University of Zurich[UZH]

Faculty of Business, Economics and Informatics – Department of Informatics

Bias and Discrimination in Machine Learning

An Overview

Bachelor's Thesis at the Institute for Informatics at the University of Zurich

By
Vincent Theus
Supervised by
Prof. Lorenz Hilty
and
Dr. Clemens Mader
Mat. Nr. 16-705-287

Date of submission:
March 24, 2020

Abstract

Phrases such as *biased AI* or *discrimination through machine learning* have started appearing more and more in mainstream and social media. The AI industry is booming and an increasing number of decisions are being made by machines. Sometimes these decisions are perceived as ethically or morally wrong by society, which causes the algorithms that made these decisions to be labeled *biased*. But what does *biased* actually mean and when is it justified to speak of discrimination through machine learning? This thesis aims to examine bias and discrimination in machine learning more closely, provide a well formulated definitions of these terms and determine possible ways through which a machine learning algorithm could become biased during its development. Finally, the fairly new topic of machine learning in education is used to illustrate the possible consequences of using biased machine learning algorithms in schools.

Contents

List of Figures

Acknowledgements

I would like to thank Prof. Lorenz Hilty for the possibility to write this thesis on this topic under his supervision and for his support during the whole process. I would also like to thank Dr. Clemens Mader for his support, his input was especially valuable in the last chapters. Also, I am very thankful to Felix Ohswald from *goStudent* for sitting down with me for an interview about his company and their use of machine learning.

Of course many thanks also go out to my friends and family for supporting me from start to finish by listening to me ramble on about biased machines, bringing in fresh perspectives and for the time they invested in proof-reading this thesis. I could not have done it without them.

1. Introduction

Machine learning algorithms have gotten considerably more powerful and consequently, more popular. The media are full of examples of new exciting technologies which use machine learning to deliver better performance, relay more information, or better adapt to the user. *Century*, an artificial intelligence learning platform by a British start-up of the same name, uses machine learning to gather information about students skill-levels in different subjects, how to help them to learn new things effectively and then uses this information to build a learning environment which is tailored to each individual student [1]. Instagram takes advantage of machine learning algorithms capabilities to process massive quantities of data to personalize each user's feed, complete with personalized advertisements which are more relevant to the user. The algorithms are also used to identify and remove spam comments and fight cyberbullying [19]. These are just two examples of how machine learning is used in every-day applications to benefit the user and/or the company behind the algorithms.

However, even though the examples mentioned above seem to have an all-around positive impact on the people using it, the increasing use of ubiquitous machine learning algorithms is not without its downsides. Several ethical and regulatory concerns have been raised, including questions of data privacy, sustainability and fears of discrimination. Since the domain of machine learning is still quite new but nonetheless already very broad, this thesis will focus on the aspect of discrimination. Especially the potentials for bias during the development, training and application of machine learning algorithms are examined in detail. The goal is to provide a review of different definitions of bias and discrimination and to evaluate what constitutes a precise and critically reflected definition of those terms. Building on these definitions, an overview is to be realized which shows possible ways through which a machine learning algorithm could become biased and what the consequences of said bias are. This overview is realized as a general process model which describes all steps in the realization of a machine learning algorithm, starting with its inception and ending with the application and its output. The model is then used to examine a machine learning-based learning platform for potential sources of bias and how its creators could deal with them. Additionally, other machine learning-powered applications in the education sector, which are mostly used in China, are analyzed for their potentials for discrimination against students.

2. Definitions

This section contains definitions of terms which are essential to the understanding of this thesis. Some of the more complex terms are explained very superficially since an in-depth technical description lies outside the scope of this thesis. The definitions are only as detailed as is necessary for the understanding of the rest of the thesis.

2.1. Algorithm

Generally speaking, an algorithm provides a step-by-step method to solve a problem. Today, it is mostly associated with computer or mathematical operations such as data processing, manipulation, calculation or sorting [32].

2.2. Machine Learning and Machine Learning Algorithms

In laymen's terms, Machine Learning (ML) algorithms process large amounts of data to find patterns. These patterns can then be used to make an educated guess of what would also fit the patterns. A good example are recommendation systems which gather information about the user's behavior and preferences and use that to recommend content which the user might also find interesting [10].

ML can be classified as a subfield of the much broader field of Artificial Intelligence (AI) which encompasses:

- Computer Vision
- Robotics
- Speech Processing
- Natural Language Processing (NLP)
- Evolutionary Computations
- Machine Learning/Neural Networks [28]

These subfields are not entirely separate disciplines but often work in conjunction to solve a problem. For example, ML tools are often used to complete NLP pro-

cessing tasks by improving an algorithm's ability to parse large amounts of plain text data and automatically get insights from it [26].

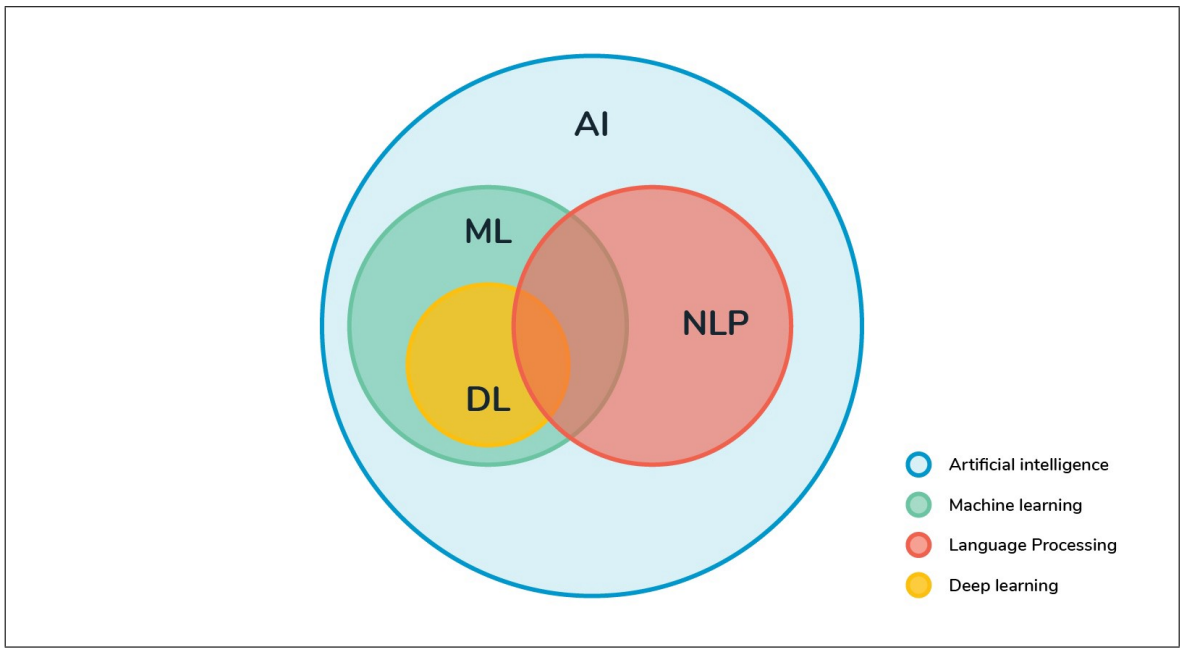

Figure 1: Intersecting subfields of artificial intelligence [26].

The subfield of machine learning can be further divided into different types of ML algorithms and techniques, with one such technique being neural networks/deep learning. ML algorithms can be broadly organized into the following four categories:

1. **Supervised learning:** The algorithm is trained on a data set with access to the correct responses. It can then generalize from this data set to respond correctly to any possible input [20].
2. **Unsupervised learning:** The algorithm is trained on a data set where the correct responses are not known and it attempts to find parallels between different inputs so that it can group similar inputs together [20].
3. **Reinforcement learning:** The algorithm is trained on a data set where the correct responses are known but it doesn't have access to them. It is informed when it gets an answer wrong, but not how to correct said answer. The algorithm then has to try different approaches to try and figure out how to get to the correct answer [20].
4. **Evolutionary learning:** An initial generation of algorithms with different characteristics is initialized and tested on a problem. The performance of each algorithm is evaluated and the top (usually top two) performers are selected as a ba-

sis for the next generation. Using the members of this "parent" generation, new "children" with a mixture of the parent's characteristics are created and tested on the same problem. This process is repeated until a predetermined performance value is reached [30].

According to Rechenberg and Pomberger, there are three fundamental techniques which are used under the aforementioned categories [24]:

Inductive reasoning: A data set of well defined, observed cases is used to try to derive universally applicable knowledge by generalizing from the observed cases. This approach is often used by banks to decide whether or not to give a loan to a customer. The system is trained on a data set of past applications for loans with many positive (loan granted) and negative (loan not granted) examples, from which it needs to identify the attributes which characterize the positive and negative cases respectively. This can be accomplished by providing the system with an attribute-value-list of the known examples from which the system can derive these typical attributes. Once this step has been completed, the system can then be fed with new cases for which there doesn't exist an attribute-value-list and the system can assess whether or not a loan should be granted based on the attributes of the case and the knowledge it has gained on the importance of these attributes [24].

Case-based reasoning: With this approach, the typical aspects of a problem and its solution are stored as a case. The basis for case-based reasoning is a large number of different cases which the ML system can use to compare new cases to. When such a new case is provided to the system, it searches its database for a similar case based on predefined similarity measures. Once a similar case has been found, the system tries to adapt it to the new case to see whether this leads to a solution. Successful attempts are stored in the database so that they can be compared with new cases in the future. Unsuccessful attempts are also stored in the database to prevent the system from making the same mistake again [24].

Neural nets (deep learning): This is by far the most complex of the three techniques and aims to simulate the human brain as a set of neurons which are linked by synapses. A neural net is made up of multiple layers of neurons with the simplest nets only consisting of an input and an output layer. The neurons of the input layer are connected to an arbitrary number of neurons in the output layer but not to neurons in the same layer. More complex neural nets have further hidden layers between the input and output layer. This is called deep learning. Looking at an in-

dividual neuron, all it does is calculate a weighted sum of all inputs – the weight of each input is defined by the synapse through which the input is fed to the neuron – which runs the sum through a function and passes the output on to a neuron in the next layer. Neural nets are trained with labelled data sets through which they learn to recognize patterns which determines the weights of the synapses delivering the inputs to the individual neurons [33]. The following figure shows a simple neural net:

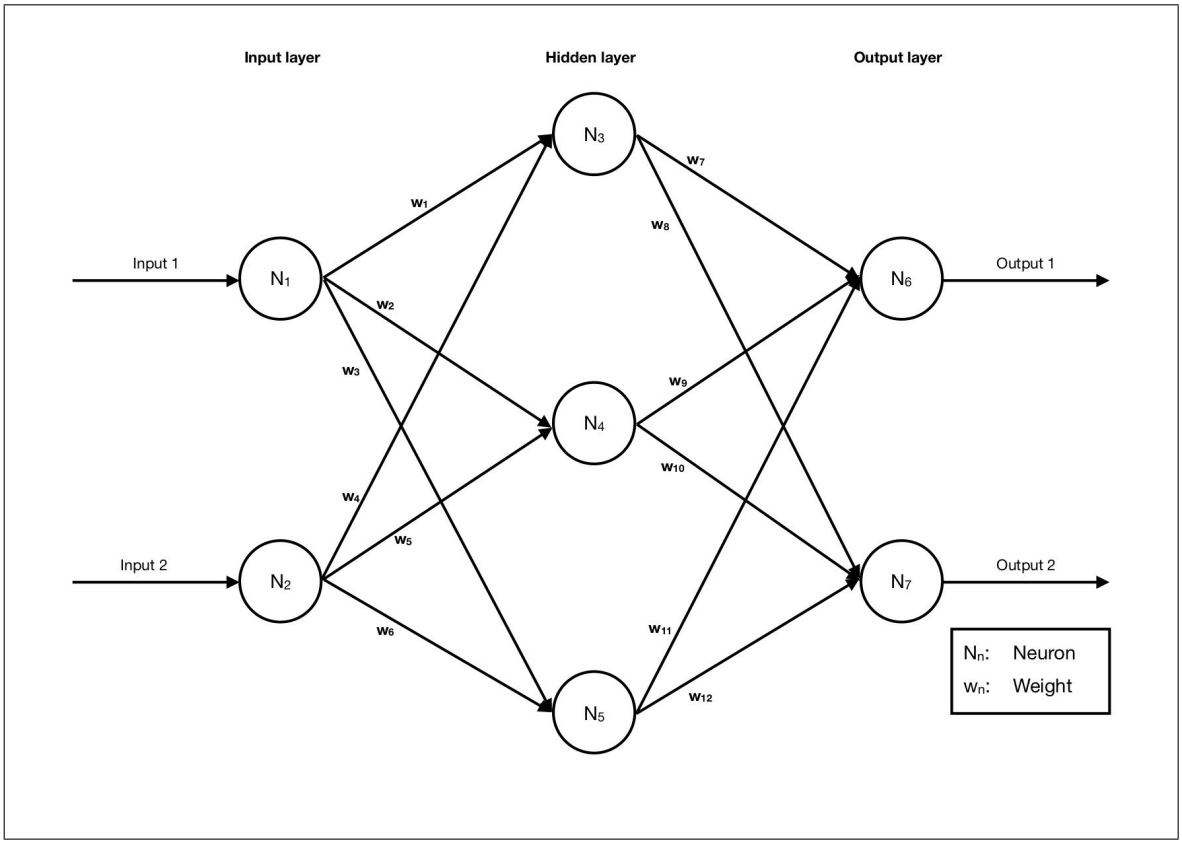

Figure 2: A simple neural net with one hidden layer (own representation based on [4]).

This is a very high-level definition of neural nets since an in-depth technical explanation is beyond the scope of this thesis. Neural nets have rapidly become one of the most widely used implementations of machine learning which is why they were used as the basis for all further analysis of ML in this thesis.

2.3. Bias

The Oxford Dictionary of English contains various definitions of the term bias, with the following two having some relevance to the context of ML algorithms:

1. *"A systematic distortion of a statistical result due to a factor not allowed for in its derivation."*
2. *"Inclination or prejudice for or against one person or group, especially in a way considered to be unfair."* [27]

From these definitions, it is already apparent that the term bias can be used to describe very different concepts. These two concepts are therefore discussed separately in the following sections.

2.3.1. Bias – The statistical view

How the concept of statistical bias has been used in the context of ML has changed significantly over the years.

According to Tom Mitchell, who was one of the first people to use the term bias in the context of ML, bias is nothing more than a *"[...]basis for choosing one generalization over another, other than strict consistency with the observed training instances."* [21]. This means that bias is used to give more weight to (or exclude) certain features in the data set which allows the algorithm to generalize beyond the initial training data set. If there is no bias, the algorithm will memorize the correct answers for the training data but will not be able to perform better on average than an algorithm which chooses an answer randomly.

More recently, bias is often used in the context of the bias-variance tradeoff (see section 2.5) which is a concept used during the training stages of a ML model. Since in the training stages, the desired output for a set of inputs is known, bias is used to refer to the difference between the desired output for a set of inputs and the average of the predictions of the model. It is basically a measure for how far off a models predictions are from the expectations of the people training the model. [29]. This concept only holds for training data since for almost all real world use-cases, there is no objectively true output for a set of inputs. This will be discussed further in section 3.

