Bachelorarbeit

Künstliche Intelligenzen und automatisierte Entscheidungsfindung

Schwachstellen und rechtliche Rahmenbedingungen in kontroversen Anwendungsfällen

Ausgeführt zum Zwecke der Erlangung des akademischen Grades eines

Bachelor of Science

im Rahmen des Studiums

Wirtschaftsinformatik

unter der Leitung von

Ao.Univ.Prof. Mag.iur. Dr.iur. Markus Haslinger
(E280 Institut für Raumplanung, Forschungsbereich: Rechtswissenschaften)

eingereicht von

Sebastian Frisch
01631851

_____	_____
Ao.Univ.Prof. Mag.iur. Dr.iur. Markus Haslinger	Sebastian Frisch

Wien, am _____

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

Weiters erkläre ich, dass ich dieses Bachelorarbeitsthema bisher weder im In- noch Ausland (einer Beurteilerin/einem Beurteiler zur Begutachtung) in irgendeiner Form als Prüfungsarbeit vorgelegt habe und dass diese Arbeit mit der vom Begutachter beurteilten Arbeit übereinstimmt.

Wien, am _____ _____

Sebastian Frisch

Abstract

Artificial Intelligence and Automated Decision Making – Vulnerabilities and Legal Conditions Relating to Controversial Use Cases

The aim of this paper is to give a broad overview of the topic of artificial intelligence (AI). Taking the rapid developments of AI performance and the multiplying application fields into account, it is attempted to discuss current questions using controversial examples. Thus, resources include online and printed reports as well as academic papers.

After a short historic overview of the development of AI, a comparison between the systems of classical AI and artificial neuronal networks (NN) is made. After this technical evaluation the focus is put on its applications, such as autonomous driving, an automated risk assessment for recidivism used by the State of California, AI in military projects, Microsoft's AI chatbot Tay, and other controversial application areas.

The next chapter reflects on AI as a black box, meaning on the intransparent usage of AI and the problem of bias. Additionally, attack strategies on neuronal networks from a technological and sociopolitical perspective will be discussed.

To conclude the topic, AI is discussed from a jurisdictional position. The main focus of this work lies in the treatment of potential discrimination through AI in current regulations as well as a depiction of the attempts to regulate so far unregulated areas.

In the light of this paper it seems that an expansion of public knowledge regarding AI is more important than ever. A continuation of advancements in this technological field seems to be very likely. Due to its projected consequences on the citizen an expansion of knowledge regarding both positive and negative aspects is paramount: a differentiation between curse and blessing seems necessary.

Inhaltsverzeichnis

Einleitung 1

1. Entwicklung der künstlichen Intelligenz 2

1.1 Geschichte der künstlichen Intelligenz 2

1.2 Definition künstlicher Intelligenz 3

2. Klassische künstliche Intelligenzen 4

2.1 Berechnung der Kreditwürdigkeit (Credit-Scoring) 4

2.2 Flugassistenzsysteme (Autopilot) 6

2.3 Sprachassistenten 6

3. Künstliche neuronale Netzwerke 7

3.1 Grundlegende Funktionsweise 7

3.2 Der Lernprozess 8

3.3 Eigenschaften 10

3.4 Grenzen neuronaler Netzwerke (das Chinese Room Argument) 14

4. Anwendungen und Brennpunkte 15

4.1 Selbstfahrende Autos 16

4.2 Anwendungen im Rechtsstaat 18

4.3 Militärische Anwendungen 19

4.4 Microsoft Tay 20

4.5 Weitere Anwendungsbereiche 21

5. Ein Blick in die Blackbox 22

5.1 Haben KIs Vorurteile? 23

5.2 Analyse und Bewertung von klassischen KIs 24

5.3 Bias in neuronalen Netzwerken 24

5.4 Analyse und Bewertung von neuronalen Netzwerken 27

5.5 Angriffe auf neuronale Netzwerke 33

6. Rechtliche Rahmenbedingungen 40

6.1 Bestehendes Recht 41

6.2 Zukünftiges Recht und Lobbying 46

Conclusio 51

Abkürzungen I

Abbildungen II

Quellen III

Bücher III

Statistiken III

Rechtsquellen III
Berichte III
Wissenschaftliche Arbeiten und Artikel IV
Journalistische Quellen VII
Sonstige Internetquellen X

Einleitung

Waren künstliche Intelligenzen vor ein paar Jahrzehnten hauptsächlich in Werken der Science-Fiction zu finden, werden diese mittlerweile in breiten Anwendungsgebieten eingesetzt. Heutzutage kommt der durchschnittliche Bürger technologisch fortgeschrittener Gesellschaften zumindest einmal täglich mit künstlichen Intelligenzen über sogenannte „Smart"-Geräte, wie dem Smartphone, dem Smart-TV oder einem Sprachassistenten, in Berührung. Entsprechend findet sich das Thema KI im Jahr 2018 gleich mehrfach im „Hype-Cycle" der Technologietrends des Marktforschungsunternehmen Gartner.[1]

Diese Arbeit verfolgt das Ziel einer breiten Einführung in das Thema. Im Hinblick auf die rasanten Fortschritte sowohl in der Leistungsfähigkeit als auch in den Einsatzgebieten der KI wurde versucht, die wesentlichen Fragen anhand möglichst aktueller Beispiele zu illustrieren. Als Quellen dienten entsprechend neben wissenschaftlichen Arbeiten auch Berichte in Internet- sowie Printmedien.

Nach einer kurzen Darstellung der historischen Entwicklung der KI wird der Unterschied zwischen Systemen der klassischen KI und künstlichen neuronalen Netzwerken (NN) herausgearbeitet. Nach diesen eher technisch orientierten Kapiteln wechselt der Fokus auf die Brennpunkte in der Anwendung: Aktuell im breiten Einsatz befindliche KIs beziehen sich vielfach auf unkritische Anwendungsfälle: Ein Fehler in der KI einer Navigations-App führt unkorrigiert schlimmstenfalls dazu, dass das Ziel später erreicht wird. KIs übernehmen aber zunehmend auch Aufgaben, in denen Fehlentscheidungen viel gravierendere Auswirkungen auf das Wohl der Nutzer haben, beispielsweise in autonomen Fahrzeugen. In Folge wechselt der Fokus auf KI als „Blackbox", d.h. die intransparente Funktionsweise der KIs. Zusätzlich werden dabei Angriffsstrategien auf neuronale Netzwerke aus technologischer und gesellschaftspolitischer Sicht behandelt.

Die vorhandenen Angriffsmöglichkeiten auf KIs sowie die dargelegten Anwendungsbeispiele zeigen, dass Fortschritte in Entwicklung und im Einsatz dieser neuen Technologien zumeist nach dem kontroversen Motto „Move Fast and Break Things"[2] erzielt worden sind und werden. Ein vorgelagerter, als bremsend empfundener tiefgehender gesellschaftlicher Diskurs kommt daher vielfach zu kurz.

Abschließend wird das Thema KI von einer rechtlichen Sichtweise behandelt. Die Schwerpunkte liegen hierbei in der Behandlung der potentiellen Diskriminierung durch KIs in derzeit geltenden Rechtssystemen sowie in einer Darstellung der Versuche für derzeit unregulierte Bereiche einen rechtlichen Rahmen zu finden.

Im Lichte dieser Arbeit erscheint eine Verbreiterung des Wissens in der breiten Bevölkerung wichtiger denn je. Aus technologischer Sicht scheint eher eine weitere Beschleunigung als ein Abflachen der Integration von KIs in den Alltag wahrscheinlich. Die damit einhergehenden Auswirkungen auf die Bevölkerung verlangen nach einer qualifizierten Auseinandersetzung, die zwischen den Polen Fluch und Segen differenziert und zweifelhaft vorhandene Risiken als Chancen begreift.

[1] Gartner. (16.8.2018). Widespread artificial intelligence, biohacking, new platforms and immersive experiences dominate this year's Gartner Hype Cycle.
[2] Business Insider. (14.10.2010). Mark Zuckerberg, Moving Fast And Breaking Things.

1. Entwicklung der künstlichen Intelligenz

1.1 Geschichte der künstlichen Intelligenz

Das Konzept der künstlichen Intelligenz hat, wie viele andere Entwicklungen auch, nicht nur einen technischen sondern einen sozialen Hintergrund. Dieser ist eng mit dem des Roboters verbunden und beginnt in der Antike mit dem Mythos des Talos (Griechisch: Τάλως), der in der am weitesten verbreiteten Version von Hephaestus, dem Schmied der Götter, erschaffen wurde, um die Geliebte des Zeus, Europa, zu beschützen.[3] Mit dem Fortschreiten mechanischer Möglichkeiten fanden im 17. Jahrhundert mithilfe von Zahnradmechanismen konstruierte Roboter -, unter anderem ein Trompeter, für den Ludwig van Beethoven sogar eine Fanfare schrieb - großen Zulauf, da sie aus immer komplexeren Mechanismen bestanden, die den Anschein intelligenten Verhaltens ermöglichten.[4]

1950 veröffentlichte Alan Turing seine Arbeit „Computing Machinery and Intelligence“, in der er den berühmten Turing-Test konzipierte, in dem künstliche Intelligenzen ihre Denkfähigkeit unter Beweis stellen müssen.[5] Dabei muss ein Proband bzw. eine Probandin durch einen Chat, also nur durch schriftliche Konversation, mit je einer anderen Person und der KI herausfinden, bei welchem der zwei Kommunikationspartner es sich um die KI handelt.[6] Alan Turings Arbeit war zu Beginn rein theoretischer bzw. philosophischer Natur, da damalige Computer noch bei weitem nicht die dafür erforderliche Rechen- und Speicherleistung aufbringen konnten.

In den 1960er Jahren wurde von Joseph Weizenbaum der Chatbot ELIZA entwickelt, der Antworten einer Therapeutin simuliert.[7] Der hinter ELIZA stehende Ansatz wirkte zu seiner Zeit derart vielversprechend, dass das US-Verteidigungsministerium im Jahr 1963 über DARPA auf künstliche Intelligenzen spezialisierte Forschungsbereiche an mehreren US-Universitäten subventionierte.[8] Mit ELIZA wandelte sich auch die öffentliche Wahrnehmung bezüglich der künstlichen Intelligenz, da diese bis dahin konzeptuell untrennbar von dem „humanoiden“ Roboter erschien, ELIZA als gesichtsloser Chatbot aber lediglich Text ausgeben konnte. Dadurch und durch die militärischen Interessen wandelte sich das Forschungsfeld: Bis dahin war, der gesellschaftlichen Sichtweise entsprechend, vor allem im Gebiet der generellen künstlichen Intelligenz, also möglichst dem Vorbild des Menschen entsprechend, geforscht worden. Das US-Militär wollte jedoch nicht gesellschaftspolitische Auswirkungen wissenschaftlich fundierter Zukunftsvisionen analysieren, sondern unmittelbar herausfordernde militärische Themen, vor allem Spracherkennung und Übersetzung, sowie Datenverarbeitung, lösen.[9]

Die Rechenleistung der damals zur Verfügung stehenden Computer reichte jedoch für diese Aufgaben nicht aus, woraufhin zwei sogenannte „Winter der künstlichen Intelligenz“ - Perioden des geringen Interesses an KI - folgten.[10] Ein Lichtblick zwischen den beiden Perioden war unter anderem die Entwicklung von „Expertensystemen“, also Programmen, in

[3] Evans L Smith, Nathan R Brown. (2008). The Complete Idiot's Guide to World Mythology. S 240-241
[4] Bruce G. Buchanan. (2005). A (Very) Brief History of Artificial Intelligence.
[5] Alan M. Turing. (1950). Computing Machinery and Intelligence.
[6] Encyclopaedia Britannica. (14.4.2019). Turing test.
[7] The Harvard Gazette. (13.9.2012). Alan Turing at 100.
[8] Harvard University. (28.8.2017). The History of Artificial Intelligence.
[9] Harvard University. (28.8.2017). The History of Artificial Intelligence.
[10] LiveScience. (25.8.2014). History of A.I.: Artificial Intelligence (Infographic).

die eine auf Expertenmeinungen basierte Lösung für jedes mögliche Szenario gespeichert wird, welches dann von den Benutzern des Systems abgefragt werden kann.[11]

Den Höhepunkt der Entwicklung rein klassischer KIs bildete IBMs „Watson", der sich bei der Quizshow „Jeopardy" 2011 gegen zwei menschliche Kontrahenten durchsetzen konnte und IBM eine Million US-Dollar Preisgeld einbrachte.[12] Watsons Antwortfindung basierte dabei auf „Knowledge-Graphs". Die Entitäten dieser Graphen bestehen aus Stichwörtern, Objekten bzw. Aktivitäten, die über mit einem Gewicht (einer Wahrscheinlichkeit) versehene Kanten verbunden sind.[13] Die Kanten des Graphen sollen „Assoziationen" zwischen den Stichwörtern ermöglichen, sodass ein Durchsuchen des Graphen mit den Stichwörtern einer Frage als Eingabeparametern, zu einer Antwort führt.[14]

Parallel zu den als klassische KIs bezeichneten Techniken entwickelte sich mit „*artificial neuronal networks*" bzw. NNs ein radikal neuer Ansatz. Zum ersten Mal 1959 zur Reduzierung des Echos in Telefonleitungen angewendet, bildet diese seit den 1980er Jahren tiefergehend erforschte Technologie die Grundlage „moderner künstlicher Intelligenz", wie dem autonomen Fahren und der Bilderkennung.[15]

1.2 Definition künstlicher Intelligenz

Der Begriff der „künstlichen Intelligenz" („*artificial intelligence*") wurde unter anderem von dem Mathematiker John McCarthy geprägt. Dieser stellte 1956 die These auf, dass jeder Aspekt der menschlichen Intelligenz von einer Maschine (bzw. einem Programm) simuliert werden kann.[16] Der Fokus der damaligen Forschung lag in der Nachbildung allgemeiner menschlicher Fähigkeiten durch Maschinen. Zwischenzeitlich umfasst der Begriff aber auch spezialisierte KIs, deren Ziel nicht die Nachbildung menschlicher Eigenschaften, sondern die Bewältigung konkreter Aufgaben ist – wobei diese Problemstellungen das menschliche Lösungsvermögen oftmals übersteigen. Eine allgemein akzeptierte Definition von KI wurde bis dato nicht entwickelt. Zusammenfassend kann jedoch festgehalten werden, dass KIs aus der Sichtweise der entwickelnden Techniker und Psychologen zumindest eine Komponente der menschlichen Intelligenz, also Lernen, Mutmaßen, Wahrnehmen, Problemlösen und Sprachverständnis erfüllen muss.[17]

Da diese Arbeit den Schwerpunkt nicht auf die Entwicklung, sondern die gesellschaftlichen Auswirkungen des Einsatzes von künstlicher Intelligenz legt, scheint hierfür die folgende Definition am zutreffendsten:

> *Eine künstliche Intelligenz ist eine Maschine bzw. ein Programm, welches sich dem Anschein des Beobachters nach intelligent verhält.*[18]

[11] E. A. Feigenbaum. (1980). Expert Systems in the 1980s.
[12] New York Times. (16.2.2011). Computer Wins on 'Jeopardy!': Trivial, It's Not.
[13] AI Magazine. (2010). The AI Behind Watson — The Technical Article.
[14] AI Magazine. (2010). The AI Behind Watson — The Technical Article.
[15] University of Toronto. (10.4.2019). 3.0 History of Neural Networks.
[16] CNBC. (17.6.2017). Everyone keeps talking about A.I.—here's what it really is and why it's so hot now.
[17] Encyclopaedia Britannica. (14.4.2019). Artificial Intelligence.
[18] Encyclopaedia Britannica. (14.4.2019). Artificial Intelligence.

Intelligentes Verhalten setzt für den Rahmen dieser Arbeit nicht voraus, dass sich der Beobachter zum Zeitpunkt der Interaktion mit der KI darüber im Klaren ist, wie diese konkret funktioniert. Für Zeitgenossen des 17. Jahrhunderts machte der mechanische Trompeter aus Kapitel 1.1 genauso den Anschein intelligenten Verhaltens wie dies heutzutage für Laien bei dessen aktueller Interpretation, dem „Teddy Bear Orchestra“[19], der Fall ist. Bekommt der Betrachter jedoch einen Einblick in die Funktionsweise der Maschine, werden grobe Schwächen offenbar oder zeigt sich die Beschränkung der Maschine durch eine schleifenhafte Wiederholung der gezeigten Aktivität, so ist diese Illusion zerstört. Da sich künstliche neuronale Netzwerke von klassischen KIs deutlich unterscheiden, werden diese beiden Arten in den folgenden Abschnitten getrennt behandelt und miteinander verglichen. Eines vorweg: Neuronale Netzwerke erfüllen bereits heute viele der oben genannten Voraussetzungen, die aus Sicht der Entwickler bzw. Psychologen eine KI ausmachen, während klassische KIs dies nicht tun.

2. Klassische künstliche Intelligenzen

Die in dieser Arbeit als klassische KIs bezeichneten Techniken haben gemein, dass die künstliche Intelligenz bis ins Detail durchgeplant und an die Problemstellung angepasst werden muss.[20] Abhängigkeiten zwischen den verschiedenen Wissenselementen - also den Fakten, die der KI zur Verfügung stehen - müssen in fast allen Fällen zuerst in Studien statistisch erhoben werden, bevor sie im Rahmen eines Modells nachgebildet werden können. Diese zum Teil sehr komplexen Modelle stellen entweder den Kern oder sogar die gesamte KI dar.[21]

2.1 Berechnung der Kreditwürdigkeit (Credit-Scoring)

Ein Beispiel dafür bilden „Credit-Scoring Algorithmen“, deren Ergebnisse vor allem in den USA öffentlich Anwendung finden, um die Kreditwürdigkeit von Bürgerinnen und Bürgern zu bestimmen. Die Institute verwenden hierbei eine Vielzahl verschiedener Parameter zur Bestimmung des Credit-Score der potentiellen Kreditnehmer und Kreditnehmerinnen. In der Regel fließen neben Bildungsgrad und Einkommen die Kredithistorie, die Anzahl der vorhandenen Bankkonten bzw. Kreditkarten sowie die Auslastung der Kreditrahmen ein.[22]

Die Berechnung von Credit-Scores stellt ein einfaches Klassifizierungsproblem dar. Die einzelnen potentiellen Kundinnen und Kunden müssen jeweils einer der beiden Gruppen „gute Kreditwürdigkeit“ bzw. „schlechte Kreditwürdigkeit“ zugeteilt werden. Für derartige Klassifizierungsprobleme gibt es aus dem Bereich der Statistik gut erforschte Ansätze der sogenannten Clusteranalyse, wie die lineare Diskriminanzanalyse oder den Entscheidungsbaum. Grundsätzlich erhält bei der linearen Diskriminanzanalyse jeder der oben genannten Faktoren ein Gewicht pro Gruppe, welches das Ausmaß der Assoziierung des

[19] Teddy Bear Orchestra. (14.4.2019).
[20] Divisio. (28.6.2019). KI leicht erklärt – Teil 2: Von klassischer KI, Neuronalen Netzen und Deep Learning.
[21] Divisio. (28.6.2019). KI leicht erklärt – Teil 2: Von klassischer KI, Neuronalen Netzen und Deep Learning.
[22] Equifax. (10.4.2019). How Are Credit Scores Calculated?

Faktors mit der jeweiligen Gruppe angibt.[23] Für die Gruppe mit einer hohen Kreditwürdigkeit wird beispielsweise der Faktor „Länge der Kredithistorie“ ein hohes Gewicht aufweisen, während der Faktor „Anzahl der versäumten Ratenzahlungen“ ein hohes Gewicht für die Gruppe mit der niedrigen Kreditwürdigkeit besitzt. Die Gewichte werden dabei automatisiert nach einem statistischen Verfahren berechnet, welches als Grundlage einen Satz bereits klassifizierter Daten verwendet und damit dem Lernprozess von NNs ähnelt.[24] Dieses relativ einfache Verfahren liefert im Vergleich zu komplexeren Modellen auch nur eine vergleichsweise geringe Genauigkeit.[25]

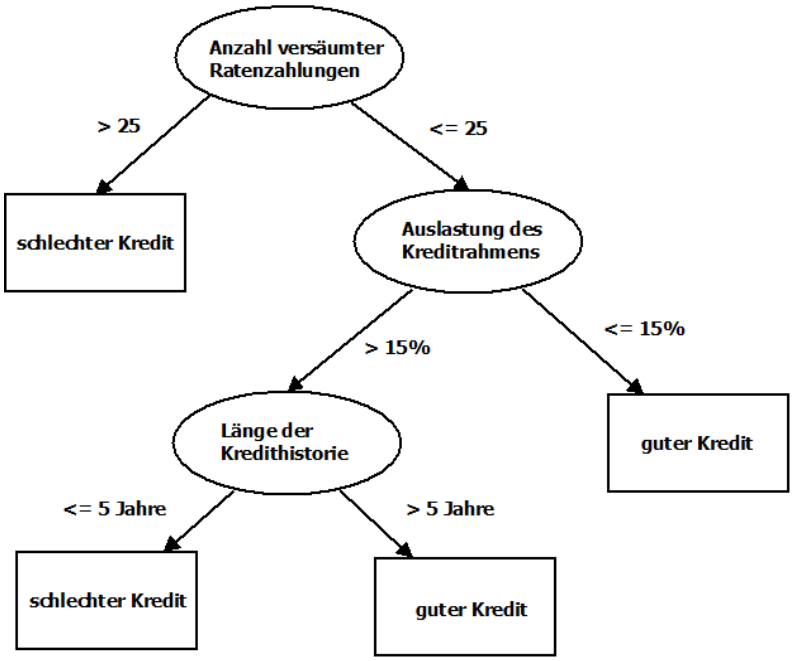

Abbildung 1: Beispiel eines einfachen Entscheidungsbaumes für das Credit-Scoring[26]

Anders der Entscheidungsbaum: Wie der Name schon suggeriert, wird hierbei ein Entscheidungsbaum gespannt, bei dem - wie in Abbildung 1 beispielhaft zu sehen - jede Entscheidung zwei mögliche Pfade eröffnet. Der Baum wird vom oberen Knoten (dem Wurzelknoten) aus analysiert. Mithilfe der bekannten Eigenschaften der Kundin bzw. des Kunden wird dem Pfad gefolgt, bis ein Endknoten, der die Einstufung bezüglich der Kreditwürdigkeit klassifiziert, erreicht ist.[27] Die Parameter der Fragestellungen können auch hier ähnlich der linearen Diskriminanzanalyse entweder statistisch erhoben oder von Experten festgelegt werden.[28] In beiden Fällen muss im Nachhinein jedoch ein „*fine-tuning*“, eine weitere Anpassung der Werte an das gewünschte Ergebnis, mithilfe von Experten der Kreditvergabe erfolgen.[29]

[23] IBM Knowledge Center. (28.6.2019). Diskriminanzanalyse.
[24] IBM Knowledge Center. (28.6.2019). Diskriminanzanalyse.
[25] David West. (2000). Neural Network Credit Scoring Models.
[26] Grafik laut Hué, Sullivan & Hurlin, Christophe & Tokpavi, Sessi. (2017). Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects. Figure 2.
[27] Joao Bastos. (2008). Credit scoring with boosted decisiontrees.
[28] Joao Bastos. (2008). Credit scoring with boosted decisiontrees.
[29] Joao Bastos. (2008). Credit scoring with boosted decisiontrees.

Mit circa 81% korrekten Zuordnungen bei einem Testdatensatz deutscher Kreditnehmer erzielen mit dieser Methode arbeitende Modelle derzeit (noch) bessere Ergebnisse als neuronale Netzwerke mit etwa 78% genereller Genauigkeit.[30] Der Grund für das gute Abschneiden der konventionellen Methode liegt vermutlich in der jahrelangen Erforschung dieser spezifischen Fragestellung, da bei großen Kreditinstituten selbst kleine Steigerungen der generellen Genauigkeit eine hohe Auswirkung auf die Ausfälle von Darlehen und damit die Gewinne der Kreditinstitute ausmachen.

2.2 Flugassistenzsysteme (Autopilot)

Während beim Credit-Scoring vor allem Experten aus dem Gebiet der Statistik benötigt werden, um gute Ergebnisse erzielen zu können, werden in diesem Anwendungsfeld primär theoretische InformatikerInnen oder LogikerInnen benötigt. Bereits 1914 kam der erste Flugregler, der Lage und Geschwindigkeit eines Flugzeuges stabilisieren sollte, zum Einsatz. In Folge wurden immer weitere Instrumente entwickelt, die dem Piloten Stabilisierungs- und Regelungsaufgaben abnahmen, bis im Jahr 1950 zum ersten Mal ein Autopilot eingesetzt wurde.[31] Mit der Digitalisierung der einzelnen Instrumente ergab sich ein zusätzliches Problem: Bei jeder sicherheitskritischen Software, wie sie z.B. auch im Auto in Form der Antiblockiervorrichtung ABS vorkommt, muss ein hoher Grad an Fehlerfreiheit nachgewiesen werden.[32] Diese aufwändige, aber zwingend erforderliche Prozedur ist der Hauptgrund, weswegen neuronale Netzwerke in diesem Bereich noch nicht eingesetzt werden: Bei diesen ist es derzeit noch nicht möglich, einen formalen Beweis der erforderlichen Fehlerfreiheit zu erbringen. Ausnahmen stellen Prototypen wie das „Sikorsky Autonomy Research Aircraft“ dar. In einem Versuchshelikopter von Sikorsky, einer Tochtergesellschaft des US-Rüstungsherstellers Lockhead Martin, wird ein auf KI basierender Autopilot in einen Helikopter integriert. Neuronale Netzwerke werden aber auch hier nur für das Erkennen naheliegender Objekte bzw. für die Bodenmessung verwendet.[33]

2.3 Sprachassistenten

Ein weiteres Beispiel klassischer künstlicher Intelligenzen ist der von Apple entwickelte Sprachassistent „Siri“. Ähnlich der in Kapitel 1.1 behandelten KI Watson von IBM besteht Siri aus mehreren Modulen, deren Kern auf der Knowledge-Graph Technik basiert.[34] Als Sprachassistent ist Siri keine generelle KI, mit der allgemeine Gespräche geführt werden können, sondern, wie der Name impliziert, ein Assistent bzw. eine Assistentin, die Befehle entgegennehmen und einfache Aufgaben wie die Terminplanung übernehmen kann.[35] Neben diversen von Apple bereitgestellten Standardfunktionalitäten können freie Entwickler ihre eigene Applikation in Siri einbinden.[36] Die Entwickler definieren dafür eigene Sprachbefehle,

[30] Vgl. Munkhdalai, Namsrai, Ho Ryu. (2018). Credit Scoring with Deep Learning. und Joao Bastos. (2008). Credit scoring with boosted decisiontrees.
[31] Rudolf Brockhaus. (2013). Flugregelung.
[32] Robert Luckner. (2006). Flugführungssysteme zur Pilotenassistenz -Was kann man aus der Luftfahrt lernen?
[33] The Verge. (5.3.2019). I flew a helicopter, and then the helicopter flew me.
[34] TechCrunch. (27.5.2009). Siri: A Powerful Virtual Assistant For The iPhone.
[35] Apple. (13.4.2019). Siri.
[36] MacRumors. (13.6.2016). Apple Opens Siri to Third-Party Developers With iOS 10.

die in genau dieser Form vom Benutzer der Applikation gesprochen werden müssen.[37] Die Transkribierung von Sprache zu Text, die zu Beginn der Entwicklung von Siri auf einer klassischen KI-Technik, den verschleierten Markov Modellen, basierte, wurde aufgrund der für die Nutzer frustrierenden Resultate im Jahr 2014 durch ein NN ersetzt.[38] Wie andere klassische KIs auch, war die Entwicklung von Siri sehr kostspielig. Vor der Übernahme durch Apple flossen 200 Millionen US-Dollar über DARPA Subventionen des US-Militärs sowie weitere 24 Millionen US-Dollar von privaten und institutionellen Investoren in die Applikation.[39]

3. Künstliche neuronale Netzwerke

3.1 Grundlegende Funktionsweise

Während bei klassischen KIs jegliches Wissen von dem Entwickler bzw. der Entwicklerin für den jeweiligen Anwendungsfall aufwändig einprogrammiert werden muss, sind NNs in der Lage, dieses weitestgehend selbstständig zu erlernen.[40] Der Fokus der Entwicklung liegt daher darauf, die bestmöglichen Rahmenbedingungen für diese Lernprozesse zu schaffen.

