in order reach what can be considered a consensus, which can then be captured and utilized it in a consistent manner.

4.2 Ethical Deployment and Consumer Trust

One of the concerns regarding the deployment of autonomous driving systems lies in the emergence of biases and inaccuracies post-release leading to a decrease in road traffic safety. Such an occurrence would likely not only have potentially devastating social consequences on its own, but also shatter long-term consumer trust in self-driving cars. Given the risk at hand and the importance of public safety, Borenstein et al. [89] argue that before any deployment of an ADS, each company should prove, in the form of passing carefully constructed trials, that their system will not reduce road safety upon release. The researchers also note that the neglect of responsibilities by designers of self-driving cars can be furthered by external pressure from manufacturers and other third parties.

This pressure could presumably manifest itself in the form of unreasonably tight release schedules. Despite these pressures and the argument of collective responsibility reducing accountability of individual designers, Borenstein et al. seek to hold designers individually accountable for danger caused by sub-optimal or dangerous ADS performances, since everyone partaking in the creation of an AV carries ethical responsibility for their own professional decisions. Among other suggestions, the researchers argue for real accountability for manufacturers and engineers of AVs instead of mainly for management and regulators. Furthermore, consumers are not to be misled about the risk involved with an AV in its presentation and instead provided with high levels of transparency.

Data privacy is another issue that requires addressing prior to release in order to sufficiently increase consumer trust in AVs. A survey recently conducted by Panagiotopoulos and Dimitrakopoulos[90] found that out of the $n = 483$ participants questioned, the vast majority of them is concerned about data privacy (95%) and cyber security (93.6%) of today's internet-enabled technologies. Furthermore, 47% of participants stated a neutral stance on system security and data privacy concerns of AVs. This issue is particularly prevalent in the concept of hyperconnected vehicles, meaning AVs which are constantly connected to the internet and able to communicate not only with infrastructure (V2I) and vehicles (V2V), but every capable entity (V2X), including pedestrians, devices and grids, according to Karnouskos and Kerschbaum[91]. In their study on data privacy and integrity concerns regarding hyperconnected vehicles, Karnouskos and Kerschbaum recommend approaches that implement either encryption or perturbation methods on data and meta-data communicated between entities, since once a vehicle's data is retrieved, information about its passengers and its environment could be inferred, even if the data is anonymized in advance. Hence, perturbing or encrypting data before sending it to other devices in the network could contribute to a "privacy by design" approach while also offering the benefits of hyperconnected autonomous vehicles. However, current encryption methods lack the performance desired by hyperconnected vehicles, whereas perturbation methods lower computation accuracy. Thus, according to the authors, development on new compilers that can efficiently process optimized cryptographic protocols without translating them into a less secure, leakable format is required. Lai et al.[92] propose relying on blockchain and multi-party private set interaction (PSI) protocols to process big data in V2X services while preserving privacy. By meaningfully incorporating privacy, security, transparency and accountability in the design and deployment of AVs, consumer trust could increase, since both the failure to deliver on promises and the misuse of sensitive data become more apparent to the public.

4.3 Safety in Tail- and Open-Class Scenarios

Due to the current state-of-the-art in computer vision datasets, significant progress has to be made in debiasing AV solutions for avoiding potentially worse driving quality in regions that are currently underrepresented in traffic datasets. Varshney and Alemzadeh[93] explain that an overall uncertainty is present in models that were trained on datasets, which do not represent real-world distributions, with regards to their true operational risk.

In order to combat this, Varshney and Alemzadeh suggest two principals for inherently safe design in machine learning models: 1) explainability in decision ruling and 2) a careful selection of only variables that are causally related to the outcome. First, by examining the output of explainable models, parts of an application that are responsible for exploiting weaknesses in a dataset and inflating a model's accuracy during testing, can be documented and consequently removed. Varshney and Alemzadeh emphasize that simply interpreting the output of a model that was not designed with interpretability in mind does not yield the same results as interpreting the output of an explainable model, because of the complexity and opaqueness in the decision making of "uninterpretable" models. Secondly, by ensuring that only causally related variables are selected from the available data, causality and underlying system laws can be more accurately captured. Since causality can not be guaranteed using standard learning constraints indicative of data-driven computer vision algorithms, theory-driven regularization for lowering the risk of unwanted bias is presumably required, likely at the cost of the algorithm's overall performance. Aside from this, the authors also mention a method capable of constraining the relative risk of protected groups compared to unprotected groups to a maximum value (in this case $\frac{5}{4}$), which ensures a lower limit for relative fairness between classes. These approaches could be relevant in the field of autonomous driving during high-quality representation learning for tail class samples of roads, traffic signs, crosswalks or cycling infrastructure.