2.3.2. Bias –The ethical view

Biased ML applications from an ethical perspective are applications which systematically discriminate (see section 2.6) or handicap certain subsets of its users or subjects. For this definition the inner workings of a ML algorithm are irrelevant, the only thing that is considered is the output of the algorithm once it has passed the training stage and is used with real-world data. In contrast to the statistical definition of bias which denotes the difference between a ML models predictions and the expected output in the training stages, here the term bias refers to the difference between the models predictions and the correct answer from a (subjective) ethical point of view. When the term bias is used in the media, they almost exclusively refer to this concept because of its relation to discrimination. Unless explicitly stated otherwise, most of the discussion on bias in the rest of this thesis will therefore also relate to this concept. Furthermore, this concept of bias will be used somewhat interchangeably with performance in certain sections. For example, an algorithm which makes subjectively biased decisions is said to be performing badly (see figure 8).

2.4. Variance

Variance is another very important term in the domain of ML. It describes the variability in the model's predictions. A model with high variance will be very accurate on its training set but perform poorly on a new, "unseen" data set. This essentially means that the model learns the noise of its training set instead of the signal which would allow it to generalize beyond the training data, also called the "true signal" of the data [29]. The term "true signal" has to be taken with a grain of salt, since there is no absolute truth to be found in most real world use-cases. The truth refers to the patterns in the data which allow the algorithm to make predictions with higher accuracy than chance for a specific use-case.

2.5. Bias-variance tradeoff

The definition of bias which is used in the bias-variance tradeoff corresponds to the second one described in section 2.3.1. An ideal ML model would be one with

both low bias and low variance which would allow it to perform well on training data and also be able to generalize beyond it to make predictions for unseen data. However, this is very difficult to achieve since bias and variance both depend on the number of parameters the model can use to tweak its predictions, i.e. the number of inputs in a neural net. If it has too many parameters it will *overfit* the training data and learn the noise of the data set, if it has too few parameters it will *underfit* the data and generalize too much [29, 9]. The following figure visualizes the effects of high/low bias and variance on the accuracy of a ML model's predictions:

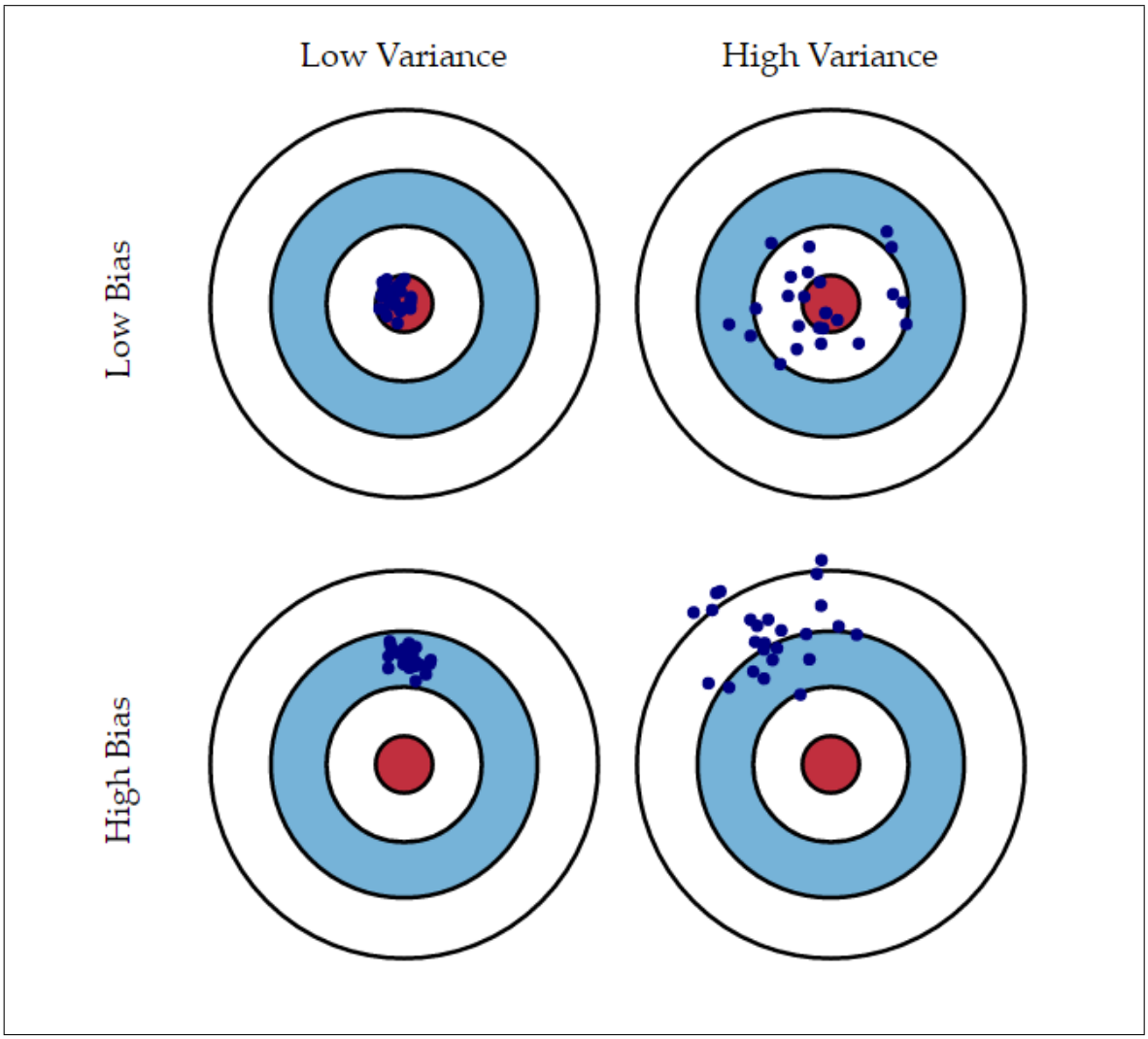

Figure 3: The effects of bias and variance on a model's predictions. Top right: overfit; Bottom left: underfit [16].

Trying to find acceptable levels of bias and variance based on the number of parameters used in the ML model is called bias-variance tradeoff.

2.6. Discrimination

In the context of civil rights law, "FindLaw" differentiate between *lawful* and *unlawful* discrimination [8].

Lawful discrimination refers to acts of discrimination which do not constitute a civil rights violation, for example if a landlord who doesn't allow pets in his building rejects an applicant because they have a dog [8].

An act of **Unlawful discrimination** constitutes a civil rights violation and is identified by unequal or unfair treatment of an individual or a group of people based solely on traits such as:

- Age
- Disability
- Ethnicity
- Gender
- Marital status
- National origin
- Race
- Religious beliefs
- Sexuality [8]

Discrimination can occur in many situations including, but not limited to:

- Education
- Employment
- Housing
- Government benefits and services
- Health care services
- Land use / zoning
- Lending and credit
- Public accomodations
- Transport
- Voting [8]

If the aforementioned landlord rejects a black applicant because they only want caucasian renters, the landlord violates the applicant's civil rights which is a case of unlawful discrimination and a punishable offense [8].

In the domain of ML, alleged discrimination through an ML algorithm is a consequence of generalization. As discussed in section 2.2, generalization is an essential part of ML since it allows the algorithm to apply knowledge of similar cases to new, unseen data and make a prediction. If such a generalization is perceived as ethically or morally wrong by humans, it is then called discrimination.

For the further discussions in this thesis, only unlawful discrimination is considered, since this is the kind of discrimination which is perceived as morally wrong.

3. A second look at bias and discrimination

With the topic of ML gaining more and more popularity in the mainstream and social media, the terms "biased AI" and "biased machine learning" have become buzzwords which are present in almost all pieces about the use of ML and AI in the real world. However, in almost all cases bias doesn't refer to the more technical definition as discussed in section 2.3.1, but to the ethical view on bias (section 2.3.2). As a consequence, discussions about discrimination by or through ML are seldom far. There are several issues with the way bias and discrimination in ML are handled in the media which are discussed in the subsequent sections.

3.1. The unbiased system

What is often ignored in the discussion about bias in ML in the media is the following question: If the output of a ML algorithm is considered biased from an ethical point of view, what would the output of an "unbiased" version of this algorithm be?

This problem can be illustrated on the example of Amazon who tried to use a ML algorithm to speed up their hiring process by filtering applicants [6]. The algorithm was trained with resumes of past applicants who were hired and the idea was that the algorithm would use the information present in those resumes to select the top candidates of a batch of new applicants. After testing the algorithm on new applications, it was discovered that the algorithm immediately discarded all female applicants since the successful applicants in the training data were mostly male [6]. The algorithm picked up on this pattern of predominantly male applicants being hired and generalized that suitable applicants must be male. Amazon stopped the development of the algorithm and claim that it was never used on real applications.

The question that remains is how an unbiased algorithm would or should have looked like in this situation. Would an unbiased algorithm have suggested male and female applicants in equal measure since that would satisfy the normative definition of equality? What would the algorithm have done if, at the end of the process, it was left with a male and a female applicant who were evaluated to be equally qualified for the position? Should the algorithm recommend the female

applicant since the majority of Amazon's employees are male and hiring a woman would help to increase diversity [6]? Couldn't this then be interpreted as a bias which favors women? It might appear that in this situation the algorithm should just recommend both applicants and leave the final decision to a human. This approach might remove this specific potential for bias from the ML system but in the end it is just outsourced to a human and humans are undoubtedly biased themselves. Besides, if the final decision were once again up to the same people who hired mostly male applicants in the past, why have an algorithm which is supposed to be free of bias in the first place?

These questions outline one of the biggest problems in the application of ML algorithms in the real world, especially if they directly impact human beings. In these cases bias just means that the algorithm delivered an output which is considered subjectively false. Defining what would be the "correct" output in these scenarios is often impossible since there is no real objective definition of the truth.

3.2. Avoiding discrimination

A common misconception in the media is that discrimination can be simply removed from ML algorithms by selecting a good, representative training set and creating an algorithm which ignores sensitive characteristics which could lead to discrimination like race or gender for example. In reality it is unfortunately not that simple. Even if such sensitive characteristics are removed, there may still exist correlated variables which can cause the algorithm to discriminate against a certain group. Such an example is provided in [39] where race is excluded from the calculations but discrimination may still occur based on the ZIP code of that person.

Also, going back to the example of the Amazon ML algorithm which was used to filter applicants; Even after the software engineers excluded the applicants' gender from the model, it still picked up on words which are statistically correlated with men or women, thus carrying on the gender bias and continuing to dismiss female applicants [6]. Zliobaite and Custers even argue that it is necessary to include sensitive characteristics like race or gender in ML models so that their influence on related variables can be isolated and removed from the model systematically [39]. However, there is no guarantee that altering the data set this much won't have

unforeseen consequences on the predictions which could result in a completely different kind of discrimination.

Even if isolating discrimination as explained above were a viable solution with no negative side-effects, using large amounts of sensitive, real world data is at odds with various data protection and data privacy regulations. These regulations have become more strict in the last few years, most recently with the EU General Data Protection Regulation (GDPR). The GDPR contains a prohibition on the use of so-called "automated decision-making" processes if the decisions are made without human intervention and have a significant effect on data subjects. This description fits most applications of ML and therefore makes the use of sensitive, personal data illegal unless the processing is required for contractual reasons, the use was explicitly green-lighted by a legal authority, or the subjects whose data is being processed have given their consent [3].

In today's society people are becoming more and more aware of the importance of data privacy and it seems highly unlikely that enough people would give explicit consent for their personal data to be used to train a ML model which makes autonomous decisions based on this data. This means that there won't be enough data available to determine the correlations of sensitive variables within the dataset which in turn means that not all potentials for discrimination can be identified and removed from the model. Balancing the need for personal data to avoid discrimination and the legal boundaries when building ML models which directly affect people seems to be an unsolvable problem.

3.3. Can bias and discrimination be defined adequately?

From the discussions in the previous few sections, as well as the concepts defined in section 2.3 it becomes clear that there is no one definition which adequately describes what bias in ML is. This is less of an issue with the technical/statistical definitions of bias than with the ethical definition. Different people have different opinions on what is morally or ethically wrong which means that these people would also have different definitions of what bias and a biased ML algorithm are. It is therefore very important to always consider the context in which the term bias is being used to evaluate which concept is being discussed.

While the exact definition of what discrimination is also differs from person to person, it is largely based on the definition provided in section 2.6 and is backed up by anti-discrimination laws in most countries.

Arguably, a precise and critically reflected definition of bias and discrimination in ML should also include a definition of what unbiased ML is and that is simply not possible in the overwhelming majority of cases. This raises the question whether and when ML should even be used.

3.4. When should machine learning be used?

It is impossible to guarantee that a ML model will be free from bias and discrimination when it is being used in a way that involves making decisions about human subjects, since bias and discrimination, as they are so often used in the context of ML, are such subjective and vague terms. Additionally, in the current mindset of society, most people probably wouldn't be comfortable with a machine deciding on important topics in their lives. This would rule out ML for all applications which have a direct impact on people and that is the case for most applications which are currently being used or developed.

There are definitely use-cases where ML could be used without the possibility of discriminating against people, for example in the domain of predictive maintenance. When manufacturing a product with any kind of machinery, it is only a matter of time before a component in the machine breaks which results in production downtime. In the worst case, the manufacturer needs to order a replacement for the broken component from a supplier and cannot continue production until the component was delivered and installed in the machinery. Depending on the time between the component breaking and its replacement being installed, the manufacturer might suffer heavy losses in revenue. It has been shown that a well trained ML model can very accurately predict the time when a component is expected to fail which enables the manufacturer to prepare and pre-order a replacement to minimize downtime [31].

Still though, arguably the most interesting, and profitable, use-cases of ML will have some kind of impact on human beings. That is why it is imperative to closely examine any ML algorithm for biases and potentials for discrimination during its

development and regularly check on its behavior post-launch. Possible sources for bias and discrimination will be analyzed in the subsequent section.

4. The discrimination pipeline

Designing, implementing, testing and deploying a ML algorithm or system is an extremely complex and time consuming process which requires large quantities of data and involves many people. Depending on the problem that is being tackled with ML, the data and people involved, as well as many other factors, the result of all this work may be a piece of software which is criticized in the media and by society because it is biased and discriminates against a group of people. Precisely because the process of creating a ML algorithm is such a complex, multi-facetted undertaking, there are many stages where bias may be introduced into the mix. To visualize this, the following process model is used in the subsequent sections to examine the most important steps in the development of a ML algorithm for their potentials for bias:

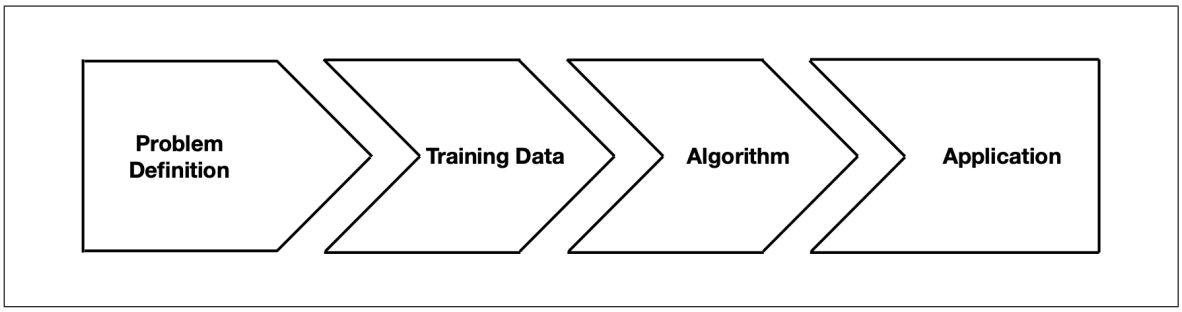

Figure 4: Key components of the machine learning development process (own representation).

In each step of this process model, I will be highlighting some questions which developers should ask themselves to ensure that they are aware of potential threats which could could cause their ML algorithm to become biased and end up discriminating against certain groups of people. The questions are also visualized in a flowchart which was designed to be used during the development of a ML algorithm and encourage developers to think about bias and discrimination during the development. It can be found in its entirety in appendix A.