Die Fähigkeit bzw. Funktionalität des derartigen Lernens basiert, wie der Name andeutet, auf der Nachbildung von Neuronen, analog den Gehirnen von Menschen und Tieren.[41]

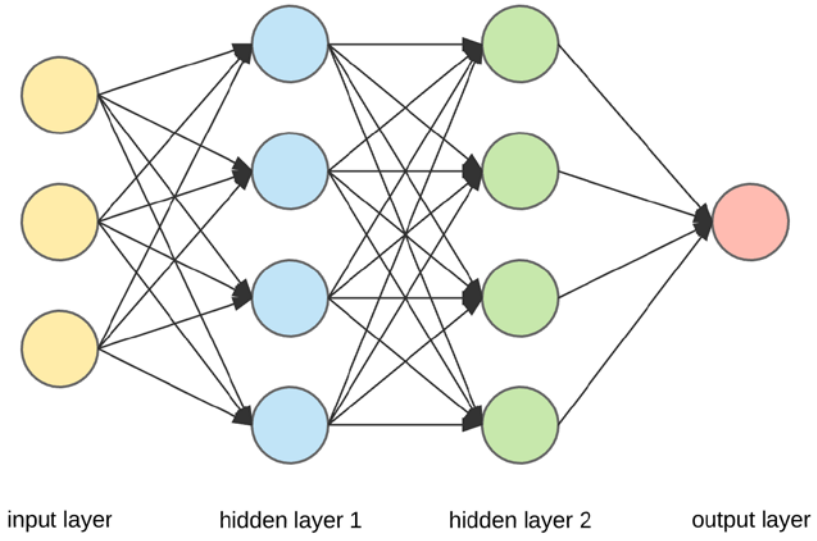

Abbildung 2: Beispiel eines einfachen Neuronalen Netzwerks[42]

[37] Alfian Losari. (26.11.2018). Building an Interactive Voice App Using Custom Siri Shortcuts in iOS 12.
[38] Wired. (24.8.2016). The iBrain Is Here—and It's Already Inside Your Phone.
[39] TechCrunch. (4.2.2010). Siri's IPhone App Puts A Personal Assistant In Your Pocket.
[40] Encyclopaedia Britannica. (15.6.2019). Neural Network.
[41] Encyclopaedia Britannica. (15.6.2019). Neural Network.
[42] Medium. (29.4.2019). Applied Deep Learning - Part 1: Artificial Neural Networks.

Diese künstlichen Neuronen werden dabei zu einem Netzwerk mit mehreren Schichten zusammengefasst. Wie in Abbildung 2 zu sehen ist, ist dabei in der grundlegendsten Variante jedes künstliche Neuron mit jedem Neuron der Folgeschicht verknüpft, wobei die Anzahl der Schichten unterschiedlich sein kann.[43] Die Aufgabe jedes einzelnen künstlichen Neurons besteht in der Gewichtung der eingehenden Daten und der Weiterleitung des daraus entstandenen Ergebnisses an alle mit ihm verbundene Neuronen.[44]

In der ersten Schicht, die Eingabeschicht bzw. „*input layer*" genannt wird, werden die zu verarbeitenden bzw. zu lernenden Daten angelegt - diese spiegelt somit die Sinnesorgane wider.[45] Die folgenden Schichten, die in Abbildung 2 als „*hidden layer*" bzw. verborgene Schichten bezeichnet werden, sind für die Verarbeitung der Eingabedaten zuständig.[46] Vereinfachend dargestellt kann man jede verborgene Schicht als eine Abstraktionsebene der Daten verstehen.[47] Dieses Konzept der Abstraktionsebenen lässt sich leicht durch das Beispiel eines NN veranschaulichen, das Fahrzeuge in Bildern erkennen soll. Jede verborgene Schicht versucht ein Muster in den eingehenden Daten zu erkennen. Der erste *hidden layer* wird so vor allem ausgeprägte Linien und Kurven in den eingehenden Bilddaten (Pixeln) analysieren, während die zweite verborgene Schicht aus diesen Linien und Kurven konkrete Formen ableitet. Die Neuronen einer weiteren späteren Schicht spezialisieren sich in Folge auf die Erkennung der allgemeinen Bestandteile von Autos, wie z.B. der Tür, der Reifen oder der Windschutzscheibe. Noch spätere Schichten sind dann für das Erkennen einzelner Aufbauarten (Sportwagen, Kombi, Lieferwagen, etc.) bzw. Automarken zuständig, bis bei der letzten Schicht, die als Ausgabeschicht bzw. „*output layer*" bezeichnet wird, die Ergebnisse ausgegeben werden. Werden Bilddaten verwendet, so können diese Abstraktionsebenen auch visualisiert werden. *(siehe Kapitel 5.4)*

Der Name „*hidden layer*" lässt sich darauf zurückführen, dass es zu Beginn der Entwicklung von NNs in dieser Phase nicht möglich war, die Inhalte dieser Abstraktionsebene darzustellen, wodurch deren Inhalt verborgen bzw. „*hidden*" war.[48] *(vgl. Kapitel 5: Ein Blick in die Blackbox)*

3.2 Der Lernprozess

Grundsätzlich arbeiten die künstlichen Neuronen nur mit Zahlen, d.h. alle zu verarbeitenden Daten, wie Bilder und Musik, müssen in Zahlenform dargestellt an die Eingabeschicht übergeben werden. Zur Veranschaulichung wird wieder von dem NN in Abbildung 2 ausgegangen, das Autos in Bildern erkennen soll. Des Weiteren wird hier angenommen, dass immer nur einzelne Zahlen - und keine Vektoren, wie häufig in der Praxis verwendet - als Daten verwendet werden. Das NN aus Abbildung 2 weist im *output layer* lediglich ein einzelnes Neuron auf und ist lediglich in der Lage, eine Fragestellung zu behandeln. Es kann daher nur aussagen, ob sich ein Auto in dem Bild befindet oder nicht. Antworten erfolgen jedoch nicht binär in Form einer Ja- oder Nein-Antwort, sondern in einer Zahl, die das

[43] IBM Developer. (30.6.2019). A neural networks deep dive.
[44] Medium. (29.4.2019). Applied Deep Learning - Part 1: Artificial Neural Networks.
[45] IBM Developer. (30.6.2019). A neural networks deep dive.
[46] University of Wisconsin Madison. (15.6.2019). A Basic Introduction To Neural Networks.
[47] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. (2011). Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks.
[48] Stand Out Publishing. (30.6.2019). Hidden Layer.

Ausgabeneuron ausgibt und der Wahrscheinlichkeit einer Ja-Antwort der Aufgabenstellung entspricht.[49] Das Ziel des Lernprozesses ist daher die Anpassung der Gewichte der einzelnen Neuronen zur Erzielung einer möglichst großen Zahl oder Wahrscheinlichkeit als Resultat, wenn sich ein Auto im Bild befindet, bzw. einer möglichst kleinen Zahl, wenn dies nicht der Fall ist.[50]

Da initial alle Gewichte mit Zufallszahlen belegt werden, ist auch das Ergebnis des ersten Bildes zufällig.[51] Ist auf dem Bild ein Auto zu sehen, wäre das richtige Ergebnis eine besonders hohe Zahl (z.B. 100 für 100%). Um nun das NN an dieses Ergebnis anzupassen - also aus dem Bild lernen zu lassen -, werden die einzelnen Gewichtungen der Neuronen von hinten nach vorne an das gewünschte Resultat von 100 angepasst.[52] Zuerst wird daher die des Ausgabeneurons so angepasst, dass den Neuronen der vorletzten Schicht (vgl. Abbildung 2 in grün dargestellt), die hohe Werte gesendet haben und die damit richtig lagen, ein höheres Gewicht und den falsch liegenden Neuronen ein niedrigeres Gewicht zugeordnet wird. Dieser Vorgang wird „Backpropagation“ genannt und schrittweise für alle Neuronen bis auf die der Eingabeschicht (sie haben keine Gewichte) durchgeführt.[53] Die Gewichtskorrektur bleibt dabei nicht immer gleich stark. Bei den ersten analysierten Bildern fällt diese sehr stark aus, um sich schnell von den zufälligen Ergebnissen weg entwickeln zu können. Im Verlauf des Lernprozesses fällt die Korrektur immer geringer aus.[54] Dies entspricht dem sinkenden Anteil der auf das aktuell zu analysierende Bild entfallenden Information bezogen auf das Wissen des NN: Nach beispielsweise tausend Lernprozessen soll der nächste Schritt geringere Abweichungen vom bereits Gelernten bewirken können, als dies beim zehnten Lernschritt der Fall ist.[55] Ohne diese Adjustierung der Gewichtung würde das letzte Auto sprichwörtlich zu stark ins Gewicht fallen, da bereits korrekt erworbene Aussagen über Autos zugunsten des aktuellsten Bildes wieder verlernt werden würden.

Damit das NN einen möglichst situationsunabhängigen Eindruck davon bekommt, was ein Auto kennzeichnet, wird dieser Vorgang mit möglichst vielen Bildern wiederholt. Da NNs fast immer mehr Daten für den Lernprozess benötigen als zur Verfügung stehen, werden diese, meist zufällig neu angeordnet, in beliebig vielen Durchgängen, die „*epoch*“ bzw. Epochen genannt werden, erneut verwendet.[56]

Zur Überprüfung des Lernfortschrittes wird immer wieder - normalerweise nach einer fixen Anzahl gelernter Datensätze - ein Test mithilfe eines Testdatensatzes durchgeführt, der die Genauigkeit des NN misst.[57] Die Testdaten sind dabei nicht Teil der Trainingsdaten um zu verhindern, dass das NN diese auswendig lernt.

[49] Stanford University UFLDL. (30.6.2019). Multi-Layer Neural Network.
[50] Jürgen Schmidhuber. (2014). Deep Learning in Neural Networks: An Overview.
[51] Jürgen Schmidhuber. (2014). Deep Learning in Neural Networks: An Overview.
[52] Stanford University UFLDL. (30.6.2019). Multi-Layer Neural Network.
[53] Michael Nielsen. (15.6.2019). Neural Networks and Deep Learning: Chapter 2
[54] Michael Nielsen. (15.6.2019). Neural Networks and Deep Learning: Chapter 1
[55] Sagar Sharma. (23.9.2017). Epoch vs Batch Size vs Iterations.
[56] Sagar Sharma. (23.9.2017). Epoch vs Batch Size vs Iterations.
[57] TensorFlow. (30.6.2019). Train your first neural network: basic classification.

3.3 Eigenschaften

Wie in dem oben genannten Beispiel dargestellt, eignen sich NNs besonders für fokussierte und exakt definierte Aufgaben. Durch das eigenständige Lernen fällt das für klassische KIs benötigte Expertenwissen größtenteils weg und die langwierige Entwicklungszeit wird durch vergleichsweise kurze Lern- bzw. Trainingsphasen des NN ersetzt.

Dieser einfache Lösungsansatz hat jedoch seinen Preis. Neben der bereits erwähnten Lernineffizienz - NNs benötigen auch für unkomplizierte Anwendungen eine überaus große Menge an Daten - besteht auch eine Redundanz bzw. eine Ineffizienz bei dem Lerninhalt. Dadurch werden viele Muster gelernt, die nichts mit der eigentlichen Problemstellung zu tun haben. *(siehe Kapitel 5.5)* So ist es beispielsweise wahrscheinlich, dass das zuvor beschriebene NN auch das Vorhandensein einer Straße erkennt, da diese bei Bildern mit Autos häufig vorkommt. Daher sind NNs in ihrer Lösung bei weitem nicht so effizient wie klassische KIs und eine manuelle Optimierung ist aufgrund der bereits angesprochenen Intransparenz der *hidden layer* nur schwer möglich. *(siehe Kapitel 5.4)* Kompensiert wird dies vor allem durch die Optimierung des „Lernmaterials", also der dem NN zur Verfügung stehenden Daten. Dabei wird versucht, Überflüssiges soweit wie möglich bereits vorab zu filtern und den Datensatz so zu wählen, dass er möglichst repräsentativ für den Anwendungsbereich ist. Bezogen auf das vorherige Beispiel werden demnach nicht nur die verschiedensten Marken, Modelle, Aufbauten und Farben von Autos verwendet, sondern auch die verschiedensten Umgebungen: Neben Bildern auf Straßen und Asphalt finden sich auch solche in der Natur, z.B. auf Wiesen. *(siehe Kapitel 5.3)*

Während der Trainingsphasen eines NN herrscht für den Entwickler ein Stillstand, da endgültige Resultate bzw. die Genauigkeit des NN erst gegen Ende bekannt sind. Daher wird versucht, diese Stunden bis Tage dauernden Phasen mit teurer Hardware zu beschleunigen. Gängige Heimcomputer reichen derzeit auch für einfache Aufgaben, wie die im Beispiel genannte Bilderkennung, nicht wirklich aus. Forscher und Hobbyisten verwenden daher für die Entwicklung und das Trainieren von NNs vor allem Hochleistungs-Grafikkarten. Diese Grafikkarten, die auch in PCs, Spielelaptops und Spielkonsolen verbaut werden, sind eigentlich für die anspruchsvolle Darstellung der Bildinhalte zuständig bzw. finden im professionellen Bereich der Videobearbeitung und in der Industrie für das detailgetreue Modellieren von komplexen Systemen - wie z.B. Autos - Anwendung. Die Preise für die Topmodelle des Marktführers Nvidia belaufen sich im Handel derzeit auf mehrere tausend Euro[58] für industrielle Anwendungen, wobei jeweils mehrere dieser Grafikkarten zusammengeschalten werden können.[59] Für einfachere NNs sind zwar auch für Konsumenten ausgelegte Grafikkarten mit Kosten unter 500€ ausreichend[60], diese müssen jedoch in einem PC verbaut werden, der neben der notwendigen Stromversorgung und der Kühlung auch leistungsfähig genug sein muss, um die Grafikkarten mit den notwendigen Datenmengen versorgen zu können. Diese Eigenschaften treffen auf herkömmliche Heimcomputer eher nicht zu.

Die Kosten für die Entwicklung von NNs sind für einen industriellen Anwender zwar deutlich geringer als das längerfristige Anstellen von Experten, welche für klassische KIs benötigt

[58] Geizhals. (1.5.2019). Grafikkarten » PCIe mit GPU Workstation nach Erscheinung: Quadro RTX 8000.
[59] Nvidia. (1.5.2019). NVIDIA NVLink High-Speed GPU Interconnect. *Für Professionelle Produkte bzw:* Nvidia. (1.5.2019). RTX. IT'S ON. GeForce RTX 2080 Ti. *Für Konsumenten-Produkte.*
[60] Towards Data Science. (1.5.2019). RTX 2060 Vs GTX 1080Ti Deep Learning Benchmarks.

werden, dennoch verhinderten die Hardwareanforderungen zu Beginn der Forschung, dass mit NNs versehene Anwendungen für den täglichen Gebrauch den Sprung von der Forschung in den Markt schaffen konnten - dies obwohl Endkunden eine weitaus weniger leistungsfähige Hardware benötigen als Entwickler, da sie die Daten nur einmalig durch das Netzwerk schleusen und den oben angeführten Lernprozess nicht mehrere hunderttausend bis Millionen Mal durchführen müssen. Ein Grund für die hohen Hardware-Kosten bzw. die hohen Anforderungen von NNs ist, dass die im herkömmlichen PC zum Einsatz gelangende Hardware, insbesondere die Grafikkarten, nicht für NNs optimiert ist, sondern für die jeweilige originär zu erfüllende Aufgabe. Die oben genannten Ende 2018 erschienen Modelle von Grafikkarten sind die erste für den Konsumentenmarkt konzipierte Serie mit inkludierten Tensor-Kernen. Derartige Kerne können die typischerweise für NNs benötigten Rechenoperationen um ein Vielfaches beschleunigt ausführen.[61] Die günstigste Grafikkarte dieser Serie (RTX 2060) ist bezogen auf die für NNs benötigten Leistungswerte gleichauf mit dem nahezu doppelt so teuren Modell der Vorserie, welches 2016 auf den Markt gebracht wurde[62].

Qualcomm, der Marktführer für Smartphone-Prozessoren, verbaut im Snapdragon 855, dem Hochleistungschip für 2019, der sich vermutlich überwiegend in den Smartphone-Topmodellen dieses Jahres finden wird, seinen ersten dedizierten Tensor-Accelerator.[63] Für Qualcomm ist dies die vierte Generation seiner NN-Beschleunigung, bisher wurden NNs jedoch auf dem Prozessor, der integrierten Grafikkarte und dem Modul für Bildverarbeitung ausgeführt.[64] Zur Anwendung kommen NNs auf Smartphones derzeit für die Nachbearbeitung von Bildern in Echtzeit, z.B. für Schönheitsfilter bzw. um die Bildqualität näher an die vollwertiger Kameras zu bringen, sowie für die Stimmerkennung und -transkribierung für Sprachassistenten. Zusätzlich werden NNs auch von Herstellern verwendet, um aus dem individuellen Nutzerverhalten zu lernen und dieses vorhersagen zu können. Dadurch können im Hintergrund laufende Apps besser für Energiesparzwecke verwaltet werden und die wahrgenommene Wartezeit beim Öffnen von Apps wird verringert durch das im Hintergrund vorgenommene Vorausladen.

Reine Tensor-Acceleratoren können seit 2017 vom Prozessorhersteller Intel bzw. seit kurzem auch von Google erstanden werden.[65] Der Kostenpunkt dieser im USB-Stick Format gehaltenen Acceleratoren liegt bei unter 100 €[66]. Diese eignen sich jedoch nur für den Einsatz simpler NNs, da deren Anwendungszweck vor allem in der Ausführung und nicht im Trainieren von NNs liegt und sie sich vor allem an Hobbyanwender im Smart-Home bzw. im Drohnen- und Robotik-Bereich richten. Tensor-Acceleratoren für Entwickler existieren zwar ebenfalls, diese werden aber hauptsächlich vom Hersteller selbst, z.B. von Google, auf Mietbasis in der Cloud zur Verfügung gestellt.[67]

Das Arbeiten mit NNs wird durch das Vorhandensein von Vorlagen bzw. bewährten Modellen für einfache Aufgaben erleichtert. Diese Basismodelle stellen abgeänderte

[61] Nvidia. (1.5.2019). Tensor Cores.
[62] Tim Dettmers. (1.5.2019). Which GPU(s) to Get for Deep Learning.
[63] Anandtech. (5.12.2018). The Qualcomm Snapdragon 855 Pre-Dive.
[64] Anandtech. (5.12.2018). The Qualcomm Snapdragon 855 Pre-Dive.
[65] Heise. (21.7.2017). 1-Watt-Rechenstick: Movidius Neural Compute Stick für maschinelles Sehen.
[66] Toms Hardware. (5.3.2019). Google's Edge TPU Machine Learning Chip Debuts in Raspberry Pi-Like Dev Board.
[67] Google. (2.5.2019). Cloud TPU Preise.

Varianten des in Abbildung 2 gezeigten Grundnetzwerkes dar, die auf das Lösen bestimmter Anwendungsbereich hin optimiert wurden.

Convolutional Neural Networks (CNN)

Convolutional Neural Networks werden vor allem bei Aufgabenstellungen verwendet, die Bilder als Rohdaten verwenden. Die Idee für dieses Modell bzw. Schema wurde in den 50er und 60er Jahren in der Hirnforschung im Rahmen der Untersuchung des visuellen Kortex von Säugetieren entwickelt.[68] Dabei stellten die Forscherinnen und Forscher fest, dass die Neuronen des visuellen Kortex nicht auf alle Bildpunkte ansprechen, wie es bei einem herkömmlichen NN-Modell der Fall wäre, sondern nur auf einen Bildausschnitt, der mit dem anderer Neuronen überlappt.[69] In CNNs wird diese Erkenntnis wie folgt umgesetzt:

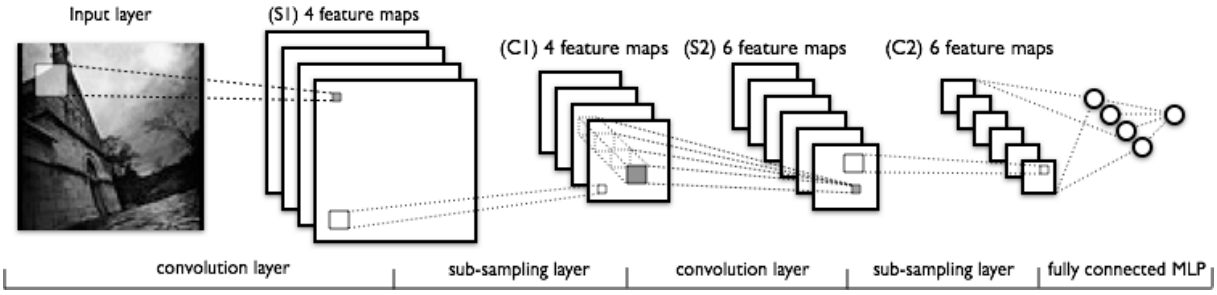

Abbildung 3: Veranschaulichung eines Convolutional Neural Network(CNN)[70]

Jedes Neuron stellt einen sogenannten „Filter" dar, der, wie in Kapitel 3.1 beschrieben, für das Erkennen eines Musters der Abstraktionsebene verantwortlich ist, z.B. das Erkennen verschiedener Kanten/Linien in der ersten Ebene. Dieses Neuron bekommt nun nur den ersten quadratischen Ausschnitt des Bildes (links oben), also z.B. die ersten 50x50 Pixel, zu sehen. Das Neuron feuert umso stärker, je größer die Übereinstimmung des Bildausschnittes mit dem Muster ist. Danach verschiebt sich der Bildausschnitt um einen Pixel nach rechts, und der Vorgang wiederholt sich, bis das gesamte Bild zeilenweise abgetastet wurde. Der Vorgang wird dabei für jedes Neuron, also jeden Filter, wiederholt.[71] Diese Filter sind in Abbildung 3 als hintereinander liegende „*feature maps*" bezeichnet.

Um diesen Prozess zu optimieren, wird zwischen jedem dieser für das Filtern zuständigen *hidden layer* ein sogenannter „*downsampling*"- bzw. „*subsampling layer*" eingefügt. Dieser reduziert die Anzahl der weitergegebenen Informationen, indem er sie zusammenfasst. Zusätzlich wird dadurch auch die Abstraktion weiter gefördert.[72]

Nach einer gewissen Anzahl von Filter-Layern und Subsampling-Layern wird nun ein herkömmliches NN eingefügt, welches in Abbildung 3 als „*fully connected MLP*" bezeichnet

[68] Daphne Cornelisse. (5.5.2019). An intuitive guide to Convolutional Neural Networks.
[69] Daphne Cornelisse. (5.5.2019). An intuitive guide to Convolutional Neural Networks.
[70] Skymind. (5.5.2019). A Beginner's Guide to Convolutional Neural Networks (CNNs).
[71] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do and Kaori Togashi. (2018). Convolutional neural networks: an overviewand application in radiology.
[72] Skymind. (5.5.2019). A Beginner's Guide to Convolutional Neural Networks (CNNs).

ist, und das schlussendlich die Zuordnung der abstrakten Informationen zur Aufgabenstellung vornimmt.[73]

Recurrent Neural Network (RNN)

Die bisher dargestellten Modelle befassen sich jedoch nicht mit einer in der Realität häufig relevanten Dimension, nämlich der Zeit. Nach Abbildung 2 modellierte NNs können daher nicht von vergangenen Eingabedaten auf die aktuellen bzw. sogar auf zukünftige Daten schließen. Dieses Wahrnehmen von Reihenfolgen ist vor allem für Problemstellungen wichtig, die sich mit Videos, Sprache und Text beschäftigen.

Um das NN mit einem gewisses Zeitgefühl auszustatten, verleihen RNNs jedem Neuron in den *hidden layers* eine auf sich selbst zeigende Verbindung (Schleife), die das aktuelle Ergebnis als zusätzlichen Eingabewert für das Neuron im nächsten Schritt, also den nächsten Daten, zur Verfügung stellt und dadurch eine Art Gedächtnis repräsentiert.[74] Dies führt in der Modellierung zu einer Erinnerungsfunktion von kurzer Dauer, also einem Kurzzeitgedächtnis entsprechend, da mehrere Schritte zurückliegende Informationen so gut wie nicht mehr ins Gewicht fallen.[75]

Auch diese Aufgabenstellung kann durch eine Änderung des Aufbaus der einzelnen Neuronen gelöst werden. Eine Möglichkeit sind sogenannte „*long short-term memory*"-Netzwerke, eine Abwandlung von RNNs, bei denen jedes Neuron eine zusätzliche Gedächtniszelle besitzt, die über Schranken bei bestimmten Ereignissen gelöscht und neu beschrieben werden kann und so ein Langzeitgedächtnis simuliert.[76]

Die „Destillier"-Methode

Zur Verwendung von NNs auch auf weniger leistungsstarken Computern bzw. mobilen Geräten ohne Tensor-Acceleratoren wurde die sogenannte „Destillier"-Methode entwickelt. Diese zielt darauf ab, ein bestehendes Netzwerk zu verkleinern, ohne maßgebliche Einbußen in dessen Genauigkeit hinnehmen zu müssen.[77]

Dazu wird zuerst ein herkömmliches, großes NN mit den bestehenden Trainingsdaten befüllt. Danach wird ein kleineres NN, welches eine geringere Anzahl von Neuronen und Schichten aufweist, mit denselben Trainingsdaten entwickelt. Dabei werden jedoch nicht die ursprünglichen, vom Entwickler festgelegten Klassifizierungen der Daten verwendet, sondern die von dem großen NN generierten.[78] Erhält ein zum Zweck der Objekterkennung in Bildern entwickeltes NN als Trainingsbild das Foto einer Katze, so ist die ursprüngliche, händisch durchgeführte Klassifizierung „100% Katze". Alle Abweichungen des NN zu dieser Klassifizierung werden dann über Backpropagation an dieses Ergebnis angepasst. Diese Anpassung erfolgt jedoch niemals perfekt, und daher ist das Ergebnis des NN nach dem Training beispielsweise „80% Katze", „10% Tiger" und „10% Leopard". Genau dies wird nun als Klassifizierungsziel für das kleine NN verwendet. Dadurch wird das kleine NN nicht

[73] Daphne Cornelisse. (5.5.2019). An intuitive guide to Convolutional Neural Networks.
[74] Suvro Banerjee. (7.5.2019). Recurrent Neural Networks and LSTM.
[75] Christopher Olah. (7.5.2019). Understanding LSTM Networks.
[76] Vgl. Christopher Olah. (7.5.2019). Understanding LSTM Networks.
[77] Geoffrey Hinton, Oriol Vinyals, Jeff Dean. (2015). Distilling the Knowledge in a Neural Network.
[78] Geoffrey Hinton, Oriol Vinyals, Jeff Dean. (2015). Distilling the Knowledge in a Neural Network.

darauf hintrainiert, die Daten den Vorgaben des menschlichen Entwicklers entsprechend zu interpretieren, sondern möglichst nahe an die Klassifizierung des großen NN heranzukommen. Daher ist das optimale Training des großen NN ein entscheidender Faktor dieser Methode.

Diese derart erstellten Modelle bzw. Methoden sind leicht verfügbar und können über Module leicht in verschiedenste Entwicklungsumgebungen eingebunden werden. Dies hat dazu geführt, dass Unternehmen wie Microsoft als Teil ihres Angebots Werkzeuge zum Training von NNs in einer Cloud Umgebungen bereitstellen, mittels derer die Auswahl des Modells sowie viele Teile des Entwickelns von NNs automatisiert vorgenommen werden können, ohne dass der Kunde eine Zeile Code schreiben muss.[79]

Generell fokussiert sich die Weiterentwicklung von NNs einerseits auf das Optimieren bereits bestehender Modelle bzw. Vorlagen und andererseits auf die Erstellung neuer NN-Vorlagen bzw. die Anpassung vorhandener NNs an spezielle Anforderungen, die meist für Großunternehmen, wie z.B. bei der Entwicklung selbstfahrender Autos, relevant sind.

3.4 Grenzen neuronaler Netzwerke (das Chinese Room Argument)

Künstliche neuronale Netzwerke sind prinzipiell dazu in der Lage, alle in Kapitel 1.2 genannten Komponenten von Intelligenz, also Lernen, Mutmaßen, Wahrnehmen, Problemlösen und Sprachverständnis, zu erfüllen. Dennoch sind manche Philosophen der Meinung, dass es nicht möglich ist, mithilfe eines Computers künstliche Intelligenzen zu erschaffen. Das wohl bekannteste Argument für diesen Standpunkt ist das des Chinesischen Raumes („Chinese Room Argument“), welches von dem Philosophen John Searle erstmals 1980 publiziert wurde und wie folgt in der „Stanford Encyclopedia of Philosophy“ beschrieben ist[80]:

Ausgangspunkt für das Argument ist der in Kapitel 1.1 geschilderte Turing-Test, der in diesem Gedankenspiel in Mandarin durchgeführt wird. Anstatt mittels eines Computers wird das Programm bzw. die KI in diesem Fall Schritt für Schritt händisch „ausgeführt“. Nun könnte diese Person, die kein Mandarin beherrscht, über das abgearbeitete Programm den Turing-Test bestehen. Searle argumentiert nun, dass aufgrund dessen, dass die abarbeitende Person kein Mandarin beherrscht, auch das Programm, also die KI, kein echtes Verständnis von Mandarin bzw. den Inhalten der Konversation besitzt, sondern nur den Anschein erweckt.

