Hochschule für angewandte Wissenschaften Würzburg-Schweinfurt
Fakultät Informatik und Wirtschaftsinformatik

Seminararbeit

Bias of Neural Networks - Security implications

David Mödl & Sebastian Lober

19. Juli 2020

Zusammenfassung

Künstliche neuronale Netze sind komplex. Wie genau die Algorithmen zu einem Ergebnis kommen, ist meist nicht mehr nachvollziehbar. Viele Entwickler beginnen sich ohne große Vorkenntnisse leichtfertig in diese Thema einzuarbeiten und kommen mit diversen KI-Bias und daraus resultierenden Sicherheitsrisiken in Berührung. Im wissenschaftlichen Bereich ist die Verbindung zwischen KI-Bias und Sicherheitsproblem nicht sehr präsent. Um diese Lücke zufüllen, setzte sich die vorliegende Seminararbeit zum Ziel über KI-Bias und dessen verbundenen Sicherheitsimplikationen aufzuklären.
Die Arbeit führt zunächst in den Grundlagen von neuronalen Netzen und Deep Learning ein. Weiter wurden die verschiedenen KI-Bias Entstehungen erklärt und dessen Sicherheitsimplikationen an konkreten Beispielen genannt. Ferner wurden die verschiedenen Präventionsmaßnahmen für die jeweiligen spezifischen Ursachen aufbereitet.
Das Grundverständnis, wie KI-Bias entstehen, kombiniert mit den Präventionsmaßnahmen, verhindern zukünftige Bias in Verbindung mit dieser Technologie und dessen daraus folgenden Sicherheitsimplikationen.

Inhaltsverzeichnis

1 Einführung

Künstliche Intelligenz(KI) oder auch artifizielle Intelligenz(AI) tritt allgegenwärtig in großen Teilen unserer Gesellschaft auf. Von Kaufvorschlägen auf Amazon, über Chat-Bots bis hin zu autonom fahrenden Autos spielt die KI eine immer größere Rolle. Ein bekanntes Beispiel ist die Software AlphaGo, die als erstes Computer Programm 2016 den weltbesten Profispieler im komplexen Brettspiel Go besiegen konnte [**alphaGo**]. Darüber hinaus ermöglicht die KI komplexe Sachverhalte zu simulieren und zu prognostizieren, wie zum Beispiel die vollautomatische Generierung hochaufgelöster, realistischer Videosequenzen auf der Grundlage simpler Eingaben [**videoToVideo**].

Einerseits gibt es viele Erfolge, die für ein KI betriebenes System sprechen. Andererseits bestärken medienwirksame Verfehlungen, wie z. B. das frauenfeindliche Bewerbungssystem von Amazon [**amazon**], die Skeptiker solcher Systeme. Ziel dieser Arbeit ist es, die unterschiedlichen Ursprünge solcher algorithmischen Verzerrungen (engl. *bias*) zu erläutern und Präventionen zu schildern, welche Bias vermeiden sollen.

Die Arbeit beginnt Grundlagen für spätere Kapitel aufzubauen. Danach wird auf die Entstehung solcher Bias eingegangen, die damit verbundenen Probleme und welche Präventionen gegen diese Fehlverhalten unternommen werden können.

2 Grundlagen

2.1 Bias

Wesentlicher Bestandteil der Arbeit ist das Erläutern der "Biases", welche durch die Nutzung von künstlicher Intelligenz auftreten können. Das Wort Bias kommt aus dem Englischen und bedeutet im Wesentlichen [**biasEffekt**]:

1. Verzerrung – im statistischen Sinn als mittlere systematische Abweichung zwischen dem erwarteten und dem tatsächlichen Modellergebnis.

2. Voreingenommenheit – je nachdem, wie wir die Welt aufgrund unserer Erfahrungen sehen, kommen wir zu unterschiedlichen Schlüssen.

Der Begriff Voreingenommenheit muss bei der Nutzung von KI vorsichtig behandelt werden, denn eine Maschine besitzt grundsätzlich keinerlei Vorurteile und weiß zu Beginn nicht was richtig oder falsch ist. Hier spricht man daher von einem Fehlverhalten oder einer Verzerrung, wie diese Entstehen wird im Kapitel 3 erläutert.

2.2 Künstliche Intelligenz

Künstliche Intelligenz(KI) oder auch artifizielle Intelligenz(AI) ist der Oberbegriff für ein Teilgebiet der Informatik. Dieses Gebiet befasst sich mit jeglicher Form von maschinelle intelligenten Verhalten und dem maschinellem Lernen, siehe Abbildung 2.1. Generell wird bei der künstlichen Intelligenz versucht biologische Intelligenz auf einen Computer zu simulieren. Die Simulation basiert meist auf simplen Algorithmen, wodurch die Begrifflichkeit 'Intelligenz' in Bezug auf Maschinen öfter in Frage gestellt wird [**Grundlagen**].

Damit ein Programm den Titel KI tragen darf, muss es zum einen die Fähigkeit zu lernen besitzen, zum anderen die Fähigkeit auch bei nicht eindeutigen Eingaben intelligent Lösungen zu finden.
KIs werden grob in zwei Kategorien aufgeteilt.

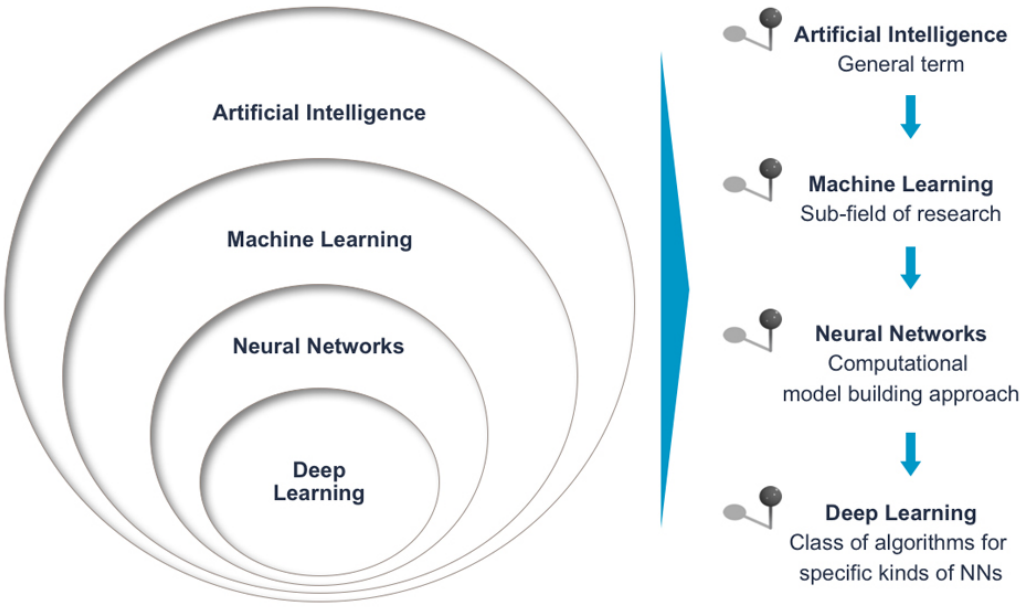

Abbildung 2.1: Verschiedene Abstraktionslevel von Artificial Intelligence in hierarchischer Ordnung

Die starke KI besitzt eine Intelligenz, die dem eines Menschens eben würdig ist bzw. sogar übersteigt. Ob eine starke KI überhaupt jemals erreicht wird, ist sehr umstritten. Schwache KI hingegen sind in unserem Alltag bereits vertreten. Schwache KI sind Algorithmen, die ganz spezielle Aufgaben lösen und die Arbeit von Menschen unterstützen.

