Hochschule für angewandte Wissenschaften Würzburg-Schweinfurt
Fakultät Informatik und Wirtschaftsinformatik

Seminararbeit

Bias of Neural Networks - Security implications

David Mödl & Sebastian Lober

7. Juni 2020

Zusammenfassung

TODO

Abstract

KI steht kurz für künstliche Intelligenz. Der Begriff KI ist jedoch irreführend. Eine 'KI' ist ein Programm, das versucht biologisches intelligentes Verhalten nachzuahmen. Die Begrifflichkeit Intelligenz in Verbindung mit Computern ist sehr umstritten, dennoch wird im Allgemeinen als auch in der Forschung das Wort 'Intelligenz' verwendet.
Aus diesem Grund und an Mangel an qualitativ hochwertigen Alternativen wird auch im Folgenden der Wortlaut KI verwendet, wohl wissend, dass die Bezeichnung nicht 100 Prozent korrekt ist.

Inhaltsverzeichnis

1 Einführung

Künstliche Intelligenz(KI) oder auch artifizielle Intelligenz(AI) tritt in großen Teilen unserer Gesellschaft auf. Von Kaufvorschlägen in Amazon, über Chat-Bots bis hin zu autonom fahrenden Autos spielt die KI eine große Rolle.

Wie beim Menschen braucht eine Maschine Erfahrungen um Intelligenz aufbauen zu können. Diese Erfahrungen werden meist zu Beginn mit Hilfe von Daten antrainiert und während dem Betrieb weiter gesammelt. Dabei greift die Maschine z.B. auf Erfahrungswerte der Menschen zu und lernt anhand dieser Muster und Gesetzmäßigkeiten kennen. Da Menschen meist bewusst oder unbewusst Vorurteile haben, lernt die Maschine diese kennen und wendet diese auf andere Fälle an. Dabei entsteht ein Fehlverhalten der KI, welches häufig als KI-Bias gekennzeichnet wird.

Diese Fehlverhalten haben verschiedene Ursprünge, welche wir in dieser Arbeit erläutern möchten. Zudem wollen wir auf einige Sicherheitsprobleme eingehen, welche dadurch entstehen können und erläutern wie man KI-Bias größtenteils vermeiden kann.

2 Grundlagen

2.1 Bias

Erklärung des Wortes in unserem Use-Case

2.2 Künstliche Intelligenz

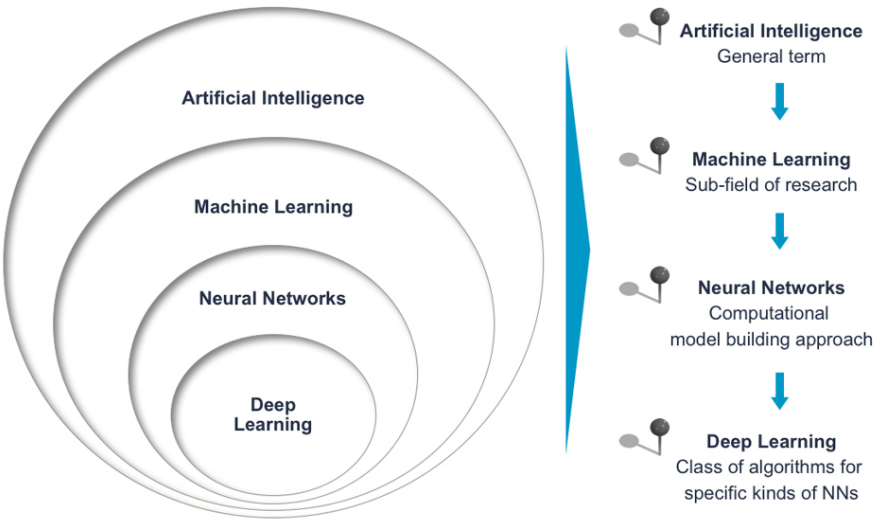

Abbildung 2.1: Verschiedene Abstraktionslevel von Artificial Intelligence in hierarchischer Ordnung

2.2.1 Maschine Learning

alt

Machine Learning (ML) ist ein Teilgebiet der künstlichen Intelligenz. Wie in 1 beschrieben, werden Systeme mit Hilfe von Machine Learning trainiert, indem durch die Trainingsdaten, Muster und Gesetzmäßigkeiten erkannt werden. Die aus den Daten gewonnenen Erkenntnisse lassen sich verallgemeinern und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwenden.
Das Ergebnis aus der Analyse von unbekannten Daten soll wiederum in den Lernprozess für weitere neue Daten integriert werden. Das Hauptziel ist, dass die Maschine automatisch lernt ohne vom Menschen Hilfe oder Anweisungen zu bekommen und daraufhin die Aktionen von alleine anpasst[3].