Für klassische KIs scheint dieses Argument schlüssig, da das Wissen und das Verhalten der KI von einem Menschen, dem Entwickler oder der Entwicklerin, „verabreicht“ wurde. Die scheinbare Intelligenz dieser KIs ist ein bis ins Detail durchdachter komplexer Algorithmus, der die in seinem Programm enthaltene Intelligenz nur widerspiegelt, nicht aber von sich aus intelligent ist.

Für NNs kann diesem Argument jedoch einiges entgegengesetzt werden. Nehmen wir beispielsweise an, dass es sich um ein NN handelt, das Katzen in Bildern erkennen soll. Die

[79] Tech Crunch. (3.5.2019). Microsoft launches a drag-and-drop machine learning tool.
[80] Stanford Encyclopedia of Philosophy. (14.4.2019). The Chinese Room Argument.

Bilder, die das NN analysiert, sind hierbei äquivalent zu den Sinneseindrücken, die das menschliche Auge beim Betrachten der Bilder an das Gehirn weiterleitet, da nicht argumentiert werden kann, dass Menschen mit eingeschränkten Sinnesorganen weniger intelligent wären. Wie in Kapitel 3.2 bzw. Kapitel 3.3 beschrieben, lernt nun das NN, Katzen in Bildern zu erkennen. Im Gegensatz zu klassischen KIs sind NNs in der Lage, die sensorischen Inputs bzw. Daten zu „erleben", d.h. das NN könnte unter dem Begriff „Katze" dasselbe „verstehen" wie ein Mensch und kennt nicht nur die an sich nichts bedeutende Buchstabenfolge „Katze", welche maximal mittels eines Knowledge-Graphs mit vereinzelten anderen Begriffen in Beziehung steht.

Diese „Erlebnisfähigkeit" lässt sich vor allem durch die in dieser Arbeit als „Halluzinieren" bzw. „Träumen" bezeichnete Methode verargumentieren *(vgl. Kapitel 5.4)*. Dabei kann ähnlich der in Kapitel 3.2 beschriebenen „Backpropagation" das NN in verkehrter Reihenfolge durchlaufen werden. Wird nun der Ausgabewert „Katze" bzw. „Ja auf dem Bild ist eine Katze" mit 100%-iger Wahrscheinlichkeit als Eingabewert der Backpropagation angelegt, kann ein Bild erzeugt werden, dass eine aus Sicht des NN typische Katze zeigt. Das NN entwickelt somit aus den Eingabebildern das Konzept einer archetypischen Katze. Es hat so wie ein Mensch zu dem Begriff „Katze" ein passendes „Bild im Kopf" und kann sich daher unter dem Begriff etwas vorstellen. Dabei sollte berücksichtigt werden, dass aktuelle NN-Modelle grundsätzlich nicht dieselbe Erlebnisfähigkeit wie Menschen entwickeln, d.h. die „Welt" anders wahrnehmen, wie die in Kapitel 5.5 beschriebenen Angriffsmöglichkeiten auf NNs zeigen.

4. Anwendungen und Brennpunkte

Die leichte und kostengünstige Entwicklung von NNs und deren besondere Eigenschaft, ähnlich wie Menschen lernen zu können, hat dazu geführt, dass diese in vielen Anwendungsbereichen, in denen die viel teureren, da empirisch von Experten entwickelten klassischen KIs keinen Einzug finden konnten, großflächig getestet werden bzw. bereits heute zum Einsatz kommen. Zu immer mehr konkreten Fragestellungen - wie z.B. bei der Vorhersage von Herzanfällen - sind NNs der menschlichen Expertise überlegen.[81] Doch bei weitem nicht alle Anwendungen sind gesellschaftspolitisch unumstritten, vor allem militärische Verwendungszwecke sorgen immer wieder für Diskussionen. Den entscheidenden Punkt stellt hier vor allem die Frage der automatisierten Entscheidungsfindung der KI dar, also ob sie einer Person nur als unterstützendes Werkzeug dienen oder selbst bereits zwischen Handlungsalternativen wählen. In der Realität scheint die Grenze zwischen diesen zwei Arten häufig zu verschwimmen, wie man an den in diesem Kapitel betrachteten kontroversen Beispielen erkennen kann.

[81] Science Magazine. (14.4.2017). Self-taught artificial intelligence beats doctors at predicting heart attacks.

4.1 Selbstfahrende Autos

2011 wurde Nevada zum ersten US-Bundesstaat, der selbstfahrende Autos explizit auf öffentlichen Straßen erlaubte.[82] Mit der Unterzeichnung eines ähnlichen Gesetzes durch den Gouverneur in der Google-Zentrale in Mountain View folgte Kalifornien ein knappes Jahr später und erlaubte somit den heimischen Technologieunternehmen das Testen autonomer Fahrzeuge vor der eigenen Haustür.[83] Die Testfahrten auf öffentlichen Straßen sorgten weltweit für Aufmerksamkeit und veranlasste Wissenschaftler und Wissenschaftlerinnen des MIT dazu, eine auf selbstfahrende Autos adaptierte Version des Trolley-Problems zu formulieren.[84] Bei dem Trolley-Problem muss ein fiktiver Weichensteller - analog der Software eines autonomen Fahrzeuges - eine schwerwiegende Entscheidung treffen: Eine nicht abbremsbare Straßenbahn rast auf ein Gleis mit fünf Personen zu, kann aber durch Umstellung einer Weiche auf ein anderes Gleis gelenkt werden, auf dem nur eine Person ums Leben kommt.[85]

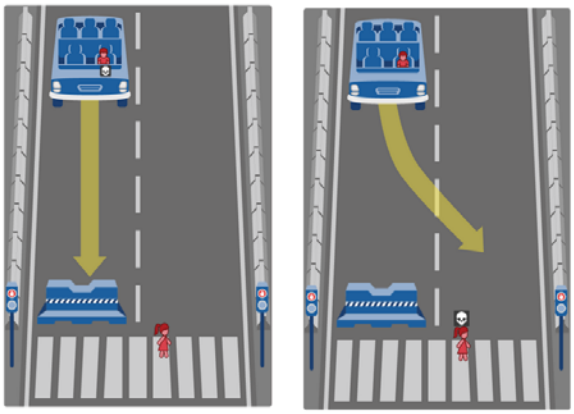

Abbildung 4: Ein Dilemma aus dem Moral Machine Projekt[86]

Wissenschaftlerinnen konnten im Rahmen der „*moral machine*“-Studie ermitteln, dass die meisten Testpersonen in diesem Fall einen utilitären moralischen Ansatz wählen, also zur Minimierung der Opfer das Gleis umstellen würden.[87] In den verschiedenen Varianten der Studie (siehe auch Abbildung 4) wurden möglichst kontroverse, vom geschilderten Basisszenario abgeleitete Dilemmata verwendet, die allesamt keine objektiv „korrekte“ Lösung besitzen. Eine Auswertung der 40 Millionen Entscheidungen, die Besucher und Besucherinnen der Website getroffen haben, bestätigt auf den ersten Blick die intuitive Erwartung: So werden wie beim Trolley-Problem mehr Personen wenigen Personen, Menschen Haustieren, gesetzestreue Personen Verbrechern und Fußgänger Passagieren vorgezogen.[88]

[82] Stanford Law School. (22.6.2011). Nevada Governor Signs Driverless Car Bill Into Law.
[83] BBC. (26.9.2012). Driverless car bill is signed in California at Google headquarters.
[84] Edmond Awad et.al. (2018). The Moral Machine Experiment.
[85] Phillipa Foot. (1967). The Problem of Abortion and the Doctrine of the Double Effect.
[86] MIT Media Lab. (28.3.2019). Moral Machine.
[87] The Guardian. (12.12.2016). The trolley problem: would you kill one person to save many others?
[88] Edmond Awad et.al. (2018). The Moral Machine Experiment.

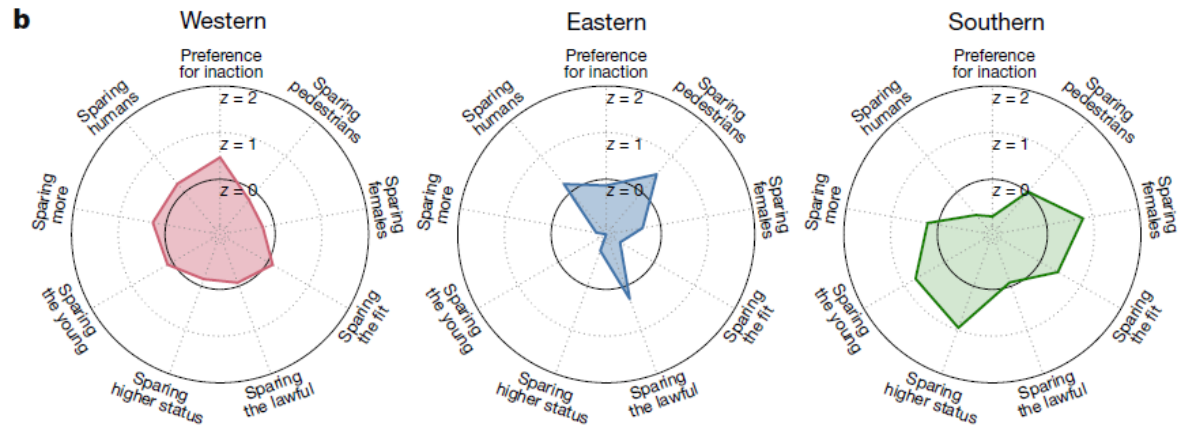

Abbildung 5: Kulturelle Unterschiede beim Moral Machine Experiment[89]

Die Autoren und Autorinnen der Studie fassten die Teilnehmer eines Herkunftslandes zusammen und wendeten auf diese Nationen eine auf neun Parameter bezogene Clusteranalyse an (siehe Abbildung 5). Mit diesem statistischen Verfahren wurden drei signifikant unterschiedliche Cluster identifiziert, die im Wesentlichen der kulturellen Zugehörigkeit einer Nation entsprechen. Der „Western"-Cluster umfasst überwiegend Teilnehmer aus Nationen der westlichen Gesellschaften, also Länder aus Europa sowie Nord-Amerika. Im „Eastern"-Cluster finden sich überwiegend Nationen aus dem islamischen bzw. konfuzianischen Kulturkreis. Zuletzt wird der „Southern"-Cluster überwiegend aus lateinamerikanischen Nationen gebildet. Im Einzelfall werden Nationen jedoch nicht dem erwarteten Cluster zugeordnet: So befinden sich z.B. Frankreich, Ungarn, Tschechien und die Slowakei im „Southern"-Cluster, Polen und Slowenien jedoch im „Eastern"- Cluster, während der Rest Europas – wie erwartet - im „Western"-Cluster liegt. Hierbei muss berücksichtigt werden, dass aufgrund der Art der Datenerhebung über eine Website die Ergebnisse im Detail nicht der Meinung der Gesamtbevölkerung der jeweiligen Nation entsprechen müssen.

Trotzdem lässt sich aus der Studie heraus vermuten, dass keine einheitliche, als ideal empfundene Lösung existiert, die für alle Kulturen moralisch zufriedenstellend ist. Somit könnten sich Autobauer gezwungen sehen, ihre autonomen Fahrzeuge mit lokal unterschiedlichen Softwareversionen auszustatten, die den jeweiligen Bedürfnissen eines kulturellen Gebietes entsprechen würde. Den Autoherstellern in dieser Frage freie Hand zu lassen wäre ebenfalls keine optimale Lösung. Zudem ist davon auszugehen, dass Kunden Fahrzeuge von demjenigen Autohersteller kaufen, der seine Software so ausgerichtet hat, dass in jeder Situation vorrangig die Passagiere des Fahrzeuges gerettet werden. Die Frage der Gestaltung der zukünftigen gesetzlichen Regulierung wird damit vermutlich für PolitikerInnen wie WissenschaftlerInnen ebenfalls zum moralischen Dilemma. Auf EU-Ebene scheint man derzeit noch weit von konkreten gesetzlichen Regelungen bezüglich autonomer Fahrzeuge entfernt zu sein: Bis jetzt konnten sich das Parlament und die Kommission lediglich auf eine Reihe von Bulletpoints einigen.[90]

[89] Edmond Awad et.al. (2018). The Moral Machine Experiment.
[90] European Parlament. (14.1.2019). Self-driving cars in the EU: from science fiction to reality.

4.2 Anwendungen im Rechtsstaat

Die USA besitzen mit mehr als 650 Häftlingen pro 100.000 Einwohner die bei Weitem größte Häftlingsquote im Vergleich zu anderen OECD Staaten.[91] Aufgrund der Größe der USA befinden sich mehr als die Hälfte der Häftlinge weltweit in US-Gefängnissen.[92] Kritikern dieses Systems zufolge ist dies auch Folge des derzeit angewendeten Kautionssystems, das aus deren Sicht bestimmte Bevölkerungsschichten benachteiligt.[93] Um dieses System fairer zu gestalten, soll es im Bundesstaat Kalifornien ab Oktober 2019 radikal geändert werden.[94] Künftig wird für jeden Beschuldigten mithilfe einer künstlichen Intelligenz eine Risikobewertung („*risk assessment*“) durchgeführt, die unter Berücksichtigung verschiedener Faktoren aus dem Lebenslauf des Beschuldigten Entscheidungen trifft.[95] Während Personen mit moderatem Risiko gegen bestimmte Auflagen bis zum Gerichtstermin freigelassen werden, verweilen diejenigen mit hohem Risiko solange in Haft, bis ein Richter oder eine Richterin im Rahmen einer Anhörung die endgültige Entscheidung trifft.[96]
Ähnliche Versuche in anderen US-Bundesstaaten haben vielversprechende Ergebnisse geliefert. In New Jersey wurden 16% und in Virginia nahezu 50% weniger Häftlinge verzeichnet, die auf ihren Gerichtstermin warten mussten - und das ohne Zunahme der Verbrechensraten.[97] Den bis dato ohne den Einsatz von künstlicher Intelligenz eingesetzten Bewertungsschemen wird vielfach eine ethnische Diskriminierung vorgeworfen[98]. Eine von der Stanford University veröffentlichte Arbeit konnte anhand von Simulationen aus Daten von New York City zeigen, dass durch den Einsatz einer künstlichen Intelligenz ohne die offensichtlich diskriminierenden Eingabewerte betreffend Ethnie und Geschlecht eine Verbrechensreduktion von bis zu 27,4% bei gleichbleibender Häftlingsrate bzw. eine Häftlingsreduktion von bis zu 41,9% bei gleichbleibender Verbrechensrate möglich ist.[99]

Obwohl demzufolge der Einsatz von künstlicher Intelligenz die Diskriminierung nach Ethnizität in der Risikobewertung verringern könnte[100], hat das Weglassen bestimmter Eingabeparameter den gegenteiligen Effekt. So gilt es als statistisch erwiesen, dass Frauen eine wesentlich geringere Strafrückfälligkeit besitzen als Männer.[101] Künstliche Intelligenzen, die das Geschlecht des bzw. der Verhafteten nicht berücksichtigen, neigen daher aufgrund des überproportionalen Männeranteils an den Häftlingen dazu, Frauen wie Männer zu behandeln und daher unverhältnismäßig einzuschätzen und zu bestrafen.[102]

Bisherige Algorithmen zur Risikobewertung sind vielfach darauf ausgelegt, als Ergebnis der Analyse immer eine Risikobeurteilung auszugeben. Als unterstützendes Werkzeug für

[91] Statista. (25.3.2019). Incarceration rates in OECD countries as of 2018.
[92] Statista. (25.3.2019). Incarceration rates in OECD countries as of 2018.
[93] The Guardian. (7.9.2018). Imprisoned by algorithms: the dark side of California ending cash bail.
[94] The Guardian. (7.9.2018). Imprisoned by algorithms: the dark side of California ending cash bail.
[95] Washington Post. (29.8.2018). California abolishes money bail with a landmark law. But some reformers think it creates new problems.
[96] Washington Post. (29.8.2018). California abolishes money bail with a landmark law. But some reformers think it creates new problems.
[97] New York Times. (20.12.2017). Even Imperfect Algorithms Can Improve the Criminal Justice System
[98] Pro Publica. (23.5.2016). Machine Bias - There’s software used across the country to predict future criminals. And it’s biased against blacks.
[99] Jon Kleinberg et.al. (2017). Human Decisions and Machine Predictions.
[100] Jon Kleinberg et.al. (2017). Human Decisions and Machine Predictions.
[101] Jennifer L. Skeem et.al., Gender, Risk Assessment, and Sanctioning: The Cost of Treating Women Like Men
[102] Jennifer L. Skeem et.al., Gender, Risk Assessment, and Sanctioning: The Cost of Treating Women Like Men

RichterInnen sind sie damit gut geeignet, in der Realität treten aber Fälle auf, bei denen die vorliegenden Faktoren zu widersprüchlichen Ergebnissen führen. Ein Richter oder eine Richterin, die das größere Gesamtbild wahrnehmen, bewerten in derartigen Situationen aufgrund ihrer Erfahrung gegenläufige Einflussfaktoren unterschiedlich stark und treffen somit vermutlich bessere und nachvollziehbarere Entscheidungen. Dieses Problem lässt sich dadurch lösen, dass Algorithmen bei derartigen Konflikten ihre Entscheidung an eine Person delegieren, anstatt immer ein eindeutiges Resultat zu liefern. Ein moralisches Dilemma bleibt jedoch bestehen: Festgenommene mit einer derartig unklaren Risikoeinschätzung werden bis zu einer Anhörung inhaftiert und damit de facto analog zu Personen mit einer hohen Risikoeinschätzung behandelt. Richter und Richterinnen werden sich zusätzlich bei Anhörungen in einem solchen System aufgrund der Vorselektion des Algorithmus mit tendenziell „schwereren" Straftaten als heute konfrontiert sehen und daher unbewusst zu härteren Maßnahmen neigen – zum Nachteil von Personen mit Parametern, für die der Algorithmus kein eindeutiges Resultat liefert.

Der kalifornische Gesetzestext lässt es den einzelnen Gerichten offen, ob NNs, klassische KIs oder Algorithmen für Risikobewertungen zum Einsatz kommen. Das Verfahren muss lediglich vom Judicial Council abgesegnet werden und wissenschaftlich nachweisbar eine hohe Genauigkeit bei geringstmöglicher Diskriminierung besitzen.[103] Aufgrund der hohen Flexibilität und Genauigkeit sowie der geringen Entwicklungskosten ist damit zu rechnen, dass sich in vielen Gerichten der Einsatz von NNs – und nicht der Einsatz von im Sinne einer klassischen KI empirisch entwickelten Scoringalgorithmen – durchsetzen wird.

Durch den Einsatz der neuen „*risk assessment*" Werkzeuge wird der Arbeitsaufwand kalifornischer RichterInnen in Fällen von Häftlingen, die mit mittlerem oder geringem Risiko eingestuft werden, deutlich reduziert werden. Diese Häftlinge können zudem durch das Fehlen eines möglicherweise voreingenommenen Richters oder Richterin eine statistisch gesehen fairere Behandlung erwarten.[104] Es bleibt jedoch offen, ob diese Arbeitsentlastung den mit hohem Risiko eingestuften Häftlingen, im Sinne einer längeren und vor allem durchdachteren Einschätzung durch den Richter oder der Richterin, zugutekommt oder den Budgetkürzungen zum Opfer fällt. Aufgrund der eher groben Formulierung des Gesetzestextes bezüglich der Qualitätsmerkmale der Risikobewertungs-Werkzeuge sowie der bisher geringen Erfahrung mit diesen, wäre erstes die anzustrebende Variante.

4.3 Militärische Anwendungen

Nicht zuletzt unter dem Einfluss von Filmen wie „2001: A Space Odyssey (1968)"[105] und „Terminator (1984)"[106], die große Zuschauermengen erreichten, führen mögliche militärische Anwendungsmöglichkeiten von KIs zu Diskussionen. Ein Beispiel dafür ist das „Project Maven" des US-Militärs. Ziel des zusammen mit der Technologiefirma Google entwickelten Projektes ist, die Flut an Videomaterial der Überwachungsdrohnen mithilfe von NNs

[103] Senate Bill 10, 2017-2018 Reg. Sess., ch. 244. (Cal . 2018)
[104] Washington Post. (29.8.2018). California abolishes money bail with a landmark law. But some reformers think it creates new problems.
[105] Box Office Mojo. (30.3.2019). 2001: A Space Odyssey.
[106] Box Office Mojo. (30.3.2019). The Terminator.

automatisiert zu analysieren.[107] Das nur für die Entscheidungsunterstützung entwickelte System soll in der Lage sein, Gebäude in Echtzeit zu identifizieren und Verbindungen zu Personen und Fahrzeugen anzuzeigen.[108] Nachdem mehrere Mitarbeiter und Mitarbeiterinnen von Google wegen ethischer Bedenken gekündigt und viele weitere gegen das Projekt protestiert hatten, leitete Google seinen Rückzug aus der Vertragspartnerschaft ein.[109] Beendet ist das Projekt dadurch jedoch nicht. Mit März 2019 wurde „Project Maven" an ein nicht bekannt gegebenes Technologieunternehmen übergeben, wobei Googles grundlegende Cloud-Technologien unverändert zum Einsatz kommen.[110] Auch wenn der Einsatz von sogenannten „Killermaschinen" - also mit künstlicher Intelligenz versehene Waffen, die ohne Steuerung durch einen Menschen tödliche Gewalt ausüben können - zur Zeit noch nicht möglich ist, bleibt zu erwarten, dass die Großmächte die Erforschung und Entwicklung derartiger Roboter aktiv betreiben werden. Ein dementsprechendes UN-Dekret wurde im September 2018 bereits von den USA und Russland blockiert[111] und auch bei einer Konferenz zur Regulierung dieser Waffen in Genua Ende März 2019 stemmten sich Großbritannien, Russland und die USA gegen jegliche regulatorische Maßnahmen.[112]

Der Rückzug großer IT-Unternehmen aus dem militärischen Anwendungsbereich bringt zusätzliche Probleme in die Thematik. So ist zu erwarten, dass die Berichterstattung bzw. die Transparenz über den technischen Fortschritt in diesem Gebiet stark zurückgehen wird. Dies wird vor allem dadurch verstärkt, dass auf den militärischen Bereich spezialisierte Unternehmen keine Endkunden, vor denen sie sich verantworten müssen, betreuen und vermutlich auch keine altruistischen Firmenvisionen besitzen, um talentierte Mitarbeiter und Mitarbeiterinnen anzuwerben. Der niedrige Grad an Transparenz sowie das Fehlen eines zu bewahrenden Image eines Unternehmens lässt befürchten, dass mit künstlicher Intelligenz versehene Waffen schon bald in Bürgerkriegsgebieten auftauchen werden bzw. auch Staaten zur Verfügung stehen, deren Staatsformen und Handlungen nicht mit den Werten westlicher demokratischer Gesellschaften vereinbar sind.

4.4 Microsoft Tay

„Tay" war ein von dem Technologieunternehmen Microsoft entwickelter Chat-Bot, mit dem Nutzer verschiedener Online-Plattformen interagieren konnten.[113] Anhand der im März 2016 in Betrieb genommenen KI sollte ein tieferer Einblick in „Gesprächsverständnis" gewonnen werden.[114] Dabei war die als 18-24 jährige Jugendliche ausgelegte Tay dazu in der Lage, von den Konversationen, die sie auf den Plattformen führte, zu lernen.[115] Die Grundversion wurde

[107] U.S. Department of Defense. (21.6.2017). Project Maven to Deploy Computer Algorithms to War Zone by Year's End.
[108] The Intercept. (04.2.2019). Google Hired Gig Economy Workers to Improve Artificial Intelligence in Controversial Drone-Targeting Project.
[109] Gizmodo. (1.6.2018). Google Plans Not to Renew Its Contract for Project Maven, a Controversial Pentagon Drone AI Imaging Program.
[110] The Intercept. (1.3.2019). Google Hedges on Promise to End Controversial Involvement in Military Drone Contract.
[111] The Independent. (3.9.2018). 'Killer robots' ban blocked by US and Russia at UN meeting.
[112] The Guardian. (29.3.2019). UK, US and Russia among those opposing killer robot ban.
[113] Zeit. (24.3.2016). Twitter-Nutzer machen Chatbot zur Rassistin.
[114] The Guardian. (24.3.2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter.
[115] Frankfurter Allgemeine Zeitung. (24.3.2016). Zum Nazi und Sexisten in 24 Stunden.

von Microsoft-Mitarbeitern unter Einbeziehung von Improvisations-Comedians – man wollte in der Zielgruppe der 18-24jährigen US-Jugendlichen gut ankommen – trainiert.[116]

Abbildung 6: Eine Nachricht der Microsoft Tay-KI auf Twitter[117]

Einige Nutzer des sozialen Netzwerks Twitter verfolgten das Ziel, Tay ein verdrehtes Gedankenbild zu vermitteln. Innerhalb weniger Stunden verbreitete Tay unter anderem Verschwörungstheorien, leugnete den Holocaust und äußerte sich anderwärtig diskriminierend geschützten Personengruppen gegenüber.[118] Viele dieser Aussagen wurden dadurch ausgelöst, dass Nutzer Tay dazu aufforderten, ihre eigenen Aussagen nachzusprechen. Dennoch sorgten Tweets wie die in Abbildung 6 dafür, dass Tay weniger als 24 Stunden nach Beginn des Experiments wieder vom Netz genommen werden musste.[119] Tay war nicht das letzte Experiment mit sozialen Chat-Bots von Microsoft. Ein paar Monate nach Tay ging die KI „Zo“ online.[120] Die ebenfalls als Jugendliche ausgelegte KI fiel jedoch bis jetzt nicht negativ auf.

4.5 Weitere Anwendungsbereiche

Die einfache Entwicklung von künstlichen Intelligenzen sowie die Kostenersparnis, die mit deren Anwendung einhergeht, macht sie für die automatisierte Verarbeitung bürokratischer Prozesse interessant. Neben dem in Kapitel 4.2 genannten und sich bereits in Anwendung befindlichen Beispiel der Behandlung von Tatverdächtigen, verwendet Kanada NNs experimentell dazu, um Prozesse im Ministerium für Immigration zu beschleunigen bzw. zu vereinfachen. [121] Neben der einfachen Zuordnung von Visa-Anträgen auf Routine- und Sonderfälle haben vor allem die Pilotversuche bei Asylanträgen für öffentliche Diskussionen gesorgt. [122] Dabei werden die Anträge automatisiert nach ihrer Legitimität sowie möglichen fälschlichen Angaben analysiert und schlussendlich eine Empfehlung generiert, die zusätzlich ein ebenfalls durch künstliche Intelligenzen bewertetes Risiko bei einer Rückkehr ins

[116] The Guardian. (24.3.2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter.
[117] Medianama. (28.3.2016). Microsoft takes down its AI chatbot which turned evil under human influence.
[118] Medianama. (28.3.2016). Microsoft takes down its AI chatbot which turned evil under human influence.
[119] Zeit. (24.3.2016). Twitter-Nutzer machen Chatbot zur Rassistin.
[120] Microsoft. (16.12.2016). Microsoft and AI: Introducing social chatbot Zo.
[121] CBC. (26.9.2018). Federal use of A.I. in visa applications could breach human rights, report says.
[122] CBC Radio. (16.11.2018). How artificial intelligence could change Canada's immigration and refugee system.