2.2.1 Maschine Learning

Machine Learning (ML) ist der Oberbegriff jeglicher Lernformen von künstlicher Intelligenz. Im Allgemeinen versucht eine KI neue Muster und Gesetzmäßigkeiten in Trainingsdaten zu erkennen, diese zu verallgemeinern und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten zu verwenden [**EliminateHumanBias**].

Diese Arbeit speziell konzentriert sich auf Deep Learning, welches eine Variante zum Trainieren von neuronalen Netzen darstellt.

Lernformen

KIs bestehen aus vielen Algorithmen. Damit eine KI lernt, müssen die Algorithmen angepasst werden. Hierbei gibt es drei übergeordnete Lernformen: überwachtes, unüberwachtes

oder bestärkendes Lernen.

Beim überwachten Lernen sind Trainingsdaten bereits kategorisiert. Nach dem Versuch die Daten eigenständig zu kategorisieren, erhält das KI-System ein positives oder negatives Feedback. Auf Basis des Feedbacks lernt die KI, auf welche Muster und Merkmale es achten muss, um die Aufgabe richtig zu lösen [**Grundlagen**].
Die andere Variante ist das unüberwachte Lernen. Hierbei sind die Trainingsdaten nicht kategorisiert und das KI-System kann somit kein Feedback bekommen. Das neuronale Netz erstellt ein statisches Modell aus den Trainingsdaten und versucht Zusammenhänge und wiederkehrende Muster zwischen den Daten zu erkennen. Zusammenhängende Daten werden einer Kategorie zugeordnet. Die lernende KI weiß zu keiner Zeit, ob das was es gruppiert auch zusammengehört [**Grundlagen**].
Bestärkendes Lernen (auch Reinforcement Learning) ist eine Lernmethode bei dem der Algorithmus durch Belohnung und Bestrafung trainiert wird. Hierbei erhält das KI-System Feedback in Form von Belohnung bei Erfolg und Bestrafung, falls das bestimmte Verhalten nicht zielführend war. Mit dieser Lernmethode wurde Google AlphaGo trainiert [**Grundlagen**].

2.2.2 Neuronale Netze

Künstliche neuronale Netze(KNN) bestehen aus künstlichen Neuronen, die untereinander verflechtet sind. Die Konstrukte sind denen der sich im Nervensystem eines Lebewesens befindenden Neuronenverbindungen nachempfunden.
KNNs sind nicht dazu da das Nervensystem von Lebewesen nachzubilden, sondern abstrakt die Eigenschaften der Informationsverarbeitung und der Lernfähigkeit zu imitieren.

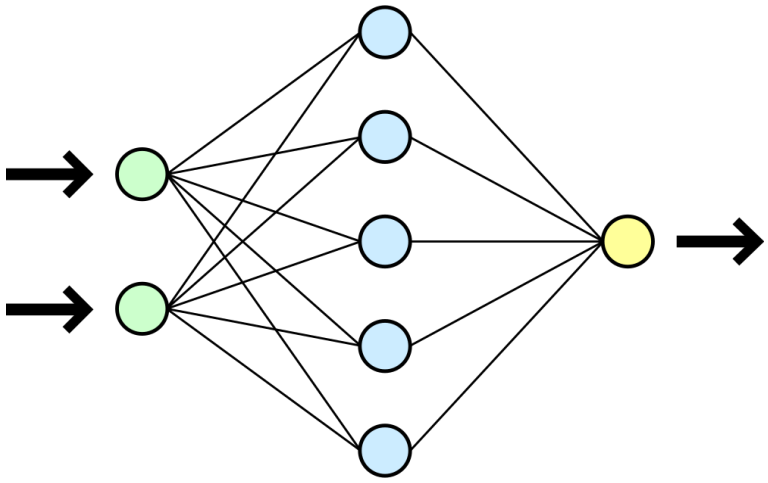

Abbildung 2.2: Vereinfachte Darstellung eines künstlichen neuronalen Netzes

KNNs sind meist in Schichten mit beliebig vielen künstlichen Neuronen aufgebaut. In der Regel besteht ein KNN aus drei Teilen der Eingangsschicht (grün), der verdeckte Schicht (blau) im Englischen Hidden Layer und der Ausgabeschicht (gelb). In der Eingangsschicht fließen die Informationen in das Netz ein und in der Ausgabeschicht das Ergebnis der Berechnungen aus. Jede Schicht besteht aus beliebig vielen Neuronen je nach Komplexität des Zieles, die Hidden Layer sogar aus beliebig viele Schichten.

In der Regel arbeiten KNNs nach dem feedforward-Prinzip, bei dem die Informationen nur in eine Richtung fließen. Es gibt jedoch auch rekurrente Netze, bei denen durch rückgerichtete Kanten Rückkopplungen im Netz entstehen.
Die einfachste Netzstruktur ist das einschichtige feedforward-Netz. Dies besteht ohne Rückkopplungen aus nur einer Schicht, der Ausgabeschicht.

Das künstliche Neuron

Künstliche Neuronen sind die Grundbestandteile eines künstlichen neuronalen Netzes. Ein künstliches Neuron ist die vereinfachte, abstrakte Version einer biologischen Nervenzelle und ist wie folgt aufgebaut.