Explainability in autonomous driving, however, remains a challenge due to the absence of substantial market regulations and research guidelines on AV explainability, the lack of interdisciplinary research on explainability in AVs as well as the high presence of bias in AV datasets affecting contemporary solutions, as mentioned by Omeiza et al. [94]. Not only are datasets, according to Omeiza et al., restricted by finiteness and deviations from real-world distributions, but also in their open-endedness towards relevant environmental variables. Hence, in order to mitigate Explanation Bias, Omeiza et al. suggest inclusivity in research on all levels and the creation of new regulations and guidelines. Furthermore, Varshney and Alemzadeh note the use of safe-fail triggers in self-driving cars for recognizing whenever no confident prediction can be made by the ADS during an event that is either rare or unprecedented to the system. In such a scenario, control can then be handed over to a different system or a human driver as backup. As the first example of such an approach, Hecker et al. present a method for detecting so-called Scene Drivability in autonomous driving scenarios. The scene drivability score is supposed to capture the likelihood of an autonomous vehicle to fail for a given scene and is derived by comparing a model's prediction against recorded human maneuvers, which are considered as ground truth for this purpose. The ADS is then able to detect probably bad model predictions and alert the human driver in time (i.e. 2 seconds in advance) to take over, thus increasing the safety and collaboration between human and AV in critical scenarios. The conclusion Hecker et al. arrive at is that it is possible to accurately predict drivability scores for particular driving scenes and use them to enhance overall safety, suggesting that their future research will aim to incorporate more modern and sophisticated driving models. Hence, methods for uncertainty and risk

estimation for tail and open class events appear to be crucial for the safe and ethical use of AVs in practice in the future.

The use of real-world driving data in autonomous driving datasets might also pose a safety risk. Research by Goddard et al. [44] indicates the existence of racial bias in drivers in the U.S.. After conducting 88 pedestrian trials, using 3 White and 3 Black male research participants as pedestrians, on 173 drivers, the results show that Black pedestrians were passed by twice as many cars and were subject to 32% longer wait times than White pedestrians. In a follow up study [95] in 2017, Goddard et al. found that drivers were more likely to stop closer to Black pedestrians, regardless of the drivers' race and gender. Hence, one could argue that the possibility of real-world datasets as a source of racial and legal bias in autonomous driving systems is to be noted and taken into consideration when preparing training data.

To sum up the answer to our third and final research question, some types of bias in automotive machine learning applications can be addressed in pragmatic ways. Moral bias in an algorithm's design, for instance, can be reduced in practice by adapting moral systems based on regional socio-cultural input. The effects of some technical biases such as algorithmic processing bias can be addressed by introducing fail-safes and adequately assessing the trustworthiness of a system's classification in real-time. Finally, emerging biases could be mitigated through more checks and industry-wide regulation on AV development prior to mass-deployment.

Conclusion

This thesis provided a general overview of algorithmic bias, its effects on recent research in computer vision and its impacts on automotive applications in practice. Algorithmic bias can emerge in a variety of ways both in the design and in the technical implementation of machine learning applications. Pre-existing biases are nowadays frequently introduced to computer vision models in the form of biased datasets, primarily caused by the long-tailed, open-ended distribution of data in the real world, while technical constraints can cause an algorithm to exhibit entirely new technical biases. Autonomous decision systems also fundamentally require morality as they are forced to make decisions guided by ethical or legal input, making them susceptible to emerging biases. Research suggests that algorithmic bias can be responsible for unfair disparities in a model's accuracy for different classes and reduce performance when testing under realistic circumstances. In the context of autonomous driving, biases can provoke dangerous behavior from self-driving vehicles in atypical or rare scenarios. Since the security of autonomous driving systems has crucial impacts on public safety, strict regulations on ethics and performance prior to market release is recommended.

Research across computer vision disciplines frequently suggests systemic debiasing measures such as increasing diversity in engineering teams, improving and standardizing datasets and higher levels of specialization for solutions. In addition, industry wide ethics guidelines as well as higher accountability for manufacturers could alleviate the impact of moral biases in practice. This will require transparency and open ethical discussions with the public. By properly eradicating harmful biases in practice, automotive machine learning algorithms can enhance road safety while acting as a driver of social equity. Without adequate measures, however, a rushed deployment of autonomous driving systems caused by conflict of interest could increase road risk particularly for vulnerable social demographics and their intersections. Hence, further research will have to focus on improving methods for imbalanced learning (i.e. explainability) and integrating ethical principles into industry practice.