Damit das System eigenständig lernen und Lösungen finden kann, ist ein vorheriges Handeln von Menschen notwendig. Beispielsweise müssen die Systeme zunächst mit den für das Lernen relevanten Daten und Algorithmen versorgt werden. Je mehr Trainingsdaten in den Maschine geladen werden, desto besser wird die Performance des Algorithmus. Zudem sind Regeln für die Analyse des Datenbestandes und das Erkennen der Muster aufzustellen. Sind passende Daten vorhanden und Regeln definiert, können Systeme mit maschinellem Lernen unter anderem folgendes:

1. Relevante Daten finden, extrahieren und zusammenfassen,
2. Vorhersagen auf Basis der analysierten Daten treffen,
3. Wahrscheinlichkeiten für bestimmte Ereignisse berechnen,
4. sich an Entwicklungen eigenständig anpassen und
5. Prozesse auf Basis erkannter Muster optimieren.

2.2.2 Neuronale Netze

alt:

Bilderkennung ist beispielsweise ein Gebiet, bei den neuronale Netzwerke heutzutage angewandt werden. Gibt man bei einer Suchmaschine ein beliebiges Wort ein, werden direkt hunderte Bilder diesbezüglich angezeigt. Auch ist es möglich Bilder hochzuladen. Ein neuronales Netzwerk interpretiert die hochgeladene Datei und ähnliche bis gleiche Bilder werden angezeigt.

Hochauflösende Bilder bestehen aus mehreren Millionen Pixel mit Position und meist RGB-Werten für die Farbe des Punktes. Nur durch Verbindung der richtigen einzelnen Punkte könnte man Konturen erkennen. Diese müssen dann der richtigen Kategorie zugeordnet werden, um sie bei einem Suchwort wie 'Haus' anzuzeigen. Doch wie kann

ein Computer aus dieser riesigen Menge an einzelnen Pixeln Bilder richtig kategorisieren?

Wir Menschen nehmen Bilder mit der Netzhaut des Auges als visuelle Reize auf und senden sie an unser Gehirn. Dieses verarbeitet die Informationen durch einen Teil der etwa 100 Milliarden Neuronen. Jedes Neuron hat 1 bis 200.000 Synapsen, also Verbindungen zu anderen Neuronen. Durch dieses Geflecht an Neuronen werden die Informationen als elektrische Reize durchgeschleust.

Diesen Prozess der Informationsverarbeitung wird in der Informatik mit künstlichen Neuronen und Synapsen versucht nachzuahmen.

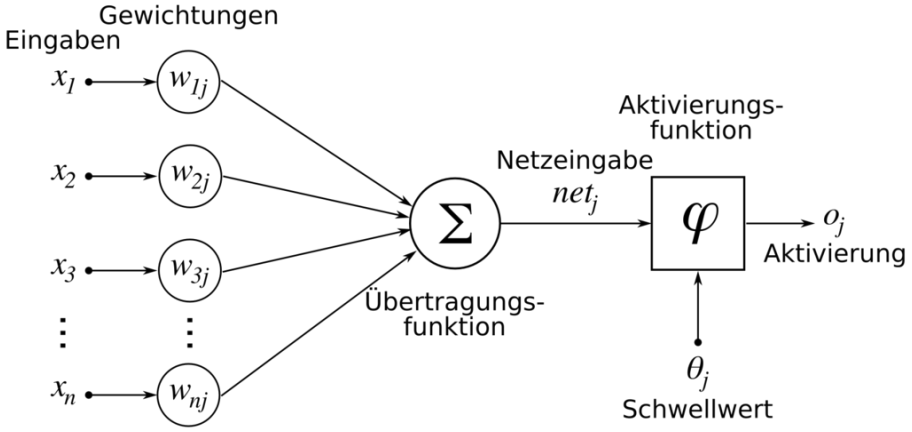

Abbildung 2.2: Künstliches Neuron

Ein Neuron besitzt mehrere Eingangskanäle und einen Ausgangskanal, das sind die Verbindungen bzw. Synapsen. Wenn das Neuron über die Eingangskanäle genügend stimuliert wird, feuert das Neuron ein Signal an den Ausgangskanal. Im Gehirn sind einige Eingangskanäle wichtiger als andere, in der Informatik wird dies mit Gewichtungen gelöst.