Ein künstliches Neuron besitzt n Eingangskanäle und einen Ausgangskanal. j repräsentiert

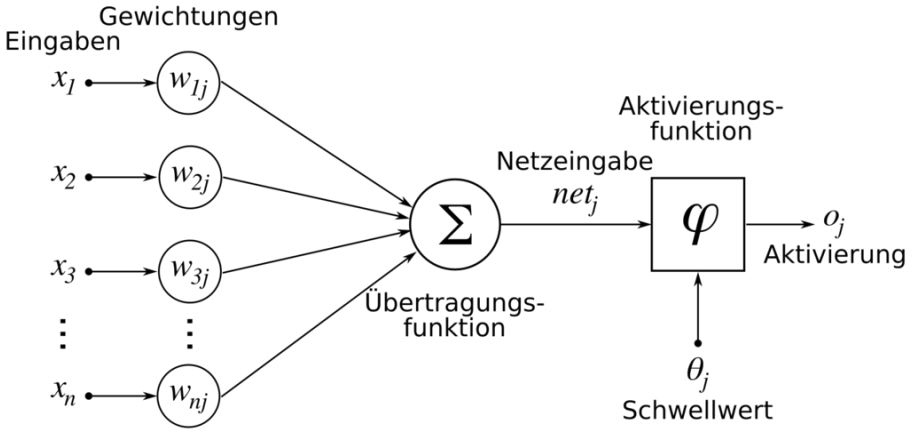

Abbildung 2.3: Ein künstliches Neuron

hier die eindeutige Nummer des Neuron. Jede Eingabe x_i besitzt ein dazugehörendes Gewicht $w_{1j}..w_{nj}$. Das Gewicht spiegelt die Wichtigkeit der Eingabe wider, welche hemmend (negativer Wert) oder erregend (positiver Wert) wirken kann. Die Übertragungsfunktion Σ summiert alle multiplizierten Eingaben mit ihrem Gewicht und geht als Netzeingabe net_j in die Aktivierungsfunktion φ ein.

Ob das Neuron "feuert" oder kein Signal sendet, wird hier berechnet. Auf die Netzeingabe net_j wird ein Schwellenwert θ_j addiert. Als mathematische Vereinfachung wird der Schwellwert θ als w_0 bezeichnet und $x_0 = 1$ eingeführt und somit in der folgenden Formel immer auf die Netzeingabe net_j addiert.

$$a = \sum_{i=0}^{n} x_i w_i$$

Die Variabel a geht in die Aktivierungsfunktion $\sigma(a)$ ein, anhängig von dem Ergebnis wird das Neuron aktiv oder bleibt inaktiv. Das Ausgangssignal eines Neurons ist gleichzeitig ein Eingangssignal eines oder mehrerer anderer Neuronen.

Dadurch dass ein Neuron mehrere Eingangskanäle besitzt, werden viele Eingangsinformationen auf ein Ergebnis reduziert. Durch mehrere Schichten und vielen Neuronen pro Schicht kann so eine große Menge an Daten schnell reduziert werden.

Jedoch muss jedes künstliche Neuron eines KNNs richtig eingestellt werden, damit das KNN dessen Ziel erfüllt. Eingestellt wird das Neuron durch Training. Trainieren bedeutet hier das Ermitteln der richtigen Werte für Gewichtungen und Schwellwerte, als auch das Einstellen der richtige Verbindungskombinationen der Neuronen untereinander. Den Prozess nennt man Deep Learning, eine Form von Machine Learning.

2.2.3 Deep Learning

Deep Learning ist eine Maschine Learning Variante, die speziell bei künstliche neuronale Netze eingesetzt wird. Beim Deep Learning werden die zahlreichen Zwischenschichten der Hidden Layer trainiert. Dabei wird eine umfangreiche und komplexe Struktur der Neuronen-Verbindungen aufgebaut. Wie das Programm endgültig die Aufgabe lösen soll, wird hierbei nicht vorgegeben, sondern bei diesem autonomen Prozess evolutionär ermittelt [**Grundlagen**].

Ein künstliches neuronales Netz wird mit dem Zweck aufgebaut, eine bestimmte Aufgabe zu lösen. Extra dafür müssen Daten aufbereitet werden. Aus den aufbereiteten Daten werden Muster und Gesetzmäßigkeiten extrahiert, um Erfahrungen zu sammeln. Aus Basis der gewonnenen Erfahrungen werden später Entscheidungen auf unbekannte Daten angewandt [**Grundlagen**].

Damit es ausgeschlossen ist, dass das Netz zufällige Zusammenhänge zwischen den Daten findet, wird eine hinreichend große Datenmenge, abhängig von der Komplexität, benötigt. Auch die Qualität der Daten ist entscheidend, da daraus der KI-System die fachlich Korrekten Schlüsse zieht, siehe Abschnitt 2.3.

Neuronale Netze sind für Menschen ab einer gewissen Größe nicht mehr nachvollziehbar. Somit können Eingaben und Ausgaben nur mit Hilfe von Testdaten verglichen werden, um so auf die Korrektheit des Ergebnis zu prüfen. Ziel ist es mit möglichst vollständigen Trainingsdaten das Netzwerk so einzustellen, dass unbekannt Daten korrekt interpretiert werden [**Grundlagen**].

Um eine Aufgabe zu trainieren, wird nicht nur ein Netz mit Zufallswerten und Verbindungen generiert sondern tausende. Beim überwachten Lernen werden alle mit den gleichen Trainingsdaten trainiert und mit den Testdaten geprüft. Für jedes Netz wird ein Mittelwert über die Korrektheit der Antworten erstellt. Da alle Netze initial mit Zufallswerten belegt sind, haben die meisten Netze eine Erfolgsrate von ca. 50 Prozent. Die Netze mit den höchsten Erfolgsraten, mit beispielsweise mehr als 60 Prozent, werden behalten, der Rest wird verworfen. Die erfolgreichsten Netze werden mehrfach kopiert und bei jedem Kopiervorgang individuell leicht verändert und erneut getestet. Die Besten werden wiederum genommen, leicht modifiziert, kopiert und schlechteren verworfen.

Nach einer gewissen Anzahl an Iterationen entscheidet das Netz nicht mehr willkürlich, sondern scheint intelligent die Aufgabe zu lösen. Dieser Iterationsschritt kann unendlich oft laufen, jedoch empfiehl es sich, je nach Anwendungsfall, ab einer gewissen Erfolgsrate das Training zu beenden oder neue Trainings- und Testdaten zu verwenden. Als Ergebnis des Prozesses erhält man durch Deep Learning ein trainiertes künstliches neuronales Netz, das im Allgemeinen als KI bezeichnet wird.

2.2.4 Loss-Funktion

Alle KKN benötigen eine Loss-Funktion(Verlustfunktion), um später trainiert zu werden. Die Verlustfunktion bestimmt die Differenz zwischen der Prognose, die das Modell liefert, und dem vorgegebenen Label. Sie muss für die Menge aller Daten berechnet werden und beschreibt damit, wie gut das Modell die Trainingsdaten abbildet.

2.3 Daten

Erst durch eine Kombination aus Algorithmen und Daten wird die Entscheidungsfindung ermittelt. Wie menschliche Entscheider liefern Algorithmen aufgrund von unvollständigen oder fehlerhaften Daten falsche Prognosen. Eine gute Datenqualität und -quantität sind ebenso wie eine richtige Unterteilung zwischen Test- und Trainingsdaten unabdingbar.