List of Figures

2.1	A merge of the taxonomies for algorithmic biases provided by Friedman and Nissenbaum and Danks and London.	8
2.2	A graph showcasing at which stages of development certain types of algorithmic biases are introduced.	8
2.3	An illustration used by Sampath et al. in [30] to describe the basic architecture of vanilla generative adversarial networks. G: Generator, D: Detective	14

Bibliography

- [1] J. K. Tsotsos, I. Kotseruba, A. Andreopoulos, and Y. Wu, “A possible reason for why data-driven beats theory-driven computer vision,” *arXiv preprint arXiv:1908.10933*, 2019.
- [2] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, “Theory-guided data science: A new paradigm for scientific discovery from data,” *IEEE Transactions on knowledge and data engineering*, vol. 29, no. 10, pp. 2318–2331, 2017.
- [3] A. Rosenfeld, *Digital picture processing*. Academic press, 1976.
- [4] O. Faugeras and O. A. Faugeras, *Three-dimensional computer vision: a geometric viewpoint*. MIT press, 1993.
- [5] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, “A deeper look at dataset bias,” in *Domain adaptation in computer vision applications*. Springer, 2017, pp. 37–55.
- [6] J. Bach, J. Langner, S. Otten, M. Holzäpfel, and E. Sax, “Data-driven development, a complementing approach for automotive systems engineering,” in *IEEE International Systems Engineering Symposium (ISSE)*. IEEE, 2017, pp. 1–6.
- [7] A. Luckow, M. Cook, N. Ashcraft, E. Weill, E. Djerekarov, and B. Vorster, “Deep learning in the automotive industry: Applications and tools,” in *IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 3759–3768.
- [8] K. O’Shea and R. Nash, “An introduction to convolutional neural networks,” *arXiv preprint arXiv:1511.08458*, 2015.
- [9] T. Wang, Z. Cao, S. Wang, J. Wang, L. Qi, A. Liu, M. Xie, and X. Li, “Privacy-enhanced data collection based on deep learning for internet of vehicles,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 6663–6672, 2019.
- [10] Y. Roh, G. Heo, and S. E. Whang, “A survey on data collection for machine learning: a big data-ai integration perspective,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, 2019.

- [11] F. Piewak, P. Pinggera, M. Schafer, D. Peter, B. Schwarz, N. Schneider, M. Enzweiler, D. Pfeiffer, and M. Zollner, “Boosting lidar-based semantic labeling by cross-modal training data generation,” in *European Conference on Computer Vision (ECCV) Workshops*, September 2018.
- [12] R. Sathya, A. Abraham *et al.*, “Comparison of supervised and unsupervised learning algorithms for pattern classification,” *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, pp. 34–38, 2013.
- [13] B. Friedman and H. Nissenbaum, “Bias in computer systems,” *ACM Trans. Inf. Syst.*, vol. 14, no. 3, pp. 330–347, 1996.
- [14] D. Danks and A. J. London, “Algorithmic bias in autonomous systems,” in *26th International Joint Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 4691–4697.
- [15] D. Samuel and G. Chechik, “Distributional robustness loss for long-tail learning,” in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9495–9504.
- [16] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, “Large-scale long-tailed recognition in an open world,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2537–2546.
- [17] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in *CVPR 2011*. IEEE, 2011, pp. 1521–1528.
- [18] T. De Vries, I. Misra, C. Wang, and L. Van der Maaten, “Does object recognition work for everyone?” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 52–59.
- [19] A. Marathe, R. Walambe, and K. Kotecha, “Evaluating the performance of ensemble methods and voting strategies for dense 2d pedestrian detection in the wild,” in *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 3575–3584.
- [20] M. Geisslinger, F. Poszler, J. Betz, C. Lütge, and M. Lienkamp, “Autonomous driving ethics: From trolley problem to ethics of risk,” *Philosophy & Technology*, vol. 34, no. 4, pp. 1033–1055, 2021.
- [21] E. Yudkowsky *et al.*, “Artificial intelligence as a positive and negative factor in global risk,” *Global catastrophic risks*, vol. 1, no. 303, p. 184, 2008.
- [22] J. Wang, X. Wang, T. Shen, Y. Wang, L. Li, Y. Tian, H. Yu, L. Chen, J. Xin, X. Wu, N. Zheng, and F.-Y. Wang, “Parallel vision for long-tail regularization: Initial results from IVFC autonomous driving testing,” *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 286–299, 2022.