4.1. Problem definition

This first step in the pipeline can already set the course for a ML model to become biased even before any data is collected and the first line of code is written. It is

important to analyze whether a problem can and should be solved with ML. If the possibility exists to solve a problem deterministically, it is probably the better approach, since a deterministic algorithm is much more stable and predictable. The impact on human beings is also something which needs to be considered when deciding whether a problem should be solved with ML. If a ML algorithm is responsible for decisions which could potentially ruin a persons life, it is questionable whether such an algorithm should ever be used.

The issue which presents itself when using ML is that real-world problems and concepts need to be translated into a numerical value which can be understood by the machine. It is often the case that these concepts are multi-facetted and/or too vague to allow for a straight forward translation. For example, if an insurance company wanted to use a ML algorithm to predict a customer's creditworthiness, the term creditworthiness would need to be defined in a way it can be computed. If creditworthiness is only defined as the highest chance to generate the maximum profit for the company, the algorithm would inevitably learn to behave in a predatory way and offer deals which will exploit people who have no other choice but to take the deal [12].

Of course this is a highly simplified example but it illustrates the problem of having to convert real-world concepts into machine-readable and understandable formats. It is very important to examine the problem from all angles to minimize the chance of introducing potentials for bias and discrimination into the pipeline before work on the algorithm has started.

Questions to ask

- Can the problem be solved without ML?
  - If a problem can be solved with deterministic algorithms, it does not make sense to use much more complex ML algorithms.
- Should the problem be solved with ML?
  - It is important to consider the impact the ML algorithm could have on human lives and whether it is ethically justifiable to have a machine making these decisions.
  - Even if a ML algorithm makes the subjectively "right" decision 99% of the time, it will come under heavy criticism for the remaining 1%.

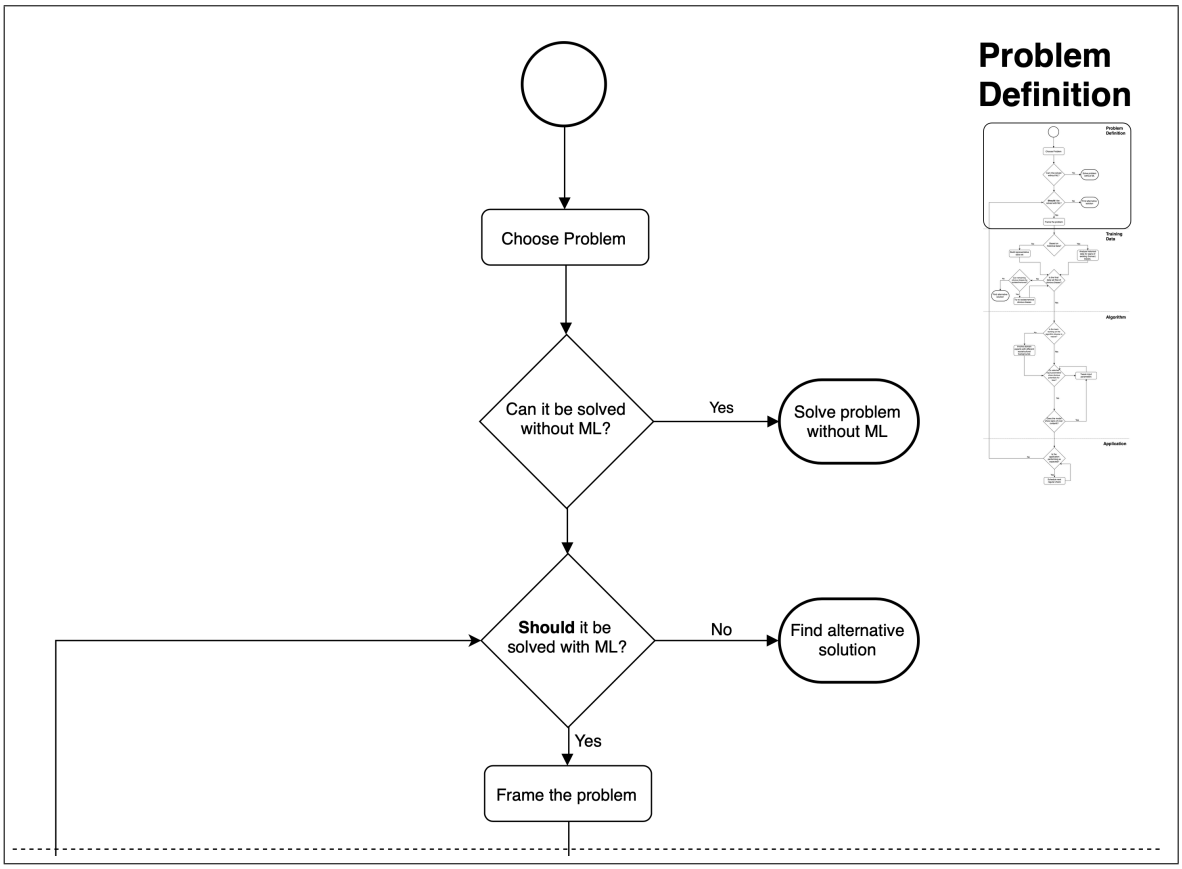

Figure 5: Flowchart depicting the most important questions in the problem definition phase (own representation).

4.2. Training data

The data used to train a ML model is arguably the most important component in the ML pipeline in terms of avoiding bias. If the training data set is flawed, the algorithm can never learn to make reasonable predictions with real-world data. There are two main ways in which a bad data set can compromise the algorithm's predictions, depending on the use case of the ML model and the origin of the data [12]. If the training data set is being built specifically for the model, it is possible to build a data set which is unrepresentative of reality. The other option is to use historical data to train a ML model which could carry innate human biases and thus influence the model's predictions. Both options are discussed in the following sections.

4.2.1. Unrepresentative data

This is not only a problem in ML models but in statistics in general. It is seldom possible to do statistical tests with data from the entire population, so most tests are done with a sample of the population. If the sample is not representative of the entire population, the conclusions drawn from that sample will not hold and the test will be biased. The same principle also holds for ML models.

A model which is trained to recognize faces cannot possibly be trained with pictures of the faces of the entire human population (and if it could, there would be no need for ML, since it would have a database with all faces and wouldn't need to recognize new, unseen faces). The developers have to build a sample data set of human faces on which the algorithm can be trained. Ideally, this data set should be representative of the human population in regards to:

- Age
- Gender
- Race
- Skin color/skin tone

If this is not the case, the algorithm will be significantly better at recognizing certain faces than others. For example, two frequently used data sets in training facial recognition models, *IJB-A* and *Adience*, were found to consist predominantly of light-skinned faces (79.6% and 86.2% respectively). Some ML models trained on these datasets subsequently performed significantly worse on determining the gender of dark-skinned woman (error rates of up to 34.7%) than light-skinned men (error rates of up to 0.8%) [2].

It is important to note at this point that a representative training data set cannot guarantee that the ML model will not be biased against some minorities. Even in a representative data set, minorities will remain minorities which means that the algorithm has less data on those minorities and will inevitably perform worse on them as a result. This source of bias cannot really be eliminated and must thus be kept in mind when relying on the predictions of ML algorithms.

Questions to ask

- Is the data set representative of the target audience?
  - A representative data set can increase the likelihood that the ML algorithm will perform well in most situations.

  - However, minorities in the target audience will therefore also be minorities in the data set. This means that the algorithm will perform worse on them, compared to the majority.
- Is the data set free of obvious biases/Can remaining obvious biases be isolated or removed?
  - It is impossible to identify all combinations of variables which could potentially result in a biased algorithm right from the start.
  - Paying special attention to combinations of sensitive characteristics, like the ones listed above, can be a good place to start finding potential sources of bias.

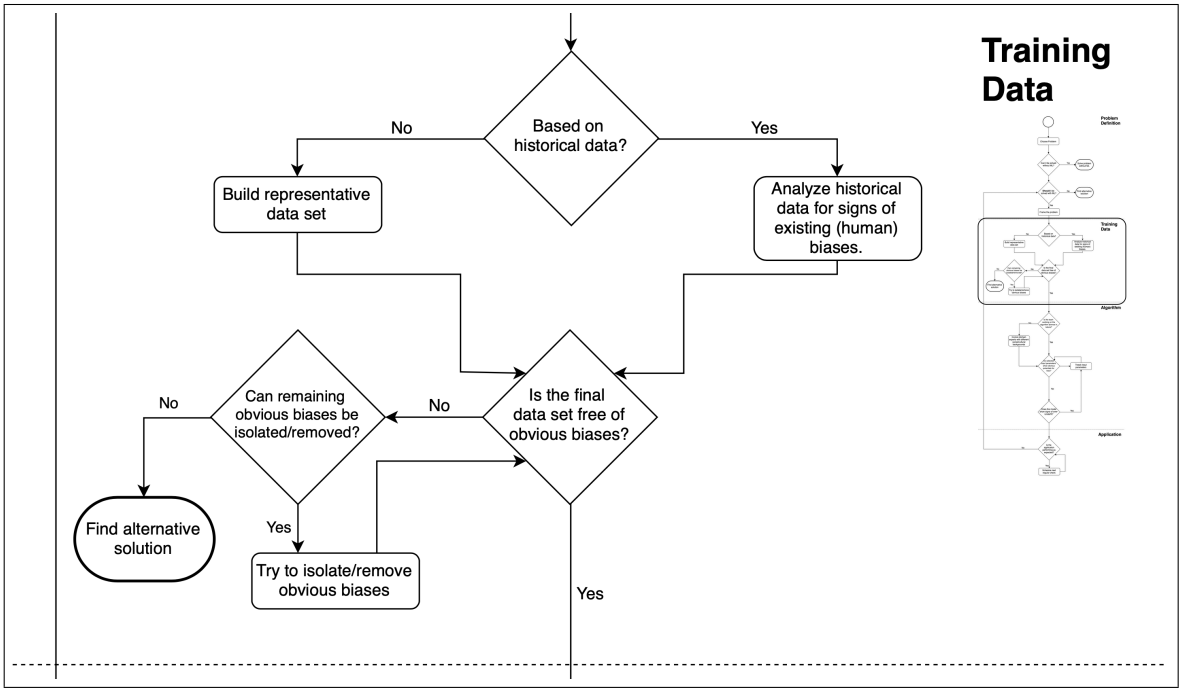

Figure 6: Flowchart depicting the most important questions during the preparation of the training data (own representation).

4.2.2. Biased historical data

When training a ML model on historical data, the goal is for the model to learn causal relationships between different variables in the data set. In this case, the data cannot be built to be a representative sample of the population since changing the historical data might compromise it in unforeseen ways and introduce new biases. This makes it very important to examine the data for pre-existing biases which could cause the ML algorithm to discriminate against certain subgroups of the population.

In England and Wales, ML systems are currently being tested with the goal of predicting possible locations of future crimes so that they may be prevented by targeted policing. The data used for training these systems consists of records of past crimes, their locations, the people involved, the nature of the crime, etc.. Concerns have been raised that the historical data may already be biased because of past over- or under-policing of certain areas and communities. For example, police officers have confirmed that young black men are much more likely to be stopped and searched than young white men due to human biases of some officers [22]. This leads to more criminal activity by young black men being recorded which skews the data and causes the ML model to more often predict crimes in areas with higher black populations.

As discussed in section 3.2, unfortunately it is very difficult if not impossible to eliminate all biases from a data set by removing sensitive characteristics. This issue raises the question whether ML models should ever be used in use cases such as the prediction of crimes since the results could have serious consequences for the people involved and the quality and fairness of the training data cannot be guaranteed.

Questions to ask

- Could the data be compromised by human biases?
  - If the historical data was largely gathered/created by humans, the chances for human biases which distort the data are very high.
  - Controversial topics, like the policing example discussed above, are particularly prone to human biases.
- Is the data set free of obvious human biases/Can remaining obvious human biases be isolated or removed?
  - Since it is not possible in most cases, to identify who exactly contributed which data points and what their personal biases are, working with historical data is always extremely dangerous.

These questions are also visualized in figure 6 in the previous section.

4.3. Algorithm

It has been established in the previous sections that a ML algorithm can never be free of bias if it is trained on bad data. However, even if an algorithm received

perfect, unbiased data, there are still many means through which it can become biased and discriminate against certain social or ethical groups.

The first potential source of bias lies in the selection of parameters through which the algorithm processes the data. Not only the number of parameters is relevant (*underfit vs. overfit*, see section 2.5) but also the selection of characteristics in the data which are used to make a prediction and how these characteristics relate to others in the data. Defining the right parameters is an extremely difficult task because it is impossible to know how the algorithm will use them to make its decisions and predictions. Judging whether a ML algorithm is biased is only possible after it has been trained and is tested on real data. If it is then determined that the algorithm engages in discriminatory behavior, it has to be retrained from scratch with different parameters in an attempt to eliminate the observed discrimination in the next iteration. But because it is not known exactly how the ML model uses the given parameters it is not as simple as identifying the *bad* parameters and removing them.

The second possible source of bias could be found in the people involved in the development of ML algorithms. The importance of representation in training data has been established previously, but could the social and cultural backgrounds of the people developing ML models be another source of bias which leads to discrimination?

The field of ML and AI is dominated by white males not only at big players in the industry like Google, Facebook or Amazon, with black people only making up 2.5% of Google's workforce, but also at universities where 80% of professors in this field are male [14]. Since technologies tend to represent the values of the people designing them, it is not unlikely that the background of developers of a ML algorithm have some influence on the algorithm's predictions, for example through the parameters they choose. The important aspect of this issue is that the bias the developers might introduce to the ML model are not added consciously but are a result of the people's origin and experiences which shape how they view and interact with the world.

It is hard to grasp the influence of the developer's social and cultural backgrounds on a ML model but as discussed in section 4.1, creating ML models requires translating abstract concepts into a machine-readable format which is no straight forward process. In recent years, the benefits of diverse teams in the workplace have been highlighted in various fields and industries. Therefore one might argue that

this general principle could also hold for the specific use-case of designing and programming a ML model. People with different backgrounds could have fundamentally different perspectives, thus giving this task to a diverse team might help to better understand the concepts and find a better way to translate them so that they can be used in a ML model.

Questions to ask

- Is the algorithm being created by a diverse team?
  - Getting people with different social/cultural, as well as professional backgrounds involved might offer new perspectives on a problem and uncover potentials for bias and discrimination which would have gone unnoticed otherwise.
  - **Important:** This is arguably the latest point in time where taking an interest in team diversity might still have an impact on the ML algorithm during its development. New perspectives might be already valuable earlier in the process and involving new people at this stage might uncover weaknesses in the earlier stages.
- Do the selected input parameters show potential for discrimination?
  - The combinations of input parameters used in neural nets have a huge influence on their predictions.
  - As described above, it is not as simple as identifying and removing problematic parameters, since there might be complex, hidden correlations in the data.
- Does the model show signs of under-/overfit during training?
  - Finding the right number of parameters is essential when trying to build a ML model.