2.3.1 Datenaufbereitung

Der Erfolg einer KI hängt von den Daten ab. Zuerst müssen Daten für den Zweck der KI Anwendung gesammelt und anschließend aufbereitet werden. Dabei müssen die Eigenschaften(z. B. eines Bildes) erfasst werden. Zusätzlich benötigt der Datensatz zur Kategorisierung beim überwachten Lernen, ein Label.

Um nun ein KNN zu trainieren, müssen die Daten in Trainings- und Testdaten unterteilt werden. Mit den Trainingsdaten wird die Maschine trainiert. Die Testdaten hingegen werden nicht für das Training benutzt, sondern dienen dazu, das fertige System gegen unbekannte Daten zu testen. Meist werden 10-20% der Daten als Testdaten reserviert.

2.3.2 Quantität

Je mehr unterschiedliche Eigenschaften in den Datensätzen existieren, umso komplexer wird das Modell. Die Komplexität der Probleme erfordert, dass die Menge an Daten entsprechend groß sein muss, damit das zu trainierende System alle Algorithmen korrekt einstellt.

Ein Beispiel hierfür findet sich in der Autoindustrie. Beim autonomen Fahren müssen Daten von Laser-, Kamera- und Radarsensoren im Auto zuverlässig und schnell verarbeitet werden. Die KI im Fahrzeug muss jederzeit über ein präzises Abbild der realen Verkehrsbedingungen verfügen, um darauf basierend in jeder Fahrsituation die richtige Entscheidung zu treffen [**autonomesFahren**].

Je nach Komplexität der Anwendungen werden mehr oder weniger Datensätze benötigt. Generell kann man sagen, dass ein KNN mit zu wenigen Trainingsdaten schwer Muster in Daten erkennt und dadurch fehleranfälliger wird.

2.3.3 Qualität

Die Qualität der Daten ist für maschinelles Lernen essentiell.
Eine hohe Datenqualität wird erreicht, wenn folgende Punkte beachtet werden [**goodDataQuality**]:

1. Widerspruchsfrei
2. Generalisiert
3. Fachlich Korrekt

4. Interpretierbar

5. Vollständig

Um die einzelnen Punkte umzusetzen, wird ein Zugang zu den Daten benötigt, damit diese kontrolliert und angepasst werden. Sind Daten nicht vollständig und fehlen bestimmte Anwendungsfälle, kann die KI die speziellen Anwendungsfälle in der Praxis nicht korrekt beantworten, siehe Abschnitt 3.1.1.

3 Bias Entstehung

Bei der Nutzung von KI Systemen können Verzerrungen(Bias) bzw. Fehlverhalten entstehen, diese sind unterschiedlicher Natur und treten an unterschiedlichen Stellen der Machine Learning Pipeline (Abbildung 3.1) auf. Die Arbeit geht auf die Daten ein, welche bei der Eingabe zu Bias führen. Anschließend werden die menschlichen Fehler erläutert, die in diesen Verarbeitungsschritten auftreten.

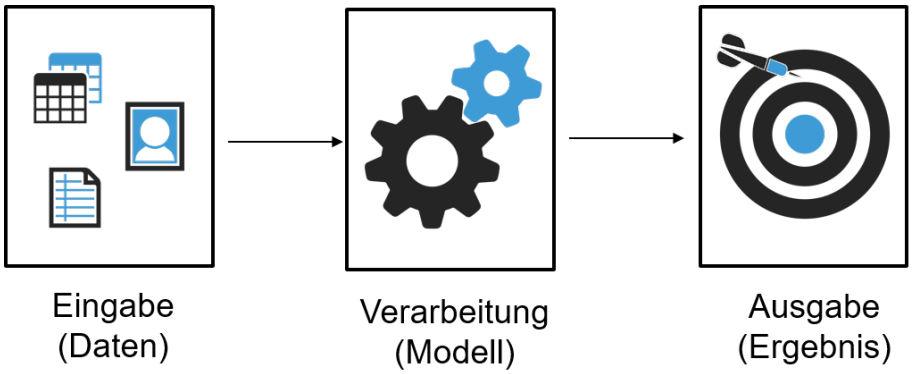

Abbildung 3.1: Machine Learning Pipeline: Eingabe, Verarbeitung, Ausgabe

3.1 Daten

3.1.1 Unvollständigkeit der Daten

Unvollständige Daten sind Trainingsdatensätze, die z. B. nur einen Teil der Bevölkerung umfasst, während ein anderer Teil unterrepräsentiert ist. Diesen Bias nennt man Stichproben-Voreingenommenheit.

Ein Beispiel dafür lieferte Apple 2017 mit der damals neuen FaceID des IPhone X. Diese Gesichtserkennungssoftware konnte zu Beginn durch Zwillinge, 3D Masken oder auch eng verwandte Personen ausgetrickst werden. Das größte Problem hierbei war allerdings, dass die Software mit Trainingsdaten trainiert wurde, bei der eine komplette

ethische Gruppe unterrepräsentiert war. So konnten nicht miteinander verwandte Chinesen das selbe Handy entsperren und machten somit die Nutzung der FaceID unsicher [**iphone**].

3.1.2 Garbage in - Garbage out

Eine Maschine kennt grundsätzlich keinen Unterschied zwischen Schwarz und Weiß, Mann und Frau oder Jung und Alt. Erst durch das Trainieren einer KI lernt eine Maschine Verhalten und Muster zu differenzieren. Hierfür werden qualitativ hochwertige Daten (siehe 2.3.3) benötigt, um das Netz richtig einzustellen.

Bleiben fehlerhafte Daten unentdeckt, wird ein System falsche trainiert. In Testreihen für autonomes Fahren in der Autoindustrie kennzeichneten Probanden auf Grund von Unachtsamkeit, Menschen am Straßenrand als Tonnen ab. Das System wurde mit solcher Fehlkennzeichnungen trainiert und wertete in einer kritischen Verkehrssituation das Überfahren einer vermeintlichen Tonne als verhältnismäßige Alternative ein, die möglichst wenig Schaden anrichtet [**trainingsDataKI**].

Bei der traditionellen Datenanalyse werden solche schlechten Daten nachträglich entfernt. Hat allerdings eine Maschine durch maschinelles Lernen etwas gelernt, wird es schwer dieses Erlernte wieder zu verlernen. Denn ähnlich wie beim menschlichen Gehirn, wird es ab einem Gewissen Grad nahezu unmöglich, herauszufinden, auf welchen Datenelementen die Vorhersagen basieren.
Baut unser erlerntes Wissen in Teilen auf falschen Informationsbausteinen auf, verliert der ganze Komplex seinen Wert und man muss alles von vorne beginnen.

Dieses Problem wird in der KI als "Garbage in - Garbage out"(Müll rein, Müll raus) bezeichnet.