Auch ist jedes Neuron mehr oder weniger sensibel. Somit muss durch die gewichteten Eingangskanäle ein gewisser Schwellwert summiert anliegen. Wenn dieser Schwellwert erreicht wird, feuert das Neuron. Der Ausgangskanal eines Neuron ist gleichzeitig ein Eingangskanal eines oder mehreren anderen Neuronen.

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

Jedoch muss ein künstliches neurales Netzwerk erst trainiert werden. Trainieren bedeutet hier das Einstellen der Gewichtungen von Verbindungen und Schwellwerte der Neuronen für eine spezielle Aufgabe. Diesen Prozess nennt man Deep Learning eine Form

von Machine Learning.

2.2.3 Deep Learning

Konzentration auf Deep Learning in dieser Arbeit.

alt

Beim Deep Learning werden die zahlreichen Zwischenschichten, eingeschlossen von Eingabe- und Ausgabe-Schicht, trainiert. Hierbei wird eine umfangreiche und komplexe Struktur der Neuronen-Verbindungen zwischen den Schichten aufgebaut. Wie das Programm endgültig die Aufgabe lösen soll, wird hierbei nicht vorgegeben, sondern wird bei diesem Prozess autonom ermittelt.

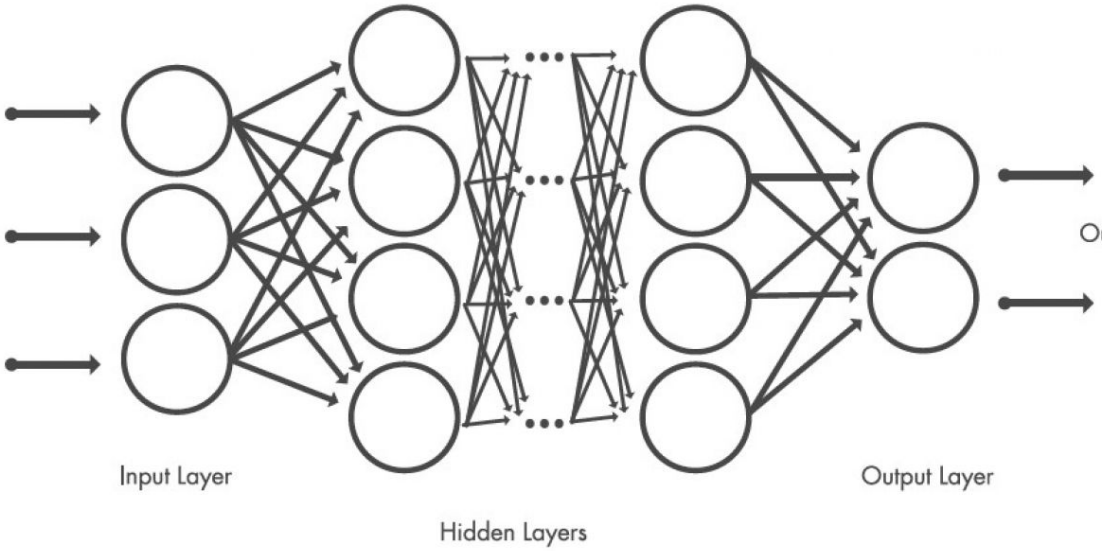

Abbildung 2.3: Hidden Layer

Ein künstliches neuronales Netz wird mit dem Zweck aufgebaut, eine bestimmte Aufgabe zu lösen. Extra dafür müssen Testdaten aufbereitet werden. Diese Art der Daten, beispielsweise Bilder, soll das fertig trainierte Netz richtig interpretieren können. Testdaten sind Daten, bei dem das korrekte Ergebnis zu jedem Test bekannt ist.