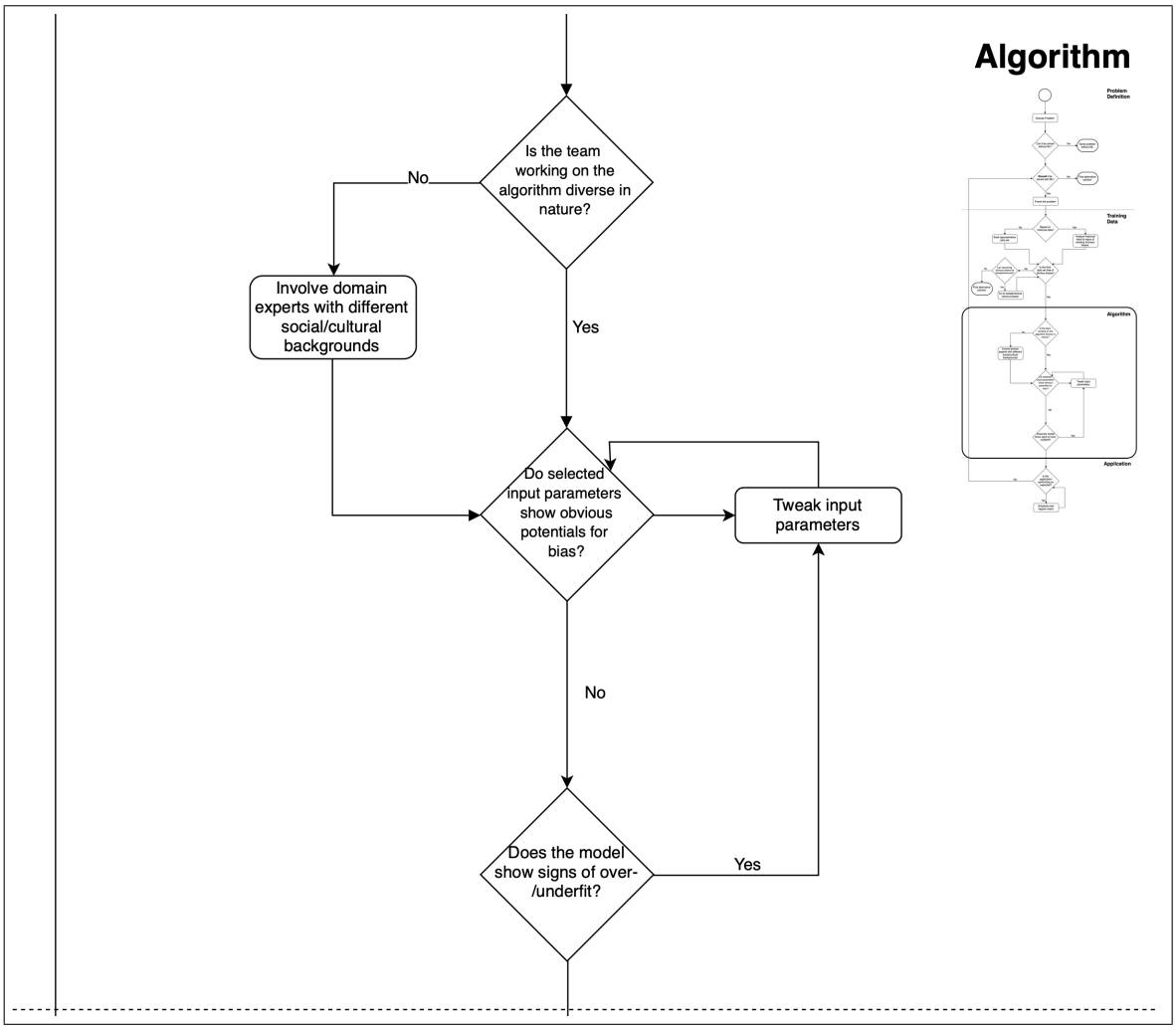

Figure 7: Flowchart depicting the most important questions when designing the ML-algorithm (own representation).

4.4. Application

The previous three steps of the discrimination pipeline have already exposed quite a lot of potentials for bias during the development of a ML model. If a model is found to be functioning as intended by its developers and free from unwanted biases, the model can be integrated into an application or be used on its own. However, this does not mean that the model is now safe from bias since the circumstances under which it is being used may change over time which might impact the adequacy of its predictions. Specifically, updates to the algorithm, the criteria

for assessing its performance and who decides on and enforces these criteria might all significantly alter an algorithms perceived performance [25].

Also, in relation to the first step in the pipeline, if the problem wasn't defined well enough and the algorithm is used in an application which it doesn't quite suit, it will not perform as expected or hoped. This can be explained by the so-called *No Free Lunch Theorem* (NFLT) [36] which, applied to ML, states that highly specialized algorithms which perform above average on a specific problem will perform below average on the remaining problems [7]. ML algorithms are one of the best examples of a highly specified algorithm. Performing above average in ML can be understood as performing better than chance, i.e. if it has to predict one of two outcomes, it will predict the correct outcome more than 50% of the time.

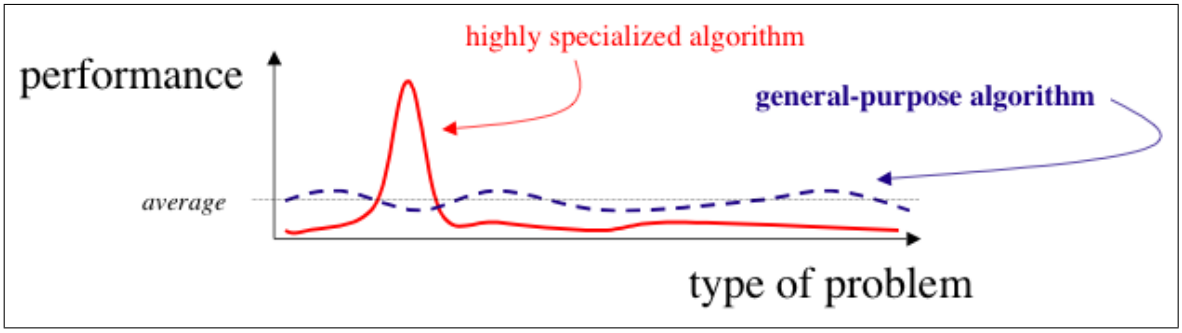

Figure 8: Illustration of the *No Free Lunch Theorem* [7].

Based on the figure above, it becomes clear that if the ML model is being used slightly differently from the way it was originally planned, its performance will suffer greatly and one of the effects could be discrimination against some of its subjects. Thus, it is very important to frequently evaluate whether the model is still suited to the task as time goes on.

Questions to ask

- Is the application performing as expected?
  - It is important to keep checking regularly whether the application is still performing as expected.
  - Since ML algorithms continue to learn once deployed, it is always possible that they might become biased over time.

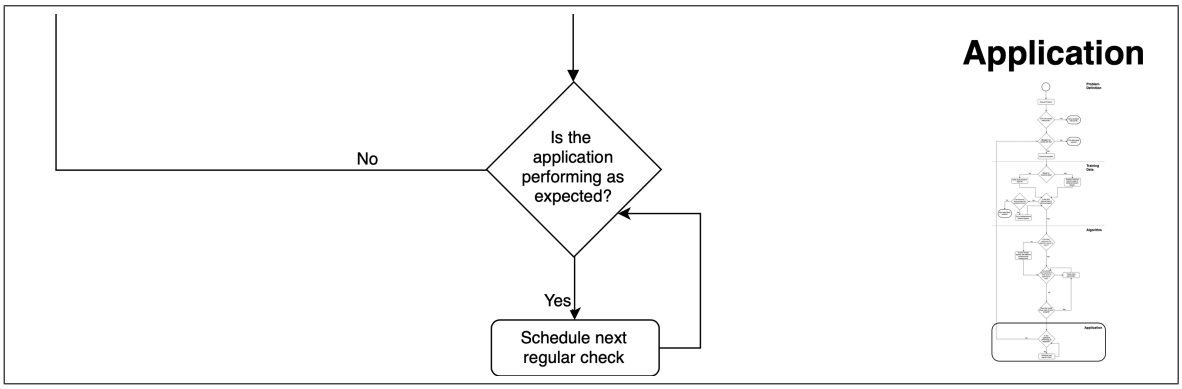

Figure 9: Flowchart highlighting the importance of regular health checks which must be done as long as the application is in use (own representation).

5. Machine learning in education

In recent years, the topic of digitalization has reached many classrooms all around the world. While many schools in Switzerland are just starting to integrate tablet computers into the classroom, there are others which are already a step further in the integration of digital technologies in the classroom. However, there are only a handful of companies which specifically target the education sector with their ML-based products. This will undoubtedly change in the next few years as AI and ML become more powerful and therefore it is quite interesting to examine the early movers in ML-powered education more closely.

On March 21, 2019, the website Quartz published a piece on a British startup, *Century*, which plans to introduce a ML platform in schools all over Belgium. The goal of this startup is to provide a personalized learning experience for each student and reduce workload for teachers by automating the grading process. This way, teachers have more time to focus on the most important part of their job: Teaching [1].

Some Chinese schools are even further ahead in the integration of computers and ML thanks to the government funding many schools' collaboration with developers. There are entire learning centers which are using ML and AI to tutor students which are struggling to keep up in a particular subject. The students sit in front of a laptop and solve tasks which are instantly graded by the ML platform the students are using. The platform learns the strengths, identifies gaps in the students knowledge and over time, with the data of many students, discovers connections between different topics which it can use to improve its tutoring. However, in contrast to the product *Century* offers, this platform does not adapt its content depending on how a student learns effectively [11].

The use of ML and AI in Chinese schools is not limited to tutoring and grading students. Some schools have started using cameras with deep-learning powered facial-recognition algorithms to identify whether students are paying attention or are distracted. Students are then assigned an attentiveness-score which can be observed over time and analyzed by teachers or even sent to parents. The cameras installed in classrooms take pictures as often as once per second to gather as much data as possible and this happens without the students knowledge or consent most of the time. According to the companies behind these technologies, the reason for using these facial-recognition systems is to ensure fairness for all students. A

teacher cannot pay attention to every student equally while the cameras can. If the ML model detects that a certain student is not participating in a specific subject, this behavior can be highlighted and proven by the data and the teacher and parents can focus on finding a solution for this specific issue [38].

An Austrian company called *goStudent*, a platform for providing tutoring online via video sessions, also uses ML and facial recognition. They track the emotional state of the student and tutor during sessions and use this information, among other like frequency of meetings and grades, as an indicator whether student and tutor work well together. If the system detects that the mood during the tutoring sessions is constantly bad, it will suggest a different tutor to try and improve the fit and hopefully the student's progress (see Appendix B, lines 148-210).

The use-cases described above all heavily rely on, and would not be possible to this extent without, ML. Most intended uses for ML in education generally seem to be going in either one or the other direction outlined above: It is planned to either use ML to help students learn by adapting class contents to the student, or at least specifically target identified gaps in knowledge, or to observe students, or in the case of goStudent tutors as well, in the classroom to analyze their behavior. Currently, the approach of classroom wide surveillance without the students knowledge is almost exclusively used in China, since the social norms and/or laws in western countries do not permit such a level of surveillance. The facial recognition system used by *goStudent* is only used with explicit consent of student and tutor. All three of these use-cases directly impact humans, which means that examining the ML models for biases that could lead to discrimination, is very important.

In the following sections, these approaches will be discussed separately since the use-cases and their implications are extremely different. To discuss the first approach, which is to adapt class contents to individual students, *Century's* ML platform will be examined by using the process model and questions defined in the previous section. The goal is to find spots where biases could have crept in during the development of the platform and what implications this could have for students working with the platform. Since there is not that much information available on the detailed implementation and use of ML to observe students in Chinese classrooms, the analysis will be done based on reports that were written about the systems instead of information directly provided by the developers/companies. The discussion of *goStudent*'s implementation and use of ML will be done based on an interview with Felix Ohswald, the founder of the platform, which took place on February 26, 2020.

5.1. Machine learning-powered human learning

Century was founded in 2013 with the goal to get all students to learn more efficiently and effectively by tailoring educational material to the individual student. Also, they aim to reduce teacher workloads by automating grading and aggregating huge quantities of data for the teachers so that they can focus on actually teaching their students. To that purpose, they built a learning platform powered by ML and AI. ML is used to identify areas where the students have gaps in knowledge or misconceptions and provides material to help them close these gaps. The platform was built in collaboration with teachers, neuroscientists and learning experts to ensure that students as well as teachers using the platform benefit to the highest degree possible. All learning content is either designed by teachers employed by *Century*, or by individual teachers at partnered schools for their own purposes. The content is then automatically broken down into micro-lessons, so-called nuggets, by the system. At the time of writing, *Century* is being used by schools in ten different countries [5]. The following discussion of their technology is based on the information available on their website and other sources from the internet, as no-one from *Century* was available for an interview to discuss their platform and methods in more detail.

5.1.1. Century – problem definition

The idea of using computer programs to teach or train specific educational material has been around for quite some time. There also already exist educational programs which adapt the content based on the user's performance. The big difference between conventional learning software and *Century's* platform is that the former use deterministic algorithms, while the latter use ML. This means that the degree to which conventional learning software can adapt the structure of the content is limited by the paths the developers have defined, whereas ML-driven software has a potentially unlimited number of paths for students to take. It is impossible for developers, even if they are supported by teachers and learning experts, to predict every optimal path for every student. Therefore, *Century's* vision of providing truly personalized learning content cannot be achieved with traditional, rule-based algorithms.

The most difficult task at this stage in the process is arguably the translation of student's different types of learning into something that the ML algorithm can

understand and compute. Learning is a very abstract concept and even though learners are generally grouped into four main types of learners (auditive, visual, haptic, and through intellect [34]), the exact preferences and strengths of each individual might almost be impossible to frame accurately. Therefore, if a student's exact needs for an optimal learning experience cannot be grasped by the platform, they will be at a severe disadvantage if a big part of the education is delivered via such a platform.

But even if the platform were able to accurately identify a student's learning type, one of the aforementioned types will not be able to learn as effectively as others since these people need to be able to interact physically with the materials [34]. Such a direct interaction cannot be provided with learning software and therefore students which learn through physical contact with materials might be at risk of being discriminated by such a platform.

This leads to the question whether this specific problem should be solved with ML. In *Century's* defense, it must be noted that the challenges with different types of learners is not a ML-related problem per se, since software based on deterministic algorithms cannot provide physical interaction with the materials either. The other issue however, where the student's optimal way of learning is not recognized accurately or even misidentified entirely could have consequences to the progress a student can make compared to the rest of the class. Since *Century* do not market their platform as a replacement for traditional education but only as an additional tool that can be used to improve learning in certain areas [5], it can be argued that any disadvantages some students receive through the ML algorithm's misidentification of their optimal way of learning, can be compensated for by teachers and pedagogical experts. In the end, there is no conclusive answer to this question since it is up to the individuals to judge whether the benefits outweigh the drawbacks.

5.1.2. Century – training data

Century do not provide a lot of information about the training of their ML model but based on a blog post on their website, they generated their own training data set, since the amount and granularity of data necessary for such a complex use-case had not existed beforehand. They trialed a beta version of their platform at 20 schools during two years which allowed them to gather data on the way students learn and train their algorithms [18]. Unfortunately, *Century* do not state which

schools they partnered with during the beta phase but since they are based in the UK, it is reasonable to assume that they partnered mostly with British schools. Integrating *Century's* platform in the schools required the schools to have an existing, quite sophisticated IT-system which was advanced enough to allow for the seamless integration of such novel technology. In addition, students would need to have access to devices with which they could access the platform. Considering that the amount of money public schools spend per student has been decreasing since 2009 [35], *Century* probably partnered mostly with well-funded (private) schools in order to trial their technology. This means that the students who participated were likely more homogenous than the British average in terms of social and cultural background, as well as more used to sophisticated learning software. This assumption is also partially backed by the testimonials listed on *Century's* website which are mostly from elite British schools [5].

Training data homogeneities, such as the one outlined above, are one of the most common ways through which bias is introduced into ML models and can often lead to discrimination down the line. As discussed in previous sections, ML algorithms need lots of data to be able to make reasonable predictions on unseen data and lack of data for certain groups in the training stage will inevitably lead to significantly worse performance on said groups. This bias could really manifest once the platform is being introduced to a wider range of schools in countries with radically different cultures and social norms.