3.1.3 Historische Verzerrung

Eine historische Verzerrung liegt vor, wenn ein Algorithmus anhand eines alten Datensatzes trainiert wird, der vergangene Werte und Moralvorstellungen etwa die Rolle der Frau in der Vergangenheit aufgreift [**kipedia**].

Ein Beispiel liefert Amazons Algorithmus aus dem Jahr 2014, der unter mehreren Bewerbungstexten automatisch die besten Bewerber herausfiltern sollte. Dabei bezog sich die Software auf voran gegangene Bewerbungen und lernte daraus welche Eigenschaften Amazon bevorzugte. Weil das Unternehmen aber Teil einer von Männern dominierten Industrie ist, sind in den zugrunde gelegten vergangenen zehn Jahren vor allem Männer

eingestellt worden. Daraus resultierte, dass Frauen grundsätzlich schlechter bewertet wurden. Selbst ohne Angabe eines Geschlechtes wurden Frauen durch Frauenvereine identifiziert und kategorisch ausgeschlossen. Das Beispiel veranschaulicht den Bias der historischen Verzerrung, welches beim maschinellen Lernen in den Daten auftreten kann [**amazon**].

Die Hoffnung solcher Anwendungen liegt eigentlich darin, Vorurteile zu vermeiden und Prozesse fairer zu gestalten, da eine Maschine wie in 3.1.2 bereits genannt, grundsätzlich keine Unterschiede kennt.

An dem Beispiel wird die zentrale Rolle der Daten für eine KI deutlich. Meist ist es nicht möglich Daten zu finden, welche nicht bereits menschliche Bias beinhalten. Solch verzerrte Trainingsdaten werden, im Bezug auf ihre Zusammensetzung, auch als WEIRD Samples (western, educated, industrialized, rich and democratic societies) bezeichnet [**BiasInKi**].

3.1.4 Under-/ Overfitting

Ein komplexer Datensatz (siehe Absatz 2.3.2) kann beim überwachten Lernen dazu führen nutzloses Wissen aufzubauen oder aus einem vorhandenen Trainingsdatensatz keine relevanten Lerninformationen zu ziehen. Wird ein Netz zu gut oder zu schlecht trainiert, tritt das Phänomene Overfitting (Überanpassung) beziehungsweise Underfitting(Unteranpassung) auf.

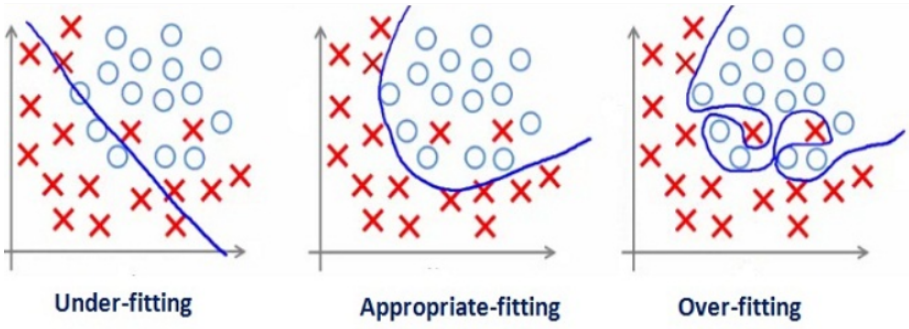

Abbildung 3.2: Over- und Underfitting

Die Abbildung 3.2 zeigt drei Graphen, die das Problem erläutern. Jeder Graph visualisiert abstrakt, wie unterschiedlich das KI-System anhand der Trainingsdaten, die einzelnen Punkte in zwei Cluster unterteilt. Die ”x” und ”o” stellen die zugewiesenen Kategorien durch den Menschen dar.

Die KI im linken Graphen ist unter angepasst, da das Modell mit den Trainingsdaten zu schlecht abgestimmt wurde. Hier wurden zu wenig Punkte richtig abgedeckt, da die Funktion zu schwach ist um eine Generalisierung zu erreichen. Solche KI-Modelle neigen zu hohen Verzerrungen der Ergebnisse und teilt die einzelnen Trainingsdaten nur oberflächlich ein.

Der rechte Graph hingegen kategorisiert alle Punkte richtig ein. Unter dieser Annahme könnte man denken, der rechte Graph wäre der beste Graph. Hier spricht man allerdings von einer Überanpassung, da die Trainingsdaten zu stark auswendig gelernt und zu wenig generalisiert wurden. Verantwortlich dafür ist die Integration der Outliner(Ausreißer), die starke Abweichungen der eigenen Kategorie beinhalten. Solche Modelle besitzen eine hohe Varianz der Vorhersagen [**overUnderfitting**].

Beispielsweise wird eine KI mit Deep Learning darauf trainiert, einen Schrank auf einem Bild zu erkennen. Ist das KI-System über angepasst, beinhaltet es zu viele Eigenschaften der Ausreißer beispielsweise Sessel ähnliche Schränke. Das KI-Programm erkennt alle Trainings- und Testdaten perfekt, sobald man jedoch anfängt mit realen Daten zu testen, ist das Programm bei überschneidenden Eingaben sehr fehleranfällig. Outliner ähnliche Daten, wie in diesem Beispiel ein normaler Sessel, werden vom Programm als Schränke fehlinterpretiert, da das Modell zuvor über angepasst wurde.

Am besten ist somit der mittlere Graph, welcher Outliner ignoriert und eine gute Balance zwischen Verzerrung und Abweichungen besitzt.

3.1.5 Aufteilung der Daten

Bei der Aufteilung von Trainings- und Testdaten können explizite Fälle nur in den Trainingsdaten auftreten. Zwar werden die Fälle antrainiert, dadurch dass sie jedoch in den Testdaten fehlen, nicht validiert. Das Netz liefert sehr gute Testergebnisse, jedoch nur da speziellen Fälle nicht getestet wurden. Der Bias fällt meist erst in der Praxis auf.

3.1.6 Ähnlichkeit der Daten

Optische Täuschungen sind Bilder, die Menschen in die Irre führen. Anhand des Beispiels aus der Abbildung 3.3 wird deutlich, dass wenn die Farbe, Struktur und die Form identisch sind, die Wahrnehmung trüben kann. Ein Bagel wird auf den ersten Blick zum Hund oder andersrum. Das gilt auch für eine KI, denn das Programm lernt anhand

Abbildung 3.3: Hund oder Bagel?

Muster, Strukturen und sonstige besondere Eigenschaften zu kategorisieren. Hier kann es auch ohne falsche Eingaben zu einer Verzerrung der Ergebnisse führen.

3.2 Menschliche Fehler

Den Faktor Mensch darf man bei der Bias Entstehung nicht vergessen. Konzeptionelle Fehler, mangels Wissen oder aus einem Missgeschick heraus, fördern Fehlverhalten. Markante Defizite können eine funktionale KI Entstehung komplett verhindern, sind jedoch in der Regel noch die besseren Missstände. Denn kleinere Mängel, die nicht sofort auffallen, können in Produktion fatale Folgen haben.