- [23] J. Zhang, M. Zheng, M. Boyd, and E. Ohn-Bar, “X-world: Accessibility, vision, and autonomy meet,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9762–9771.
- [24] W. Yu, T. Yang, and C. Chen, “Towards resolving the challenge of long-tail distribution in UAV images for object detection,” in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3258–3267.
- [25] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes *et al.*, “Argoverse 2: Next generation datasets for self-driving perception and forecasting,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [26] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, “A comprehensive survey on transfer learning,” *IEEE*, vol. 109, no. 1, pp. 43–76, 2021.
- [27] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [28] L. A. Jeni, J. F. Cohn, and F. De La Torre, “Facing imbalanced data—recommendations for the use of performance metrics,” in *Humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 245–251.
- [29] B. H. Zhang, B. Lemoine, and M. Mitchell, “Mitigating unwanted biases with adversarial learning,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.
- [30] V. Sampath, I. Mourtua, J. J. Aguilar Martín, and A. Gutierrez, “A survey on generative adversarial networks for imbalance problems in computer vision tasks,” *Journal of big Data*, vol. 8, no. 1, pp. 1–59, 2021.
- [31] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R. W. Lau, and M.-H. Yang, “Vital: Visual tracking via adversarial learning,” in *IEEE conference on computer vision and pattern recognition*, 2018, pp. 8990–8999.
- [32] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, “Adversarial sensor attack on lidar-based perception in autonomous driving,” in *ACM SIGSAC Conference on Computer and Communications Security*. New York, NY, USA: Association for Computing Machinery, 2019, p. 2267–2281.
- [33] A. Boloor, X. He, C. Gill, Y. Vorobeychik, and X. Zhang, “Simple physical adversarial examples against end-to-end autonomous driving models,” in *IEEE International Conference on Embedded Software and Systems (ICESS)*, 2019, pp. 1–7.
- [34] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger, “Explainable AI: the new 42?” in *International cross-domain*

- conference for machine learning and knowledge extraction*. Springer, 2018, pp. 295–303.
- [35] L. Merrick and A. Taly, “The explanation game: Explaining machine learning models using shapley values,” in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds. Cham: Springer International Publishing, 2020, pp. 17–38.
 - [36] G. Joshi, R. Walambe, and K. Kotecha, “A review on explainability in multimodal deep neural nets,” *IEEE Access*, vol. 9, pp. 59 800–59 821, 2021.
 - [37] A. Wang, A. Liu, R. Zhang, A. Kleiman, L. Kim, D. Zhao, I. Shirai, A. Narayanan, and O. Russakovsky, “Revise: A tool for measuring and mitigating bias in visual datasets,” *International Journal of Computer Vision*, pp. 1–21, 2022.
 - [38] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, “Addressing bias in big data and AI for health care: A call for open science,” *Patterns*, vol. 2, no. 10, p. 100347, 2021.
 - [39] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.
 - [40] K. Stathoulopoulos and J. C. Mateos-Garcia, “Gender diversity in AI research,” 2019.
 - [41] J. García-González, P. Forcén, and M. Jimenez-Sanchez, “Men and women differ in their perception of gender bias in research institutions,” *PloS one*, vol. 14, no. 12, pp. 5–11, 2019.
 - [42] S. Leavy, “Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning,” in *1st international workshop on gender equality in software engineering*, 2018, pp. 14–16.
 - [43] M. Livingston, “Preventing racial bias in federal AI,” *Journal of Science Policy Governance*, vol. 16, no. 02, pp. 1–5, 2020.
 - [44] T. Goddard, K. B. Kahn, and A. Adkins, “Racial bias in driver yielding behavior at crosswalks,” *Transportation research part F: traffic psychology and behaviour*, vol. 33, pp. 1–6, 2015.
 - [45] M. Whittaker, M. Alper, C. L. Bennett, S. Hendren, L. Kaziunas, M. Mills, M. R. Morris, J. Rankin, E. Rogers, M. Salas *et al.*, “Disability, bias, and AI,” *AI Now Institute*, p. 6, 2019.