Counteracting this potential source of bias is the fact that the actual learning content is not generated by ML algorithms but rather by teachers at *Century* and the partnered schools. This removes the possibility of actual biased learning materials from the equation since the teachers at the individual schools have years of experience and know which kind of materials work best for their students. Also, *Century* have come a long way since their initial beta version of the platform and their platform is now being used by a large number of schools in different countries which helps the ML model by providing it with fresh data based on which it can generalize in the future.

5.1.3. Century – algorithm

The exact input parameters *Century* use to train their ML algorithms are not known to anybody outside their organization. In a blog post by the company's CEO, Priya Lakhani, she states that *Century* strictly separate all data into three partitions: Per-

sonal data, learning data and content. Training and improving the ML algorithms is only done with learning data, i.e. data about the student's interaction with the platform which is anonymized before use [17]. Since the goal of data anonymization is only to eliminate the possibility that a specific person can be identified with the data [15], Lakhani's statement about anonymizing their data does not mean that *Century* remove or anonymize **all** personal data like age or gender of students using their platform. This means that at least some sensitive characteristics, as they were defined in section 2.6, could be used to train and improve the ML algorithms. If that is the case, the use of these characteristics in combination with the aforementioned homogeneity of the training data could have lead to biases in the algorithms. These biases could once again manifest once a more diverse group of students with different social and cultural norms are working with the platform. However, as discussed in section 3.2, it is not even guaranteed that the removal of all sensitive characteristics would help combat this source of bias since the complex correlations in the data are never easily identifiable.

The fact that *Century* is not only made up of computer scientists but also neuroscientists, learning experts and teachers ensures that at least from a professional perspective, the team behind the platform is very diverse. Working with schools in different countries also ensures that they regularly get input from people with different cultural and social backgrounds as well, which could help with the issues outlined above.

5.1.4. Century – application

Like with any other ML-driven application, the algorithms behind *Century's* platform will continue to learn and develop over time. With every new school that integrates this tool, new students with different needs and preferred ways of learning will start generating more and more data. It is therefore crucial that the performance of all algorithms is assessed regularly to ensure their continued excellence. If this is not done, the algorithms could start developing biases toward certain students' ways of learning which would result in a deteriorating experience for others. This could occur if, for example, a large amount of new schools in countries with different social norms and learning cultures started integrating *Century's* platform and generating large quantities of new data. The algorithms would start finding patterns in this new data and begin adapting to better deal with these patterns. If this transformation was left unchecked, it could result in a worse learning expe-

rience for earlier users even though they have not changed the way they interact with the platform.

Admittedly, the expected negative effects of such a transformation would arguably be quite limited in their significance in this specific use-case. As mentioned before, the task of the ML-algorithms on this platform is limited to the finding of gaps in knowledge and creation of an optimal path through various micro-lessons to effectively plug these gaps. The AI does not change the actual learning content and therefore students using the platform will still always receive tasks which their teachers found to be appropriate. In the absolute worst case, a gap in knowledge would not be recognized and the student would receive tasks which they can't solve properly. Since the platform is not designed as a replacement for class or teachers, and the teacher can see how each student is performing, it would not take long for the teacher to notice that a student is struggling. Together they could find the root cause for the student's struggles and possibly even provide feedback to *Century* that their algorithms failed to identify a gap in knowledge.

5.1.5. Century – conclusion

Based on the information available and the insights gained by applying the process model from section 4, the potential for biases which lead to severe discrimination of some users is rather small. Because the platform is designed as an additional tool to support teachers and students during class and at home and not as a school replacement, the damage it could cause because of bias is very limited. If the system works as intended, it helps target and improve specific weaknesses more efficiently than before, and if it doesn't work, these weaknesses just have to be identified in a different way.

It will be interesting to follow *Century* and their platform in the future to see how their AI and ML-algorithms develop and whether any signs for bias start to appear as they roll out their platform in more schools and new countries. Their currently ongoing project in Belgium will bring *Century* closer to the public eye since an entire region will be rolling out the platform in all its public schools [1].

5.2. Machine learning-powered schools

The approach Chinese schools have chosen to take arguably pursues the same goal as *Century*; They also aim to increase their students' performance. Their methods however, are substantially different and aim more for pressure through surveillance and less towards the creation of an optimal learning experience for students. Along with the aforementioned ML-powered evaluation of student's concentration levels with the help of surveillance cameras (see section 5.1), some schools employ even more drastic means to get their students to focus in class.

In 2019, the Wall Street Journal published an article about the use of headbands which measure electric signals from the brains of students. With the help of ML and AI these signals, along with other vital signs such as heart rate and blood pressure, are evaluated and can supposedly be used to determine a students' concentration level. If students fail to uphold high concentration values, they might be reprimanded by their teachers who also frequently send this data to students' parents who often put high pressure on their children to perform in school [13]. Apart from the obvious psychological stress this puts on students, the other big issue with these headbands is the reliability of the data recorded and fed to the ML algorithms. Is it actually possible to gather meaningful data on children's concentration levels?

5.2.1. Concentration levels – analyzing brainwaves with machine learning

Measuring brain activity with electrodes, also called electroencephalogram (EEG), is known for being highly susceptible to interference, or noise from many different sources. Minimal muscle movements and even cardiac signals can cause the recorded signals to become unusable for accurately determining brain activity [23]. In addition, EEGs are usually recorded in a controlled environment, with more than the three electrodes used in the headbands, to improve accuracy and minimize various external sources of noise, which is not the case for the classrooms these headbands are used in. This means that it is questionable whether the data captured by the headbands and used by the ML models is of adequate quality to allow for reasonable predictions of students concentration levels. As discussed in detail in section 4.2.2, when data with intrinsic biases is used to train ML algorithms, it is impossible to avoid bias in the application.

The consequences for students could be quite severe. In China, the pressure to perform in school is considerably higher than in most western countries. The students not only face pressure from their parent's high expectations but also from the importance of some tests which could make or break their future [38]. The additional psychological stress put on these children by even tracking their brain activity is enormous and if the algorithms evaluating the data are biased, the long-term consequences on the students mental health could be devastating.

In summary, the use of EEGs which are analyzed with ML algorithms to determine concentration levels of students is ethically highly questionable, at least in the eyes of the western world, since the chance for bias is extremely high due to the inherent susceptibility of EEGs to noise which distorts the data used to train the ML algorithms. Also, the algorithms' predictions can have serious consequences on the lives and futures of the affected subjects. These are typical signs that this problem **should not be solved with ML** in this way.

5.2.2. Concentration levels – facial recognition

The recognition of students' concentration levels through facial recognition might seem less concerning and error-prone than measuring their brainwaves, but it still shares most of the problems with said approach. Facial recognition systems have become quite powerful in identifying people and recognizing some basic emotions such as happiness, sadness, disappointment, anger, fear and surprise. Still, in some Chinese schools and universities, where facial recognition is also used across the campus to gain access to buildings and even pay for food, students are reporting that these systems frequently fail to identify them. Especially female students who often change their hairstyle and accessories are frequently not being recognized [37]. This indicates that the training data was not representative and/or not diverse enough to adequately prepare the algorithms. The detection of levels of concentration and whether a student is staring out of a window because they are distracted, or because they are thinking hard and do not really pay attention to where they are looking, seems even more complicated than identifying basic emotions. It is justified to question whether ML algorithms can perform well enough to ensure fairness for all students.

The consequences of this complexity could cause ML algorithms to misjudge wether a student is focused or not and permanently damage the students records, by flagging them as not paying attention in class. This results in the same psychological

stress for students described in the previous section, at least under the assumption that the students know that they are being watched and their attentiveness is being evaluated. However, as multiple sources confirm [37, 38], in some schools that is not the case. This fact raises several other ethical questions in regards to privacy, but discussing these is outside the scope of this thesis, since the issue of constant surveillance is not limited to, much less caused by ML.

Once again, whether the students know that they are being watched or not, ML algorithms which misidentify a student's concentration levels can have a huge impact on their life in such a performance-oriented society.

What must be mentioned at this stage is that the information about the use of ML-powered facial recognition in Chinese schools might not be covering the entire truth about the actual practices in China. According to Felix Ohswald, who visited a school in China which uses such a facial recognition system in December 2019, the goal of facial recognition is not necessarily only on identifying and punishing students who do not pay attention. In Chinese society, teachers are seen as very important since they oversee the education of the country's future workforce. Therefore, the focus of using ML-driven facial recognition is apparently much more on the teacher. If the algorithm detects that in a certain class, one teacher consistently has the attention of most students but with different teacher, 40% of students are not paying attention, the problem will be sought with the second teacher and not with the class (see Appendix B, lines 501-529).

Nonetheless, I will reiterate that managing concentration levels in classrooms **should not be solved with ML** in this way either, regardless of whether the focus lies on the students or the teachers.

5.3. goStudent – analyzing the emotional state with machine learning

While *goStudent* also use ML and facial recognition to observe their (virtual) classrooms, they do so on a smaller scale and an arguably more controlled environment. Felix Ohswald describes *goStudent* as a virtual school building with virtual classrooms where students meet one on one with their tutors to target specific issues they have in regular class. The process of matching a new student to their first tutor happens without the use of ML. *goStudent* first conduct an interview with the parents of a potential new student to find out as much about the child as possible

and then use that information in deterministic algorithms to match that student to one of their tutors based on their profile and preferences. Then, a trial session is held between student and prospective tutor and if both parties give positive feedback, the parents can buy a subscription from *goStudent* (see Appendix B, lines 9-122).

If a subscription is purchased, the regular tutoring begins in an agreed upon frequency via video conferences where students can bring material they are struggling with. These sessions can be recorded and stored on *goStudent*'s servers. The recordings capture the content the two participants are working on, as well as the video feed from either participant's computer camera which shows their faces. A ML algorithm can then be used to analyze the footage of both students and tutors faces for their emotional state and overall mood during the sessions. Things they look for are whether the student, or tutor, is looking happy, distracted, satisfied, or bored. This is being done because *goStudent* have the hypothesis that an overall positive mood and emotional state during tutoring sessions leads to higher frequency of sessions and a more sustainable, long-term performance improvement for the student (see Appendix B, lines 148-197). Thus, if the ML algorithm detects that the mood during sessions is constantly bad, or either participant is distracted often, it alerts *goStudent*'s employees which can then decide to suggest a different tutor to parents and students. So, does that mean that a tutor could potentially lose a student and therefore a source of income, if the ML algorithm finds the mood during sessions to be bad, regardless of whether it actually is?

When analyzing this use of ML for potentials for bias and discrimination, some of the same concerns as discussed in section 5.2 can be raised. Especially the detection of concentration levels during the tutoring session is something that can easily flag a student as not being attentive even though they might be lost in thought and thinking really hard about a problem. In contrast to the use in Chinese schools however, the ML algorithm is being used in a more controlled environment where there are fewer external factors and the images being analyzed are more consistent since they all come from webcams which show little else but the person's faces.

Also, since according to Felix Ohswald, the goal is not to meticulously record episodes where the student was distracted but the overall mood during the entire session (see Appendix B, lines 184-191), such mislabeling should not lead to any discrimination of students or tutors. Additionally, the decision whether or not to suggest a new tutor is not made based solely on the judgements of the ML al-

gorithm. If the algorithm identifies a possible bad fit, employees of *goStudent* first have a look at the other data that is being collected without the use of ML, like how the student's performance has developed, how often sessions are being held, etc.. If, and only if, this data also suggests that the collaboration is not working optimally, a new tutor will be suggested, with the final decision being up to the parents and students (see Appendix B, lines 537-565).

Overall, since the the influence of the ML algorithm on people's lives is so limited and it does not make impactful decisions on its own, without human involvement, the potentials for bias and discrimination are almost non-existent in this use-case. The answer to the question whether the mood during tutoring actually has a significant impact on a student's improvement in school will only be revealed after the continued analysis of the data being collected by *goStudent*. Felix Ohswald is adamant about the this correlation and argues that the data they have collected so far seems to support that hyptothesis.

5.4. Future of machine learning in education

The uses of ML discussed in the last few sections vary greatly in their potentials for bias and the consequences these biases could have on students. While the use of ML to watch over students to pressure them into focusing – whether that even works in the long term be put aside – is ethically highly questionable and would probably not be allowed in Europe under the new GDPR guidelines [3], using ML to support learning by identifying gaps in knowledge and provide a personalized learning experience certainly has potential. Public opinion and acceptance of a relatively new technology such as ML plays a very important part in the future of the technology. If ML in schools of the future is used to encourage and support individual learning and good habits, it is much more likely to be embraced than if it is being used to discourage and punish students for every little moment of imperfect behavior.

6. Conclusion and reflexion

The topic of bias and discrimination in machine learning is rapidly becoming more and more relevant, as ever more applications start to include ML in some capacity. As I have discussed in this thesis, it is therefore crucial that the discourse about bias and discrimination is built upon a solid understanding of these concepts. This understanding does not only entail the difference between the technical, statistical concept of bias and the more subjective, ethical one, but also awareness of the distinction between lawful and unlawful discrimination. Furthermore, I have shown that it is not always possible to find a definitive answer to the question how an unbiased version of a biased application would look like.

By examining the design and development process of ML algorithms, I have highlighted some of the most common ways through which an algorithm might become bias in such a way that it will discriminate against certain groups of people once it is being put to use. What I have also shown is that some influences on the ML algorithm are easier to quantify, such as the use of biased historical data in the training stage, while others are more obscure and harder to grasp, such as the development team composition in terms of social and cultural backgrounds. The *questions to ask* that were listed in each step of the process model provide a good guide to start moving towards a more bias-aware development process.

Using the aforementioned process model, I have analyzed the some of the most prominent approaches to using ML in education for their potentials of bias and discrimination, with very different conclusions. Using ML to support learning by identifying knowledge gaps and generating a path through learning materials to close those gaps effectively is very promising and its potentials for, and the consequences of bias are limited. In contrast however, using ML in combination with facial recognition or even to analyze students' brain activity might have dramatic consequences on its subjects when affected by bias, especially under the cultural circumstances in China.

What I would like to highlight once again at this point is the fact that the effort to prevent bias never stops as long as the ML algorithm is in use. Due to their nature, ML algorithms will inevitably change their behavior over time and this slow mutation must always be watched closely so that an intervention is possible at the earliest signs of bias or discrimination. This is not only in the interest of the ML algorithms subjects, but also the company and developers behind it. In the

current situation in the mainstream and social media, automated decision-making is being watched very carefully and judged very quickly. If a ML algorithm or a ML-powered application gets a reputation for being biased, it is almost impossible to get rid of this reputation again.

Finally, the most important question that has come up during this thesis and to which I want to refer once again is whether any given problem **should** be solved with ML. The fact that there is an official paragraph in the GDPR specifically geared towards limiting the use of ML in areas where the effects on people are considerable, shows that government entities are also aware of the potentially devastating effects of automated decision making. Whether this paragraph in the GDPR proves sufficient to protect people from the dangers which could come through biased ML and AI remains to be seen.

7. Outlook and future work

As ML and AI become more and more capable, high quality data will become more valuable than ever before, especially in use-cases which require personal data of the subjects. With this need for more high quality personal data, issues of privacy and data security will become more prevalent. As discussed in this thesis, the use of personal, sensitive data can be a source of bias, even when the more explicit characteristics like gender or race are removed. I believe that implicit, hidden biases in training data sets will be one of the biggest issues in future applications of ML. It is questionable whether it will ever be possible to detect and remove all sources of bias without altering the data set so much that it is no longer suited to train a ML algorithm on it.

It will therefore also become very important in the future to provide guidelines to ML companies to combat preventable biases. The process model provided in this thesis could be extended further to include acceptance criteria in each step and form the foundation of an industry standard which compels companies to adapt a bias-aware development process. A similar approach could be taken by law makers and other governing entities to enforce certain rules and regulations which ML companies have to follow in order to be allowed to release their algorithms.