Heimatland einbezieht.[123] Ähnlich wie in Kapitel 4.2 trifft auch hier eine Person die endgültige Entscheidung, es ist jedoch nicht öffentlich bekannt, in welchem Umfang diese Experimente kontrolliert werden bzw. ob potentiell diskriminierende Eingabewerte bei den künstlichen Intelligenzen zur Anwendung kommen.[124]

In der Versicherungsbranche existieren ebenfalls viele mögliche Anwendungen für künstliche Intelligenzen. Neben der gesellschaftspolitisch unkritischen Kommunikation mit Kundinnen und Kunden über Chat-Bots und der automatischen Schadensbewertung mithilfe selbst aufgenommener Fotos bzw. Videos, wird auch mit Anwendungen experimentiert, die als kontrovers angesehen werden können.[125] Zum Beispiel führte das US-Versicherungsunternehmen StateFarm 2016 einen Wettbewerb zur Erkennung abgelenkter Autofahrer anhand von einer „Dash-Cam", also einer Kamera in der Fahrkabine, mithilfe von künstlichen Intelligenzen durch.[126] Mit ihrem „*Drive Safe & Save*"-Programm setzt State-Farm dieses Konzept zwischenzeitlich in die Praxis um.[127] Neben einem Prämienrabatt von ca. 5% für die Teilnahme an dem Programm erhalten Fahrerinnen und Fahrer im Falle einer vorausschauenden defensiven Fahrweise, die durch die sensorischen Daten ihres Mobilfunkgeräts beobachtet wird, weitere Nachlässe in Höhe von bis zu 50% der Versicherungsprämie.[128] Auch in Österreich gibt es bereits ähnliche Produkte. Bei dem SafeLine Tarif der Versicherung Uniqa gibt es für die Teilnahme sogar 10%, es muss jedoch neben der Smartphone-App ein verpflichtender GPS-Sensor ins Fahrzeug verbaut werden.[129] Dash-Cams werden in beiden Fällen noch nicht verwendet, neuere Fahrzeugmodelle hingegen haben diese jedoch oft bereits verbaut, da sie zum Zwecke des teilautonomen Fahrens abgelenkte Fahrer und Fahrerinnen erkennen können müssen. Diese Kameras und die bereits im Fahrzeug verbauten Sensoren werden abgelenkte Fahrer wesentlich genauer bestimmen können als ein Mobilfunkgerät. Nicht zuletzt aufgrund der immer häufiger integrierten, entfernt ausführbaren Softwareupdates für Fahrzeuge scheint es wahrscheinlich, dass in naher Zukunft verstärkt ähnliche Programme angeboten werden, welche die vollen technischen Möglichkeiten moderner, an das Internet angebundener Fahrzeuge ausschöpfen.

5. Ein Blick in die Blackbox

Aufgrund des rapiden technischen Fortschritts sind Bürger westlicher Gesellschaften im Alltag mit immer mehr Technologien und Konzepten konfrontiert, die sie nicht im Detail verstehen und die somit eine Art „Blackbox" für sie darstellen. Das kann einfache Anwender betreffen, die mit Privatsphäre-Themen in großen sozialen Netzwerken wie Facebook nicht umzugehen wissen[130], aber auch Experten, wie die jüngsten Abstürze der Boeing

[123] The Citizen Lab. (September 2018). Bots at the Gate A Human Rights Analysis of Automated Decision Making in Canada's Immigration and Refugee System.
[124] CBC Radio. (16.11.2018). How artificial intelligence could change Canada's immigration and refugee system.
[125] McKinsey. (April 2018). The industry is on the verge of a seismic, tech-driven shift. A focus on four areas can position carriers to embrace this change
[126] Kaggle. (1.4.2019). State Farm Distracted Driver Detection
[127] StateFarm. (1.4.2019). You're in the Driver's Seat When It Comes to Your Discount.
[128] StateFarm. (1.4.2019). Have Drive Safe & Save™ Questions? We've Got Answers.
[129] Uniqa. (1.4.2019). SafeLine - die Autoversicherung, die Leben retten kann.
[130] Wired. (20.12.2018). The 21 (and Counting) Biggest Facebook Scandals of 2018.

Flugzeugreihe 737 MAX[131] zeigen. KIs und im besonderen NNs stellen die Basis der nächsten Technologiewelle dar und werden - wie im vorherigen Kapitel dargelegt – zunehmend in vielfältigen Anwendungsgebieten wie der Rechtsprechung und dem autonomen Fahren eingesetzt.

Dieses Kapitel versucht, einen Einblick in die Blackbox von KIs zu geben und die Werkzeuge zu beschreiben, die ein Verständnis der Funktionsweise einer individuellen KI ermöglichen. Aufgrund des viel höheren Grades an Intransparenz bei NNs wird auf diese Unterkategorie von KIs schwerpunktmäßig eingegangen. Zusätzlich werden auch Fragen wie die des Bestehens von Vorurteilen (Bias) in KIs und aktuelle Angriffsmöglichkeiten behandelt. Diese ermöglichen nicht nur ein tieferes Verständnis für die Funktionsweise von NNs, sondern auch für die zukünftige Sicherheit von NN-Anwendungen, z.B. der des autonomen Fahrens. Dadurch sind diese ein wesentliches Hilfsmittel für die Schaffung zukünftiger rechtlicher Rahmenbedingungen.

5.1 Haben KIs Vorurteile?

Die Frage nach dem Bestehen von „Vorurteilen“ von KIs ist kein neues Thema. Größere öffentliche Diskussionen entstanden bereits zu Beginn dieses Jahrtausends. Beispielsweise argumentierten Wissenschaftlerinnen und Wissenschaftler um Bernard E. Harcourt von der University of Chicago im Jahr 2005 gegen das zwei Jahre zuvor in ihrer Stadt eingeführte „*Predictive Policing*“ System[132],[133]. Bei dieser Anwendung handelte es sich um eine klassische KI. Diese versuchte unter Zugrundelegung von Algorithmen und Expertenwissen aus vorhandenen Statistiken Schlüsse auf zukünftige Verbrechen zu ziehen und diese durch gezielte Polizeipräsenz zu verhindern.[134] Den Kritikern der Anwendung dieser KI zufolge führte dies zu einer unsachgemäßen Fokussierung der polizeilichen Maßnahmen auf gesellschaftliche Randgruppen.

In einer aktuellen Studie aus dem Jahr 2018 stellten Wissenschaftler in einer Kooperation des MIT und Microsoft grobe Schwächen aktueller Gesichtserkennungssysteme in der Zuordnung des Geschlechts von Personen ethnischer Minderheiten fest.[135] Dies führt beispielsweise dazu, dass diese Personengruppen Funktionen moderner elektronischer Geräte wie z.B. das Entsperren eines Smartphones mittels Gesichtserkennung nur eingeschränkt nutzen können.

In der Frage, ob das Bestehen von Vorurteilen inhärent für KIs ist, muss zwischen den zwei in dieser Arbeit in Kapitel 2 und 3 behandelten KI-Ansätzen unterschieden werden:

Klassische KIs spiegeln das Wissen der Entwickler wider. Da die Auswahl der von der KI zu analysierenden Parameter von den Entwicklern definiert wird, ist das Auftreten von Vorurteilen im Entscheidungsprozess der KI wahrscheinlich. Dies ist z.B. der Fall, wenn bei dem in Kapitel 2.1 beschriebenen Credit-Scoring Verfahren in einem Entscheidungsbaum das

[131] BBC. (5.4.2019). Boeing 737 Max: What went wrong?
[132] CIO. (10.5.2019). Chicago Police Department Uses IT to Fight Crime, Wins Grand CIO Enterprise Value Award 2004
[133] Bernard E. Harcourt. (2005). Against Prediction: Sentencing, Policing, and Punishing in an Actuarial Age.
[134] CIO. (10.5.2019). Chicago Police Department Uses IT to Fight Crime, Wins Grand CIO Enterprise Value Award 2004
[135] Joy Buolamwini, Timnit Gebru. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.

durchschnittliche Vermögen bzw. Gehalt der Bewohner eines Bezirkes mit einbezogen werden.

NNs hingegen besitzen keinen vordefinierten Entscheidungsprozess und damit auch keine inhärenten Vorurteile – die Entscheidungsprozesse und damit auch die Vorurteile eignet sich das NN über die Trainingsdaten selbst an. Aber auch die Daten, die für das Trainieren von NNs verwendet werden, müssen von Menschen beurteilt bzw. „klassifiziert“ werden, um für das automatisierte Training verwendbar zu sein. Daher können sich in beiden Kategorien, den klassischen KIs und den NNs, ungewollte (oder auch gewollte) gesellschaftliche Vorurteile wiederfinden.

5.2 Analyse und Bewertung von klassischen KIs

Klassische KIs werden aufgrund ihres hohen Entwicklungsaufwands in den meisten Fällen möglichst einfach gehalten. Dadurch können auch die Ansätze für eine Analyse dieser Algorithmen auf Vorurteile simpel aufgebaut werden.

Stellen wir uns im Folgenden vor, ein Unternehmen führt auf Ersuchen der Rechtsabteilung hin ein „Audit“ des in Verwendung befindlichen Entscheidungsbaums eines Credit-Scoring *(siehe Kapitel 2.1)* durch. Dabei würden die verwendeten Parameter bzw. Fragestellungen auf eine mögliche Diskriminierung überprüft und die Einhaltung der ethischen Richtlinien bzw. des Leitbilds des Unternehmens untersucht werden. Neben der Formulierung der Fragestellung an sich würden dabei auch die Position der Fragestellung (Wie weit oben im Baum befindet sich diese?) und deren Gewichtung untersucht werden.

Schlussendlich muss auch die Zielgenauigkeit dieser KI erneut getestet und mit vorherigen Versionen abgeglichen werden. Hierbei ist natürlich essentiell, dass die KI durch die im Laufe der Zeit stattfindenden Veränderungen des gesellschaftspolitischen Umfeldes möglichst wenig an Genauigkeit einbüßt. Man kann jedoch davon ausgehen, dass geringe Einbußen im Allgemeinen in Kauf genommen werden, da die Daten, anhand deren die KI getestet wird, ebenfalls einen gewissen Bias enthalten.

Sind die Algorithmen komplexer, so muss ein zusätzlicher Schritt unternommen werden: Dabei muss der Entscheidungsfindungsprozess der KI, also jeder Schritt, den die KI unternimmt, aufgezeichnet werden. Danach wird jeder einzelne Vorgang händisch analysiert und im Falle von erkanntem Verbesserungspotential nachjustiert. Da die Korrektur solcher Fehler - vor allem im Falle von Änderungen der Gewichtungen - durch vorhandene Korrelationen auch einen Einfluss auf andere Fragebeantwortungen hat, entsteht dadurch eine aufwändige Testprozedur. In sehr komplexen Fällen lohnt es sich daher, die Aufzeichnungen aller Durchgänge statistisch zu analysieren, um bei etwaigen Änderungen der Rahmenbedingungen ungefähr abschätzen zu können, wie groß deren Auswirkungen sind.

5.3 Bias in neuronalen Netzwerken

Ursachen von Bias

Unrechtmäßige Diskriminierung bzw. ein Bias kann in NNs auf mehrere Arten entstehen. Dieser kann bereits bei der Formulierung der Problemstellung bzw. bei der Zielvorgabe für das NN auftreten. Grund dafür ist, dass die bestmögliche Zielerreichung nicht unbedingt die

fairste bzw. gerechteste ist.[136] Besteht das Ziel einer in Kapitel 4.2 beschriebener KI zur Entscheidung der Kautionsbedingungen für Häftlinge darin, die Straftaten zu reduzieren, welche durch diese Häftlinge in der durch die Erfüllung der Kautionsauflagen gewonnen Freiheit begangenen werden, wird eine derart instruierte KI dazu neigen, einen höheren Anteil der Beschuldigten in Haft zu belassen. Besteht die Zielsetzung der KI jedoch darin, eine möglichst geringe Gefängnispopulation zu erreichen, so wird die KI einer größeren Anzahl von Beschuldigten Kaution gewähren.

NNs eigenen sich Vorurteile hauptsächlich während der Trainingsphase im Zuge der Analyse der Trainingsdaten an. Hier kann sowohl die quantitative Verteilung der Daten als auch deren Qualität ausschlaggebend dafür sein, dass Vorurteile entstehen. Spiegeln die Trainingsdaten in ihrer Mengenverteilung nicht alle Facetten der Realität adäquat wider, so werden „Sonderfälle“ häufig falsch behandelt.[137] In einer in Kapitel 5.1 kurz angesprochenen Studie stellten Wissenschaftler des MIT und Microsoft bei aktuellen Gesichtserkennungssystemen fest, dass die Erkennung des Geschlechts bei Personen aus Minderheitsgruppen deutlich ungenauer ist. Eine mögliche Ursache ergibt sich aus der Struktur der standardisierten Testdatensätze für Gesichtserkennungs-KIs, die veröffentlicht wurden, um die Effizienz verschiedener Systeme miteinander vergleichen zu können. In zwei von drei untersuchten Testdatensätzen, „Adience“ und „IJB-A“, stellen Datensätze von Angehörigen von Minderheiten nur ca. 20% oder weniger der verwendeten Daten – und somit weniger als dies ihrem Anteil an der Gesamtbevölkerung entsprechen würde.[138]

Ist die Datenqualität nicht ausreichend, spiegeln also die Trainingsdaten die Realität nicht in ihrer vollen Vielfältigkeit wider bzw. sind diese sogar falsch, so wird ein NN unverhältnismäßig viele diskriminierende Ergebnisse liefern.[139] Ein Beispiel dazu lässt sich leicht für das NN zur Bewertung des Kautionsrisikos angeben: Wird das NN an den Fällen eines rassistischen Haftrichters trainiert, so wird das daraus resultierende NN dieses Entscheidungsmuster als gewünscht erlernen und neue Fälle im gleichen Maße diskriminieren.

Schlussendlich kann ein Bias auch über die Auswahl der Inputfaktoren bzw. Eingabewerte entstehen.[140] Da es aus diversen Gründen – vor allem bedingt aus monetären und Datenschutzgründen - nicht möglich ist, alle notwendigen Daten für die Entscheidungsfindung zu erfassen und zu verarbeiten, müssen NNs mit einer sehr eingeschränkten Sichtweise auf die jeweilige Situation agieren. Ein Beispiel dafür sind bereits in Kapitel 4.2 dargelegte Fälle, bei denen für die Bewertung des Kautionsrisikos diskriminierende Inputfaktoren, wie die Hautfarbe, verwendet wurden. Stehen detaillierte Informationen über das soziale Umfeld der Personen nicht zur Verfügung, so kann das NN Scheinkorrelationen nicht erkennen und seine Entscheidungsmuster nicht auf Basis der kausalen Einflussfaktoren aufbauen. Statistisch nur scheinbar relevante Faktoren führen somit dazu, dass das NN auf Basis dieser ein in der Realität nicht existierendes Muster sucht, dieses

[136] Bryce Goodman, Seth Flaxman. (2016). European Union regulations on algorithmic decision-makingand a “right to explanation”
[137] Time Magazine. (7.2.2019). Artificial Intelligence Has a Problem With Gender and Racial Bias. Here’s How to Solve It.
[138] Joy Buolamwini, Timnit Gebru. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.
[139] Technologyreview. (4.2.2019). This is how AI bias really happens—and why it’s so hard to fix.
[140] Technologyreview. (4.2.2019). This is how AI bias really happens—and why it’s so hard to fix.

aufgrund der eingeschränkten Sichtweise auf die Realität aber feststellt und auf dessen Basis Entscheidungen trifft, die unverhältnismäßig diskriminierend sind. So könnte ein NN z.B. eine Scheinkorrelation von 95% zwischen verliehenen Mathematik-Doktoraten und eingelagertem Uran in US-Atomkraftwerken[141] nicht als irrelevant erkennen.

Erkennen von Bias

Bias bzw. Vorurteile sind in NNs aufgrund des Blackbox-Prinzips viel schwerer zu erkennen als in klassischen KIs. Einfache „Audits", also das Schritt für Schritt Durchspielen des Entscheidungsprozesses und das Analysieren von Entscheidungskriterien, sind daher nur sehr schwer möglich. Während ein Bias in der Zieldefinition oder bei der Wahl der Inputfaktoren relativ offensichtlich ist, entzieht sich ein durch schlecht kalibrierte Daten entstehender Bias einer leichten Erkennung.

Eine logische Schlussfolgerung aus der Entstehung dieses Bias wäre, die Datenqualität und Datenquantität der Trainingsdaten statistisch zu erfassen. In manchen Fällen, wie bei der Bewertung des Kautionsrisikos, ist dies - zumindest hinsichtlich der Verteilung der Datenmenge bzgl. des auf Angehörige von Minderheiten entfallenden Anteiles - trivial. In anderen Fällen jedoch - wie bei der Erkennung des Geschlechts auf Basis von Bildern - ist dies schwer möglich, da die Fotos unter Umständen nur nach dem eigentlichen Ziel, dem Geschlecht, klassifiziert sind. Um hier die statistische Verteilung des Datensatzes bezüglich Minderheiten festzustellen, wird für diese zusätzliche Klassifizierung entweder ein zusätzliches NN benötigt oder es werden die einzelnen Daten (Bilder) händisch zugeordnet. Beide Lösungen sind in der Praxis nur schwer zu rechtfertigen, da einerseits das NN zur Klassifizierung, falls noch nicht existent, ebenfalls auf einem Datensatz trainiert werden müsste, der womöglich voreingenommen ist, und andererseits, da das händische Klassifizieren der oftmals in die Millionen gehenden Datensätzen wirtschaftlich unrentabel ist.

Zur Auffindung von Vorurteilen werden bei einem sich bereits im Einsatz befindlichen NNs daher zuerst die Statistiken der Ergebnisse analysiert. Dabei wird vor allem auf die typischen Fehler geachtet, die durch unzureichende Datenquantität und unausgeglichene Datenqualität entstehen. Typisch für unausgeglichene Datenquantität wären z.B. schwere Fehleinschätzungen bei Sonderfällen, die in den Trainingsdaten nicht ausreichend repräsentiert sind. Hingegen sind Abweichungen bezüglich Faktoren, die keine oder nur eine geringe Auswirkung haben sollten – wie beispielsweise die Hautfarbe bei der Einschätzung des Kautionsrisikos – ein Indikator für eine nicht ausreichende Datenqualität.

Falls bei einem bereits operativ eingesetzten NN ein Verdacht auf Voreingenommenheit besteht, können fiktive Vergleichs- und Sonderfälle erstellt werden, die diese erkennen sollen. Im Beispiel des Kautions-NN würden hierbei fiktive Häftlings- bzw. Personenprofile erstellt werden, die sich nur in den zu testenden Faktoren – also jenen, die vermutlich zu der unrechtmäßigen Diskriminierung beitragen - ändern. Wie stark sich die Bewertung für die einzelnen Profile unterscheidet, gibt schlussendlich einen Hinweis darauf, ob und wie groß der Einfluss dieser Faktoren auf die endgültige Entscheidung ist.

[141] Tyler Vigen.(3.6.2019). Spurious Correlations.

Gegenmaßnahmen

Doch auch diese Technik der fiktiven Testdaten wird bei vielen Inputfaktoren oftmals aufwändig. Es wäre daher als logische Schlussfolgerung sinnvoller, Faktoren, die potentiell ungerechtfertigt diskriminieren, von Anfang an zu streichen. Doch damit setzt man unter Umständen wieder eine weitere Ursache für einen Bias. Beispielsweise würde das Streichen des Faktors „Geschlecht“ bei der Risikobewertung für Kautionszwecke zu einer unverhältnismäßigen Benachteiligung von Frauen führen, da diese eine statistisch signifikante, kausal bedingte wesentlich geringere Strafrückfälligkeit besitzen als Männer.[142]

Die einzige Möglichkeit einen vorhandenen Bias zu entfernen ist, diesen mit einem neuen bzw. veränderten Trainingsdatensatz „abzutrainieren“. Dabei kann der neue Datensatz entweder möglichst bias-frei sein oder aber gezielt auf die Bekämpfung dieser auszumerzenden Voreingenommenheit ausgelegt sein. Bei ersterem sinkt die vorhandene Voreingenommenheit langsam und kontinuierlich ab, während bei letzterem die Angleichung durch den inversen Bias viel schneller erfolgt und - wenn nicht rechtzeitig gestoppt - sogar umgekehrt wird.

In beiden Fällen kann jedoch das Vorhandensein eines Rest-Bias nicht ausgeschlossen werden, da dieser nur mit einem beschränkten Testdatensatz überprüft werden kann. Daher ist es meist sinnvoller, das NN von Grund auf mit Bias-bereinigten Datensätzen zu trainieren.

Bedauerlicherweise können in der praktischen Anwendung all diese Gegenmaßnahmen das Verhindern von Voreingenommenheit oftmals nicht gewährleisten *(siehe auch [143])*: So hatte das Unternehmens Amazon im Bereich der Personalverwaltung ein NN zur Vorsortierung von Bewerbungen im Einsatz. In dieser wurden alle direkt das Geschlecht der Aspiranten identifizierenden Wörter aus den Bewerbungsunterlagen eliminiert, um die Möglichkeit einer unrechtmäßigen Diskriminierung durch das NN ausschließen zu können.[144] Dennoch wurde das NN eingestellt, nachdem sich herausgestellt hatte, dass dieses Frauen implizit über Adjektive, die eher zur Beschreibung von Frauen als von Männern verwendet werden, identifizieren konnte und so weiterhin das Merkmal Geschlecht in die Entscheidungsfindung einfloss.[145]

5.4 Analyse und Bewertung von neuronalen Netzwerken

Im vorherigen Kapitel wurde mit der Technik der fiktiven Testdaten ein Werkzeug behandelt, das bis zu einem gewissen Grad einen Einblick in die Blackbox eines NN ermöglicht. Neben diesem sehr einfachen Mittel existieren jedoch mittlerweile weitere, mächtigere Werkzeuge, die in diesem Kapitel näher erläutert werden.

Topologische Datenanalyse

Die topologische Datenanalyse ist ein im ersten Jahrzehnt dieses Jahrtausends entstandenes statistisches Werkzeug bzw. eine klassische KI, die es erlaubt, Gemeinsamkeiten in

[142] Jennifer L. Skeem et.al., Gender, Risk Assessment, and Sanctioning: The Cost of Treating Women Like Men
[143] Wired. (12.2.2019). The Real Reason Tech Struggles With Algorithmic Bias.
[144] Reuters. (10.10.2018). Amazon scraps secret AI recruiting tool that showed bias against women.
[145] Reuters. (10.10.2018). Amazon scraps secret AI recruiting tool that showed bias against women.

Datensätzen zu erkennen, diese zu gruppieren und in einer Topologie bzw. einem Netzwerk miteinander zu verknüpfen.[146]

Führt man eine topologische Datenanalyse auf dem Testdatensatz eines NN durch, so kann man einen Einblick in die Bereiche erhalten, die der KI Probleme bereiten.

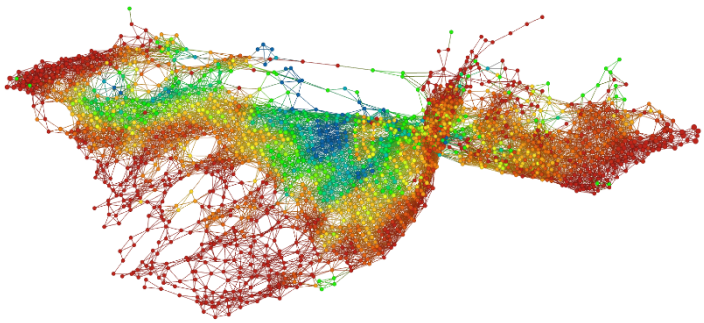

Abbildung 7: Topologische Datenanalyse eines Teilbereichs des ImageNet Testdatensatzes[147]

Abbildung 7 zeigt eine Visualisierung der topologischen Datenanalyse eines Teilbereichs des „ImageNet"-Testdatensatzes. ImageNet ist als Bilddatenbank mit über 14 Millionen indexierten, d.h. beschrifteten, frei verfügbaren Bildern eine willkommene Quelle für Wissenschaftler, die ein kostengünstiges und möglichst umfassendes Datenmaterial für NN-Trainings- und Testzwecke suchen.[148] Daniel Goldfarb von der Cornell University hat aus dieser Datenbank für den von ihm eingesetzten Testdatensatz etwa 5.000 Bilder, die sich gleichmäßig auf fünf Hunde- und fünf Katzenarten verteilen, ausgewählt und damit „VGG16", ein NN von Wissenschaftlern der Oxford University, getestet.[149]

Die Visualisierung in Abbildung 7 zeigt die Ergebnisse dieses Tests: Rot gekennzeichnet sind Bilder, die vollständig richtig erkannt wurden, während die Farbe Blau falsch erkannte Bilder kennzeichnet. Durch diese Farbkodierung lassen sich die Schwächen des NN leicht erkennen, indem einfach ein Blick auf die sich in diesen Clustern befindlichen Bilder und deren Gemeinsamkeiten geworfen wird. In diesem Fall befinden sich im Cluster im Zentrum von Abbildung 7 Bilder von Großkatzen hinter Gittern und im Cluster links oben Bilder mit vielen zusätzlichen, möglicherweise ablenkenden Objekten.[150]

Mit dieser Information können die im vorherigen Kapitel 5.3 beschrieben Gegenmaßnahmen ergriffen werden. Um dabei keine neuerlichen Fehler zu begehen, kann mit „*heat map*s" ein weiteres Werkzeug verwendet und damit der Fehlerursache auf den Grund gegangen werden.

---

[146] Frédéric Chazal and Bertrand Michel. (2017). An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists

[147] Daniel Goldfarb. (2018). Understanding Deep Neural Networks Using Topological Data Analysis.

[148] ImageNet. (12.5.2019). ImageNet.

[149] Daniel Goldfarb. (2018). Understanding Deep Neural Networks Using Topological Data Analysis.

[150] Daniel Goldfarb. (2018). Understanding Deep Neural Networks Using Topological Data Analysis.

Heat maps

Um festzustellen, welcher Bereich eines Datensatzes für die konkrete Entscheidung eines NN ausschlaggebend ist, kann analysiert werden, welche Neuronen in welchen Schichten in dieser Aufgabenstellung feuern. Handelt es sich bei diesen Datensätzen um Bilder und visualisiert man die feuernden Neuronen, so erhält man eine „*heat map*“.

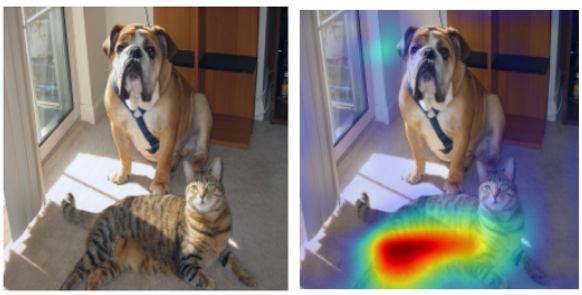

Abbildung 8: Vergleich des Originalbilds (links) und Heat Map (rechts)[151]

Dafür wird eine Momentaufnahme des NN - d.h. alle zwischen den Neuronen weitergegebenen Werte und deren Gewichte - analysiert. Im Vordergrund stehen hier die Werte und Gewichte des letzten „*convolutional layer*“ und der darauffolgenden Schicht, welche die erste Schicht des für die Ergebnisfindung zuständigen NN ist[152] (siehe Kapitel 3.3 bzw. Abbildung 3). In den *feature maps* des letzten *convolutional layer* befinden sich die Formen bzw. Objekte der letzten Abstraktionsschicht. Die *feature maps* werden gewichtet und übereinandergelegt. Dadurch wird schlussendlich die *heat map* erzeugt, die in Abbildung 8 rechts transparent über das Originalbild gelegt wurde.[153] Angenommen das in Abbildung 8 dargestellte NN diene dazu, Katzen auf Bildern zu identifizieren. Dann werden die *feature maps* der letzten Schicht typische Teile von Katzenkörpern darstellen. Die in Abbildung 8 präsenten Teile werden nun von den jeweiligen *feature map*s erkannt, welche darauf hin feuern.

In der praktischen Anwendung werden NNs häufig nicht nur zur Identifizierung eines einzigen Objekts auf Bildern verwendet, sondern es sollen mehrere verschiedene Formen erkannt werden. Am Beispiel der Abbildung 8 muss das NN in der Lage sein, sowohl Katzen als auch Hunde zu identifizieren und entsprechend ein mehrstufiges Ergebnis zu liefern: Auf dem Bild befinden sich eine Katze und ein Hund. Um bei solchen Aufgabenstellungen *heat map*s für jedes einzelne Objekte zu erstellen, wird ähnlich der für Lernprozesse verwendeten und in Kapitel 3.2 näher beschriebenen Technik der Backpropagation vorgegangen: Hierbei wird einfach der im Fokus stehende Teil des Ergebnisses über die feuernden Neuronen bis zu den *convolution layers* zurückverfolgt, in denen wie beschrieben die *heat map* auf Basis der Neuronen erstellt wird.[154]

[151] Ramprasaath R. Selvaraju, et.al. (2016). Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization
[152] Bolei Zhou, et.al. (2015). Learning Deep Features for Discriminative Localization.
[153] Bolei Zhou, et.al. (2015). Learning Deep Features for Discriminative Localization.
[154] Bolei Zhou, et.al. (2015). Learning Deep Features for Discriminative Localization.