- [46] N. B. Chetan, J. Gong, H. Zhou, D. Bi, J. Lan, and L. Qie, “An overview of recent progress of lane detection for autonomous driving,” in *6th International conference on dependable systems and their applications (DSA)*. IEEE, 2020, pp. 341–346.
- [47] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [48] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine learning with oversampling and undersampling techniques: overview study and experimental results,” in *11th international conference on information and communication systems (ICICS)*. IEEE, 2020, pp. 243–248.
- [49] Q. Shi and H. Zhang, “Fault diagnosis of an autonomous vehicle with an improved svm algorithm subject to unbalanced datasets,” *IEEE Transactions on Industrial Electronics*, vol. 68, no. 7, pp. 6248–6256, 2020.
- [50] Y. Li and N. Vasconcelos, “Repair: Removing representation bias by dataset re-sampling,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9572–9581.
- [51] C. Drummond, R. C. Holte *et al.*, “C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling,” in *Workshop on learning from imbalanced datasets II*, vol. 11. Citeseer, 2003, pp. 1–8.
- [52] H. He, Y. Bai, E. A. Garcia, and S. Li, “Adasyn: Adaptive synthetic sampling approach for imbalanced learning,” in *IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 1322–1328.
- [53] A. Sarkar and K. Czamecki, “A behavior driven approach for sampling rare event situations for autonomous vehicles,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6407–6414.
- [54] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [55] J. Phillion, “Fastdraw: Addressing the long tail of lane detection by adapting a sequential prediction network,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 582–11 591.
- [56] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.
- [57] W. Ouyang, X. Wang, C. Zhang, and X. Yang, “Factors in finetuning deep model for object detection with long-tail distribution,” in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 864–873.

- [58] S. Gong, X. Liu, and A. K. Jain, “Mitigating face recognition bias via group adaptive classifier,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3414–3424.
- [59] C. for Disease Control and Prevention, “Global road safety,” Dec 2020. Available: <https://www.cdc.gov/injury/features/global-road-safety/index.html> [Online; accessed: 2022-09-30]
- [60] I. Model and L. Shamir, “Comparison of data set bias in object recognition benchmarks.” *IEEE Access*, vol. 3, no. 1, pp. 1953–1962, 2015.
- [61] V. Prabhu, R. R. Selvaraju, J. Hoffman, and N. Naik, “Can domain adaptation make object recognition work for everyone?” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3981–3988.
- [62] P. Kohli and A. Chadha, “Enabling pedestrian safety using computer vision techniques: A case study of the 2018 uber inc. self-driving car crash,” in *Future of Information and Communication Conference*. Springer, 2019, pp. 261–279.
- [63] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, “Towards reaching human performance in pedestrian detection,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 973–986, 2017.
- [64] N. Jaipuria, X. Zhang, R. Bhasin, M. Arafa, P. Chakravarty, S. Shrivastava, S. Manglani, and V. N. Murali, “Deflating dataset bias using synthetic data augmentation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [65] D. Leslie, “Understanding bias in facial recognition technologies,” *Understanding bias in facial recognition technologies: an explainer*. The Alan Turing Institute., vol. 4050457, 2020.
- [66] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.
- [67] P. Majumdar, S. Mittal, R. Singh, and M. Vatsa, “Unravelling the effect of image distortions for biased prediction of pre-trained face recognition models,” in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3786–3795.
- [68] J. Dastin, “Amazon scraps secret AI recruiting tool that showed bias against women,” Oct 2018. Available: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> [Online; accessed: 2022-09-29]
- [69] W. Knight, “Job screening service halts facial analysis of applicants,” Jan 2021. Available: <https://www.wired.com/story/job-screening-service-halts-facial-analysis-applicants/> [Online; accessed: 2022-09-30]