The domain of ML-supported education will undoubtedly become increasingly relevant in the future as well. The long-term effects of integrating ML into the classroom will need to be examined for how they affect students, teachers and other stakeholders in the domain. Considering the rapid improvement of ML and AI, it is likely that their presence and influence in the classroom will only grow as time goes on. It might supercharge education in countries which can afford the technology, allowing first world countries to pull even further ahead of developing and third world countries. As a first step however, the performance of students using a ML-driven platform such as *Century* could be observed and compared to students with more traditional teaching styles in a long-term study which could clearly show the potential of the technology as well as possibly uncovering some unexpected drawbacks.

References

[1] J. Anderson, "A british start-up will put ai into 700 schools in belgium," 2019. [Online]. Available: https://qz.com/1577451/century-tech-signs-deal-to-put-ai-in-700-classrooms-in-belgium/ [Accessed: July 25, 2019]

[2] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, 2018, pp. 77–91.

[3] A. Burt, "How will the gdpr impact machine learning?" 2018. [Online]. Available: https://www.oreilly.com/radar/how-will-the-gdpr-impact-machine-learning/ [Accessed: May 16, 2018]

[4] J. Castañón, "10 machine learning methods that every data scientist should know," 2019. [Online]. Available: https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9 [Accessed: September 25, 2019]

[5] Century, "Century homepage," 2020. [Online]. Available: https://www.century.tech/ [Accessed: January 7, 2020]

[6] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [Accessed: October 3, 2019]

[7] L. Fedden, "The no free lunch theorem (or why you can't have your cake and eat it)," 2017. [Online]. Available: https://medium.com/@LeonFedden/the-no-free-lunch-theorem-62ae2c3ed10c [Accessed: November 27, 2019]

[8] FindLaw, "What is discrimination?" n.d. [Online]. Available: https://civilrights.findlaw.com/civil-rights-overview/what-is-discrimination.html [Accessed: September 27, 2019]

[9] P. Gupta, "Balancing bias and variance to control errors in machine learning," 2017. [Online]. Available: https://towardsdatascience.com/balancing-bias-and-variance-to-control-errors-in-machine-learning-16ced95724db [Accessed: September 27, 2019]

[10] K. Hao, "What is machine learning," 2018. [Online]. Available: https://www.technologyreview.com/s/612437/what-is-machine-learning-we-drew-you-another-flowchart/ [Accessed: August 29, 2019]

[11] K. Hao, "China has started a grand experiment in ai education. it could reshape how the world learns." 2019. [Online]. Available: https://www.technologyreview.com/s/614057/china-squirrel-has-started-a-grand-experiment-in-ai-education-it-could-reshape-how-the/ [Accessed: January 6, 2020]

[12] K. Hao, "How ai bias happens," 2019. [Online]. Available: https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/ [Accessed: November 7, 2019]

[13] S. Hong, C. Tai, and Y. Wang, "China's efforts to lead the way in ai start in its classrooms," 2019. [Online]. Available: https://www.wsj.com/articles/chinas-efforts-to-lead-the-way-in-ai-start-in-its-classrooms-11571958181 [Accessed: January 29, 2020]

[14] A. I. Index, "Artificial intelligence index - 2018 annual report," 2018. [Online]. Available: http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf [Accessed: November 27, 2019]

[15] Informatica, "What is data anonymization?" n.d. [Online]. Available: https://www.informatica.com/nl/resources/articles/what-is-data-anonymization.html#fbid=iYVKHdepKqO [Accessed: January 9, 2020]

[16] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[17] P. Lakhani, "The 'no nonsense' guide to artificial intelligence," 2018. [Online]. Available: https://www.century.tech/news/no-nonsense-guide-to-ai/ [Accessed: January 9, 2020]

[18] M. Ma, "How does century's ai work?" 2019. [Online]. Available: https://www.century.tech/news/how-does-centurys-ai-work/ [Accessed: January 7, 2020]

[19] B. Marr, "The amazing ways instagram uses big data and artificial intelligence," 2018. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2018/03/16/the-

amazing-ways-instagram-uses-big-data-and-artificial-intelligence/ [Accessed: March 16, 2018]

[20] S. Marsland, *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2014.

[21] T. M. Mitchell, *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research . . . , 1980.

[22] M. Oswald and A. Babuta, "Data analytics and algorithmic bias in policing," 2019.

[23] R. Paranjape, J. Mahovsky, L. Benedicenti, and Z. Koles, "The electroencephalogram as a biometric," in *Canadian Conference on Electrical and Computer Engineering 2001. Conference Proceedings (Cat. No. 01TH8555)*, vol. 2. IEEE, 2001, pp. 1363–1366.

[24] P. Rechenberg, G. Pomberger *et al.*, *Informatik-Handbuch*. Hanser, 2002, vol. 3.

[25] J. Sandhu, "Understanding and reducing bias in machine learning," 2019. [Online]. Available: https://towardsdatascience.com/understanding-and-reducing-bias-in-machine-learning-6565e23900ac [Accessed: November 27, 2019]

[26] B. Sathiyakugan, "Learn natural language processing from scratch," 2018. [Online]. Available: https://blog.goodaudience.com/learn-natural-language-processing-from-scratch-7893314725ff [Accessed: September 25, 2019]

[27] J. Simpson and E. Weiner, *The Oxford English Dictionary*, ser. 18. Oxford: Clarendon Press, 1989.

[28] N. Singh, "Artificial intelligence and it's sub-fields," 2018. [Online]. Available: https://medium.com/@neha49712/artificial-intelligence-and-its-sub-fields-a5a63d8263e8 [Accessed: September 25, 2019]

[29] S. Singh, "Understanding the bias-variance tradeoff," 2018. [Online]. Available: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229 [Accessed: September 26]

[30] D. Soni, "Introduction to evolutionary algorithms," 2018. [Online]. Available: https://towardsdatascience.com/introduction-to-evolutionary-algorithms-a8594b484ac [Accessed: September 25, 2019]

[31] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 812–820, 2014.

[32] Techopedia, "Definition - what does algorithm mean?" 2019. [Online]. Available: https://www.techopedia.com/definition/3739/algorithm [Accessed: August 29, 2019]

[33] N. Udyavar, "A beginner's guide to neural networks: Part one," 2017. [Online]. Available: https://towardsdatascience.com/a-beginners-guide-to-neural-networks-b6be0d442fa4 [Accessed: September 26, 2019]

[34] F. Vester, *Denken, Lernen, Vergessen : was geht in unserem Kopf vor, wie lernt das Gehirn, und wann lässt es uns im Stich? [Thinking, learning, forgetting: What happens in our head, how does our brain learn, and when does it let us down?]*, updated reissue, 37 ed., ser. dtv. Munich: dtv, 2016, vol. 33045.

[35] S. Weale, "Funding for 80% of schools in england 'worse next year than 2015'," 2019. [Online]. Available: https://www.theguardian.com/education/2019/sep/30/funding-80-percent-schools-england-worse-next-year#maincontent [Accessed: January 8, 2020]

[36] D. H. Wolpert, W. G. Macready *et al.*, "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation*, vol. 1, no. 1, pp. 67–82, 1997.

[37] E. Xie, "Artificial intelligence is watching china's students but how well can it really see?" 2019. [Online]. Available: https://www.scmp.com/news/china/politics/article/3027349/artificial-intelligence-watching-chinas-students-how-well-can [Accessed: January 29, 2020]

[38] X. Yujie, "Camera above the classroom," 2019. [Online]. Available: https://www.sixthtone.com/news/1003759/camera-above-the-classroom [Accessed: January 6, 2020]

[39] I. Zliobaite and B. Custers, "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models," *Artificial Intelligence and Law*, vol. 24, no. 2, pp. 183–201, 2016.

Appendix

Appendix A Bias flowchart

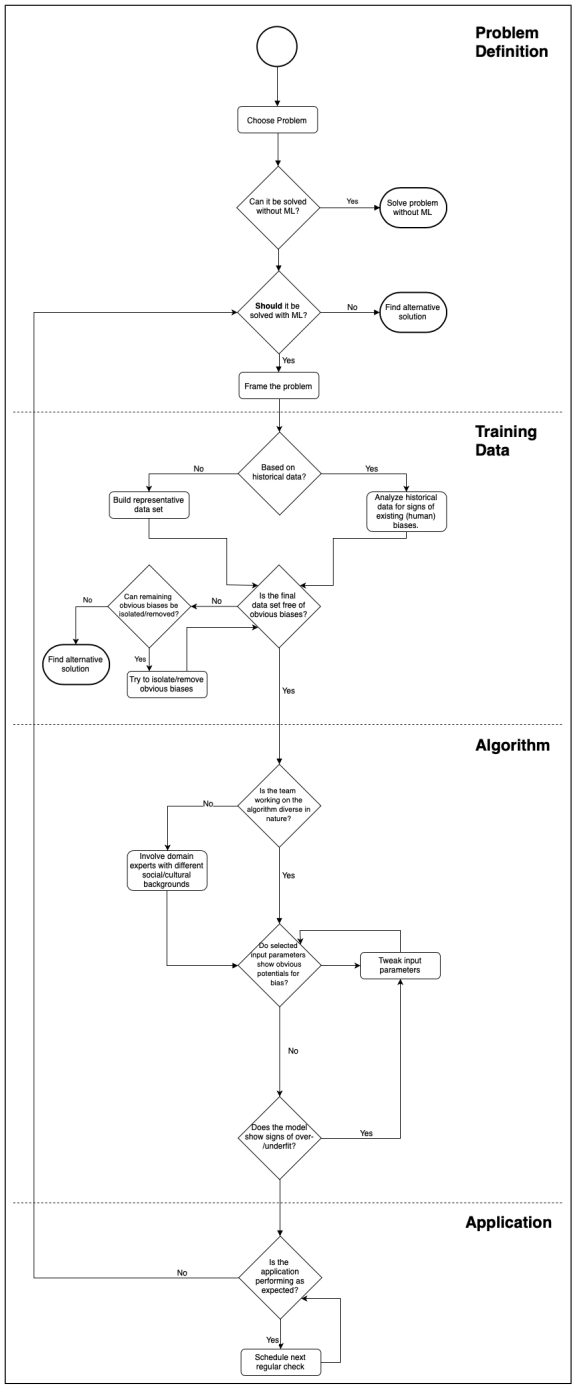

Figure A.1: Bias flowchart to highlight important questions during a bias-aware development process.

Appendix B Interview with Felix Ohswald from goStudent

Speakers: Vincent Theus (V), Felix Ohswald (F)

{Disclaimer: Original interview in German, transcribed and simultaneously translated into English.}

V: Alright, so I have already had a look on your website goStudent and Clemens Mader made me aware of you guys and you mediate, among other things, between tutors and students and how is this happening at the moment, do you do that by hand, do the tutors get a selection of whom they can coach, or how does that work?

F: We as the company goStudent are something like a global school. That means we connect students from the ages of 6 to 19 years with world class teachers. And we connect them in a virtual classroom, which means that students and teachers don't see each other physically in the same room, but in a virtual class rom via video calls like we are doing right now, specifically tailored for class. And what we offer as a platform, we provide the infrastructure which enables long-term, successful tuition. That means we take the load off parents so that they don't have to spend time to find a suitable tutor for their child, but we do that work for them, we take care of a frictionless payment process, we take care of the quality, we make sure that the child is actually getting better at school, we make sure that the child enjoys the lessons, that the child has fun with the tutor, that there is a transparent overview over all lessons, so everything which belongs to tuition we provide and support. To answer the question: How much is done by hand, it's like this: If parents send an enquiry, for example if they go to our home page send an enquiry there, they leave their phone number and contact info and then as a first step there is a real phone call. That means one of our distribution team calls the parents and conducts a so-called "diagnosis-chat". This diagnosis-chat is the first step towards successful tuition, because this diagnosis-chat is like a chat with a doctor. It's about finding the symptoms, yes, where do parents have difficulties with the child, which subjects does the child struggle with? Has the child had a private tutor after school before? What are the parents' expectations for the child for the rest of the school year? This is is... in this chat we're figuring out what the status quo is. Where are we at the moment? And now it's like this that the sales person, the distribution person fills this data into our contact management tool and based on the data the person entered, our system proposes tutors who can fit well. This is the first step

which is already partially automated. Because the system gets the... the input, okay this is a student visiting 10th grade, from Bavaria, German "Realgymnasum", has problems in maths and English, has an availability which is rather on Saturdays and Sundays, prefers male tutor, female tutor... has already had a tutor in the past. We collect this data and in this first chat... the system can say, based on this data, okay we have two tutors at the moment in the system where we know they fit well to this profile. And then the distribution person makes the decision and says okay, I'll take tutor A because I believe they fit even better. Then it's like this.. the student and the tutor are brought together in a first trial unit that... the students can try out for free and is being used to get to know each other where the tutor can have a look... can get a better picture of the student, yes, and can better figure out where the student is at. Also conduct a kind of assessment and based on this assessment can figure out, okay, where are the screws which need to be turned to bring the child to success. After this first chat, after this acquaintance lesson, a second phone call takes place where the distribution person sells one of our memberships to the parents. Everything else, the further communication usually takes place between tutor and parents on our platform, on our infrastructure which we provide. ... That's how you can imagine the process.

V: Exactly.. And, uhm, when you say the system automatically suggests candidates... uhm ... is that happening, from the point of view of the algorithms, are these all deterministic things, do you already have something with machine learning or AI in use or is that still being planned?

F: I can show you something, hang on.

V: Yes?

F: There is one part which is just... I'm gonna say classically deterministic, yes? Uhm, ... let's have a look at that one first. One moment. ... [Pause 20 seconds] ... So... I'm gonna share my screen here.

V: Mhm.

F: That should give you an overview over.. over data that is being collected.

V: Mhm.

F: Mhm. So we structure this into a diagnosis part, uhm, this means, this are for example data "before trial session" data that are being collected in the phone call.

V: Mhm.

F: There are the things after the trial lesson. There are also things being recorded like the feedback of the tutor on the subject of motivation, uhm, level of understanding, proposed frequency. Then there is the purchase information this then when the customer has bought, what the customer actually bought. Then here there is the category "tutor match". These are exactly the subjects, there is data stored here like gender, over what are they studying or what kind of education does the person have. What are interests, how many lessons has this tutor already had with other students. Uhm... which subjects is this tutor currently teaching, uhm, how are for example the ... no-show rates, that means how often does the tutor not show up or there are delays. How many of the trial lessons is the tutor converting, for example the "trial-session conversion rate". That means we can see if the tutor has a unit with a student, does the student buy afterwards, or not, yes? That is for example a very essential information about this ... this first success. So there is a multitude of, that is classic now, no? These are just the single data points that we measure.

V: Mhm.

F: ... And, uhm, based on these ... based on these data points, we can ... Do you now also see the other page, or not? One moment.. Do you still see the excel?

V: No, I see the, ... the slides.

F: Yes, alright. ... This means that we create here in the background, we create a profile for each tutor.

V: Mhm.

F: Yes, that's how you can imagine that. And at the same time we look at, if we then look at the tutor's profile, how does at the same time improve the performance of the child.

V: Mhm.

F: Is there ... is there a positive ... is there a positive development in the grades the child is writing, in the frequency, in the level of understanding. That means we take a look here whether there is an improvement. And here it is so that from this, from this combination if there was an improvement with previous students, we can ... we create a score, we average a value that tells us okay, uhm, with a high

probability this tutor matches this new student. Because with a similar profile they have worked together successfully before.

V: Mhm.

F: That's the first, ... that's the first time where you really... uhm ... where you really have something that is just calculated algorithmically.

V: Yes.