Neben dieser grundlegenden Technik zur Erstellung von *heat map*s existieren darauf aufbauende Verbesserungen, welche genauere *heat map*s, die vor allem für kleinere Objekte notwendig sind, erzeugen können.[155, 156] Durch genaue *heat map*s kann mit der Lokalisierung der gesuchten Objekte auf den Bildern eine zusätzliche Problemstellung gelöst werden.

Feature map Visualisierungen

Mithilfe von *heat map*s kann nun festgestellt werden, wo in Datensätzen bzw. Bildern die Neuronen bzw. *feature maps* feuern. Man kann mithilfe derselben Technik jedoch auch das umgekehrte Ergebnis erhalten und approximieren, wie die einzelnen *feature maps* aussehen.

Anstatt das NN auf Basis eines konkreten Beispiels zu analysieren, erhält man im Fall von CNNs einen generellen Überblick über die einzelnen Abstraktionsschichten (*convolutional layers*) und deren Neuronen (*feature maps*). Dabei werden möglichst viele Testdaten bzw. Bilder durch das NN geschickt und für alle Neuronen aufgezeichnet, wie stark diese bei jedem Bildausschnitt feuern. Abschließend bekommt jedes Neuron denjenigen Bildausschnitt bzw. Datensatz zugeordnet, bei dem dieses am stärksten feuert.[157] (Für Visualisierungen, siehe [124] Figure: 2)

Diese Bildausschnitte stellen jedoch nur eine Approximation der *feature maps* dar, deren Genauigkeit von der Anzahl der verwendeten Bilder abhängt. Es ist jedoch auch möglich, die Neuronen direkt zu visualisieren.

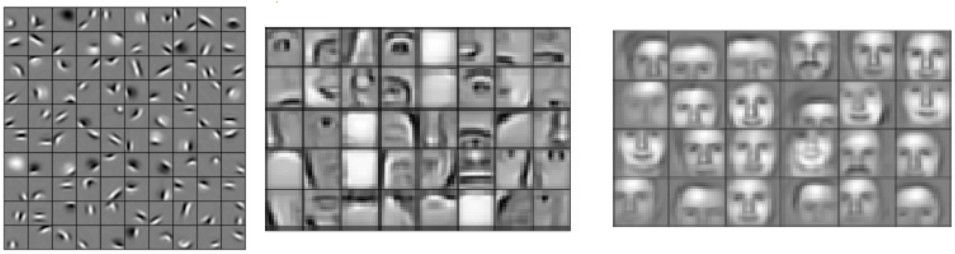

Abbildung 9: Feature Map Visualisierungen eines Convolutional Deep Believe Network[158]

Abbildung 9 zeigt eine solche Visualisierung der unterschiedlichen Abstraktionsebenen. Die erste Abstraktionsebene (Abbildung 9 links) stellt den ersten *hidden layer* des CNN dar. Hier werden noch keine klaren Strukturen erkennbar. In der zweiten Abstraktionsebene (Abbildung 9 Mitte) sind bereits einzelne Gesichtsausschnitte zu erkennen. Schlussendlich werden in der dritten Abstraktionsebene (Abbildung 9 rechts) vom CNN als repräsentativ eingestufte Gesichter dargestellt. Jedes in Abbildung 9 dargestellte Quadrat entspricht dabei einer *feature map* der CNN auf der jeweiligen Abstraktionsebene. Die Visualisierungen der Neuronen

[155] Ramprasaath R. Selvaraju, et.al. (2016). Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization
[156] Jianming Zhang, et.al. (2016). Top-down Neural Attention by Excitation Backprop.
[157] Matthew D. Zeiler, Rob Fergus. (2013). Visualizing and Understanding Convolutional Networks.
[158] Nvidia. (13.5.2019). Deep Learning in a Nutshell: Core Concepts.

werden erzeugt, indem alle vorherigen verbundenen *feature maps* mit ihren jeweiligen Gewichten übereinandergelegt werden.[159]

Neben ihrer Eignung als Veranschaulichungsmaterial besitzen die Visualisierungen dieser Abstraktionsebenen auch praktischen Nutzen: Sind diese Visualisierungen noch von einem starken Rauschen betroffen, so ist dies ein Indiz dafür, dass das NN noch zu wenig trainiert wurde.[160]

Mehrfach vorkommende Muster – erkennbar in Form von optisch kaum unterscheidbaren Visualisierungen - sind ein Anzeichen dafür, dass das NN für die Aufgabenstellung falsch dimensioniert wurde und sich zu viele Neuronen auf den jeweiligen Abstraktionsebenen der *hidden layer* befinden. Gibt es nämlich einen Überschuss an Neuronen in einer Schicht, d.h. gibt es mehr Neuronen als „*features*“, so benützt das NN in dieser Ebene die überschüssigen Kapazitäten zum Auswendiglernen.

Aktivierungsmaximierung bzw. „Träumen“

Neben der oben dargelegten Methode zur Visualisierung von *feature map*s, die auf der Kombination der *feature map*s der verbundenen vorhergehenden Schichten basiert, existiert noch eine weitere Methode: Die der Aktivierungsmaximierung bzw. des „Träumens“.

Die Methode beruht darauf, dass sich das Feuern der Neuronen als mathematische Funktion darstellen lässt. Wird diese Funktion maximiert, so erhält man denjenigen Datensatz bzw. dasjenige Bild, welches das Neuron absolut am stärksten feuern lässt. Dadurch kann man eine alternative Darstellung dieser *feature map* erhalten.[161]

Abbildung 10 zeigt drei mit diesem Ansatz visualisierte *feature maps* der letzten Abstraktionsschicht. Da das zugrunde liegende NN auf die Objekterkennung ausgelegt ist, können die jeweiligen *feature maps* direkt benannt werden. Dazu wird wie bei der Erstellung von *heat maps* vorgegangen: Der gesuchte Begriff (z.B. „Zitrone“) wird am Ende des NN angelegt und über die einzelnen Neuronen bis hin zum letzten *convolutional layer* zurückverfolgt, in dem sich die *feature map*s der höchsten Abstraktionsschicht befinden.[162] Das mit Abstand am stärksten feuernde Neuron stellt dann diejenige *feature map* dar, die für das Erkennen dieses Objekts in den Daten verantwortlich ist.

---

[159] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. (2011). Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks.
[160] Stanford University. (13.5.2019). Visualizing what ConvNets learn.
[161] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. (2009). Visualizing Higher-Layer Features of a Deep Network.
[162] Karen Simonyan, Andrea Vedaldi, Andrew Zisserman. (2014). Deep Inside Convolutional Networks: VisualisingImage Classification Models and Saliency Maps.

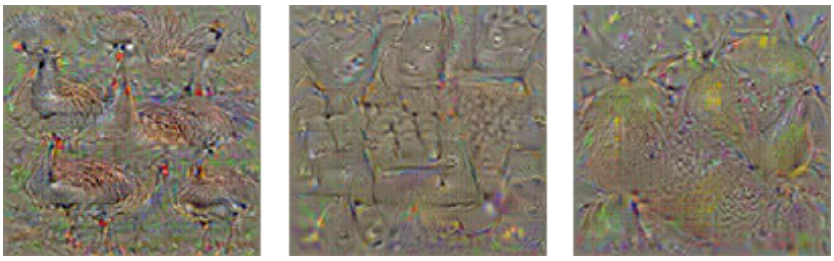

Abbildung 10: Darstellung von Feature Maps mithilfe der Aktivierungsmaximierung[163]
(links Gans, Mitte Tastatur, rechts Zitrone)

Zu beachten ist, dass die Darstellungen in Abbildung 10 die Objekte aus unterschiedlichen Perspektiven zeigen. Aufgrund des hohen Rauschgehalts der Visualisierungen können in der Praxis vermutlich nur grobe Missstände in der Objekterkennung festgestellt werden.

Die an Traumbilder erinnernden Darstellungen haben dieser Methode zu großer Popularität verholfen, da diese zur Erschaffung neuer „Kunstwerke" oder zur Interpretation existierender Werke verwendet werden können. Dabei werden die am stärksten feuernden Neuronen auf die jeweiligen Bildausschnitte angewendet. Die resultierende Visualisierung wird danach wieder durch das NN geschickt, wodurch eine Feedbackschleife entsteht, die zu den psychedelisch bzw. traumartig wirkenden Bildern führt.[164] Diese Methode kann nicht nur bei bereits vorhandenen Bildern angewendet werden, sondern auch auf „nicht vorhandene Bilder", also Bildern, die nur aus Rauschen bestehen.[165] Die Interpretation des NN entdeckt so scheinbar verborgene Botschaften.

Mathematische Analyse- und Bewertungsmethoden

Neben den bisher behandelten Methoden zur Visualisierung von NNs existieren rein statistische bzw. mathematische Methoden zur Analyse von NNs. Mit diesen soll erkannt werden, ab wann z.B. ein „*overfitting*" auftritt:

Wie bei der Beschreibung des Lernprozesses für NNs in Kapitel 3.2 dargelegt, benötigen NNs für das Erkennen von Zusammenhängen viel mehr Informationen als Menschen, wodurch Daten mehrmals verwendet werden müssen. Werden diese Datensätze jedoch zu oft wiederholt berücksichtigt, so entsteht ein sogenanntes „*overfitting*"[166]. Anstatt die Zusammenhänge zwischen den Daten zu finden, lernt das NN die bestehenden Daten quasi auswendig. Dieses oft beobachtete Phänomen kann in der Anwendung neben mathematischen Analysemethoden auch während des Lernprozesses erkannt werden: Das Auswendiglernen des Trainingsdatensatzes hat nämlich zur Folge, dass das NN bei dem Testdatensatz immer schlechter abschneidet.

Mathematische Methoden sind jedoch aufgrund ihrer hohen Komplexität weitaus weniger zugänglich und eine sachgemäße Behandlung wäre im Rahmen dieser Arbeit nicht

[163] Karen Simonyan, Andrea Vedaldi, Andrew Zisserman. (2014). Deep Inside Convolutional Networks: VisualisingImage Classification Models and Saliency Maps.
[164] Google. (2015). Inceptionism: Going Deeper into Neural Networks.
[165] Google. (2015). Inceptionism: Going Deeper into Neural Networks.
[166] Jean Francois Puget. (30.6.2019). Overfitting In Machine Learning.

zielführend. Daher wird an dieser Stelle lediglich auf die weiterführende Literatur verwiesen, wie z.B. S.W. Ellacotts *„Techniques for the mathematical analysis of neural networks“*.[167]

5.5 Angriffe auf neuronale Netzwerke

Adversarial-Bilder

Im vorhergehenden Kapitel wurde beschrieben, wie NNs bzw. CNNs in Bildern bzw. sogar in reinem Rauschen nicht existente Objekte erkennen. Forscher und Forscherinnen der University of Wyoming haben auf Basis dieses Phänomens eine Methode zur Bilderzeugung bzw. Bildgenerierung entwickelt, die NNs mittels simpler Muster täuscht und diese zum fälschlichen Erkennen von Objekte verleitet.[168] Dabei wurden klassische KIs nach einem „evolutionären“ Prinzip verwendet: Dieses verändert ein generiertes Rauschbild bzw. Muster zufällig und baut bei jedem iterativen Schritt auf demjenigen Ergebnis auf, das dem NN bzw. CNN ein nicht existentes Objekt vorgaukelt. Mit dieser Methode gelang es den Forschern und Forscherinnen, dass das NN Objekte in reinen Rauschbildern zu erkennen glaubte und diesen Ergebnissen eine 99,6% Sicherheit zuschrieb.

Es können aber auch bereits Bilder, die tatsächliche Objekte darstellen, mithilfe von Rauschen so abgeändert werden, dass NNs darin komplett andere Körper wahrnehmen. Dafür wird die geringstmögliche Rauschfunktion (die Änderung der Daten bzw. des Bildes durch Rauschen) definiert, die benötigt wird, um das derzeitige Resultat in ein gewähltes anderes Resultat zu ändern.[169]

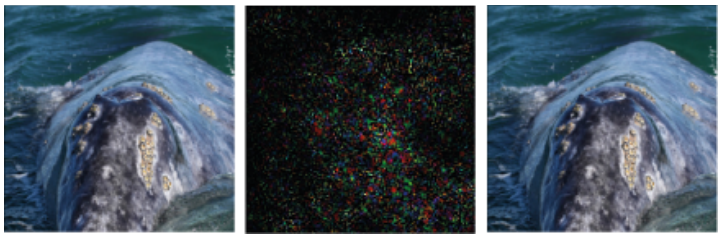

Abbildung 11: Beispiel eines "adversarial" Bildes[170]
(links) Originalbild (Wal), (Mitte) minimale Rauschfunktion, (rechts) „adversarial“-Bild (Schildkröte)

Die minimale Rauschfunktion kann mit einfachen mathematischen Methoden nicht präzise ermittelt, sondern nur approximiert werden. Verschiedene wissenschaftliche Arbeiten haben daher darauf abgezielt, mit standardisierten Verfahren möglichst minimale Rauschfunktionen zu erzeugen.[171] [172] Abbildung 11 zeigt eine dieser Techniken, bei denen das zusätzliche

[167] Vgl. S.W. Ellacott. (1992). Techniques for the mathematical analysis of neural networks.
[168] Anh Nguyen, Jason Yosinski, Jeff Clune. (2015). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images.
[169] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard. (2016). DeepFool: a simple and accurate method to fool deep neural networks.
[170] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard. (2016). DeepFool: a simple and accurate method to fool deep neural networks.
[171] Vgl. Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy. (2015). Explaining and harnessing adversarial examples.
[172] Vgl. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. (2014). Intriguing properties of neural networks.

Rauschen in dem „adversarial“-Bild (rechts) für das bloße Auge nicht erkennbar ist. In einer anderen Studie entwickelten Jiawei Su, Danilo Vargas und Sakurai Kouichi eine „One-Pixel Attack“ genannte Methode, bei der das NN mittels der Änderung eines einzigen Pixels der Testbilder aus der Bahn geworfen werden kann.[173]

Eine Rauschfunktion ist üblicherweise individuell auf das jeweilige NN zugeschnitten. Es hat sich jedoch gezeigt, dass mit einer derartigen Rauschfunktion auch andere, nicht verwandte NNs getäuscht werden können.[174] [175] Verfolgt man jedoch das Ziel, eine möglichst breite Vielzahl unterschiedlicher NNs zu täuschen, so muss den Daten ein hohes Quantum an Rauschen hinzugefügt werden.

Eine Anwendung dieser Angriffstechnik wären beispielsweise elektronische Visaanträge. Bei diesen werden Gesichtsportraits hochgeladen, die anschließend von Gesichtserkennungssystemen geprüft werden. Das Rauschen in dem hochgeladenen Portrait würde in diesem Fall dafür sorgen, dass ein NN keine Beziehung zwischen dem hochgeladenen Bild und der Referenzaufnahme der gesuchten Person herstellen kann. Umgekehrt kann es jedoch der Fall sein, dass das NN auf dem Portrait fälschlicherweise andere Personen als den Gesuchten identifiziert. In diesem Fall würde der Störversuch vermutlich aber einfach bemerkt werden - vor allem dann, wenn die scheinbar identifizierte Person und das Portrait keine Gemeinsamkeiten zeigen.

Diese Angriffsmethoden, insbesondere der „One-Pixel Attack“, zeigen zudem ein weiteres Problem auf. Der Automobilhersteller Tesla setzt bei seinem „Autopilot“-Feature in der Hardware-Ausstattung nicht auf den teuren Industriestandard „LIDAR“ - ein auf Laser basierendes Radar -, sondern nur auf herkömmliche Kameras.[176] In den Bildern dieser kostengünstigen Kameras treten jedoch unvermeidlich Pixelfehler, z.B. durch Verschleiß, auf. Dies kann im Extremfall dazu führen, dass das das Fahrzeug steuernde NN die zu analysierenden Objekte, wie z.B. Personen oder Straßenschilder, nicht korrekt zuordnen kann. Verstärkt wird diese Befürchtung dadurch, dass gerade der Einsatz von „LIDAR“-Sensoren, die bei autonomen Fahrzeugen anderer Hersteller verwendet werden, Kamerasensoren anderer Autos beschädigen können. Fahren ein Fahrzeug der Firma Tesla und ein anderes autonomes Fahrzeug auf einer Autobahn parallel zueinander, so könnten die Seitenkameras des Tesla-Fahrzeugs daher permanenten Schaden davontragen, der die Insassen des Tesla und andere Verkehrsteilnehmer von einem Moment auf den anderen gefährdet.

Dabei bleibt es jedoch nicht bei einzelnen Pixelfehlern, sondern es können ganze Regionen des Kamerasensors zerstört werden. Ein Teilnehmer an der Consumer-Electronics-Show in Las Vegas im Januar 2019 berichtete davon, wie die „LIDAR“-Sensoren an einem Vorführwagen seine ca. 2.000 US-Dollar teure Kamera merkbar beschädigten.[177]

---

[173] Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi. (2019). One Pixel Attack for FoolingDeep Neural Networks.
[174] Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi. (2019). One Pixel Attack for FoolingDeep Neural Networks.
[175] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. (2014). Intriguing properties of neural networks.
[176] The Verge. (24.4.2019). It’s Elon Musk vs. everyone else in the race for fully driverless cars.
[177] ArsTechnica. (11.1.2019). Man says CES lidar’s laser was so powerful it wrecked his $1,998 camera.

Zusätzlich wurde gezeigt, dass auch bei ausgedruckten Versionen dieser Adversarial-Bilder NNs zu einer Fehlinterpretation neigen.[178] WissenschaftlerInnen entwickelten zusätzlich eine „robuste“ Version einer Rauschfunktion, die auch aus verschiedenen Blickwinkeln ihr Ziel erfüllt.[179] Damit konnten sie dreidimensionale - in diesem Fall 3D-gedruckte - Objekte herstellen, die aufgrund der Rauschfunktion in der Textur von NNs falsch klassifiziert wurden und dennoch für den menschlichen Betrachter ununterscheidbar sind.

In der Theorie könnten Aufkleber entworfen werden, die Gegenstände wie Schusswaffen für NNs, die Videos von Überwachungskameras analysieren, unsichtbar machen und so die Reaktionszeit auf sich im Gang befindende illegale Aktivitäten deutlich verlängert wird. In der Praxis würde dies jedoch vermutlich durch die niedrige Qualität der Überwachungsvideos scheitern: Vor allem nachts überdeckt das natürliche, durch den Mangel an ausreichender Beleuchtung hervorgerufene Rauschen den Störversuch. Generell machen Adversarial-Bilder daher in Angriffen auf NNs nur dort Sinn, wo der Angreifer die Daten direkt, also digital, bereitstellen kann.

Adversarial-Töne

Adversarial-Attacken können prinzipiell mit jeder Form von Daten ausgeführt werden. So entwickelten Forscherinnen und Forscher der Universität Bochum eine Angriffstechnik, mit der sich verborgene, für den Menschen so gut wie unhörbare Sprachbefehle in Tonaufnahmen und Musik einfügen lassen.[180] Die Töne werden dabei nach dem psychoakustischen Prinzip versteckt, das auch bei dem bekannten Audio-Codec MP3 zur Komprimierung von Audio-Dateien zur Anwendung kommt. Die Arbeit des Teams wurde jedoch nur mit demjenigen NN getestet, auf das die Adversarial-Töne zugeschnitten wurden. Zudem wurden die Tonspuren digital in das verwendete Spracherkennungssystem „Kaldi“[181] eingespeist und nicht wie in der Praxis der Fall über ein Mikrofon aufgenommen. Aufgrund der Ähnlichkeit zu Bild-basierten Adversarial-Attacken kann jedoch davon ausgegangen werden, dass diese Attacke mit geringfügigen Modifikationen der Rauschfunktion zukünftig auch bei NNs funktioniert, auf die die Attacke nicht perfekt zugeschnitten ist. In der Studie beschriebene mögliche realitätsnahe Situationen beschreiben Scherze wie das ungewollte Bestellen von Produkten, über das Sammeln persönlicher Daten durch Sprachassistenten bis hin zum Deaktivieren von Sicherheitssystemen bzw. Überwachungskameras im Smart-Home.[182] Als Hilfsmittel für solche Angriffe können laut dem Forscherteam unter anderem Fernsehwerbungen oder auch Smartphone-Apps verwendet werden.

Es ist davon auszugehen, dass jede Form von Sprachaufzeichnungen, also z.B. Sprachdurchsagen auf Flughäfen bzw. abgespielte Anweisungen bei Sicherheitskontrollen, für Angriffe verwendet werden können, wodurch derartige Techniken zu gefragten Werkzeugen für geheimdienstliche Akteure werden, aber auch zur Massenüberwachung missbraucht

[178] Alexey Kurakin, Ian Goodfellow, Samy Bengio. (2017). Adversarial examples in the physical world.
[179] Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok. (2018). Synthesizing Robust Adversarial Examples.
[180] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, Dorothea Kolossa. (2018). Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding.
[181] Github. (26.5.2019). Kaldi Speech Recognition Toolkit.
[182] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, Dorothea Kolossa. (2018). Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding.

werden können. Verstärkt wird die Gefahr solcher Angriffe durch die „*always on*“ Funktionalität von Sprachassistenten auf Smartphones, die zu jedem Zeitpunkt - auch im Standby - Sprachbefehle entgegennehmen und deren etwaige Sicherheitslücken seitens potentieller Angreifer ausgenutzt werden könnten.

Adversarial-Text

Ein Forscherteam der Zhejiang University in China entwickelte eine effektive Adversarial-Angriffsmethode für textbasierte NNs.[183] Die Vorgehensweise gleicht auch hier den bereits behandelten Varianten: Zuerst wird der zu ändernde Textabschnitt durch ein von dem Forscherteam kontrolliertes NN geschickt. Die Wörter, die für das Ergebnis des NN ausschlaggebend sind, werden durch absichtliche Rechtschreibfehler so geändert, dass diese für das NN unscheinbar wirken bzw. ignoriert werden.

> *I watched this movie recently mainly because I am a Huge fan of Jodie Foster's. I saw this movie was made right between her 2 Oscar award winning performances, so my expectations were fairly high. ~~Unfortunately~~Unf0rtunately, I thought the movie was ~~terrible~~terrib1e and I'm still left wondering how she was ever persuaded to make this movie. The script is really ~~weak~~wea k.*[184]

In diesem Beispiel aus der Studie, einer auf der populären Website IMDB veröffentlichten Filmkritik[185], wird der Originaltext von der Textverständnis-KI des Technologiekonzerns Amazon als 100% negativ eingestuft. Nach Anwendung der Angriffstechnik des Forscherteams, hier rot markiert, stuft dieselbe KI die Filmkritik als 89% positiv ein. Wie an dem Beispiel zu sehen ist, behält der Text dabei für menschlichen Leser seine ursprüngliche Bedeutung – es wurden lediglich Zeichen durch optisch ähnliche ersetzt. Den Tests des Forscherteams zufolge erzielt diese Methode derzeit durch die Bank hohe Erfolgsquoten bei KIs bekannter Technologieunternehmen wie Microsoft, IBM, Google und Facebook. In mindestens 70% der Fälle aus dem IMDB-Datensatz konnte die Einschätzung der Filmkritik in das jeweils Gegenteilige geändert werden, ohne mehr als 10% der Wörter zu ändern. Dabei ist zu berücksichtigen, dass das Team keinen Zugang auf den Quellcode bzw. die Parameter der KIs dieser Unternehmen hatte und der Algorithmus der Angriffstechnik daher nicht auf diese hin optimiert wurde.

Primäre Angriffsziele dieser Methode scheinen – wie das bewusst gewählte Beispiel illustriert - vor allem Online-Zensur bzw. Filtersysteme zu sein. Mithilfe der anderen adversarial-Angriffstechniken ist es vermutlich möglich, auf NNs basierte digitale Zensur bzw. Upload-Filter, wie sie im Rahmen des umstrittenen Artikel 17 der EU-Urheberrechts-Richtlinie 2019 eingeführt wurden[186], zu umgehen. Dabei ist es nicht nur möglich, illegale „Propagandainhalte“ - wie jüngst das Video des Attentäters von Christchurch[187] - sondern auch urheberrechtlich relevante Inhalte - wie beispielsweise unveröffentlichte, von „*leaks*“

---

[183] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, Ting Wang. (2018). TextBugger: Generating Adversarial Text Against Real-world Applications.
[184] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, Ting Wang. (2018). TextBugger: Generating Adversarial Text Against Real-world Applications.
[185] IMDB. (19.5.2019). Ratings and Reviews for new Movies and TV-Shows.
[186] Art 17 lit c. EU-Richtlinie über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt.
[187] The Guardian. (15.3.2019). Social media firms fight to delete Christchurch shooting footage.

betroffene Serienepisoden und Filme - für NNs unkenntlich zu machen, ohne dabei die Bildqualität wesentlich zu verschlechtern. Benötigt würden dafür jedoch ein vorhandenes NN bzw. die allgemeine Funktionsweise solcher Systeme, um ein eigenes NN zu trainieren, anhand dessen die Rauschfunktion für den Angriff approximiert werden kann.

Somit ist der Einsatz dieser Techniken nicht nur für Bürgerrechtler und Dissidenten interessant, sondern auch für Cyberkriminelle, die mit diesen moderne Spam-Filtersysteme vor neue Herausforderungen stellen könnten.

Adversarial-Patch

Bei der Angriffstechnik auf NNs mithilfe von „Stickern“ werden diese meist physikalisch in eine Szene mit dem Ziel eingefügt, die Reaktion des NN auf die Interpretation zu beeinflussen.

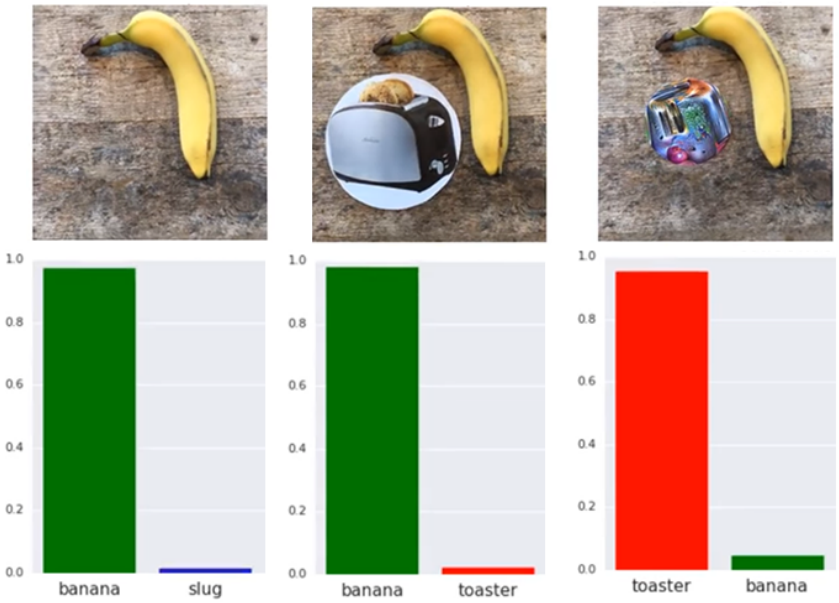

Abbildung 12: Demonstration der adverserial-Patch Angriffsmethode[188]

Abbildung 12 zeigt eine Demonstration dieser Angriffsmethode, wie sie von einem Team des Unternehmens Google entwickelt wurde.[189] Der rechte Bildausschnitt zeigt, wie das NN beeinflusst durch den psychedelischen Sticker einen Toaster wahrnimmt und dabei die danebenliegende Banane - im Gegensatz zum Kontrollbild (Mitte) - mehr oder weniger ignoriert. Der Sticker stellt dabei eine für diesen Zweck bearbeitete Form eines Toasters dar, der wesentlich weniger dem von einem Menschen als typisch empfundenen Bild eines Toasters ähnelt als der Sticker im mittigen Beispiel. Erzeugt wird dieses Phänomen ähnlich der in Kapitel 5.4 beschriebenen Methode des „Halluzinierens“ bzw. „Träumens“: Der Sticker besteht dabei zu Beginn aus reinem Rauschen und wird in jedem Durchgang durch die

[188] Zusammengeschnitten und Bearbeitete Inhalte aus dem Demonstrationsvideo. Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer. (2018). Adversarial Patch.
[189] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer. (2018). Adversarial Patch.