3.2.1 Falsche Zielsetzung

Die Zielsetzung eines KNNs ist ein nicht zu unterschätzender Teil bei der KI Entwicklung. Setzt man hier den falschen Grundstein, entwickeln sich vermeidbare Fehlverhalten einer KI.

Subjektive Ziel Definition

Eine Zielvariable spiegelt eine Antwort einer Frage wider, die eine KI lösen soll. Wenn die Aufgabe lautet, “finde den besten Mitarbeiter“, muss die Zielsetzung in messbare Werte übersetzt werden. Bei der Definition kann bewusst oder unbewusst Bias eingepflegt werden. Anhand welcher Werte der ’Beste’ Mitarbeiter erkannt wird, liegt allein an der subjektiven Meinung des Entwicklers. Ist die Meinung ’männlich, jung, intelligent’ würde die KI Frauen diskriminieren [**KIPediaBias**].

Zu viele Ziele

Viele Köche verderben den Brei. Diese Weisheit kann auch auf die Ziele von KNNs umgemünzt werden. Denn je mehr Ziele ein KNN hat, desto größer und komplexer muss ein künstliches neuronales Netz sein, um alle Fälle zu abdecken. Je komplexer ein KNN ist, desto größer ist die Wahrscheinlichkeit, dass Fehler passieren. Auch ist sich das Netz deutlich unsicherer bei seinen Entscheidungen.

Wenn man eine nahezu perfekte KI mit einer Aufgabe erweitern möchte, hat das meist zu Folge, dass die KI nach der Erweiterung zwar mehr kann, jedoch seine ehemalige Hauptaufgabe nicht mehr so gut meistert wie zuvor.

3.2.2 Bewusste Manipulation der Trainingsdaten

Vorsätzliches Manipulieren von Trainingsdaten provoziert bewusst Fehlverhalten. Bei geschickter Manipulation ist der Betrug in den Datensätzen nicht erkennbar. Auch die KI funktioniert erwartungsgemäß. Nur bei bestimmten Eingaben tritt ein unerwünschtes Verhalten auf [**KIPediaBias**].

Vor allem bei öffentlichen Datensätzen kann man nicht hundertprozentig sicherstellen, dass keine Manipulation an den Daten vorgenommen wurde .

4 Sicherheitsprobleme durch BIAS

4.1 Gefahren für Maschinen

Computer, die von Computern gesteuert werden, sind heutzutage keine Science Fiction mehr. In der Industrie 4.0 nimmt KI eine immer größere Rolle ein. Somit können Fehlverhalten nicht nur Menschen sondern auch Maschinen schaden.

Kühlung von Rechenzentren

Seit 2016 setzt Google eine KI zur Optimierung des Kühlungsprozesses ein. Anfangs lieferte die KI nur Konzepte zur Kühlungsoptimierung, welche Mitarbeiter auswerten und einsetzen konnten. Heutzutage nimmt die KI Änderungen autonom an der Kühlung vor. Bis zu 40 Prozent an Energie konnte somit eingespart werden, was sich einerseits aus ökonomischer, als auch aus finanzieller Sicht lohnt [**GoogleKI**].
Unterläuft der KI hier ein Fehler, hat das direkte Folgen auf die Langlebigkeit der Computerelemente des Rechenzentrums und kann sogar zu Brandherden in Gebäuden führen.

4.2 Gefahren für Menschen

In Zukunft werden KIs immer mehr zu unseren Alltag gehören und greifen immer stärker in unsere Leben ein. Dies hat viele Vorteile, jedoch ebenso Nachteile, insbesondere für ärmere Teile der Bevölkerung sowie Minderheiten.

4.2.1 Diskriminierung von bestimmten Personengruppen oder Minderheiten

Durch die #BlackLivesMatters Bewegung 2020 ist die Forderung nach Gleichberechtigung aktuell wieder omnipräsent. KI könnte hier Abhilfe schaffen, da es sich um Programme handelt die keine Vorurteile besitzen. Leider ist die Theorie nur eine Halbwahrheit. Natürlich sind Algorithmen von sich aus nicht rassistisch motiviert, jedoch spiegeln sie diskriminierendes Verhalten aus Daten oder sonstigen Fehlern wider.

Alltagsrassismus

Amazon entwickelte 2014 ein Bewerbungsauswahlsystem, wie bereits in 3.1.3 beschrieben. Dort hat die KI empfohlen nur Männer einzustellen.

In Finnland wurde Svea Ekonomi ein Onlinekreditvergabe Unternehmen auf 100.000 Euro Strafe verurteilt, aufgrund eines Falles von direkter Mehrfachdiskriminierung. Der Antragssteller wurde ein Verlängerung des Kredits verweigerten ohne Berücksichtigung des Kreditverhaltens und Kreditwürdigkeit.
Allein basierend auf statistische Zusammenhänge von Faktoren wie Geschlecht, Muttersprache, Alter und Wohnort wurde ein Score berechnet, auf dem die Kreditwürdigkeit beruht.
Männer wurden schlechter bewertet als Frauen, auch wurden schwedische Muttersprachler besser bewertet als finnische. Da der finnisch sprechende Antragssteller aus einem für das System unbekannten Gebiet kam, welches ebenso eine Benachteiligung darstellt, wurde der Kredit verweigert. Statistisch gesehen, ist es korrekt, dass junge finnische Männer, die aus einer schlechteren Gegend kommen, häufiger Rückzahlungsprobleme haben. Jedoch diese Daten zu verallgemeinern und ohne auf individuelle Berücksichtigung des Kreditverhaltens und Kreditwürdigkeit der einzelnen Person einzugehen, ist schlicht weg diskriminierend. Wäre nur ein Faktor, arithmetisch gesehen, günstiger für den männlichen Antragssteller gelegen, hätte er den Kredit bekommen [**DiskriminierungKredit**].

Risikobeurteilung

In Amerika wird eine KI namens COMPAS eingesetzt, um Risikobeurteilung von Angeklagten und Häftlingen zu erstellen. Benachteiligt werden eindeutig Minderheiten. Nachweislich lag die Software doppelt so oft mit den Prognosen zukünftiger Verbrechen von Afroamerikanern falsch im Vergleich zu weißen Angeklagten. Durch schlechte Sozialprognosen haben afroamerikanische Angeklagte eine geringe Chance auf frühzeitige

Entlassung oder auf Bewährung. Bedeutet, afroamerikanische Straftäter müssen durchschnittlich länger Inhaftieren als Weiße, dank KIs.