Am Anfang ist das künstliche neuronale Netz meist mit relativ zufälligen Werten und Verbindungen vorbelegt. Testdaten werden dem zu trainierenden Netzwerk an die Eingangsschicht übergeben. Diese durchlaufen das Netz. Das Ergebnis wird an der Ausga-

beschicht überprüft. Die Ausgabeschicht besteht im einfachsten Fall aus zwei Neuronen, beispielsweise 'Gesicht erkannt' oder 'kein Gesicht', an dem gemessen wird, wie viel gewichtete Signale ankommen. Diese summiert, ergeben die Endergebnisse der Berechnungen und das Neuron mit dem höchsten Gewicht, ergibt die Antwort.

Neuronale Netze sind für Menschen ab einer gewissen Größe nicht mehr nachvollziehbar. Somit kann nur die Eingabe mit der Ausgabe verglichen werden, um auf die Korrektheit der Aufgabenlösung zu prüfen. Ziel ist es durch möglichst viele und umfangreiche Testdaten das Netzwerk so einzustellen, dass diese nicht nur die Testdaten richtig beantwortet, sondern auch unbekannt Daten korrekt interpretiert.

Um eine Aufgabe, wie 'Gesicht in Bild erkennen', zu trainieren, wird nicht nur ein Netz mit Zufallswerten und Verbindungen generiert sondern tausende. Alle werden mit den gleichen Testdaten geprüft und für jedes Netz 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 wieder genommen und 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 Testdaten zu verwenden. Als Ergebnis des Prozesses erhält man durch Deep Learning ein trainiertes künstliches neuronales Netz, das im Allgemeinen als KI bezeichnet wird.

Die KI kann nun die Testdaten nahe zu perfekt interpretieren. Trotzdem muss nun noch weiter überprüft werden, ob die KI auch unbekannte Daten korrekt interpretiert. Fehlverhalten einer KI kann leicht übersehene werden, da es fast unmöglich ist, jeden Testfall abzudecken. Beispiele von KIs mit Fehlverhalten und wie sie genau entstehen, werden im nächsten Abschnitt genauer betrachtet.

Lernarten

i. Supervised
ii. Unsupervised
iii. Controlevisional
iv. Reinforcement

2.2.4 Loss-Funktion

Performanz

2.2.5 Informationsverlust zwischen Schichten

1. Erste Schicht verbunden mit letzter Schicht
a. Eingabe hoher Einfluss auf Endergebnis
2. Jede Schicht nur Verbindung zu der Nächsten
a. Hohe Informationsverlust

2.3 Neuronen und Features

2.3.1 Arten von NN

Es gibt drei Arten, wie neuronale Netze Daten verarbeiten können.

3D

Gewichtete Netze

Features

2.4 Architekturen

Arten
1. Full Connected
2. CNN
3. ResNet
4. Natural Network Connection
5. Dropout
6. ...

2.5 Daten

2.5.1 Datenaufbereiten

Daten sollen Generalisiert sein
a. Beispiel Bilder
i. Sammeln
ii. Drehen
iii. Spiegeln
iv. Verzerren
v. Etc.

2.5.2 Daten und Architektur

Daten müssen zur Architektur passen. Generell
Je Komplexer mehr Neuronen
Mehrfälle desto Fehleranfälliger = mehr Bias

3 Bias Entstehung

3.1 Daten

3.1.1 Unvollständigkeit der Daten

3.1.2 Bias in Test-/Trainingsdaten

3.1.3 Under-/ Overfitting

3.1.4 Ähnlichkeit der Daten

Selbst für Menschen auf ersten Blick schwierig zu differenzieren
ii. Bild Croissant vs Katze
iii. https://distill.pub/2019/activation-atlas/

3.2 Menschliche Fehler

3.2.1 Falsches Ziel

3.2.2 Falsche Architektur

3.2.3 Falsches Lernen

KI lernt einfachste Unterschiede
i. Nicht Unterscheid zwischen Auto und Boot sondern Untergrund(Wasser/Land)
ii. Sehr Fehleranfällig z.B. Auto fährt durch flaches Wasser (KI -¿ Boot)

Abbildung 3.1: Hund oder Bagel?