- [70] A. Najibi, “Racial discrimination in face recognition technology,” *Harvard Online: Science Policy and Social Justice*, vol. 24, 2020.
- [71] C. Garvie, *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.
- [72] Congress.gov, “All info - h.r.4008 - 116th congress (2019-2020): No biometric barriers to housing act of 2019,” July 2019. Available: <https://www.congress.gov/bill/116th-congress/house-bill/4008/text?r=11&s=1> [Online; accessed: 2022-09-29]
- [73] “George Floyd: Timeline of black deaths and protests,” Apr 2021. Available: <https://www.bbc.com/news/world-us-canada-52905408>
- [74] J. Greene, “Microsoft won’t sell police its facial-recognition technology, following similar moves by amazon and ibm,” Jun 2020. Available: <https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/> [Online; accessed: 2022-09-29]
- [75] K. Hill and R. Mac, “Facebook, citing societal concerns, plans to shut down facial recognition system,” Nov 2021. Available: <https://www.nytimes.com/2021/11/02/technology/facebook-facial-recognition.html> [Online; accessed: 2022-09-29]
- [76] “Greece: New biometrics policing program undermines rights,” Jan 2022. Available: <https://www.hrw.org/news/2022/01/18/greece-new-biometrics-policing-program-undermines-rights> [Online; accessed: 2022-09-29]
- [77] S. Shaily, S. Krishnan, S. Natarajan, and S. Periyasamy, “Smart driver monitoring system,” *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 25 633–25 648, 2021.
- [78] Y. Toyoda, G. Lucas, and J. Gratch, *Predicting Worker Accuracy from Nonverbal Behaviour: Benefits and Potential for Algorithmic Bias*. New York, NY, USA: Association for Computing Machinery, 2021, p. 25–30.
- [79] M. Shirpour, S. S. Beauchemin, and M. A. Bauer, “A probabilistic model for visual driver gaze approximation from head pose estimation,” in *IEEE 3rd Connected and Automated Vehicles Symposium (CAVS)*. IEEE, 2020, pp. 1–6.
- [80] E. Murphy-Chutorian and M. M. Trivedi, “Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness,” *IEEE Transactions on intelligent transportation systems*, vol. 11, no. 2, pp. 300–311, 2010.
- [81] M. Ngxande, “Correcting inter-sectional accuracy differences in drowsiness detection systems using generative adversarial networks (gans).” Ph.D. dissertation, University of KwaZulu-Natal, 2020.

- [82] F. Hauer, I. Gerostathopoulos, T. Schmidt, and A. Pretschner, “Clustering traffic scenarios using mental models as little as possible,” in *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1007–1012.
- [83] J. Zhao, J. Fang, Z. Ye, and L. Zhang, “Large scale autonomous driving scenarios clustering with self-supervised feature extraction,” in *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 473–480.
- [84] J.-F. Bonnefon, A. Shariff, and I. Rahwan, “The social dilemma of autonomous vehicles,” *Science*, vol. 352, no. 6293, pp. 1573–1576, 2016.
- [85] H. S. M. Lim and A. Taeihagh, “Algorithmic decision-making in avs: Understanding ethical and technical concerns for smart cities,” *Sustainability*, vol. 11, no. 20, p. 5791, 2019.
- [86] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, “The moral machine experiment,” *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.
- [87] V. Vakkuri, K.-K. Kemell, J. Kultanen, M. Siponen, and P. Abrahamsson, “Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study,” *arXiv preprint arXiv:1906.07946*, 2019.
- [88] J.-F. Bonnefon, A. Shariff, and I. Rahwan, “The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view],” *IEEE*, vol. 107, no. 3, pp. 502–504, 2019.
- [89] J. Borenstein, J. Herkert, and K. Miller, “Self-driving cars: Ethical responsibilities of design engineers,” *IEEE Technology and Society Magazine*, vol. 36, no. 2, pp. 67–75, 2017.
- [90] I. Panagiotopoulos and G. Dimitrakopoulos, “An empirical investigation on consumers’ intentions towards autonomous driving,” *Transportation research part C: emerging technologies*, vol. 95, pp. 773–784, 2018.
- [91] S. Karnouskos and F. Kerschbaum, “Privacy and integrity considerations in hyper-connected autonomous vehicles,” *IEEE*, vol. 106, no. 1, pp. 160–170, 2017.
- [92] C. Lai, R. Lu, D. Zheng, and X. Shen, “Security and privacy challenges in 5g-enabled vehicular networks,” *IEEE Network*, vol. 34, no. 2, pp. 37–45, 2020.
- [93] K. R. Varshney and H. Alemzadeh, “On the safety of machine learning: Cyber-physical systems, decision sciences, and data products,” *Big data*, vol. 5, no. 3, pp. 246–255, 2017.
- [94] D. Omeiza, H. Webb, M. Jirotko, and L. Kunze, “Explanations in autonomous driving: A survey,” *IEEE Transactions on Intelligent Transportation Systems*, 2021.

- [95] K. B. Kahn, J. McMahon, T. B. Goddard, and A. Adkins, “Racial bias in drivers’ yielding behavior at crosswalks : Understanding the effect,” 2017, NITC-RR-869. Portland, OR: Transportation Research and Education Center (TREC).