F: This means calculating based on the ... the ... the success or failure that the student experiences during the lessons we can calculate, okay, based on this improvement, the student matches this tutor very well. This means that if a new student comes into the system who has a similar profile ... then this tutor is suited for this for sure.

V: Mhm.

F: At the same time, if a tutor does not yet have a customer, that means a tutor freshly joins the platform, then a comparison is being done with tutors that have a similar profile and are already on the platform for a longer time.

V: Yes.

F: That means we compare, okay, here we have the tutor Sebastian. The tutor Sebastian tutors already .. uhm .. I don't know ... 5-6 hours per week, uhm, or 10-15 hours per week, has this background profile, has these students, that means our system knows, okay, this new tutor that's joining that has a similar profile.. these are probably the students with an advantage.

V: Mhm.

F: Mhm. But this is something, that if you break it down mathematically, uhm, uhm, it's basically something that's not... there's no complex science behind it.

V: Yes.

F: The complexity or the peculiarity behind it is rather that we collect all this data and have all this data. Once you have this data mix available, you can play a lot with it and can look, okay, are there matches or are there no matches. There you don't need ... for this we don't have enough tutor-student matches, to say, okay, we're gonna use a deep learning algorithm and use that there to train it ourselves.

V: Mhm.

F: At the moment we have about 4000 students that are receiving tutoring with us with about 30000 learning units that are being held per month. That is already a large number but it's still not enough to get something reasonable, or promising with deep learning algorithms.

V: Mhm.

F: That's why the algorithms we use here to build these correlations are structured very simply.

V: Yes.

F: The topic that we do also have, that we already tested, where we work together with another company ... I can show you that here ... This is an, ... an, ..., an additional layer of information that we can get from a lesson.

V: Mhm.

F: That means, you can see here an extract, this is a screenshot from a lesson about accounting.

V: Mhm.

F: In the top right you can see the tutor and below you can see the student.

V: Yes.

F: The lessons with us can be recorded, like we're recording the Zoom call right now. And these Zoom lessons are being evaluated afterwards, where it's about measuring emotions. Uhm, here it's like this, we're working with a tool that's called ... I'll send it to you via chat ... iMotions / Affectiva. There on the page of iMotions you can load there booklet, which shows a detailed overview what exactly is being measured, what is being tracked, what algorithm is being used in the background. Then we can, locally on our servers, we can analyze that. This means that I can, uhm, I can analyze the recording, for different, uhm, uhm, emotion, uhm, measurements. That means, for example, I can determine, uhm, is the child looking happy. Is the child looking attentive. Uhm, is the child looking distracted, uhm, satisfaction, is the student in a ... in a ... in a ... positive expression, or a bored expression. The same for the tutors. Uhm, because we have the hypothesis that if a tutor and student match each other well, and they are having a good lesson, then you can recognize that in the emotions on their faces. That's our hypothesis. That's our assertion. And, uhm, we believe in that, that means if the program shows us, okay, there are positive emotions here that we can measure, uhm, then student and

teacher match each other well. If we can see that the emotional mood is rather negative, then we see, ... we know okay, they don't match each other well. And if this information, uhm, is created, then we can, in our contact management tool, we can recommend a new tutor to the parent. With the feedback; We have seen that the, uhm, the engagement on an emotional level is not working optimally with each other. We suggest trying out the alternate tutor, to then be able to compare; Was the feedback positive or negative.

V: Mhm.

F: And that's something where, in the background, there is already running, uhm, uhm, a lot of AI, if you want, these are also only these emotional topics, it's just ... they took a database of, of, of, a couple of million faces and defined emotional expressions of these faces. And, uhm, the system measures, uhm 32 points in the face and can then make a comparison. And that is very accurate of course, because it relies on a large data sample based on which it can then make the, ..., the calculation.

V: Mhm.

F: And this combination is of course very interesting, because you can, for the first time, uhm, make lessons between a student and a tutor emotionally analyzable. That did not exist in this form prior to this.

V: Yes.

F: Because this is something that, can be concerning, or alarming at first when you think about it, okay, we're measuring the face of the student and the tutor, uhm, but in the end, it's only about finding out, is there... is the mood positive, or is the mood negative. It's not about pointing fingers at the student and saying, okay, in minute 23 you were distracted or checked your phone and for that you shall be punished. That's not the idea at all. It's about whether the base mood in this, in the learning unit is positive or not and if it would work better or worse with a different tutor. And at the same time, that's what we're currently working on, that's going to be super interesting once we have more data points, to see whether there is a direct correlation between ... emotional mood and improved performance and frequency. What we believe ... I am convinced that if the mood is more positive between student and teacher, that the lesson is more fun and correspondingly, the units are being done more frequently and therefore the performance of the child improves sustainably.

V: Mhm.

F: And the tutor is way more content with his job, because the tutor doesn't want students that are not attentive during lessons either.

V: Yes.

F: So, students and tutors always have to be coordinated optimally. And that is something that in, in, in the educational system has never happened before. It was always about, okay, we need, uhm, enough teachers to meet the demand, but it's totally irrelevant what kind of educational profile the teacher possesses, what exactly qualifies the person in the first place, uhm, to teach and does this person match the student with, uhm, with whom I bring them together. Because in the offline world, you could never make this measurable. That's what we do here and that is surely something, with the facial recognition, which is technologically, I'll say, the most complex ... uhm ..., but delivers a, an extremely interesting outcome.

V: Mhm. Yes, that, that sounds very interesting. Uhm, now, I don't recall exactly what I told you in my Email, but my thesis is for the most part about bias und discrimination by machine learning or AI. That means, in my thesis I first looked at what ... at how you define bias in the first place, that was already a first challenge because there are a lot of different views and different uses of this term. Uhm, and then I looked at the development of a ML algorithm, what can, I'll say, go wrong. Which in the end leads to an algorithm which uhm, which discriminates against a certain subset of the target audience. So, I have a very classic example in there from Amazon, who wanted to filter job applications with machine learning and then, based on some characteristics make a decision who fits this and this position better and they then had the problem that the algorithm just rejected all female applicants because they had trained the thing on a data set which consisted of 80% males. They were past job applications which were successful and, uhm, because the company hired 80% men in this department ...[audio cuts out for 5 seconds]...

F: Uhm, yes, that, that, that is a good example, that you give, you can even make it more extreme. You can say what if I only had male applicants in the past and have therefore only brought men into the company, then a women doesn't have any chance at all. Uhm, there you have to add that of course the, the, the interpretation of the results of these algorithms has to be done carefully and you always have to look closely at the data, what actually are the raw data on which this is based.

V: Exactly.

F: Because there is ... there is ... you can ... you can basically with this algorithms, you can go two ways. Way A, there are already raw data available that you understand and you know, okay, these are for example raw data that describe moves in chess. Very simply put. And now now you fill the algorithm with data about chess moves. The outcome is that you find the best possible chess move. There can't be any discrimination there. Because whether i now move the pawn or the knight, they won't feel like they are being treated unfairly. It's about the outcome of the game. The second option is that I, that I use these algorithms when I don't know yet what I am searching for, when I'm looking at whether there might be a pattern which I didn't see at first. Of course that is always very ... there you often have, I don't know, uhm, you're analyzing tissue samples, uhm, to find a tumor, but don't know how exactly that looks yet and you fill the system with these samples and have a look at whether sometime there is an output the computer recognizes which you don't know yet. But of course if it's about topics such as job application management in companies, then, then, I think you have to watch out, uhm, what is left at the end of the day and, and, and does it makes sense what the output is.

V: Exactly.

F: Now in the domain of education, or concretely, what we do here. Uhm, ... is that emotion recognition is basically something that, I think you have to take a closer look with them, it is basically something that the raw data it is based on, that it is basically weighted the same. That means the male emotions, or the female emotions you can recognize in the facial expression, ... [audio cuts out for 5 seconds]..., or a 50/50 data sample if you will.

V: Mhm.

F: Uhm, ..., uhm, there you put, uhm, It would be interesting, that is a good point, I haven't even though about that actually, uhm, to think about, to look at the differences if I, uhm, if I uhm compare 100 male tutor and 100 female tutors, if you for example see some emotions with men more often or stronger then with women and if that in turn has an influence on the student or not. Very interesting to ... to think about... but ... a discrimination, or I think that's probably the question you're leading to, if there is something which with us ... uhm... channels this risk, no?

V: Yes, that's the direction I'm going in. [laughs]

F: Yes. [laughs]. Uhm, I have to think about that for a second ... [15 seconds pass] ... I mean, that which you have to say or what you have to look out for is ... we, uhm, in the system we bring together the student and the tutor. And uhm, it is a possibility that uhm, only male tutors are bringing the student quicker to success than female tutors. Does that now mean that the system in the future with a match should automatically favor the man or not? Hard to answer. Uhm, in principle I believe with such topics, that it is sensible to always look at it from the point of view of gender neutrality. That means you don't even give the system the information male or female. That's how I would think about it for our case.

V: Mhm.

F: That you don't even tell the system, this was the male tutor or the female tutor, but that you only profile, the educational profile, the performance profile, the matching with the student, that you, uhm, that you feed that into the system and leave out the gender. Because if you would, with job application management, simply remove the information, the data point male or female, then the outcome would probably be a different one.

V: I'm glad you say that. Because uhm, [laughs] they obviously also noticed that with the algorithm.

F: Mhm.

V: And then they actually removed the gender from the equation... with the application so that it isn't considered in the rating of the application. Uhm, they retrained the thing from scratch.

F: And what was the outcome? [smiles]

V: It still rejected all women. Because: In job applications there are apparently certain terms, how people describe themselves, that correlate strongly with men or women. That means, the male applicants described themselves with adjectives which better matched the ones used by applicants which were hired and the women used different adjectives and terms to describe themselves. And based on that they were still filtered out.

F: Mhm. I mean, you have to add that the outcome is not really surprising, if you look at the fact that in the past, probably companies always were male dominated. That means probably, if you look at the data then you can see in management 90%, I can see that in the speed of promotions, men were faster than women, and

probably, based on this bad data, probably the application management is biased this way.

V: Yes, that's also one of the main quintessences of my thesis that, if you, if you want to train a machine learning algorithm and you train it with fundamentally biased data, that doesn't match the target audience at all, then it can only go wrong.

F: Yes. Makes sense, totally comprehensible. Okay ... I do believe that there are examples in education that can also be extremely biased. To give you a simple example, uhm, if you look ad kindergarden educators, that you find in kindergarden. I don't know the exact numbers but, uhm, from what I've seen it's that, only looking at Europe, in Europe in most cases, the kindergarden educators are female. But, which also correlates with the fact that we always, also in the past have said that the upbringing and education of children in the early stages between the ages of, let's say uhm, 2 and 5, that has to be done by a woman. Uhm, and you only get face to face with the male teacher in higher grades for the first time. Even in primary schools you tend to find more women than men. If you now probably use such an algorithm to find out what makes an ideal kindergarden educator, that this system is probably gonna filter al male educators even if you exclude the gender solely based on description and other attributes. Even though, even though, often a male kindergarden educator can even be much better than a female kindergarden educator in some cases.

V: Yes.

F: I am sure that something, if you have historically viewed, uhm, yes, basically had this biased view, that if you use this data that it will lead to bad results.

V: Mhm.

F: So that would be an example where I say, I can imagine it could lead to weird results in education.

V: Yes if you actually ... if someone actually has the idea to do something like that.

[both laugh]

F: With us, it's actually like this, it's like this because we, because we haven't existed for long, the company has existed for 4 years, but the business model in it's current form has only existed since 1.5 years, that we, from the start, ... went into the

process very unbiased. With recruitment of tutors, as well as ... uhm, we don't discern between male and female... we do discern, we do have that information, this data point, but it doesn't enter into the decision whether they are accepted as a tutor or not.

V: Mhm.

F: Because a female tutor can give virtual lessons as well as a male tutor can. And we do have at the moment on the platform a very balanced level of tutors, in terms of male and female. What would be interesting to see, that I can't tell you ad hoc, that I have to, have to, have to observe more closely, are the differences, the differences in the performance development of the students that are being tutored.

V: Mhm.

F: So is there a measurable difference between genders.

V: Mhm.

F: Uhm, but a question to you:

V: Yes?

F: Is it ... at what point, at what point is this ... bias present, so at what point... I mean the example you gave with Amazon that makes sense to me. There, the population says: Makes total sense, women should work, therefore the result is wrong and there, there is a mistake in there somewhere. But I'm sure there are examples that are more complex where maybe this decision is not immediately ... or is it always based on gender, is that the classic use-case of bias in that area?

V: So, I am also asking this question in my thesis, so... There are, like you say these unequivocal cases, but because this bias, depending on the topic and person can be interpreted differently ... this bias is not the same in all contexts and not everybody says that the same things are right or wrong, uhm, that often you can't ... uhm ... you can't say what the right decision should be in a specific case, uhm, so if we take this example once again, I also ask these questions afterwards in my thesis... if you were to actually use the tool and it were at a point where for a position it rates a man and a woman the same, so they are, in the eyes of the system, equally qualified, uhm... who should be suggested then? Do you say: Well there are already 80% men in this company, therefore we suggest the woman, solely based on reasons of gender equality, or we want a better female quota... and then you can again make the argument, well now the man is being discriminated... he's not chosen because

he's a man or because the other person is a woman, or do you say we leave the decision to someone else.. a person? Uhm..

F: Yes, okay, that would be my answer to that, yes. So I myself, we in the company we are just above 40 people, we're currently hiring many new people, so until end of year, our team will be tripled, we're interestingly enough in the company in a phase where we often ask the question do we take candidate A or candidate B?

V: Mhm.

F: Uhm ... also the reality is often a bit different, in reality you often don't have to equally qualified candidates in front of you, uhm, but, but, what you do have often at the en of the day is that in some way you have decide based on a gut feeling. There is, if you do have that situation, that you have a male and female candidate for the same job, and they have a similar profile, similar strengths, similar weaknesses, then me personally, I decide based on a gut feeling with which of the two people I feel like we're on the same wavelength. Maybe there's just more common ground with that person. And that is something I look at, on principle, by nature, gender neutral. Others maybe don't do that, other managers possibly don't. But, but, but there I would actually act based on gut feeling, because sometimes you have to make decisions based on a gut feeling because you don't have the time to over-analyze.

V: Mhm, yes, exactly. So I then just went in the direction... I thought in the direction that if you do that in a company which previously hired 80% men, and you have an algorithm which says: Here, these two are equally qualified and then you give that to the person to decide, who hired 80% men previously, what was the point of having a machine learning algorithm that is free of bias towards a gender if you then give it back to the person who only hired men in the first place. If I can exaggerate a bit.

F: Yes. Hmm. Yes.

V: And then I get a bit to the point where there are problems which are not easily solvable. There is no clear answer, what is the right thing to do. And to then make decisions with machine learning in these situations, ... for most people this is a black box, there's just some output and it's then very difficult for people who have nothing to do with it, maybe even people who do have something to do with that, to accept that.

F: Yes well, I ... I think generally the recruitment process is, in terms of this data analysis, it is a touchy subject. Also, a colleague of mine owns a company which, uhm, do that, they offer to analyze CVs for companies and based on the CVs they make recommendations whether person A is better suited than person B. And basically, broken down, this algorithm reacts to, or with what it was trained, this, this, this algorithm, only with certain keywords. This means that if you have certain keywords, that means if you have a certain degree from a university, that means that you have a certain skill set, that means that you write certain hobbies into your CV, that they are better qualified than people who don't put these keywords in there. Uhm ... I see that basically only ever as a kind of additional tool. That can help the person or the recruiter or support them, maybe draw attention to things that the recruiter or human wouldn't have noticed but it can't replace them in this case.

V: Yes, exactly, I also think that.