3.3 Angriff auf KI

Adversarial Attacks

4 Sicherheitsprobleme durch BIAS

4.1 Gefahren für Maschinen

Google KI -¿ Kühlung von Maschinen

4.2 Gefahren für Menschen

Tesla Autopilot
Etc.

5 Prävention

5.1 Passende Architektur zu Daten

5.2 Nur ein Ziel

Viele Ziele = Komplex -¿ Fehleranfällig -¿ Bias
Beispiel: Baidu Gesichtserkennung (erkennt nur Asiaten)

5.3 Verfahren zum Validieren

Unterschiedliche Personen (Entwickler/Tester)

An echte Daten Testen (Überwachtes Demo Live Betrieb)

Beispiel:
Polizei Berlin Gesichtserkennung bei Überwachungskamera
3 verschiedene KIs

5.4 Test-/Trainingsdaten Aufbereiten

Vollständigkeit prüfen
-¿ Fehler hier = Bias

Bias aus Daten entfernen

6 Fazit

Thema ist größer als hier beschreibbar
Evtl. Deep Fake http://iphome.hhi.de/samek/pdf/LapNCOMM19.pdf https://ujjwalkarn.me/2016/08/
explanation-convnets/

7 Alt: Problemstellung Fehlverhalten von künstlichen neuronalen Netzen

7.1 Was sind Fehlverhalten von künstlichen neuronalen Netzen?

Ca. 4 Jahre entwickelte Amazon einen Algorithmus, welcher unter mehreren Bewerbungstexten automatisch die besten Bewerber herausfiltern sollte. Dabei bezog die Software sich auf voran gegangene Bewerbungen, verdeutlichte dabei aber ein grundlegendes Problem des maschinellen Lernens in seiner aktuellen Form.

Der Algorithmus hatte mit den Datensätzen der angenommen Bewerber trainiert und lernte daraus welche Eigenschaften Amazon bevorzugt. Weil das Unternehmen aber Teil einer von Männern dominierten Industrie ist, waren in den zugrunde gelegten vergangenen zehn Jahren vor allem Männer eingestellt worden. Daraus resultierte, dass Frauen grundsätzlich schlechter bewertet wurden, selbst wenn gar kein Geschlecht angegeben wurde und dieses z.B. nur durch Frauenvereine erkennbar wurde. Die KI blieb diesen Auswahlkriterien treu und bevorzugte vorwiegend Männer.[1]

Dieses Fehlverhalten der KI führte dazu, dass die Software nicht genutzt werden konnte, bzw. nach Anpassungen nur eine beratende Funktion besaß.

Unbekannt ist dieses Verhalten allerdings nicht. Es gibt weitere Fälle in der eine KI so wirkt, als hätte sie Vorurteile gegenüber manchen Gruppen/Geschlechter/Religionen, welche allerdings Fehlverhalten sind und durch den Menschen antrainiert wurden.

7.2 Garbage in - Garbage Out - Wie entstehen Fehlverhalten von künstlichen neuronalen Netzen?

Wie in vorangegangen Beispielen bereits erwähnt wurde, wird Künstliche Intelligenz in vielen Bereichen eingesetzt. Die Hoffnung solcher Anwendungen, liegt eigentlich darin,

Vorurteile zu vermeiden und Prozesse fairer zu gestalten, da eine Maschine grundsätzlich keinen Unterschied zwischen Schwarz und Weiß, Mann und Frau oder Jung und Alt erkennt.

Anhand des Beispiels aus 7.1 sieht man, dass dies nicht der Fall ist, da die Daten, mit welchen die KI lernt, zentralen Einfluss auf das Ergebnis hat. Meist ist es nicht möglich Daten zu finden, welche nicht bereits Vorurteile, enthalten. Solch verzerrte Trainingsdaten, werden unter Bezug auf ihre Zusammensetzung, auch als WEIRD Samples(western, educated, industrialized, rich and democratic societies) bezeichnet.

Ein weiteres Problem ist die fehlende Diversität auf Seiten der/die Entwickler/innen. Nur 15% bei Facebook und 10% bei Google in der KI Entwicklung sind Frauen. Für dunkelhäutige Menschen ist es noch schlimmer. Bei Google z.B. sind nur 2,5% und bei Facebook und Microsoft 4% dunkelhäutige Mitarbeiter[4].