4.2.2 Lebensgefahr

Tesla und Volvo prognostizierten 2015, dass 2017 vollständig autonome Fahrzeuge marktreif wären. Das Fraunhofer-Institut prophezeit 2025 als frühesten Zeitpunkt, während die Deutsche Akademie der Technikwissenschaften 2030 nennt [**AutonomAuto**].
Trotzdem schickt Tesla seine Kunden mit einer unausgereiften KI autonom auf die Straßen. Allein mit dem Zweck Daten zu sammeln. Dass in Verbindung mit autonomen Fahren immer wieder Unfälle bis hin zu Todesfällen gemeldet werden, zeigt, dass künstliche Intelligenz unmittelbare Konsequenzen für den Menschen mit sich bringt.

Ein Beispiel für die Unausgereiftheit ist das Verkehrsschild aus Abbildung 4.1. Die Ge-

Abbildung 4.1: 35 oder 85 mph?

schwindigkeitsbegrenzung wurde bewusst manipuliert und indem das mittele Teil der Drei verlängert wurde. Teslas KI erkannte die Drei als Acht und beschleunigte automatisch auf mehr als 80 km/h. Solche bewusste Manipulationen nennt man Adversarial Attacks [**TeslaHack**].

4.3 Angriff auf KI

Ist eine Künstliche Intelligenz sehr gut trainiert, erzielt sie in ihrem Aufgabebereich überragende Ergebnisse. Doch selbst nahezu perfekte KIs sind nicht unfehlbar. Durch methodisch gestaltete Störungen der Eingaben kann eine KI bewusst getäuscht werden und damit die Computer-Wahrnehmung in die Irre führen. Diese Art der Ausnutzung von KI Schwächen nennt man Adversarial Attacks, zu deutsch gegensätzlicher Angriff.

Adversarial Attacks

Bildererkennungssoftware ist besonders anfällig für solche Angriffe. Hierbei legt man über das Bild bestimmte Pixelmuster, die für das menschliche Auge in der Regel im Gesamtbild untergehen. KIs hingegen registrieren jegliches Muster und nehmen sie in ihre Berechnungen mit auf. Als Ergebnis werden Bilder zum Teil massiv fehlinterpretiert. Im folgenden Angriff wird mit einer Brille und deren speziellen Muster, eine KI überlistet.

Abbildung 4.2: Reese Witherspoon wird mit Adversarial Attack Brille zu Russel Crowe

Im linken Bild ist die amerikanische Schauspielerin Reese Witherspoon zu sehen. Das linke Bild wird auch korrekt als sie selbst von der KI erkannt. Setzt man ihr eine spezielle Brille auf, beginnt das neuronale Netz Fehler zu begehen. Plötzlich erkennt die KI Reese Witherspoon mit der Brille als Russel Crowe (Bild rechts) [**Attack**].

5 Prävention

5.1 Passender Algorithmus zu Daten

Wie im Abschnitt 3.1.2 ist es ab einem gewissen Grad bei einem bestehenden Modell schwer, Fehlverhalten zu erkennen und zu eliminieren.

In der Annahme, ein KI-System wurde richtig trainiert, aber die zeitliche Komponente nicht beachtet, kann es zu einer historischen Verzerrung 3.1.3 kommen. Wie Stöcker 2019 [**stoecker**] bereits erwähnte, benötigt eine sich ändernde Gesellschaft auch kontinuierliche weiterlernende Algorithmen. Solch ein Lernprozess wird auch als *adaptive learning* bezeichnet, womit dem Effekt sich ändernder Datenumgebungen(concept drift) Rechnung getragen wird [**gama**].

Auch bei einem für Diskriminierung anfälligem Umfeld muss besondere Sorgfalt auf die Wahl des passenden Algorithmus gelegt werden. In der Regel reicht es einzelne Attribute aus den Eingangsdaten auszuschließen, um eine algorithmische Diskriminierung bestimmter Personengruppen zu verhindern. Das Ausschließen einzelner Attribute genügt jedoch oftmals nicht, denn wie das Beispiel von Amazon aus 3.1.3 zeigt, kann der Algorithmus Rückschlüsse anhand anderer Attribute ziehen(z. B. durch Frauenvereine).

In diesem Kontext sollten KI-System bevorzugt werden, welche in der Entwicklung zusätzliche Einschränkungen einführen, um eine Diskriminierung aufgrund bestimmten Attributen zu verhindern [**kamiran**].

5.2 Nur ein Ziel

Generell ist es zu empfehlen mehrere spezialisierte KIs für einzelne Aufgaben zu verwenden, anstatt einer KI immer mehr Aufgaben aufzulasten. Je konkreter und einfacher die Aufgabe und dessen Zielsetzung gehalten wird, desto fehlerfreier kann eine KI die Aufgabe lösen.

Setzt man sich zu hohe Ziele, kann das komplexe System leicht unter- oder übertrainiert

werden, siehe Abschnitt 3.1.4. Man braucht dementsprechend exorbitant viele Datensätze zum Trainieren des Netzes. Hier kann es jedoch passieren das leicht historische Verzerrungen 3.1.3 oder Manipulationen 3.2.2 übersehen werden.

5.3 Verfahren zum Validieren

Damit menschliche Fehler (Abschnitt 3.2) oder fehlerhafte Daten (Abschnitt 2.3) frühzeitig erkannt werden und nicht zu Bias führen, sollten diese zuvor validiert werden. Folgende Verfahren wirken menschlichen Fehlern entgegen [**trainingsDataKI**]:

1. Lektorat oder Peer-Review

2. Das Vier-Augen-Prinzip

3. Mehrheitsentscheide bei unterschiedlichen Ergebnissen

Die oben genannten Verfahren bekämpfen Flüchtigkeitsfehler und bewusste Manipulation. Beim Peer-Review kontrollieren nachträglich Experten das KI-System auf dessen Korrektheit. Beim Vier-Augen-Prinzip und bei Mehrheitsentscheiden passiert die Validierung während des Arbeitsprozesses.

Zudem sollte das KI-System ausreichend auf Fehleranfälligkeit geprüft werden. Vor allem in kritischen Bereichen, wie der Medizindiagnostik, sollte die Fehleranfälligkeit gegen null gehen. Anhand des Problems aus Abschnitt 3.1.5, können ungeprüfte Fälle fehlerhaft sein.

Eine Möglichkeit um das zu verhindern, bietet die Kreuzvalidierung.