F: Hmm... It's also, like I said, at the end of the day, also an emotional topic, like you say, if you reject a person based on the result from the black box, it's just not comprehensible and harder to process for the person. And you also won't be so stupid, as the company who uses the algorithm to say okay, look these are the main ten keywords, write those into your CV, because then the whole thing becomes obsolete again.

V: Mhm, yes.

F: A game of chess is simpler in this, there are no emotions.

V: Yes exactly. [both laugh]. It obviously gets complicated as soon as people are involved, or rather as soon as the decisions of this black box have a strong influence on the lives of a person. And here it get's difficult very quickly.

F: You also have, in principle, these uhm, trained algorithms are there to make binary decisions, yes?. Either you do A, or you, you do B. You take, take, take, one of two ways.

V: Mhm.

F: Uhm ... but it's interesting ... uhm.. so what is the ... the conclusion of your thesis or the conclusion of the results you take from such examples?

V: Uhm... so a large part of the thesis is about how such a bias enters into such an algorithm and what consequences that can have on people. And ... on a high

level it works out that in cases where effectively the consequences for a person are considerable, that of course already depends again on whom you ask, what a considerable consequence even is, that in today's surroundings and today's culture, I don't know if that will change in the future, that it is difficult to justify such a machine learning algorithm exactly because people don't see inside and they only see a piece of technology which discriminates against them and that is then perceived as extremely unfair.

F: But where would you say, where would you say does discrimination start? I want to give you a different example. Uhm, let's say that your company has a policy that you only want people who have finished there studies with an average of 5.5.

V: Mhm.

F: And there are companies, also in Switzerland, which actually do this. They just say okay, you apply, you upload your final evidence of achievement and then you see that 80% of those who are above 5.5 are men. Without complicated algorithms, simply: Do they have the threshold or not? Is it then already discrimination that you say okay, I now have a pool remaining, where 80% are men and 20% are women. This obviously means that I will hire 80% men and 20% women statistically speaking. Is that then discrimination? Do you then say ... you're not allowed to evaluate based on who is above 5.5, you have to do that differently? What do you think about that?

V: So the question whether you are or aren't allowed is very difficult to answer in general. And, I'm sure that if you ask 100 people you're gonna get I don't know how many different answers to that question.. uhm... There .. there... I can't give you a good answer to this question and that's something that is becoming more and more clear in this thesis I am writing. That there are many problems, where there isn't a clear answer. That makes it even more difficult to deploy such an algorithm that makes such decisions. Because no matter what it outputs, you will always find someone who says that's not okay at all, that is discrimination because of this, this and this. And... that's why I think that there are certain problems which you should not try to solve with such algorithms. Simply because they are too controversial and we are today still in the mindset that decisions made by machines are still harder to accept than if there is a person behind it. But ...

F: So if you... if you... I believe especially when you, when you're not 100% convinced that the machine's decision has turned out to be better. Because if you know

for example, when analyzing an X-Ray image, uhm, the machine tells me that I don't have a tumor then I trust the machine if I know that this has turned out to be true in many cases in the past.

V: Mhm.

F: If I know beforehand, okay the machine is only a lottery, whether it has found something or not, then I would rather go to a doctor and have this personal opinion but... even when the machine is statistically better as the doctor because the machine has saved and seen more of these images than a human can ever look at and remember.

V: Yes.

F: But of course, ... especially with the recruitment of people it's very difficult.

V: Yeah, and I think that... if a machine makes a mistake, if I carry on with your example of identifying a tumor, I think if, if the machine says: No, no tumor and there actually is one, then it is, or so I think, it is received much worse...

F: Yeah sure, but you can't point a finger at someone.

V: Yes.

F: If the doctor makes a mistake, which statistically probably happens way more often, that the doctor makes the mistake. You can always say well they made the mistake. Or they diagnosed it wrong. Whereas... and you put them on the pillory.. but if it's the machine.. you can't put the developer behind it on the pillory.

V: Yes, yes exactly.

F: Mmh, I also believe that this mindset is something people haven't gotten used to yet. Or that it's something that probably only the next generation will be used to.

V: Yes, yes exactly.

F: Okay, but that is, that is interesting, I understand. So these are the questions you're working on.

V: Mhm. Yes and for many questions in this are, there are no universally valid answers that are always right for everybody.

F: Obviously, yes. Uhm ... it's always a matter of whom you ask. I mean if, for example you ask in India where women's rights are being spurned the answer is obvious. [laughs]

V: Yes [laughs]

F: Why on earth hire a woman? [laughs]

V: Yes exactly. I am also looking at the use of machine learning in education more closely even though there isn't too much there yet. I have found two big things. One is a company, I don't know if you have heard that name before, Century. It's a startup from the UK. They are building a platform where the students can solve tasks on and based on the data the platform gathers, how the students solve these tasks, uhm, the machine learning algorithm finds gaps in knowledge and, and problems in understanding and then proposes certain exercises thereby supporting learning. And the second big direction is being used mostly in China. They started mounting cameras everywhere in their classrooms that do facial recognition and then evaluate whether the student is paying attention or not. And then based on that a report is generated ...

F: Although here, I have to add, I just was in China in December and had a look at some of the things there live. It is about ... it is often reported a bit one-sided in the western world. For the Chinese person or for the, for the, for the, society there, the teacher plays a very important role. Uhm, the teacher is at the center, they are the central authority that is responsible that the children, at the end of the day, perform well. And the facial recognition that is being used in many classrooms also has the purpose, often to find out whether the students pay attention better with this teacher or don't they? So less like: I am going to punish the student, uhm, Maxi Huber, now with extra homework because he didn't pay attention in class, but rather the other way around, I am punishing the teacher because 40% of his class obviously didn't pay attention. And that is often here from this ... of course we feel, we are all strong individuals in the western world, we feel personally attacked when somebody is filming us, for example. Whereas in China this collective thinking is much stronger, uhm, and it then means as much as: If a couple of students in this collective don't pay attention, apparently the lesson is not good, the lesson doesn't work well. And less that the louts, the evildoers, are those that never pay attention. That was also... I found that very interesting when I was there. But it is of course repulsive at first. That was also with us, the emotion recognition, that

is basically exactly the same, always less with the thought: I am going to punish one or the other and more: Does it work together?

V: Mhm.

F: Positive facial expressions, or negative ones. And based on that make the conclusion, it works well or it works not so well.

V: Mhm... So how is it.. do you actually already use that?

F: Yes, exactly.

V: Uhm, how is the reaction to, let's say the worst case scenario for one of your tutors. They are having lessons with a student and your facial recognition notices: Ahh, something doesn't fit well, and uhm, the tutor then looses the student based on the assessment of this algorithm that something's not working there?

F: Yes, as I have said, it is part of the equation, measuring the emotion is not the, the, the only criterion if we say: Okay, you can't tutor them anymore. We also look at, okay, was there a performance increase with the student, how regularly are sessions happening, is there maybe an additional feedback we got that was negative. So there must be many things at the same time that it actually comes to a separation, but, what we have seen in first tests is, that the expressiveness of the emotional analysis is pretty good. So if we see that, uhm, that there is little euphoria in the virtual classroom, if you like, uhm, and we then make a different assignment, a different match with a student and this then works better, with better results and higher frequency, uhm, and that there is more revenue being generated.

V: Mhm.

F: So that is already very interesting to see that there is apparently a, a, a, a correlation.

V: Mhm.

F: But as I've said, it can't be one-sided. Only based on this decision we do that. And because of that... the nice thing with us is that the entire lesson is mapped with us, the students, the tutors, the parents always do everything via our platform, via our ecosystem, on our infrastructure. That means we're really collecting all data points. From first contact to the student finishing school, or stopping lessons before that or they don't show up, etc.

V: Mhm.

F: So... you can really imagine it like this: You have this school building, it's not there physically but it is a virtual building, that consists of the tutors, the classrooms, the students, the organization.. uhm... and uhm, that's what we're measuring and building here.

V: Mhm.

F: Uhm, That's something that ... I believe that the facial recognition, the emotion recognition topic is less controversial than creating personalized exercises. Uhm, i can recommend you there, or I can give you the contact of Sana labs in Scandinavian countries, they are also quite good, they consult, for example, publishers, text-book makers... they consult them to offer the text books more personalized. That goes exactly in that direction. The students are getting the text books digitally, yes, and can uhm solve the exercises there and based on the solutions the students give, other exercises are recommended. That goes in the same direction as Century that you mentioned.

V: Yes.

F: There you have, I would presume, from the biasedness ... I can't think of an example where you could see such a bias, right? Do you have something, for example with Century, is there something that you say is critical or you have to look at more closely.

V: So, from what I have analyzed and examined, I came to a similar conclusion as you just did. I see the potential damage through bias as not that high, this has various reasons. The exercises are, so it's not like the machine learning algorithm creates the exercises on its own, the exercises that the students solve are created by their teachers and packaged in little micro-lessons. The only thing the algorithm does is change the order how the students step through the material. And ... the only thing you can think about in this case, this goes a bit towards pedagogics with different learning types, so how children and students learn the best. There are different kinds of types, there are those that have to interact physically with the material to learn most effectively, those that need to hear something, those that need to read something and on such a digital platform you can't provide something where the children can physically interact with. That means ... but that's not a problem of machine learning, but of of digital teaching aids that you can't serve all these different learning types.

F: Mhm.

V: But since Century isn't meant for completely replacing traditional class, but only as additionally a couple hours per week, uhm, possible disadvantages of people with a learning type which isn't well supported by the platform can be egalized by working in the classroom with the teacher, with physical objects. So, bias especially, where you have to say if certain students use this platform then they will have a disadvantage for life or, a little exaggerated ... I don't see that either.

F: Mhm. Yes, makes sense.

V: It's just a less big advantage.

F: Yes, yes, sure.

V: Yes.

F: No, I understand. Interesting. There is, there is... for example with text books, I also saw that, a Chinese company, they actually offer text books, it's a physical text book, you solve tasks, for example you have multiple choice tasks in that book and you have a pen which is delivered alongside the text book. If you, for example, circle answer d, this is then directly synced to a screen, so the application knows that you circled answer d and let's say d is wrong. Then the computer tells you okay, the next task you're going to do in your book, please on the next but one page.

V: Yes.

F: It tries to alleviate that a little bit that you can get the people who need the physical text book and can still work with pen and paper, but the order how they work in the book is determined by the computer. Was quite cool and interesting to see that live. Yeah so, there is quite a lot going on.

V: Yes, I am interested to see how that develops in the future.

F: Yes.

V: Century now has, last year or the year before that, signed a contract with Belgian officials. And they are now rolling out their system in all public schools in a region in Belgium, to really test that. Until now they only used it in private schools in the UK and Dubai and so on. That's going to be interesting to see how that thing works once a larger number of students have access to that.

F: Mhm. That's true. That's really interesting. Like I said, let's keep talking, I can also as soon as we have more data ... I find these correlations in the data very inter-

esting , like what is being measured during class and what is the outcome outside, outside our virtual classroom... that's very interesting for me and the company to observe and see. You notice, we're still at a very early stage... we're collecting loads of data, there are many treasures hidden there but to then interpret, to leverage, to use that is a big challenge.

F: Is there anything else you would like to talk about?

V: I'm just having a look at my questionnaire ... but I think we have talked about pretty much all the points I wanted to talk about... that's not bad. ... Mmh yes, maybe something that is a difficult topic that is hard to measure, but uhm... what is generally becoming more important in business and in general is a team that consists of many different people with different cultural, personal and professional backgrounds ... that they can maybe more effectively, that they can better analyze a problem from different perspectives and can uncover problems that a more homogenous team wouldn't have found. And that because of that you could think about who is working on such a machine learning algorithm .. uhm .. if I use the example of Century again, they have software engineers, teachers, neuroscientists, learning experts on their team ... so that you can minimize the risk of missing some important aspect.

F: Mhm... So me personally, I am am more ... I see that more relaxed... I believe that one of the biggest mistakes that has been done in education in the past 100 years is that ... is that we always want to do everything right. I believe that exactly this has led to the fact that we in education have not innovated as much in the last 100 years as would have been possible. Everybody surely does... also if you talk with your family and friends.. you even poke fun at it and say: Yes, the classroom of my grandparents, 50-60 years ago, has looked exactly the same as today's classroom. Frontal lessons with the teacher up front who educates the children is still the same. The text books have maybe changed a bit so that they are now gendered and male and female are visible in text books, that there are now maybe ... uhm ... uhm ... more modern terms being used than before... because we are always scared of what terrible things would happen if you make a mistake just once, yes? That's the motto... if I make a mistake in education, then ... then I only breed idiots who are socially inadequate and I can't take that risk. But I am still very convinced that you have to try things, that maybe nobody before you has ever dared to try, where you maybe don't have to listen to the self-proclaimed education-experts, yes? But you just say okay, I'm gonna do it like I think it could work and we'll see what happens. Uhm ... that's just ... that's just, I'm gonna say ... that's a basic philosophy

that I have personally. Uhm ... because education is something that is highly emotional because everyone has had something to do with an educational institution in their lives at some point, yes? Every human was once in kindergarten, primary school, occupational school, grade school, whatever, everybody has experienced some kind of education. Uhm ... and thus everybody has an opinion on it. Everyone can voice their opinion. Uhm... therefore it's very difficult to realize things in that domain because you're always gonna have a group of people who will say: No that doesn't work at all for them, that's out of the question uhm ... I myself am a big fan of the approach, especially in education to say: Let's do things the way nobody before us has ever done them. Or look at it from a point of view ... that is totally different. Let's see what happens. Uhm ... I don't know if that is where you were going with this question.. or the topic you wanted to discuss... but because you said... that makes sense to me. Century can advertise with that and Century can sell to the Belgian officials ... it's important to be able to say we have the neuroscientist, we have the education expert, we can provide the seal of approval ... that the kids actually learn something with this. Uhm ... from the point of view of sales, this obviously makes total sense ... Only it can theoretically be possible that a different algorithm, where the educational expert first says: Hmm, that will more likely make the students more stupid ... that could lead to the opposite. ... This is just something that me personally, I believe you often have to, for example ... uhm ... in education, we occupy ourselves a lot with the training of tutors. Because the heart of our platform are our tutors. We only accept... in the tutor application process, we only accept about 10% of candidates. That means 90% don't make it through our entrance examination. And we have to be this strict because we want to provide the best possible service for our clients. At the same time we see that, if you take a public educational system as a comparison, that they don't do a quality examination when hiring teachers. The only thing you need are completed studies, that you have the diploma for teaching, that gives you the right to teach at a school. What we see in our data is that our best tutor is for example the electrical engineering student, or the mechanical engineering student who tutors in Maths and Physics. It's not necessarily the person who studied Maths with a teaching degree.

V: Mhm.

F: That's what we can see in our data. That's where the students are the happiest, that's where there is the fastest improvement, that's where we have the highest engagement, the highest frequency, the parents are super happy, the teachers are

super happy... Uhm ... Why? That's where you would have to say as a next step: A country doesn't only take teachers who have studied Maths with a teaching degree, but you can also teach Maths if you just studied Maths or electrical engineering, or mechanical engineering.. but if you want to teach you have to go through an entrance examination. Finland are doing it this way, for example. And Finland has the lowest expenses which are a burden on families in education in the entirety of Europe. So... the, the, the, these are interesting points to start from, to say how the status quo is, if we continue like this.. it's obviously not the right way to go, if we do it differently, it works better. Like I said, education is always a super difficult topic because it is so highly emotional. Uhm .. yes.

V: Super. Let's see where we're heading there. [laughs]

F: Absolutely, absolutely.

Appendix C Files on the accompanying CD

- Thesis in PDF-format
- Zsfsg.txt: Abstract in German
- Abstract.txt: Abstract in English