Traummethode von dem NN auf das Zielobjekt (in diesem Fall „Toaster“) hin optimiert – dies unter dem Zusatzziel, dass das Gesamtbild ebenfalls als das Zielobjekt identifiziert wird.[190] In Folge dieser Optimierungen feuern die Neuronen bzw. *feature map*s bei der Interpretation des Stickers so stark, dass sie die restlichen Objekte auf dem Bild verdrängen bzw. komplett in den Hintergrund rücken. Bei Tests des Mitarbeiterteams von Google stellte sich heraus, dass derartige Sticker mit einer Wahrscheinlichkeit von mehr als 80% das NN erfolgreich verwirren können, auch wenn die Sticker nur 20% der Bildfläche ausmachen. Dieses Ergebnis wird wohlgemerkt bei NNs erzielt, die nicht zur Entwicklung des Stickers verwendet wurden. Dies und die Tatsache, dass mit dieser Methode erzeugte Sticker robust sind, d.h. aus allen Blickwinkeln den gewünschten Effekt erzielen, macht sie zu einer praktischen und effektiven „Waffe“ für generelle Angriffszwecke. Ähnliche Methoden wurden bereits von verschiedenen Forscherteams verwendet, um psychedelische Brillenrahmen gegen Gesichtserkennungssysteme[191] oder auch T-Shirts bzw. Tafeln gegen Personen-Erkennungssysteme[192] zu konstruieren.

Es gibt jedoch weitere Beispiele dieser Herangehensweise an die Störung der Funktionsweise von NNs: Auf Straßenschilder geklebte Sticker können das Schild für autonome Fahrzeuge unsichtbar machen[193] und auf die Fahrbahn geklebte Sticker können Spurassistenzsysteme von autonomen Fahrzeuge dazu bringen, die Spur zu wechseln bzw. sogar in den Gegenverkehr zu steuern.[194]

Im oben genannten Fall des Fahrspurwechsels wurden simple gräulich gefärbte Sticker, die eine teilweise Spurmarkierung andeuten, verwendet. Für den Autofahrer sind diese im Kontrast zu den echten Fahrbahnmarkierungen nahezu unsichtbar. Für das getestete Fahrzeug des Automobilherstellers Tesla - dessen Spurhalteassistent bzw. Autopilot wie bereits erwähnt nur auf herkömmlichen Kameras basiert - wirken diese wie eine Aufforderung zu einem Spurwechsel.[195] Eine logische Erklärung dafür wäre, dass der Autopilot des Tesla im Alltagseinsatz auch auf sehr verbleichte Markierungen reagieren muss und die dadurch erforderliche Fehlertoleranz das Erkennen dieser Störversuche erschwert.

Derartige Sticker sind für den Laien wie den Experten gut getarnt und bleiben den Behörden somit länger unbekannt. Doch selbst wenn die zuständigen Behörden über die Existenz von Stickern informiert werden würden, würden diese für einige Zeit eine Gefahr darstellen, da der Aufwand für das Entfernen doch beachtlich ist.[196] Diese psychedelischen Sticker können zudem getarnt werden, indem sie z.B. durch ein Peace-Symbol(☮) durchscheinen und so wie einfache Werbesticker erscheinen.[197]

---

[190] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer. (2018). Adversarial Patch.
[191] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter. (2016). Accessorize to a Crime: Real and Stealthy Attacks onState-of-the-Art Face Recognition.
[192] Simen Thys, Wiebe Van Ranst. (2019). Fooling automated surveillance cameras: adversarial patches to attack person detection.
[193] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, Kai Chen. (2019). Seeing isn’t Believing: Practical Adversarial Attack Against Object Detectors.
[194] Tencent Keen Security La. (2019). Experimental Security Research of Tesla Autopilot.
[195] Tencent Keen Security La. (2019). Experimental Security Research of Tesla Autopilot.
[196] SafeUm. (10.8.2017). You can confuse self-driving cars by altering street signs.
[197] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer. (2018). Adversarial Patch.

Gegenmaßnahmen

Da das Lernprinzip das Grundkonzept der Entwicklung von NNs darstellt, kann die Robustheit auf Adversarial-Angriffe in das Trainingsmaterial eines NN aufgenommen werden. Weiters zeigt dies zusätzlich auf, dass NNs die Rauschfunktion in den Daten als wichtigen Bestandteil dieser sehen und deshalb falsche Schlussfolgerungen lernen. Vorläufige wissenschaftliche Ergebnisse zeigen, dass mit solchem robusten Trainingsmaterial trainierte NNs besser generalisieren können, d.h. diese können sich besser auf das Wesentliche in den Daten konzentrieren und ein Rauschen als ein solches erkennen.[198]

Diese Methode hat jedoch praktische Nachteile: Um mit dieser Methode gute Ergebnisse zu erzielen, müssten für einen Großteil der Daten zusätzliche Adversarial-Versionen eingefügt werden, im Extremfall sogar mehrere unterschiedliche Versionen. Dies würde die Anzahl der Trainingsdaten vervielfachen und so zu langen bzw. aufwändigen Trainingsprozessen führen.

Alternativ dazu entwickelte ein Forscherteam der Penn State University eine Möglichkeit,mithilfe der „Destillier"-Methode (siehe Kapitel 3.3) NNs vor Adversarial-Attacken zu schützen.[199] In deren publizierter Methode sind beide NNs, das ursprüngliche und das destillierte, jedoch gleich groß. Deren Ergebnisse zeigten, dass die Erfolgswahrscheinlichkeit von Adversarial-Attacken mit Bildern nach einem „Destillieren" des NN in einem Testdatensatz von 95,89% auf 0,45% bzw. in einem weiteren von 87,89% auf 5,11% sank und dabei weniger als 1,37% an Genauigkeit einbüßte.[200]

Gegen Adversarial-Töne existiert zusätzlich eine triviale Maßnahme. Wie bereits beschrieben filtert der Audio-Codec MP3 - um Speicherplatz zu sparen - alle psychoakustischen Töne aus Aufnahmen heraus. Daher können NNs gegen Adversarial-Töne gewappnet werden, indem deren Eingabedaten zuvor mithilfe dieses Codec gefiltert werden.

Zur Verhinderung von textbasierten Adversarial-Attacken gibt es aktuell keine ausreichenden Gegenmaßnahmen. Eine im April 2019 erschienene Arbeit fasst hierfür jedoch alle derzeit versuchten Ansätze zusammen.[201]

Die bereits erwähnte Technik des Einbindens von Adversarial-Angriffen in die Trainingsdaten von textinterpretierenden NNs erzielt laut den Forscherinnen und Forschern nicht immer den gewünschten Effekt, da es davon abhängt, welche Art der Erzeugung von Adervsarial-Daten verwendet wird. Dies ist auch logisch nachvollziehbar, da es viele verschiedene Möglichkeiten gibt, um die eigentliche Aussage des Textes vor NNs möglichst zu verstecken – so können Zeichen vertauscht, Rechtschreibfehler eingebaut oder die Semantik verändert werden.

Eine andere von dem Forscherteam genannte Möglichkeit ist das Verwenden von automatischer Rechtschreibkorrektur als Filter von Adversarial-Text. Sie stellen jedoch auch hier fest, dass dies nur für Rechtschreibfehler gilt und nicht für das geschickte Verändern von

---

[198] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. (2014). Intriguing properties of neural networks.
[199] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, Ananthram Swami. (2015). Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks.
[200] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, Ananthram Swami. (2015). Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks.
[201] Wenqi Wang, Lina Wang, Benxiao Tang, Run Wang, Aoshuang Ye. (2019). Towards a Robust Deep Neural Network in Text Domain A Survey.

Semantik bzw. das Verwenden von Synonymen. Außerdem ist in manchen Sprachen die Anwendung dieser Methode aufgrund deren Grammatik bzw. Verwendung von Zeichen nicht möglich.

Eine neue Abwehrmaßnahme gibt es gegen Adversarial-Patches: Im ersten Schritt der als „Local-Gradient-Smoothing“ bezeichneten Methode werden über das NN die Bildbereiche erkannt, die für ein „verdächtig“ starkes Feuern der Neuronen in diesem spezifischen Bild verantwortlich sind.[202] Im Gegensatz zu „natürlichen“ Objekten, die von dem NN erkannt werden sollen, besitzen Sticker möglichst viel „Reizkraft“ pro Fläche, da diese trotz ihrer relativ geringen Größe eine möglichst hohe Auswirkung haben sollen. Durch das Herausfiltern dieser Bereiche wird jedoch meistens nicht nur der Sticker selbst entfernt, sondern auch Teile des gewünschten Bildes. In Experimenten mit 1.000 Testbildern mit Stickern (vgl. Abbildung 12), konnte diese Methode die Erfolgsrate des Angriffs von 88% auf 18% senken, wobei die Sticker hierbei 2% der Bildfläche einnahmen.[203]

Dabei muss bedacht werden, dass 2% der Bildfläche ein verschwindend kleiner Bereich ist - ein Adversarial-Patch auf einer Stopptafel kann circa 10% der Fläche einnehmen, ohne für den Menschen in einem hohen Maß irritierend zu sein. Die Forscher, die dieses Experiment durchführten, erwiderten diesem Argument, dass es in Fällen von größere Flächenanteile einnehmenden Stickern leichter sei, das Vorhandensein eines Stickers zu erkennen und so das gesamte Bild zu verwerfen, d.h. zu ignorieren.[204] Im Fall von Verkehrsschildern bei autonomen Fahrzeugen wäre dies aber natürlich eine unzureichende Lösung.

6. Rechtliche Rahmenbedingungen

Künstliche Intelligenzen sind vom Prinzip her diskriminierend, da es ihre Aufgabe ist, Objekte und Personen aus Rohdaten in bestimmte Kategorien einzuordnen und dementsprechend zu bewerten.[205] Die Gefahr einer Kategorisierung aufgrund von diskriminierenden Merkmalen ist umso größer, je geringer und weniger vielfältig die im Rahmen des Trainierens des NN zur Verfügung stehenden Daten sind. Zwischenzeitlich kann bei der Entwicklung von NNs auf die Erkenntnisse einer breiten Forschung aufgesetzt werden, daher liegt der Großteil der Kosten nun in der Bereitstellung von Trainingsdaten. Um KIs nun möglichst günstig zu entwickeln, wird daher vor allem bei der Qualität der Daten und in der Qualitätssicherung gespart, wodurch unrechtmäßige Diskriminierung wahrscheinlicher wird. Die Verwendung potentiell diskriminierende Informationen an sich zu verbieten, stellt keine Lösung dar - so führt, wie in Kapitel 4.2 näher beschrieben, das Weglassen der

[202] Muzammal Naseer, Salman H. Khan. (2018). Local Gradients Smoothing: Defense against localized adversarial attacks.
[203] Muzammal Naseer, Salman H. Khan. (2018). Local Gradients Smoothing: Defense against localized adversarial attacks.
[204] Muzammal Naseer, Salman H. Khan. (2018). Local Gradients Smoothing: Defense against localized adversarial attacks.
[205] Bryce Goodman, Seth Flaxman. (2016). European Union regulations on algorithmic decision-makingand a “right to explanation”

Geschlechtsinformation bei KIs zur Kautionsrisikobewertung dazu, dass Frauen unverhältnismäßig diskriminiert werden.

In diesem Kapitel werden die bestehenden Rechtslagen, nach denen KIs derzeit reguliert werden, sowie Ansätze und Empfehlungen behandelt, die die Grundlage für zukünftige Regulierungen darstellen.

6.1 Bestehendes Recht

In den in Kapitel 4 beschriebenen konkreten Anwendungsfällen - vor allem in denen des Militärs und des Rechtsstaates - scheint es eine moralische Grenze zu geben, ab der die Anwendung von KIs kontrovers wird bzw. für einen öffentlichen Diskurs sorgt. Dabei handelt es sich oft um die Grundsatzfrage, inwiefern KI unabhängig vom Menschen Entscheidungen treffen sollen. Während KIs meinen Recherchen zu folge als entscheidungsunterstützende Werkzeuge bzw. als „Berater" gesellschaftlich akzeptiert zu sein scheinen, sind autonom agierende KIs dies nicht.

Maßgebend für die lokale Regulierung von NNs ist die am 27.4.2016 veröffentlichte und am 25.5.2018 in Kraft getretene EU-Datenschutzgrundverordnung (DSGVO).[206] Diese regelt vor allem den Schutz personenbezogener Daten sowie deren Verarbeitung und Analyse, wodurch im Besonderen KIs bzw. NNs aufgrund der für sie benötigten Daten betroffen sind. Darüber hinaus ist die DSGVO und deren nationale Implementierung bzw. Anpassung im Rahmen des Datenschutz-Deregulierungsgesetz 2018[207] derzeit der einzige in Österreich rechtlich bindende Gesetzestext, der KIs bzw. NNs direkt betrifft.

Recht auf Diskriminierungsfreiheit

Von besonderer Bedeutung für KIs ist Artikel 22, der sich mit dem Thema „*Automatisierte Entscheidungen im Einzelfall einschließlich Profiling*" befasst und jeder Person grundsätzlich das Recht gibt, „*nicht einer ausschließlich auf einer automatisierten Verarbeitung — einschließlich Profiling — beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt*"[208]. Diese Regelung ermöglicht es den EU-Bürgern, grundsätzlich gegen jede automatisierte Entscheidung einer KI Einspruch zu erheben, falls diese sie „erheblich" beeinträchtigt. Wird eine solche automatisierte Entscheidung jedoch für einen Vertragsabschluss benötigt, ist dies ebenso erlaubt wie in Fällen, die diesen Einsatz nach nationalem Recht explizit erlauben.[209] Zusätzlich ist dies zulässig, wenn die betroffene Person dem Einsatz automatisierter Entscheidungssysteme ausdrücklich zustimmt.[210]

In Erwägungsgrund 71 werden dazu einige Beispiele genannt: Nicht zulässig ist demnach die automatisierte Prüfung von Online-Kreditanträgen, aber auch das automatisierte Profiling. Bei diesem werden persönlichen Daten von einem NN analysiert, um Prognosen zur zukünftigen Arbeitsleistung oder der wirtschaftlichen Lage einer Person treffen zu können. Beispiele für

[206] Verordnung (EU) 2016/679 (Datenschutz-Grundverordnung). (27.4.2016)
[207] Datenschutz-Deregulierungsgesetz 2018. (15. Mai 2018).
[208] Art 22 Abs 1 Datenschutz-Grundverordnung (DSGVO).
[209] Art 22 Abs 2 Datenschutz-Grundverordnung (DSGVO).
[210] Art 22 Abs 3 Datenschutz-Grundverordnung (DSGVO).

erwünschte Ausnahmen werden ebenfalls dargelegt: So soll automatisiertes Profiling erlaubt bleiben, wenn es die Steuerfahndung unterstützt.

In Fällen, in denen die betroffene Person nun eine Einwilligung gibt bzw. dies für die Vertragserfüllung erforderlich ist, wurden zusätzlich weitere Einschränkungen definiert: So hat die betroffene Person das Recht darauf, dass eine Person in den Prozess eingreift, darauf, ihren eigenen Standpunkt darlegen zu können, und darauf, jede durch automatisierte Entscheidungsfindung entstandene Entscheidung anzufechten.[211]

Weiters gibt es verbindliche Einschränkungen in Bezug auf die im Rahmen eines automatisierten Profiling verwendeten personenbezogenen Informationen. Auf Daten, aus denen laut Artikel 9 *„die rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen, sowie die Verarbeitung von genetischen Daten, biometrischen Daten zur eindeutigen Identifizierung einer natürlichen Person, Gesundheitsdaten oder Daten zum Sexualleben oder der sexuellen Orientierung einer natürlichen Person"*[212] dürfen Entscheidungen nur dann beruhen, wenn die betroffene Person dem einwilligt bzw. dies von großem öffentlichen Interesse ist. Doch auch hier kann der Staat der Person das Recht auf Einwilligung entziehen.[213]

Die Regelung des Artikel 22 verfolgt das legistische Ziel, die unverhältnismäßige Diskriminierung, bzw. die in Kapitel 5.1, 5.3 und 4.2 dargelegte Problematik des Bias in KIs zu verbieten. Die Arbeit der Wissenschaftler Bryce Goodman und Seth Flaxman der Oxford University - die sich ausführlich mit den Folgen dieses Artikels befasst und im Laufe dieses Kapitels mehrmals als Quelle dient – behandelt die dadurch entstehenden Auswirkungen auf NNs bezogen auf zwei verschiedene Interpretationsvarianten.[214]

Nach der minimalen Interpretation dürfen alle Daten verwendet werden, die nicht explizit in eine der oben genannten Kategorien fallen. Das Problem hierbei ist, dass NNs auch ohne explizite Informationen über diese Kategorien in den Trainingsdaten dennoch nach diesen diskriminieren könnten, z.B. indem implizit aus einer Kombination anderer Merkmale Schlüsse auf die unzulässigen Kategorien gezogen werden. (Für Beispiele siehe Kapitel 5.3*)*

Die maximale Interpretation wählt daher den strengeren Ansatz, wonach auch alle mit den oben genannten Kategorien korrelierten Daten unter diesen Absatz fallen. Die Oxford-Wissenschaftler merken hier jedoch an, dass es bei großen Datensätzen nur sehr schwer bzw. gar nicht mehr möglich ist, alle möglichen relevanten Korrelationen zu erkennen. Zudem schadet das Entfernen dieser korrelierenden Daten potentiell der Genauigkeit des NN. Im Beispiel des Credit-Scoring (siehe Kapitel 2.1) würde das bewusste Ignorieren der Wohnadresse, die in vielen Fällen einen Hinweis auf die ethnische Herkunft und andere geschützte Kategorien gibt, vermutlich zu einer wesentlich geringeren Genauigkeit der Ergebnisse führen. Schlussendlich merken die Wissenschaftler an, dass nach der maximalen Interpretation auch die in Kapitel 5.3 behandelten Diskriminierungen durch Datenquantität

[211] Art 22 Abs 3 Datenschutz-Grundverordnung (DSGVO).
[212] Art 22 Abs 4 Datenschutz-Grundverordnung (DSGVO).
[213] Art 9 Abs 2 lit a,g Datenschutz-Grundverordnung (DSGVO).
[214] Bryce Goodman, Seth Flaxman. (2016). European Union regulations on algorithmic decision-makingand a "right to explanation"

(ungleichmäßige Verteilung von Trainingsdaten) und Datenqualität (Trainingsdaten spiegeln nicht die Realität wider) unrechtmäßig wären.

Zusammenfassend bleibt die Verwendung von KIs bzw. NNs auch in Anwendungen erlaubt, die große persönliche Auswirkung auf einzelne Personen haben, sofern diese dem ausdrücklich zustimmen und nationale Gesetze dem nicht widersprechen. Sollte sich in der Rechtsprechung jedoch die maximale Interpretation durchsetzen, so müssten betroffene Personen der Verwendung potentiell diskriminierender Merkmale explizit zustimmen, da wie oben erwähnt eine unrechtmäßige Diskriminierung über korrelierende Daten nicht ausgeschlossen werden kann.

Recht auf Auskunft

Die DSGVO geht aber über die bis jetzt erwähnten, auf Datenschutz fokussierten Bestimmungen hinaus. In Kapitel 5 wurde bereits das Blackbox Phänomen moderner Technologien beschrieben. Diesem zufolge werden im Alltag viele Technologien verwendet, die für die Anwender - nicht nur aufgrund deren Komplexität - undurchsichtig sind und so eine Blackbox bilden.

Der Artikel 13, 14 und 15 der DSGVO versuchen hier mehr Transparenz zu schaffen: Für Datenverantwortliche gilt eine Informationspflicht, für Bürgerinnen und Bürgern der EU ein Recht auf Auskunft. Diese Artikel befassen sich großteils mit datenschutzrelevanten Inhalten: Welche persönlichen Daten werden gesammelt, wofür werden sie verwendet und an wen weitergegeben.

Artikel 13 Abs. 2 lit. f bzw. Artikel 14 Abs. 2 lit. g und Artikel 15 Abs. 1 lit. h befassen sich mit der in Artikel 22 definierten und hier bereits behandelten automatisierten Entscheidungsfindung sowie dem automatisierten Profiling. So muss die betroffene Person darüber informiert werden, dass eine automatisierte Entscheidung getroffen wird bzw. Profiling angewendet wird. Zusätzlich müssen *„aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung“* für die betroffene Person offengelegt werden.

Bezieht man diese Aussage auf klassische KIs, so ist klar, welche Informationen der betroffenen Person mitgeteilt werden müssen. Im Falle des Credit-Scoring (siehe Kapitel 2.1) wären das die Themen „Wie gelangt der Algorithmus zu seiner Entscheidung?“ und „Wie stark wird jede Information gewichtet?“, also „Wie stark fließt die einzelne Information in die Entscheidung ein?“. Zusätzlich müsste natürlich auch mitgeteilt werden, dass auf Basis des Ergebnisses angestrebte Leistungen - wie z.B. ein Kredit - abgelehnt werden können.

Bezogen auf NNs hat derzeit aufgrund der unklaren Formulierung des Gesetzestextes der Datenverantwortliche einen großen Interpretationsspielraum, wie weit er „aussagekräftige“ Informationen mitteilt. Ein Vorgehen gemäß der maximalen Interpretation ist hier nicht möglich, da die Frage des „Warum?“ im Einzelfall nicht geklärt werden kann. Über die in Kapitel 5 beschriebenen Analysemethoden kann der betroffenen Person nur dargestellt werden, wie das NN in dem konkreten Fall zu der Lösung gekommen ist. Die Antwort auf diese Frage ist aber vor allem im Zusammenhang mit dieser Rechtsvorschrift wichtig, da diese nur Fälle behandelt, die große Auswirkungen auf die betroffene Person haben. Nehmen wir an, die in Kapitel 4.2 beschriebene Kautions-KI wird in einem EU-Mitgliedsstaat

implementiert. Auf die Frage eines Verhafteten, warum er nun aufs erste in Haft bleiben muss, kann in diesem Fall beispielsweise nur geantwortet werden, dass seine Wohnadresse und sein Verhaftungsort maßgeblich waren und sein bisheriger Strafverlauf so gut wie gar nicht in die Entscheidung eingeflossen sind. Auf die Frage, warum genau diese Informationen die Entscheidung maßgeblich beeinflusst haben, kann die Antwort nur lauten: Nicht, weil dies auf menschlichen Erfahrungswerten beruht, sondern weil es das NN so gelernt hat.

In Österreich wurde Artikel 15 der DSGVO zusätzlich eingeschränkt. So gilt das Auskunftsrecht unter anderem dann nicht, wenn dieses von einem „hoheitlich", also staatlich tätigen Verantwortlichen eingefordert wird, und dies seine gesetzlich übertragene Aufgabe gefährden würde.[215] Eine mögliche Begründung für diese Einschränkung ist, dass das Verhalten der Anwendungsunterworfenen nicht im Wissen um die Funktionsweise des NN geändert werden darf, z.B. indem kriminelle Personen einen Wohnsitz „vorausschauend" wählen. Eine öffentlichkeitswirksame Diskussion über die Funktionsweise eines NN soll auch deshalb vermieden werden, da ansonsten bewährte, im Einsatz befindliche NNs aufgrund politischen Drucks oder einer gerichtlichen Anweisung hin eventuell nicht mehr verwendet werden dürften. Das könnte kurzfristig zu Verzögerungen und Einschränkungen bei alltäglichen Behördengängen und Anträgen führen. Sollte der rechtliche Rahmen längerfristig nicht präzisiert werden, so könnte dies zu einer ernstzunehmenden Gefahr werden, da in Folge des Einsatzes von NNs für diverse Aufgaben die Anzahl an Beamten bzw. Verantwortlichen mit staatlichem Auftrag aus Kostengründen reduziert wäre und die verbleibenden aufgrund fehlender Erfahrung und Einschulung stark auf entscheidungsunterstützende und automatisch entscheidende NNs angewiesen sind.

Artikel 15 wird in Österreich zusätzlich auch zum Vorteil für Unternehmen eingeschränkt, nämlich wenn durch das Bereitstellen der Informationen ein *„Geschäfts-oder Betriebsgeheimnis des Verantwortlichen bzw. Dritter gefährdet würde"*.[216] Das Geschäftsmodell von Unternehmen, die auf die Verarbeitung großer Datenmengen (BIG-Data) spezialisiert sind, basiert oft auf den selbst entwickelten Algorithmen. Ein Beispiel dafür ist das Technologieunternehmen Google und sein bekannter Such-Algorithmus. Eine Offenlegung dieses Algorithmus über das Recht auf Auskunft würde ein Kopieren durch Konkurrenten ermöglichen. Zusätzlich wäre davon auszugehen, dass Unternehmen und Privatpersonen öffentlich zugängliche Informationen über den Suchalgorithmus missbrauchen würden, um die eigene Website ungerechtfertigt in den Suchergebnissen zu priorisieren und dadurch mehr Kunden anzulocken.

Recht in der Praxis

Wie erwähnt lassen die rechtlichen Bestimmungen zum Diskriminierungsverbot und dem Recht auf Auskunft viel Interpretationsspielraum zu. Einerseits erlaubt dieser Freiraum erst die praktische Verwendbarkeit von NNs, andererseits kann dieser auch gezielt ausgenutzt werden.

In Kapitel 4.2 wurde beispielhaft dargelegt, wie Kalifornien KIs einsetzt, um Kautionsrisiken von Häftlingen zu bewerten. Diskriminierung kann hier bereits in der Zieldefinition entstehen:

[215] §4 Abs. 5 Datenschutzgesetz (DSG). Fassung vom 26.5.2019.
[216] §4 Abs. 6 Datenschutzgesetz (DSG). Fassung vom 26.5.2019.

Ein simuliertes System konnte eine Verbrechensreduktion von bis zu 27,4% bei gleichbleibender Häftlingsrate bzw. eine Häftlingsreduktion von bis zu 41,9% bei gleichbleibender Verbrechensrate erzielen - welches der beiden Ziele für den praktischen Einsatz verwendet wird, ist eine politische Entscheidung. In Bezug auf staatliche Aufgabenstellungen sorgen regelmäßig Wahlen dafür, dass Zieldefinitionen von NNs mit dem Willen der Bevölkerung übereinstimmen – ansonsten würde eine Regierung ihre Wiederwahl gefährden.

Der Austausch von KIs könnte sich aber nicht nur wegen den dadurch entstehenden Kosten als problematisch herausstellen: Während davon auszugehen ist, dass die Bevölkerung trotz der Verpflichtung der Staatsdiener zu einer objektivierten Entscheidungsfindung ein gewisses Ausmaß an menschlicher Subjektivität akzeptiert, könnte dies bei KIs nicht der Fall sein. An die Ergebnisse der KIs wird vermutlich ein gewisser Perfektionsanspruch gestellt werden, d.h. es wird erwartet, dass die Entscheidungen diskriminierungsfrei und wiederholbar sind. Von menschlichen Entscheidern wird zudem gefordert, dass unter regelbasierten Vorgaben der Austausch des Sachbearbeiters keinen Einfluss auf die Entscheidung aufweist. Bei einem Wechsel von KIs würde es aber unvermeidbar zu geänderten Falleinschätzungen kommen - einem Häftling würde die Kaution durch das neue System beispielsweise gewährt werden, durch das alte jedoch nicht. Derartige Fälle wären bezogen auf das für staatliche Stellen gebotene Gleichbehandlungsprinzip problematisch und müssten öffentlich diskutiert werden, um das Bewusstsein zu schaffen, dass bei dem Einsatz von KIs in Einzelfällen selbst große Abweichungen eine untergeordnete Rolle spielen, solange das Gesamtergebnis die Erwartungen erfüllt. Die verantwortlichen Stellen würden daher vermutlich vermeiden, KIs zu rasch auszuwechseln, wodurch politische Entscheidungen bezüglich der eingesetzten KI-Systeme Auswirkungen haben, die über die jeweilige Legislaturperiode hinausgehen könnten.

Die betroffene Person kann zwar immer von ihrem Recht laut Artikel 22 der DSGVO Gebrauch machen und Einspruch gegen das Resultat erheben, dadurch entsteht jedoch ein weiteres Dilemma: Die verantwortliche Person, die sich mit dem Einspruch befasst (z.B. der Richter bzw. die Richterin), sieht sich mit der hohen, für sich sprechenden, Erfolgsquote der KI konfrontiert. Diese könnte daher - im Sinne eines Einschlagens des bequemeren Weges - dazu neigen, die Entscheidung der KI zu bestätigen, selbst wenn sie formal das letzte Wort hat und sich nicht durch die Entscheidung der KI beeinflussen lassen darf. In dem Beispiel der Kautionsrisikobewertung scheint es wahrscheinlich, dass das Gewähren einer Kaution gegen die diesbezügliche individuelle Empfehlung der KI mit einer wahrgenommenen Verantwortung für das Verhalten des Beschuldigten und einem damit eingehenden Risiko verbunden ist. Im Beispiel eines Bankberaters bei der Kreditvergabe könnte dieser dazu neigen, den Entscheidungen der KI auch in als nicht gerechtfertigt empfundenen Fällen zu folgen um dadurch den eigenen Arbeitsaufwand – Abweichungen von den Empfehlungen der KI müssten vermutlich begründet und dokumentiert werden – zu minimieren.