Daher kommt der Ausdruck "Garbage in - Garbage Out", dass heißt benutzt man fehlerhafte Daten oder welche, die Vorurteile beinhalten, erhält man fehlerhafte oder mit Vorurteilen belastete Ergebnisse.

7.3 Probleme durch Fehlverhalten von künstlichen neuronalen Netzen

Durch die bereits genannten Beispielen in den vorherigen Kapiteln, werden Probleme deutlich. Das Fehlverhalten einer KI kann zu Diskriminierung einzelner Geschlechter/-Gruppen oder Kulturen führen. Ein anderes Problem wird aber deutlich, wenn man ein Beispiel aus der Medizin anschaut.

In der USA existiert eine KI, welche die Gesundheitsversorgung möglichst effektiv gestalten soll. Diese soll eine Person mit besonderem Pflegebedarf identifizieren. Eine im Oktober 2019 veröffentlichte Studie zeigt allerdings auf, dass Menschen mit Afroamerikanischen Wurzeln bei gleicher Krankheitsschwere, seltener für extra Pflege vorgeschlagen wurden als Weiße[2].
Dies führt dazu das Afroamerikaner eine niedrigere Gesundheitsversorgung haben als andere und dies kann zu großen Gesundheitliche Problemen führen.

Ein weiteres Beispiel nannte das Heise Magazin 2017[5]. Das Online Magazin erwähnte die Software COMPAS, welche auch in der USA verwendet wird und dort in mehreren Bundesstaaten zum Einsatz kommt. COMPAS steht für „Correctional Offender Management Profiling for Alternative Sanctions“ und gibt vor, das Rückfallrisiko von Straftätern verlässlich berechnen zu können.

COMPAS errechnet für jeden Delinquenten einen individuellen Risk Score, welcher auf das spätere Strafmaß Auswirkungen hat. In die Risikobeurteilung des Algorithmus fließen eigene Vorstrafen, eventuelle Vorstrafen naher Verwandter, Alkohol- und Drogenmissbrauch, soziale Bindungen, usw..

Falls der Risk Score eines Verurteilten nun zwischen 1 und 10 ist, lässt der Richter keine Bewährung mehr zu sondern schickt ihn präventiv hinter Gitter. Das Problem hierbei ist, dass der Algorithmus nicht offen gelegt wird und somit wird dem Algorithmus blind vertraut, ohne überprüfen zu können, wie dieser Score zustande kam.

Wie Heise Magazine weiter berichtete, belegte allerdings 2016 eine Studie der Investigativ-Plattform ProPublica, dass die COMPAS-Algorithmen beispielsweise schwarzen Angeklagten grundsätzlich ein höheres Risiko attestieren, erneut straffällig zu werden, als dies tatsächlich der Fall ist. Bei weißen Angeklagten ist es hingegen genau umgekehrt.

Falls solche Algorithmen blind vertraut werden, kann dies zu gravierenden Folgen eines Verdächtigen führen, welcher zu unrecht ein ganzes Leben hinter Gitter sitzen könnte.

Tabellenverzeichnis

Literatur

[1] Dastin Jeffrey. *Amazon scraps secret AI recruiting tool that showed bias against women*. 10. Okt. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (besucht am 24. 05. 2020).

[2] Obermeyer Z. und Powers B. und Vogeli C. und Mullainathan S. *Dissecting racial bias in an algorithm used to manage the health of populations*. 2019, S. 447–453.

[3] E. Sengupta u. a. „Techniques to Elimenate Human Bias in Machine Learning". In: *2018 International Conference on System Modeling Advancement in Research Trends (SMART)*. 2018, S. 226–230.

[4] West S.M. und Whittaker M. und Crawford K. *Discriminating Systems: Gender, Race and Power in AI*. 2019, S. 3. URL: https://ainowinstitute.org/discriminatingsystems.html (besucht am 24. 05. 2020).

[5] Peter-Michael Ziegler. *Im Namen des Algorithmus*. 2017. URL: https://www.heise.de/select/ct/2017/25/1512700333136715 (besucht am 24. 05. 2020).