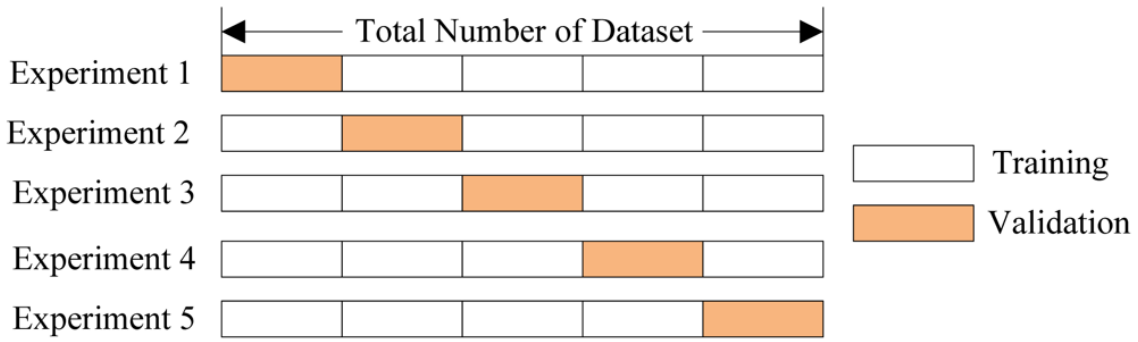

Abbildung 5.1: fünf fache Kreuzvalidierung

Anstelle eines fixen Testdatensatzes wird der komplette Datensatz in viele kleine Blöcke

unterteilt, in Abbildung 5.1 Fünf. In diesem Beispiel werden fünf Durchläufe ausgeführt, in denen jeweils abwechselnd mit vier unterschiedlichen Datensätzen trainiert und mit einem validiert wird. Bei jedem Durchlauf entsteht ein Score über die Fehleranfälligkeit. Am Ende wird ein durchschnittlicher Score über diese fünf Durchläufe gebildet, welcher die gesamte Fehleranfälligkeit für alle Datensätze wiedergibt [**towardData**].

Mit Hilfe der Kreuzvalidierung lässt sich eine Aussage treffen, wie gut das Modell mit unbekannten Daten trainiert. Das Prinzip nennt man Generalisierung.

5.4 Over- und Underfitting

Die Kunst der Modellierung besteht darin, ein optimales Modell(siehe mittleren Graphen 3.2) zu finden, dass eine perfekte Balance zwischen Verzerrung und Abweichung der Daten besitzt. Das Modell sollte eine möglichst geringe Komplexität aufweisen und gut beschriebene, generalisierte Daten beinhalten. Da das nicht immer einfach zu realisieren ist, treten Underfitting und Overfitting auf, siehe Abschnitt 3.1.4.

5.4.1 Underfitting

Underfitting ist die Folge eines zu einfachen Modells und liefert somit schlechte Vorhersagen. Eine Möglichkeit um Unterangepasstheit zu verhindern ist das Erhöhen der Flexibilität. Das heißt die Anzahl an Merkmalen die einen Datensatz auszeichnen und die Menge der Trainingsdaten müssen erhöht werden. Eine weitere Möglichkeit bietet das Erhöhen der Durchläufe beim Trainieren der Trainingsdaten [**amazonOverUnderfitting**].

5.4.2 Overfitting

Overfitting stellt in der KI eine große Herausforderung dar und ist die Folge aus einem zu komplexen Modell. Eine Möglichkeit um eine Überanpassung zu vermeiden, bietet eine Resampling Methode, wie z. B. die Kreuzvalidierung, siehe Abschnitt 5.3. Eine andere Methode ist die Vereinfachung des Ki-Systems. So kann mit Hilfe von Pruning die Größe eines Baumes reguliert werden [**overUnderfittingNovu**]. Bei Regressionsmodellen hingegen können Regulierungstechniken angewandt werden.

Bei der Regulierung wird ein Strafterm eingeführt, der die Anzahl an Variablen und deren Wechselwirkungen berücksichtigt(siehe Abbildung 5.2).

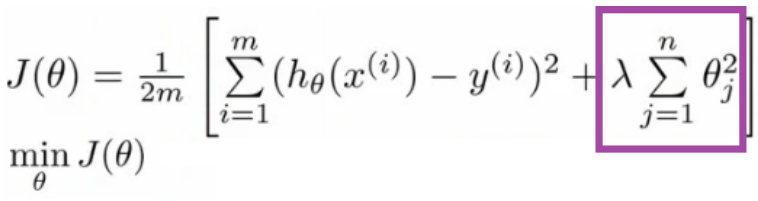

Abbildung 5.2: Regularisierung in Loss Funktion

Wenn in der Abbildung 5.2 die Komplexität steigt, wird eine Strafe hinzugefügt. Das hat zur Folge, dass die Wichtigkeit solcher langen Terme sinkt und zugleich die Komplexität des Modells weniger wird.

Beispiele für solche Techniken sind folgende:

1. L1 Regulierung (Lasso Regulierung)
2. L2 Regulierung (Ridge Regulierung)
3. Elastic net

6 Fazit

Die künstliche Intelligenz ist weiterhin auf dem Vormarsch. So wundert es nicht, dass sich immer mehr Unternehmen mit der Technologie beschäftigen. Die Nutzung birgt jedoch das Risiko, KI Bias zu erzeugen, welche im schlechtesten Fall zu hohen Kosten führen. Die Arbeit beschäftigte sich daher mit der Entstehung solcher Bias und wie sie verhindert werden.

Die Datenaufbereitung ist ein zentraler Baustein für den Erfolg einer KI. Ohne die Aufbereitung der Daten entstehen zahlreiche Bias. Die Folgen von Verzerrungen sind Sicherheitsrisiken für einzelne Gruppen oder Geschlechter oder auch Maschinen. Die Folgen solcher Fehlimplementierungen sind teilweise sehr hohe Kosten oder Imageschäden, können jedoch mit guter Datenqualität vermieden werden.

Es gibt viele Forscher, die sich mit solchen Themen beschäftigen und gezielter auf einzelne Bias eingehen. Das Hauptaugenmerk der Arbeit lag allerdings auf der allgemeine Entstehung von Bias und die dadurch auftretenden Sicherheitsprobleme.

Noch mehr gefährliche Auswirkungen auf Gesellschaft und Demokratie beschreibt Norbert Lossau in seiner Analyse über Deep Fakes [**deepFake**]. Deep Fakes sind manipulierte Videos, in denen Personen erfundene Aussagen in den Mund geschoben werden oder in denen sie scheinbar Handlungen begehen, die in Wirklichkeit nie stattgefunden haben. Fakes werden immer besser, so dass sie kaum als Fakes zu erkennen sind. Die Sicherheitsrisiken für die Menschen sind somit durchaus größer als hier beschrieben. Daher ist es sinnvoll Untersuchungen auf solche Thematiken auszuweiten.

Eidesstattliche Erklärung

Hiermit versichern wir, dass wir die vorgelegte Seminararbeit selbstständig verfasst und noch nicht anderweitig zu Prüfungszwecken vorgelegt haben. Alle benutzten Quellen und Hilfsmittel sind angegeben, wörtliche und sinngemäße Zitate wurden als solche gekennzeichnet.

David Mödl Sebastian Lober, am 19. Juli 2020