Durch rechtliche Regelungen lässt sich dieses Dilemma vermutlich nicht ganz lösen. Wahrscheinlicher scheint, dass Schulungen zur Sensibilisierung der Verantwortlichen sowie ein besseres Verständnis über die Funktionsweisen von KIs eine größere Wirkung erzielen können. Wird beispielsweise klar gemacht, dass auch KIs mit einer – deutlich über der menschlichen liegenden – Erfolgsquote von z.B. 89% anfällig für grobe Fehlentscheidungen, beispielsweise unerwartete schwere Fehler bei der Steuerung eines autonomen Fahrzeuges,

sind, so würde dies dem Verantwortlichen eine mögliche Entscheidung gegen die KI-Empfehlung erleichtern.

Konkretisierungen

Über die DSGVO wurde ein Europäischer Datenschutzausschuss ins Leben gerufen, der unter anderem dafür sorgen soll, dass die Verordnung einheitlich in allen EU-Mitgliedsstaaten angewendet wird.[217] Konkret sollen laut Artikel 70 Absatz 1, „*Leitlinien, Empfehlungen und bewährte Verfahren*" bezüglich Artikel 22 Absatz 2 - also den Ausnahmen bezüglich automatisierter Entscheidungsfindung und Profiling - bereitgestellt werden. Diese Leitlinien würden somit konkretisieren, inwiefern diskriminierende Daten bei KIs verwendet werden dürfen, und auch eine Grundlage für nationale Datenschutzbehörden darstellen, auf denen dann Ermittlungen aufgrund vermuteter Verstöße durchgeführt werden würden.

Derzeit herrscht jedoch noch Unsicherheit darüber, wie Artikel 22 genau auszulegen ist, da der Europäische Datenschutzausschuss derzeit noch kein einziges Dokument bezüglich „künstlicher Intelligenzen"[218] und „automatisierter individueller Entscheidungsfindung"[219] veröffentlicht hat.

6.2 Zukünftiges Recht und Lobbying

Vereinigte Staaten

Im April 2019 brachten zwei US-Senatoren einen Gesetzesvorschlag zur Regulierung automatisierter Entscheidungssysteme ein.[220] Der „Algorithmic Accountability Act" soll - ähnlich wie die DSGVO - Systeme regulieren, bei denen durch automatisierte Entscheidungen ein großer Einfluss auf die persönlichen Lebensumstände von Menschen ausgeübt wird.[221] Neben Datenschutz und Sicherheit stehen hier vor allem Bias und Diskriminierung im Vordergrund.[222] Es werden zwar auch hier Kategorien von sensiblen, potentiell diskriminierenden Daten angegeben, deren Anwendung wird jedoch in keiner Weise eingeschränkt.[223] Stattdessen werden in Section 2 Studien zur Folgenabschätzung („*impact assesments*") definiert, die für neue Entscheidungssysteme vor deren Einführung bzw. für bestehende Systeme zwei Jahre nach Inkrafttreten des Gesetzes (wenn möglich durch unabhängige Dritte) durchgeführt werden müssen.

Im Gegensatz zu den Auskünften nach der DSGVO sind die Inhalte der Folgenabschätzung in dem Algorithmic Accountability Act konkreter definiert: So müssen diese laut Section 2, 2 „*eine genaue Beschreibung des automatisierten Entscheidungssystems, des Designs, Trainingsablaufs, Daten und Verwendungszweck*" sowie eine weitreichende Kosten/Nutzen Analyse beinhalten, die sich unter anderem auch mit Datenminimierung befasst, also mit Vorgehensweisen zur Reduzierung der Anzahl der verwendeten Informationen. Zusätzlich

[217] Art 70 Abs 1 Datenschutz-Grundverordnung (DSGVO).
[218] Europäischer Datenschutzausschuss. (31.5.2019). Artificial Intelligence
[219] Europäischer Datenschutzausschuss. (31.5.2019). Automated individual decision-making
[220] Ron Wyden US-Senator for Oregon. (10.4.2019). Wyden, Booker, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms.
[221] Algorithmic Accountability Act. Section 2, 7
[222] The Verge. (10.4.2019). A new bill would force companies to check their algorithms for bias.
[223] Algorithmic Accountability Act. Section 2, 7, lit. C

muss auch eine Risikoanalyse erstellt werden, die auf die Sicherheit persönlicher Daten und der Möglichkeit unverhältnismäßig diskriminierender Resultate des Systems fokussiert ist und die mögliche „*technische und physikalische Sicherheitsmaßnahmen*“ beinhaltet, um dies zu verhindern.[224]

Schlussendlich betrifft der Algorithmic Accountability Act nicht alle Anbieter dieser Systeme, sondern nur jene, die mindestens 50 Millionen US-Dollar jährlichen Umsatz aufweisen bzw. Kundeninformationen über eine Millionen Individuen oder eine Millionen Endgeräte besitzen.[225]

Europäische Union

Auf europäischer Ebene existieren aktuell keine Gesetzesentwürfe zur weiteren Regulierung von KIs bzw. NNs. Es werden jedoch laufend Kommunikationspapiere veröffentlicht, die den Standpunkt der europäischen Kommission sowie deren weitere geplante Schritte darlegen.

Den Grundstein legt der im Dezember 2018 veröffentlichte „Koordinierte Plan für künstliche Intelligenz“.[226] Neben Planungen zur Förderung und erhöhten Koordination von KI-Tätigkeiten in der EU behandeln der Plan und dessen Anhang auch die „Integrierte Ethik“ und den regulatorischen Rahmen. In diesen Kapiteln hält die Kommission daran fest, Ethik als zentrales Element in jeder Entwicklungs- und Anwendungsphase einer KI integrieren zu wollen. Zusätzlich kündigt sie an, über *„mögliche Lücken in dem Sicherheits- und Haftungsrahmen“* der derzeitigen Gesetzeslage bezüglich KIs berichten, Lösungsansätze zur rechtlichen Klärung der Frage der Haftung für das Fehlverhalten von KIs entwickeln und Sicherheitsstandards festlegen zu wollen. Zukünftiges Recht soll laut dem Papier unter Berücksichtigung der Beiträge von Mitgliedsstaaten und einer eigens eingerichteten Expertengruppe[227] vorgeschlagen werden.

Eine der Aufgaben dieser Expertengruppe war es, Ethik-Richtlinien für „vertrauenswürdige KIs“ zu verfassen, welche im April 2019 veröffentlicht wurden.[228] Die Expertengruppe nennt dabei drei Komponenten, die eine vertrauenswürdige KI ausmachen und in jeder Phase der KI von der Entwicklung bis hin zur Ausmusterung eingehalten werden sollten: *(i) gesetzliche Konformität*, *(ii) ethische Konformität* und *(iii) Robustheit* (Robustheit bezieht sich hierbei auf die Fehleranfälligkeit der KI). In den Richtlinien wird angemerkt, dass gesetzliche und ethische Konformität alleine nicht ausreichen, um eine KI vertrauenswürdig zu machen, da diese auch trotz guter Absicht schwere Fehlentscheidungen treffen könnte.

Auf Basis dieser drei Komponenten wurden sieben Anforderungen formuliert, die eine vertrauenswürdige KI gewährleisten muss:

- Menschliche Kontrolle und Aufsicht
- Technische Robustheit und Sicherheit
- Privatsphäre und Datenkontrolle
- Transparenz

[224] Algorithmic Accountability Act. Section 2, 2, lit. C-D
[225] Algorithmic Accountability Act. Section 2, 5
[226] Europäische Kommission. (7.12.2018). Coordinated Plan on Artificial Intelligence.
[227] Europäische Kommission. (4.6.2019). High-Level Expert Group on Artificial Intelligence.
[228] Europäische Kommission. (8.4.2019). Ethics guidelines for trustworthy AI.

- Diversität, Nicht-Diskriminierung und Gerechtigkeit
- Umwelt- und soziale Verträglichkeit
- Verantwortung

Diese Grundsätze wurden von der EU-Kommission übernommen und in einem weiteren Kommunikationspapier[229] als Empfehlungen an das EU-Parlament und weitere Institutionen weitergeleitet.

Zur Gewährleistung menschlicher Kontrolle soll laut dem Entwurfspapier jederzeit ein Mensch die Kontrolle über die KI übernehmen, eine Übersicht über die Aktivitäten dieser erlangen und falls nötig durch KIs getroffene Entscheidungen ändern können. Zusätzlich soll sichergestellt werden, dass die Exekutive und andere zuständige Behörden im Rahmen ihrer Aufgaben ebenfalls diese Berechtigungen erwirken können. Schlussendlich empfiehlt die Kommission höhere Autonomiegrade, wie beispielsweise bei autonomen Fahrzeugen, an striktere Regelungen zu binden, um das Vertrauen in die KI nicht zu gefährden.

Unter technischer Robustheit und Sicherheit versteht die EU-Kommission konkret, dass KIs sowohl offenen als auch versteckten Angriffen standhalten können müssen und im Notfall auf eine Art Plan B zurückfallen sollen. Zur Erlangung der notwendigen Robustheit sollen alle Entscheidungen der KI reproduzierbar sein und Sicherheit in jedem Schritt des Lebenszyklus einer KI verankert werden.

Dies soll auch für den Datenschutz gelten. Daher soll zusätzlich zur Einhaltung bestehender und zukünftiger Datenschutzbestimmungen (u.a. der DSGVO) auch die Datenqualität der Trainingsdaten (siehe Kapitel 5.3) in der Entwicklung berücksichtigt und dokumentiert werden.[230]

Transparenz soll sichergestellt werden, indem jeder Schritt in der Entwicklung einer KI dokumentiert wird. Zusätzlich dazu soll auch die Nachvollziehbarkeit (wie in der DSGVO verlangt) der Ergebnisfindung soweit wie möglich sichergestellt werden. Das bezieht sich nicht nur auf die Funktionsweise der KI, sondern auch warum diese eingesetzt wird und welche Rolle sie in dem Gesamtprozess spielt. Weiters muss sich eine KI immer als eine solche zu erkennen geben und involvierte Personen müssen über deren Grenzen bzw. Schwächen aufgeklärt werden.[231]

Die EU-Kommission setzt sich in dem Papier auch dafür ein, Gerechtigkeit und Fairness in KIs sicherzustellen. Gelingen soll dies, indem beispielsweise Vertreter von potentiell von diskriminierender Behandlung betroffenen Gruppen in den Entwicklungsprozess eingebunden werden und im besten Fall sogar Teil des Entwicklerteams sind.

Weiters soll laut dem Kommunikationsdokument ein verantwortungsvoller Umgang mit der Umwelt, mit demokratischen Prozessen und der Gesellschaft als Ganzes in KI-Entscheidungen verankert werden.

[229] Europäische Kommission. (8.4.2019). Communication: Building Trust in Human Centric Artificial Intelligence.
[230] Europäische Kommission. (8.4.2019). Communication: Building Trust in Human Centric Artificial Intelligence.
[231] Europäische Kommission. (8.4.2019). Communication: Building Trust in Human Centric Artificial Intelligence.

Unter dem letztgereihten der sieben Punkte - dem Punkt „Verantwortung“ - versteht die EU-Kommission eine Folgenabschätzung über mögliche trade-offs und potentielle negative Auswirkungen der KI im weitesten Sinne. Derartige Analysen, die von internen und externen Prüfern durchgeführt werden sollte, würden laut EU-Kommission zu einer erhöhten Vertrauenswürdigkeit von KIs führen.

Diese sieben Punkte stellen aber keine explizite Forderung nach direkten regulatorischen Maßnahmen der Mitgliedsstaaten dar. Als vorgelagerte Maßnahme wurden ab Juni 2019 Interessensgruppen aus dem privaten und öffentlichen Sektor dazu eingeladen, die Umsetzbarkeit dieser Punkte in der Praxis zu prüfen und Feedback bzw. Veränderungswünsche an die KI-Expertengruppe weiterzuleiten.[232] Nach Ende dieser Pilotphase per Anfang 2020 wird die Expertengruppe eine aktualisierte Version der Richtlinie publizieren, auf deren Basis die EU-Kommission dann weitere Schritte setzen kann.[233]

Österreich

In Österreich veröffentlichten das Bundesministerium für Verkehr, Innovation und Technologie sowie das Bundesministerium für Digitalisierung und Wirtschaftsstandort 2018 einen ersten Entwurf für eine nationale KI-Strategie.[234] Der als „Artificial Intelligence Mission Austria 2030” bezeichnete Bericht[235] stellt jedoch eher eine Broschüre dar, in der Beispiele von in Österreich entwickelten KIs gelistet sind und KI-Zukunftsfelder genannt werden. Konkrete Absichten fehlen in dem Dokument, unter dem Zukunftsfeld „Gesellschaft, Ethik und Arbeitsmarkt“ wird jedoch zum Dialog zwischen Vertretern verschiedener Interessensgruppen aufgerufen, um die Herausforderungen dieses Feldes meistern zu können.

International (außerhalb Europas)

Auf außereuropäischer Ebene existierten bis vor kurzem keine multinationalen Abkommen bzw. Regelungen zum Einsatz von KIs. Im Mai 2019 unterzeichneten die 36 OECD-Staaten zusammen mit sechs weiteren Nationen ein erstes Dokument zur Regulierung von KI.[236] Die fünf in dem rechtlich nicht bindenden Dokument[237] enthaltenen Prinzipien decken sich dabei mehr oder weniger mit denen der EU-Kommission: Diese sind die Prinzipien *der (i) Umwelt- und sozialen-Verträglichkeit, (ii)Diversität*, *Nicht-Diskriminierung und Gerechtigkeit*, *(iii)Transparenz*, *(iv)Technische Robustheit und Sicherheit* und *(v)Verantwortung*.

Bei einem G20-Treffen der Wirtschaftsminister in Japan wurde eine Erklärung abgegeben, die einige Prinzipien des OECD-Dokuments widerspiegelt.[238] [239] Hierbei hat man sich darauf verständigt, dass eingesetzte oder entwickelte KIs die Menschenrechte, demokratische Werte

[232] Europäische Kommission. (8.4.2019). Communication: Building Trust in Human Centric Artificial Intelligence.
[233] Europäische Kommission. (8.4.2019). Communication: Building Trust in Human Centric Artificial Intelligence.
[234] Bundesministerium für Digitalisierung und Wirtschaftsstandort. (9.6.2019). Künstliche Intelligenz (KI).
[235] Bundesministerium für Verkehr, Innovation und Technologie und Bundesministerium für Digitalisierung und Wirtschaftsstandort. (2018). Artificial Intelligence Mission Austria 2030.
[236] OECD. (22.5.2019). Forty-two countries adopt new OECD Principles on Artificial Intelligence.
[237] OECD. (2019). Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449
[238] Futurezone. (9.6.2019). "Robust und sicher": G20-Staaten einigen sich auf KI-Prinzipien.
[239] The Japan Times. (8.6.2019). G20 ministers agree on guiding principles for using artificial intelligence.

und Rechtsgrundsätze respektieren müssen. Zusätzlich wurde auch die Sicherheit behandelt: KIs sollen laut der Minister „robust, gesichert und sicher“ sein.[240] Neben diesen Prinzipien soll bei KI-Anwendungen generell der Mensch im Mittelpunkt stehen.[241]

Kritik an Lobbying

Nach Erscheinen des EU-Ethikrichtlinien-Dokuments trat eines der Mitglieder des KI-Expertenrates, Thomas Metzinger[242] von der Universität Mainz, an die Öffentlichkeit, um den Einfluss großer Technologieunternehmen auf die Entstehung der Ethikrichtlinien zu kritisieren. In einem Artikel des Magazins „Wired“[243] gibt Metzinger an, im Rahmen seiner Tätigkeit im Expertenrat die Aufgabe übernommen zu haben, eine Liste von Anwendungsgebieten zu erstellen, die zukünftig verboten werden sollten. Diese Verbotsliste - auf der sich ursprünglich auch autonome Waffen und soziale Kreditwürdigkeitssysteme befanden - wurde später in eine Liste von Anwendungen reduziert, die als „besonders besorgniserregend“ einzustufen sind. Die Verhinderung einer Verbotsempfehlung ist laut Metzinger dem Einfluss der Technologiekonzerne geschuldet. Zwar sitzen keine direkten Vertreter großer Technologieunternehmen in der Expertengruppe, wohl jedoch Vertreter der Lobbygruppe „DigitalEurope“, die Interessen dieser Unternehmen vertreten.[244]

Neben Thomas Metzinger kritisierte auch der Rechtsprofessor Yochai Benkler von der Harvard University den hohen Einfluss der Technologieunternehmen auf die KI-Gesetzgebung.[245] Seiner Meinung nach können sich Unternehmen in Sachen KI nicht selbst regulieren, da von auf Profitmaximierung ausgelegten Unternehmen entwickelte KIs notwendigerweise öffentliche Interessen verletzen würden. Zusätzlich kritisiert er, dass Forschung im KI-Bereich überwiegend in den Händen dieser großen Unternehmen liege und dass von diesen Unternehmen finanzierte Studien - vor allem bezüglich Ethik und Moral in KI - als Basis für zukünftige gesetzliche Regelungen dienen würden.

Zuletzt warnt Yochai Benkler aber auch vor einer Überregulierung von KIs. In seinen Augen ist Technikskepsis genauso wenig zielführend wie die Technikeuphorie großer Unternehmen.[246]

[240] Futurezone. (9.6.2019). "Robust und sicher": G20-Staaten einigen sich auf KI-Prinzipien.
[241] Futurezone. (9.6.2019). "Robust und sicher": G20-Staaten einigen sich auf KI-Prinzipien.
[242] Univ.-Prof. Dr. Thomas Metzinger.
[243] Wired. (16.5.2019). How Tech Companies Are Shaping the Rules Governing AI.
[244] Wired. (16.5.2019). How Tech Companies Are Shaping the Rules Governing AI.
[245] Nature. (1.5.2019). Don’t let industry write the rules for AI.
[246] Nature. (1.5.2019). Don’t let industry write the rules for AI.

Conclusio

Künstliche Intelligenzen sind leichter zu implementieren als je zuvor: Bei klassischen KIs wird sowohl auf technischer Ebene als auch auf der Ebene der Anwendung ein komplettes Expertenteam monatelang benötigt – ein neuronales Netzwerk kann ein einziger Programmierer mit dem entsprechenden Know-How, einem leistungsstarker Computer und einer großen Datenmenge in viel kürzerer Zeit entwickeln. Selbst wenn vielfach bei Qualität und Quantität der (teuren) benötigten Daten gespart wird, können die Ergebnisse dieses NN auf den ersten Blick in vielen Anwendungsgebieten überzeugen. Bei genauerem Hinsehen zeigt sich jedoch, dass ohne ausreichende Einbeziehung von Experten des jeweiligen Anwendungsgebietes entwickelte NNs in vom Regelfall abweichenden Situationen zu Fehleinschätzungen und unrechtmäßiger Diskriminierung neigen. Auch wenn NNs unmöglich auf jede in der Realität auftretende Situation getestet werden können, ist es mit neuen Analysemethoden, wie der topologischen Datenanalyse und *heat maps*, dennoch möglich, grobe Diskriminierung und Fehler zu erkennen bzw. deren Ursachen im Einzelfall festzustellen. Das Fehlen von Experten auf technischer Ebene erhöht die Wahrscheinlichkeit, dass die entwickelten KIs anfälliger für verschiedene Arten von Attacken sind, die den Nutzer im Extremfall in lebensgefährliche Situationen bringen können. Derzeit werden diese Angriffsmöglichkeiten vor allem unter wissenschaftlichen Aspekten erforscht. Mit der weitergehenden Integration von NNs in den gesellschaftlichen Alltag kann man aber annehmen, dass das aus der Internetsicherheit bekannte Katz-und-Maus-Spiel zwischen Angreifern und Sicherheitsforschern in naher Zukunft mit entsprechend gravierenden Auswirkungen auch auf NNs erweitert wird.

Es ist davon auszugehen, dass unserer Gesellschaft durch den zu erwartenden technologischen Fortschritt in diesem Bereich nachhaltig verändert werden wird. Entsprechend gilt es bei der regulatorischen Begleitung dieser Entwicklung Chancen und Risiken ausgewogen zu berücksichtigen. Die derzeit geltenden rechtlichen Regulierungen, insbesondere Artikel 22 der DSGVO, bieten keine Lösung für das breite Spektrum an Herausforderungen, die KIs mit sich bringen. Zwar wurden die ersten Schritte für zukünftige Regelungen bereits gesetzt, da bei vergleichbaren Regulierungen wie der DSGVO jedoch sechs Jahre vom ersten offiziellen Dokument bis zum Inkrafttreten vergingen, kann wohl erst ab ca. 2025 mit rechtlichen bindenden Standards bzw. Gütesiegeln für KIs auf europäischer Ebene gerechnet werden.[247]

[247] EUR-Lex. (25.6.2019). Verfahren 2012/0011/COD

Abkürzungen

CNN……Convolutional Neural Network

DARPA……Defense Advanced Research Projects Agency

DSGVO……...……EU-Datenschutzgrundverordnung

EU……...Europäische Union

KI……. künstliche Intelligenz

MIT……..……Massachusetts Institute of Technology

NN……... Neuronales Netzwerk

PC……..….Personal Computer

US……Vereinigte Staaten

Abbildungen

Abbildung 1: Beispiel eines einfachen Entscheidungsbaumes für das Credit-Scoring 5
Abbildung 2: Beispiel eines einfachen Neuronalen Netzwerks 7
Abbildung 3: Veranschaulichung eines Convolutional Neural Network(CNN) 12
Abbildung 4: Ein Dilemma aus dem Moral Machine Projekt 16
Abbildung 5: Kulturelle Unterschiede beim Moral Machine Experiment 17
Abbildung 6: Eine Nachricht der Microsoft Tay-KI auf Twitter 21
Abbildung 7: Topologische Datenanalyse eines Teilbereichs des ImageNet Testdatensatzes 28
Abbildung 8: Vergleich des Originalbilds (links) und Heat Map (rechts) 29
Abbildung 9: Feature Map Visualisierungen eines Convolutional Deep Believe Network 30
Abbildung 10: Darstellung von Feature Maps mithilfe der Aktivierungsmaximierung 32
Abbildung 11: Beispiel eines "adversarial" Bildes 33
Abbildung 12: Demonstration der adverserial-Patch Angriffsmethode 37

Quellen

Bücher

Evans Lansing Smith, Ph.D., Nathan Robert Brown. (2008)
The Complete Idiot's Guide to World Mythology

Rudolf Brockhaus. (2013)
Flugregelung

Statistiken

Statista. Incarceration rates in OECD countries as of 2018.
https://www.statista.com/statistics/300986/incarceration-rates-in-oecd-countries/ - zuletzt abgerufen am 25.3.2019

Rechtsquellen

Algorithmic Accountability Act. OLL19293.
https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202019%20Bill%20Text.pdf – zuletzt abgerufen am 2.6.2019

Bundesgesetz zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten (Datenschutzgesetz –DSG).
https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=bundesnormen&Gesetzesnummer=10001597 – zuletzt abgerufen am 15.6.2019

Datenschutz-Deregulierungsgesetz 2018.
ELI: https://www.ris.bka.gv.at/eli/bgbl/I/2018/24/20180515 - zuletzt abgerufen am 15.6.2019

Richtlinie (EU) 2019/790 des Europäischen Parlaments und des Rates vom 17. April 2019 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt und zur Änderung der Richtlinien 96/9/EG und 2001/29/EG.
ELI: http://data.europa.eu/eli/dir/2019/790/oj – zuletzt abgerufen am 15.6.2019

Senate Bill 10, 2017-2018 Reg. Sess., ch. 244. (Cal . 2018)
https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB10 - zuletzt abgerufen am 15.6.2019

Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung).
ELI: https://eur-lex.europa.eu/eli/reg/2016/679/oj – zuletzt abgerufen am 15.6.2019

Berichte

Bundesministerium für Verkehr, Innovation und Technologie und Bundesministerium für Digitalisierung und Wirtschaftsstandort. (2018). Artificial Intelligence Mission Austria 2030.

https://www.bmdw.gv.at/DigitalisierungundEGovernment/Strategien/Documents/AIM_2030.pdf - zuletzt abgerufen am 9.6.2019
Europäische Kommission. (8.4.2019). Communication: Building Trust in Human Centric Artificial Intelligence.
https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence - zuletzt abgerufen am 4.6.2019

Europäische Kommission. (7.12.2018). Coordinated Plan on Artificial Intelligence.
https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence - zuletzt abgerufen am 4.6.2019

Europäische Kommission. (8.4.2019). Ethics guidelines for trustworthy AI.
https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai - zuletzt abgerufen am 4.6.2019

The Citizen Lab. (9.2018). Bots at the Gate A Human Rights Analysis of Automated Decision Making in Canada's Immigration and Refugee System.
https://citizenlab.ca/wp-content/uploads/2018/09/IHRP-Automated-Systems-Report-Web-V2.pdf - zuletzt abgerufen am 31.3.2019

OECD. (2019). Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449.
https://legalinstruments.oecd.org/api/print?ids=648&lang=en – zuletzt abgerufen am 10.6.2019

Wissenschaftliche Arbeiten und Artikel

AI Magazine. (2010). The AI Behind Watson — The Technical Article.
Abrufbar unter: http://www.aaai.org/Magazine/Watson/watson.php - zuletzt abgerufen am 15.6.2019

Alan M. Turing. (1950). Computing Machinery and Intelligence.
DOI: https://doi.org/10.1093/mind/LIX.236.433 - zuletzt abgerufen am 15.6.2019

Alexey Kurakin, Ian Goodfellow, Samy Bengio. (2017). Adversarial examples in the physical world.
arXiv: https://arxiv.org/abs/1607.02533 - zuletzt abgerufen am 15.6.2019

Anh Nguyen, Jason Yosinski, Jeff Clune. (2015). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images.
arXiv: https://arxiv.org/abs/1412.1897v4 - zuletzt abgerufen am 15.6.2019

Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok. (2018). Synthesizing Robust Adversarial Examples.
arXiv: https://arxiv.org/abs/1707.07397 - zuletzt abgerufen am 15.6.2019

Bernard E. Harcourt. (2005). Against Prediction: Sentencing, Policing, and Punishing in an Actuarial Age.
DOI: http://doi.org/10.2139/ssrn.756945 - zuletzt abgerufen am 15.6.2019

Bolei Zhou, et.al. (2015). Learning Deep Features for Discriminative Localization.
arXiv: https://arxiv.org/abs/1512.04150 - zuletzt abgerufen am 15.6.2019

Bruce G. Buchanan. (2005). A (Very) Brief History of Artificial Intelligence.
DOI: https://doi.org/10.1609/aimag.v26i4.1848 – zuletzt abgerufen am 15.6.2019

Bryce Goodman, Seth Flaxman. (2016). European Union regulations on algorithmic decision-makingand a “right to explanation”
arXiv: https://arxiv.org/abs/1606.08813 - zuletzt abgerufen am 15.6.2019
Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. (2014). Intriguing properties of neural networks.
arXiv: https://arxiv.org/abs/1312.6199 - zuletzt abgerufen am 15.6.2019

Daniel Goldfarb. (2018). Understanding Deep Neural Networks Using Topological Data Analysis.
arXiv: https://arxiv.org/abs/1811.00852 - zuletzt abgerufen am 15.6.2019

David West. (2000). Neural Network Credit Scoring Models.
DOI: https://doi.org/10.1016/S0305-0548(99)00149-5 – zuletzt abgerufen am 15.6.2019

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. (2009). Visualizing Higher-Layer Features of a Deep Network.
DOI: https://doi.org/10.1145/2001269.2001295 – zuletzt abgerufen am 15.6.2019

E. A. Feigenbaum. (1980). Expert Systems in the 1980s.
Abrufbar unter: https://stacks.stanford.edu/file/druid:vf069sz9374/vf069sz9374.pdf - zuletzt abgerufen am 15.6.2019

Edmond Awad, Sohan Dsouza, Richard Kim et.al. (2018). The Moral Machine experiment.
DOI: https://doi.org/10.1038/s41586-018-0637-6 - zueltzt abgerufen am 15.6.2019

Frédéric Chazal and Bertrand Michel. (2017). An introduction to Topological Data Analysis: fundamentaland practical aspects for data scientists.
arXiv: https://arxiv.org/abs/1710.04019 - zuletzt abgerufen am 15.6.2019

Geoffrey Hinton, Oriol Vinyals, Jeff Dean. (2015). Distilling the Knowledge in a Neural Network.
arXiv: https://arxiv.org/abs/1503.02531 - zuletzt abgerufen am 15.6.2019

Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. (2011). Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks.
arXiv: http://www.cs.utoronto.ca/~rgrosse/cacm2011-cdbn.pdf

Hué, Sullivan & Hurlin, Christophe & Tokpavi, Sessi. (2017). Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects. Figure 2.
Abrufbar unter: https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=IAAE2018&paper_id=185 – zuletzt abgerufen am 28.6.2019

Jianming Zhang, et.al. (2016). Top-down Neural Attention by Excitation Backprop.
arXiv: https://arxiv.org/abs/1608.00507v1 - zuletzt abgerufen am 15.6.2019

Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi. (2019). One Pixel Attack for Fooling Deep Neural Networks.
arXiv: https://arxiv.org/abs/1710.08864 - zuletzt abgerufen am 15.6.2019

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, Ting Wang. (2018). TextBugger: Generating Adversarial Text Against Real-world Applications.
arXiv: https://arxiv.org/abs/1812.05271 - zuletzt abgerufen am 15.6.2019

Joao Bastos. (2008). Credit scoring with boosted decisiontrees.
Abrufbar unter: https://mpra.ub.uni-muenchen.de/8156/1/MPRA_paper_8156.pdf - zuletzt abgerufen am 15.6.2019

Jon Kleinberg et.al. (2017). Human Decisions and Machine Predictions.
DOI: http://doi.org/10.3386/w23180 - zuletzt abgerufen am 15.6.2019

Joy Buolamwini, Timnit Gebru. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.
Abrufbar unter: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf - zuletzt abgerufen am 15.6.2019

Jürgen Schmidhuber. (2014). Deep Learning in Neural Networks: An Overview.
arXiv: https://arxiv.org/abs/1404.7828 - zuletzt abgerufen am 15.6.2019

Karen Simonyan, Andrea Vedaldi, Andrew Zisserman. (2014). Deep Inside Convolutional Networks: VisualisingImage Classification Models and Saliency Maps.
arXiv: https://arxiv.org/abs/1312.6034v2 - zuletzt abgerufen am 15.6.2019

Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, Dorothea Kolossa. (2018). Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding.
arXiv: https://arxiv.org/abs/1808.05665 - zuletzt abgerufen am 15.6.2019

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter. (2016). Accessorize to a Crime: Real and Stealthy Attacks onState-of-the-Art Face Recognition.
DOI: http://doi.org/10.1145/2976749.2978392 - zuletzt abgerufen am 15.6.2019

Matthew D. Zeiler, Rob Fergus. (2013). Visualizing and Understanding Convolutional Networks.
arXiv: https://arxiv.org/abs/1311.2901v3 - zuletzt abgerufen am 15.6.2019

Munkhdalai, Namsrai, Ho Ryu. (2018). Credit Scoring with Deep Learning.
Abrufbar unter:
https://www.researchgate.net/publication/325071650_Credit_Scoring_with_Deep_Learning#pf3 – zuletzt abgerufen am 15.6.2019

Muzammal Naseer, Salman H. Khan. (2018). Local Gradients Smoothing: Defense against localized adversarial attacks.
arXiv: https://arxiv.org/abs/1807.01216 - zuletzt abgerufen am 15.6.2019

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, Ananthram Swami. (2015). Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks.
arXiv: https://arxiv.org/abs/1511.04508 - zuletzt abgerufen am 15.6.2019

Phillipa Foot. (1967). The Problem of Abortion and the Doctrine of the Double Effect.
DOI: https://doi.org/10.1093/0199252866.003.0002 - zuletzt abgerufen am 15.6.2019

Ramprasaath R. Selvaraju, et.al. (2016). Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization.
arXiv: https://arxiv.org/abs/1610.02391v2 - zuletzt abgerufen am 15.6.2019

Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do and Kaori Togashi. (2018). Convolutional neural networks: an overviewand application in radiology.
DOI: https://doi.org/10.1007/s13244-018-0639-9 - zuletzt abgerufen am 15.6.2019

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard. (2016). DeepFool: a simple and accurate method to fool deep neural networks.
arXiv: https://arxiv.org/abs/1511.04599 – zuletzt abgerufen am 15.6.2019

Simen Thys, Wiebe Van Ranst. (2019). Fooling automated surveillance cameras: adversarial patches to attack person detection.
arXiv: https://arxiv.org/abs/1904.08653 - zuletzt abgerufen am 15.6.2019

Skeem, Jennifer L. and Monahan, John and Lowenkamp, Christopher (2016). Gender, Risk Assessment, and Sanctioning: The Cost of Treating Women Like Men.
DOI: http://doi.org/10.2139/ssrn.2718460 - zuletzt abgerufen am 15.6.2019

S.W. Ellacott. (1992). Techniques for the mathematical analysis of neural networks.
DOI: https://doi.org/10.1016/0377-0427(94)90307-7 – zuletzt abgerufen am 15.6.2019

Tencent Keen Security La. (2019). Experimental Security Research of Tesla Autopilot.
Abrufbar unter:
https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf - zuletzt abgerufen am 15.6.2019

Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer. (2018). Adversarial Patch.
arXiv: https://arxiv.org/abs/1712.09665 - zuletzt abgerufen am 15.6.2019

Wenqi Wang, Lina Wang, Benxiao Tang, Run Wang, Aoshuang Ye. (2019). Towards a Robust Deep Neural Network in Text Domain A Survey.
arXiv: https://arxiv.org/abs/1902.07285 - zuletzt abgerufen am 15.6.2019

Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, Kai Chen. (2019). Seeing isn't Believing: Practical Adversarial Attack Against Object Detectors.
arXiv: https://arxiv.org/abs/1812.10217 - zuletzt abgerufen am 15.6.2019

Journalistische Quellen

Anandtech. (5.12.2018). The Qualcomm Snapdragon 855 Pre-Dive.
https://www.anandtech.com/show/13680/snapdragon-855-going-into-detail/2 - zuletzt abgerufen am 1.5.2019

ArsTechnica. (11.1.2019). Man says CES lidar's laser was so powerful it wrecked his $1,998 camera.
https://arstechnica.com/cars/2019/01/man-says-ces-lidars-laser-was-so-powerful-it-wrecked-his-1998-camera/ - zuletzt abgerufen am 16.5.2019

BBC. (26.9.2012). Driverless car bill is signed in California at Google headquarters.
https://www.bbc.com/news/technology-19726951 - zuletzt abgerufen am 27.3.2019

BBC. (5.4.2019). Boeing 737 Max: What went wrong? https://www.bbc.com/news/world-africa-47553174 – zuletzt abgerufen am 22.5.2019

Business Insider. (14.10.2010). Mark Zuckerberg, Moving Fast And Breaking Things.
https://www.businessinsider.com/mark-zuckerberg-2010-10?IR=T – zuletzt abgerufen am 12.6.2019

CBC. (26.9.2018). Federal use of A.I. in visa applications could breach human rights, report says.
https://www.cbc.ca/news/politics/human-rights-ai-visa-1.4838778 - zuletzt abgerufen am 1.4.2019

CBC Radio. (16.11.2018). How artificial intelligence could change Canada's immigration and refugee system. https://www.cbc.ca/radio/thesundayedition/november-18-2018-the-sunday-edition-1.4907270/how-artificial-intelligence-could-change-canada-s-immigration-and-refugee-system-1.4908587 - zuletzt abgerufen am 1.4.2019

CNBC. (17.6.2017). Everyone keeps talking about A.I.—here's what it really is and why it's so hot now. https://www.cnbc.com/2017/06/17/what-is-artificial-intelligence.html - zuletzt abgerufen am 14.4.2019

Frankfurter Allgemeine Zeitung. (24.3.2016). Zum Nazi und Sexisten in 24 Stunden. https://www.faz.net/aktuell/wirtschaft/netzwirtschaft/microsofts-bot-tay-wird-durch-nutzer-zum-nazi-und-sexist-14144019.html - zuletzt abgerufen am 15.6.2019

Futurezone. (9.6.2019). "Robust und sicher": G20-Staaten einigen sich auf KI-Prinzipien. https://futurezone.at/netzpolitik/robust-und-sicher-g20-staaten-einigen-sich-auf-ki-prinzipien/400518757 - zuletzt abgerufen am 9.6.2019

Gizmodo. (1.6.2018). Google Plans Not to Renew Its Contract for Project Maven, a Controversial Pentagon Drone AI Imaging Program. https://gizmodo.com/google-plans-not-to-renew-its-contract-for-project-mave-1826488620?rev=1527878336532&utm_campaign=socialflow_gizmodo_twitter&utm_source=gizmodo_twitter&utm_medium=socialflow – zuletzt abgerufen am 30.3.2019

Heise. (21.7.2017). 1-Watt-Rechenstick: Movidius Neural Compute Stick für maschinelles Sehen. https://www.heise.de/newsticker/meldung/1-Watt-Rechenstick-Movidius-Neural-Compute-Stick-fuer-maschinelles-Sehen-3780324.html - zuletzt abgerufen am 2.5.2019

MacRumors. (13.6.2016). Apple Opens Siri to Third-Party Developers With iOS 10. https://www.macrumors.com/2016/06/13/apple-siri-api-third-party-developers/ - zuletzt abgerufen am 13.4.2019

New York Times. (17.2.2011). Computer Wins on 'Jeopardy!': Trivial, It's Not. https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html - zuletzt abgerufen am 10.4.2019

New York Times. (20.12.2017). Even Imperfect Algorithms Can Improve the Criminal Justice System. https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html - zuletzt abgerufen am 26.3.2019

Pro Publica. (23.5.2016). Machine Bias - There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing - zuletzt abgerufen am 26.3.2019

Reuters. (10.10.2018). Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G - zuletzt abgerufen am 12.5.2019

SafeUm. (10.8.2017). You can confuse self-driving cars by altering street signs. https://safeum.com/blog/2664-you-can-confuse-self-driving-cars-by-altering-street-signs.html - zuletzt abgerufen am 21.5.2019

Science Magazine. (14.4.2017). Self-taught artificial intelligence beats doctors at predicting heart attacks. https://www.sciencemag.org/news/2017/04/self-taught-artificial-intelligence-beats-doctors-predicting-heart-attacks - zuletzt abgerufen am 1.4.2019

Tech Crunch. (3.5.2019). Microsoft launches a drag-and-drop machine learning tool. https://techcrunch.com/2019/05/02/microsoft-launches-a-drag-and-drop-machine-learning-tool-and-hosted-jupyter-notebooks/ - zuletzt abgerufen am 7.5.2019

TechCrunch. (27.5.2009). Siri: A Powerful Virtual Assistant For The iPhone. https://techcrunch.com/2009/05/27/siri-the-virtual-assistant-that-will-make-everyone-love-the-iphone-even-more/ - zuletzt abgerufen am 13.4.2019

TechCrunch. (4.2.2010). Siri's IPhone App Puts A Personal Assistant In Your Pocket. https://techcrunch.com/2010/02/04/siri-iphone-personal-assistant/ - zuletzt abgerufen am 13.4.2019

Technologyreview. (4.2.2019). This is how AI bias really happens—and why it's so hard to fix. https://www.technologyreview.com/ - zuletzt abgerufen a, 12.5.2019

The Guardian. (07.9.2018). Imprisoned by algorithms: the dark side of California ending cash bail. https://www.theguardian.com/us-news/2018/sep/07/imprisoned-by-algorithms-the-dark-side-of-california-ending-cash-bail - zuletzt abgerufen am 1.4.2019

The Guardian. (15.3.2019). Social media firms fight to delete Christchurch shooting footage. https://www.theguardian.com/world/2019/mar/15/video-of-christchurch-attack-runs-on-social-media-and-news-sites - zuletzt abgerufen am 18.5.2019

The Guardian. (24.3.2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter - zuletzt abgerufen am 10.6.2019

The Guardian. (12.12.2016). The trolley problem: would you kill one person to save many others? https://www.theguardian.com/science/head-quarters/2016/dec/12/the-trolley-problem-would-you-kill-one-person-to-save-many-others - zuletzt abgerufen am 28.3.2019

The Guardian. (29.3.2019). UK, US and Russia among those opposing killer robot ban. https://www.theguardian.com/science/2019/mar/29/uk-us-russia-opposing-killer-robot-ban-un-ai - zuletzt abgerufen am 30.3.2019

The Harvard Gazette. (13.9.2012). Alan Turing at 100. https://news.harvard.edu/gazette/story/2012/09/alan-turing-at-100/ - zuletzt abgerufen am 9.4.2019

The Independent. (3.9.2018). 'Killer robots' ban blocked by US and Russia at UN meeting. https://www.independent.co.uk/life-style/gadgets-and-tech/news/killer-robots-un-meeting-autonomous-weapons-systems-campaigners-dismayed-a8519511.html - zuletzt abgerufen am 30.3.2019

The Intercept. (1.3.2019). Google Hedges on Promise to End Controversial Involvement in Military Drone Contract. https://theintercept.com/2019/03/01/google-project-maven-contract/ - zuletzt abgerufen am 30.3.2019

The Intercept. (4.2.2019). Google Hired Gig Economy Workers to Improve Artificial Intelligence in Controversial Drone-Targeting Project. https://theintercept.com/2019/02/04/google-ai-project-maven-figure-eight/ - zuletzt abgerufen am 30.3.2019

The Japan Times. (8.6.2019). G20 ministers agree on guiding principles for using artificial intelligence. https://www.japantimes.co.jp/news/2019/06/08/business/g20-ministers-kick-talks-trade-digital-economy-ibaraki-prefecture/ - zuletzt abgerufen am 10.6.2019

The Verge. (10.4.2019). A new bill would force companies to check their algorithms for bias. https://www.theverge.com/2019/4/10/18304960/congress-algorithmic-accountability-act-wyden-clarke-booker-bill-introduced-house-senate - zuletzt abgerufen am 2.6.2019

The Verge. (5.3.2019). I flew a helicopter, and then the helicopter flew me. https://www.theverge.com/transportation/2019/3/5/18250996/sikorsky-autonomous-helicopter-flying-taxi-lockheed - zuletzt abgerufen am 13.4.2019

The Verge. (24.4.2019). It's Elon Musk vs. everyone else in the race for fully driverless cars. https://www.theverge.com/2019/4/24/18512580/elon-musk-tesla-driverless-cars-lidar-simulation-waymo – zuletzt abgerufen am 17.5.2019

Time Magazine. (7.2.2019). Artificial Intelligence Has a Problem With Gender and Racial Bias. Here's How to Solve It. https://time.com/5520558/artificial-intelligence-racial-gender-bias/ - zuletzt abgerufen am 30.6.2019

Toms Hardware. (5.3.2019). Google's Edge TPU Machine Learning Chip Debuts in Raspberry Pi-Like Dev Board. https://www.tomshardware.com/news/google-edge-tpu-coral-dev-board-usb-accelerator,38750.html – zuletzt abgerufen am 2.5.2019

Washington Post. (29.8.2019). California abolishes money bail with a landmark law. But some reformers think it creates new problems. https://www.washingtonpost.com/news/morning-mix/wp/2018/08/29/california-abolishes-money-bail-with-a-landmark-law-but-some-reformers-think-it-creates-new-problems/ - zuletzt abgerufen am 25.3.2019

Wired. (16.5.2019). How Tech Companies Are Shaping the Rules Governing AI. https://www.wired.com/story/how-tech-companies-shaping-rules-governing-ai/ - zuletzt abgerufen am 9.6.2019

Wired. (20.12.2018). The 21 (and Counting) Biggest Facebook Scandals of 2018. https://www.wired.com/story/facebook-scandals-2018/ - zuletzt abgerufen am 22.5.2019

Wired. (24.8.2016). The iBrain Is Here—and It's Already Inside Your Phone. https://www.wired.com/2016/08/an-exclusive-look-at-how-ai-and-machine-learning-work-at-apple/ - zuletzt abgerufen am 13.4.2019

Wired. (12.2.2019). The Real Reason Tech Struggles With Algorithmic Bias. https://www.wired.com/story/the-real-reason-tech-struggles-with-algorithmic-bias/ - zuletzt abgerufen am 30.6.2019

Zeit. (24.3.2016). Twitter-Nutzer machen Chatbot zur Rassistin. https://www.zeit.de/digital/internet/2016-03/microsoft-tay-chatbot-twitter-rassistisch - zuletzt abgerufen am 10.6.2019

Sonstige Internetquellen

Alfian Losari. Building an Interactive Voice App Using Custom Siri Shortcuts in iOS 12. https://medium.com/appcoda-tutorials/building-custom-siri-shortcut-intent-ui-extension-to-display-remote-data-alfian-losari-efe891a44a70 - zuletzt abgerufen am 13.4.2019

Apple. Siri. https://www.apple.com/siri/ - zuletzt abgerufen am 13.4.2019

Box Office Mojo. 2001: A Space Odyssey. https://www.boxofficemojo.com/movies/?id=2001.htm – zuletzt abgerufen am 30.3.2019

Box Office Mojo. The Terminator.
https://www.boxofficemojo.com/movies/?id=terminator.htm – zuletzt abgerufen am 30.3.2019

Bundesministerium für Digitalisierung und Wirtschaftsstandort. Künstliche Intelligenz (KI).
https://www.bmdw.gv.at/DigitalisierungundEGovernment/Strategien/Seiten/K%C3%BCnstliche-Intelligenz.aspx - zuletzt abgerufen am 9.6.2019

Christopher Olah. Understanding LSTM Networks.
https://colah.github.io/posts/2015-08-Understanding-LSTMs/ - zuletzt abgerufen am 7.5.2019

CIO. Chicago Police Department Uses IT to Fight Crime, Wins Grand CIO Enterprise Value Award 2004.
https://www.cio.com/article/2439813/chicago-police-department-uses-it-to-fight-crime--wins-grand-cio-enterprise-value-aw.html – zuletzt abgerufen am 10.5.2019
Daphne Cornelisse. An intuitive guide to Convolutional Neural Networks.
https://medium.freecodecamp.org/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050 - zuletzt abgerufen am 5.5.2019

Divisio. KI leicht erklärt – Teil 2: Von klassischer KI, Neuronalen Netzen und Deep Learning.
https://divis.io/2019/03/ki-fuer-laien-teil-2-klassischer-ki-neuronalen-netzen-und-deep-learning/ - zuletzt abgerufen am 28.6.2019

Equifax. How Are Credit Scores Calculated?
https://www.equifax.com/personal/education/credit/score/how-is-credit-score-calculated/ - zuletzt abgerufen am 10.4.2019

Encyclopaedia Britannica. Artificial Intelligence.
https://www.britannica.com/technology/artificial-intelligence - zuletzt abgerufen am 14.4.2019

Encyclopaedia Britannica. Neural Network.
https://www.britannica.com/technology/neural-network - zuletzt abgerufen am 15.6.2019

Encyclopaedia Britannica. Turing test.
https://www.britannica.com/technology/Turing-test - zuletzt abgerufen am 14.4.2019

EUR-Lex. Verfahren 2012/0011/COD.
https://eur-lex.europa.eu/legal-content/DE/HIS/?uri=celex%3A32016R0679 – zuletzt abgerufen am 25.6.2019

Europäischer Datenschutzausschuss. Unsere Arbeit und Hilfsmittel -> Unsere Dokumente -> Artificial Intelligence.
https://edpb.europa.eu/our-work-tools/our-documents/topic/artificial-intelligence_de - zuletzt abgerufen am 31.5.2019

Europäischer Datenschutzausschuss. Unsere Arbeit und Hilfsmittel -> Unsere Dokumente -> Automated individual decision-making.
https://edpb.europa.eu/our-work-tools/our-documents/topic/automated-individual-decision-making_de zuletzt abgerufen am 31.5.2019

Europäische Kommission. High-Level Expert Group on Artificial Intelligence.
https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence - zuletzt abgerufen am 4.6.2019

European Parlament. Self-driving cars in the EU: from science fiction to reality. http://www.europarl.europa.eu/news/en/headlines/economy/20190110STO23102/self-driving-cars-in-the-eu-from-science-fiction-to-reality - zuletzt abgerufen am 28.6.2019

Gartner. Widespread artificial intelligence, biohacking, new platforms and immersive experiences dominate this year's Gartner Hype Cycle. https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/ - zuletzt abgerufen am 12.6.2019

Geizhals. Grafikkarten » PCIe mit GPU Workstation nach Erscheinung: Quadro RTX 8000. https://geizhals.at/?cat=gra16_512&asuch=quadro+8000&v=e&hloc=at&filter=aktualisieren&sort=p&bl1_id=30&xf=9807_11808+-+Quadro+RTX+8000#gh_filterbox – zuletzt abgerufen am 1.5.2019

Github. Kaldi Speech Recognition Toolkit. https://github.com/kaldi-asr/kaldi - zuletzt abgerufen am 26.5.2019

Google. Cloud TPU Preise. https://cloud.google.com/tpu/docs/pricing - zuletzt abgerufen am 2.5.2019

Google. Inceptionism: Going Deeper into Neural Networks. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html - zuletzt abgerufen am 14.5.2019

Harvard University. The History of Artificial Intelligence. http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/ - zuletzt abgerufen am 9.4.2019

IBM Developer. A neural networks deep dive. https://developer.ibm.com/articles/cc-cognitive-neural-networks-deep-dive/ - zuletzt abgerufen am 30.6.2019

IBM Knowledge Center. Diskriminanzanalyse. https://www.ibm.com/support/knowledgecenter/de/SSLVMB_sub/statistics_mainhelp_ddita/spss/base/idh_disc.html - zuletzt abgerufen am 28.6.2019

ImageNet. ImageNet. http://image-net.org/index - zuletzt abgerufen am 12.5.2019

IMDB. Ratings and Reviews for new Movies and TV-Shows. https://www.imdb.com/ - zuletzt abgerufen am 19.5.2019

Jean Francois Puget. Overfitting In Machine Learning. https://www.ibm.com/developerworks/community/blogs/jfp/entry/Overfitting_In_Machine_Learning?lang=en – zuletzt abgerufen am 30.6.2019

Kaggle. State Farm Distracted Driver Detection. https://www.kaggle.com/c/state-farm-distracted-driver-detection#description - zuletzt abgerufen am 1.4.2019

LiveScience. History of A.I.: Artificial Intelligence (Infographic). https://www.livescience.com/47544-history-of-a-i-artificial-intelligence-infographic.html - zuletzt abgerufen am 9.4.2019

McKinsey. The industry is on the verge of a seismic, tech-driven shift. A focus on four areas can position carriers to embrace this change.
https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance - zuletzt abgerufen am 1.4.2019

Medianama. Microsoft takes down its AI chatbot which turned evil under human influence.
https://www.medianama.com/2016/03/223-microsoft-tay-chatbot/ - zuletzt abgerufen am 10.6.2019

Medium. Applied Deep Learning - Part 1: Artificial Neural Networks.
https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6 - zuletzt abgerufen am 29.4.2019

Michael Nielsen. Neural Networks and Deep Learning.
http://neuralnetworksanddeeplearning.com/index.html - zuletzt abgerufen am 15.6.2019

Microsoft. Microsoft and AI: Introducing social chatbot Zo, Cortana takes on new roles and more.
https://blogs.microsoft.com/blog/2016/12/13/microsoft-ai-introducing-social-chatbot-zo-cortana-takes-new-roles/ - zuletzt abgerufen am 10.6.2019

MIT Media Lab. Moral Machine.
http://moralmachine.mit.edu/ - zuletzt abgerufen am 28.3.2019
Nvidia. Deep Learning in a Nutshell: Core Concepts.
https://devblogs.nvidia.com/deep-learning-nutshell-core-concepts/ - zuletzt abgerufen am 13.5.2019

Nvidia. NVIDIA NVLink High-Speed GPU Interconnect.
https://www.nvidia.com/de-de/design-visualization/nvlink-bridges/ - zuletzt abgerufen am 1.5.2019

Nvidia. RTX. IT'S ON. GeForce RTX 2080 Ti.
https://www.nvidia.com/de-de/geforce/graphics-cards/rtx-2080-ti/ - zuletzt abgerufen am 1.5.2019

Nvidia. Tensor Cores.
https://www.nvidia.com/de-de/data-center/tensorcore/ - zuletzt abgerufen am 1.5.2019

OECD. Forty-two countries adopt new OECD Principles on Artificial Intelligence.
https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm - zuletzt abgerufen am 10.6.2019

Robert Luckner. Flugführungssysteme zur Pilotenassistenz -Was kann man aus der Luftfahrt lernen?
https://mediatum.ub.tum.de/doc/1145165/1145165.pdf - zuletzt abgerufen am 13.4.2019

Ron Wyden US-Senator for Oregon. Wyden, Booker, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms.
https://www.wyden.senate.gov/news/press-releases/wyden-booker-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms- - zuletzt abgerufen am 2.6.2019

Sagar Sharma. Epoch vs Batch Size vs Iterations.
https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9 - zuletzt abgerufen am 15.6.2019

Skymind. A Beginner's Guide to Convolutional Neural Networks (CNNs).
https://skymind.ai/wiki/convolutional-network - zuletzt abgerufen am 5.5.2019

Stand Out Publishing. Hidden Layer.
http://standoutpublishing.com/g/hidden-layer.html - zuletzt abgerufan am 30.6.2019

Stanford Encyclopedia of Philosophy. The Chinese Room Argument.
https://plato.stanford.edu/entries/chinese-room/ - zuletzt abgerufen am 14.4.2019

Stanford Law School. Nevada Governor Signs Driverless Car Bill Into Law.
http://cyberlaw.stanford.edu/blog/2011/06/nevada-governor-signs-driverless-car-bill-law - zuletzt abgerufen am 27.3.2019

Stanford University UFLDL. Multi-Layer Neural Network.
http://deeplearning.stanford.edu/tutorial/supervised/MultiLayerNeuralNetworks/ - zuletzt abgerufen am 3.6.2019

Stanford University. Visualizing what ConvNets learn.
http://cs231n.github.io/understanding-cnn/ - zuletzt abgerufen am 13.5.2019

StateFarm. Have Drive Safe & Save™ Questions? We've Got Answers.
https://www.statefarm.com/customer-care/faqs/drive-safe-save - zuletzt abgerufen am 1.4.2019

StateFarm. You're in the Driver's Seat When It Comes to Your Discount.
https://www.statefarm.com/insurance/auto/discounts/drive-safe-save - zuletzt abgerufen am 1.4.2019

Suvro Banerjee. Recurrent Neural Networks and LSTM.
https://towardsdatascience.com/recurrent-neural-networks-and-lstm-4b601dd822a5 – zuletzt abgerufen am 7.5.2019

Teddy Bear Orchestra.
http://www.teddybearorchestra.com/ - zuletzt abgerufen am 14.4.2019

TensorFlow. Train your first neural network: basic classification.
https://www.tensorflow.org/tutorials/keras/basic_classification - zuletzt abgerufen am 30.6.2019

Tim Dettmers. Which GPU(s) to Get for Deep Learning.
https://timdettmers.com/2019/04/03/which-gpu-for-deep-learning/ - zuletzt abgerufen am 1.5.2019
Towards Data Science. RTX 2060 Vs GTX 1080Ti Deep Learning Benchmarks.
https://towardsdatascience.com/rtx-2060-vs-gtx-1080ti-in-deep-learning-gpu-benchmarks-cheapest-rtx-vs-most-expensive-gtx-card-cd47cd9931d2 - zuletzt abgerufen am 1.5.2019

Tyler Vigen. Spurious Correlations.
http://www.tylervigen.com/spurious-correlations - zuletzt abgerufen am 3.6.2019

Uniqa. SafeLine - die Autoversicherung, die Leben retten kann.
https://www.uniqa.at/versicherung/cms/privatkunden/lebenssituationen/SafeLine.de.html - zuletzt abgerufen am 1.4.2019

University of Toronto. 3.0 History of Neural Networks.
http://www.psych.utoronto.ca/users/reingold/courses/ai/cache/neural4.html - zuletzt abgerufen am 10.4.2019

University of Wisconsin Madison. A Basic Introduction To Neural Networks.
http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html - zuletzt abgerufen am 15.6.2019

Univ.-Prof. Dr. Thomas Metzinger.
https://www.philosophie.fb05.uni-mainz.de/arbeitsbereiche/theoretische/thmetzinger/ - zuletzt abgerufen am 9.6.2019

U.S. Department of Defense. Project Maven to Deploy Computer Algorithms to War Zone by Year's End.
https://dod.defense.gov/News/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/ - zuletzt abgerufen am 30.3.2019